Les Perceptrons Multicouches: de la régression non-linéaire aux problèmes inverses *

Fouad BADRAN et Sylvie THIRIA

Laboratoire d'Océanographie DYnamique et de Climatologie

4, Place Jussieu

75005 PARIS, FRANCE

CEDRIC, Conservatoire National des Arts et Métiers

292, rue Saint-Martin

75003 PARIS, FRANCE

Abstract

The problem of nonlinear least square regression using multi-layered perceptrons is addressed in this paper. A general formulation using maximum likelihood is given and a particular attention is paid to algorithm accounting for uncertainties. The generalization to density estimation is given. In the second part of the paper, a general framework allows to modelize general inverse problems.

Keywords : *Non linear regression. Multi-layered Perceptrons. Back-propagations . Inverse problem. Density estimation.*

*Acknowlegment : Ce travail a été réalisée dans le cadre du programme NeuroSAT (env94-CT96-0314) soutenu par la CEE.

1 Introduction

Avec le développement de nouvelles méthodes qui permettent d'effectuer des mesures à distance, la géophysique accumule des quantités gigantesques de données. Ces données sont hétérogènes, elles proviennent de capteurs, d'appareils de mesure sophistiqués parfois embarqués à bord de satellites. Bien qu'elles soient le plus souvent très fortement bruitées, il est cependant possible d'en extraire des informations précises sur les phénomènes observés en faisant l'étude de certains paramètres, qui les caractérisent. Un grand nombre de méthodes sont mises en oeuvre par la géophysique pour atteindre cette information, elles dépendent de domaines comme la physique ou la statistique. La physique essaie de représenter la réalité à l'aide de lois, la statistique essaie d'inférer certaines caractéristiques de la réalité à partir de données empiriques. L'information extraite va dépendre de l'ensemble des modèles physiques ou statistiques utilisés pour décrire le problème et des différentes hypothèses qui seront choisies au départ. Nous aborderons dans cet article les modélisations les plus classiques de la statistique, qui permettent de traiter le problème de la détermination des fonctions de transfert ou de la restitution des paramètres physiques des phénomènes sous jacents aux mesures observées. Nous traiterons donc des problèmes de la régression, de la résolution de problèmes inverses et de la classification.

La modélisation à partir de données empiriques est une tâche difficile, en effet les relations sous-jacentes aux mesures peuvent être fortement non linéaires et non univoques, elles sont de plus comme nous l'avons rappelé, fortement bruitées. Les méthodes conventionnelles sont souvent insuffisantes face à ce type de données, c'est pourquoi les recherches s'orientent vers le développement de nouveaux outils et de nouvelles méthodologies capables de traiter des données de plus en plus complexes. Les réseaux de neurones de type Perceptrons Multicouches (PMC) ont montré leur efficacité en tant qu'outil de modélisation appliqué aux données empiriques, ils permettent d'apporter des solutions aux problèmes que nous venons d'évoquer.

Dans la suite de l'article, nous présentons le modèle neuronal le plus connu qui est le Perceptron MultiCouches (PMC). Nous rappelons dans le premier paragraphe les propriétés théoriques associées au PMC, ce sont elles qui expliquent le mieux pourquoi l'utilisation d'un PMC permet souvent d'obtenir des performances intéressantes. Le second paragraphe introduit les PMC, l'algorithme d'apprentissage et les propriétés d'approximateur universel. L'apprentissage d'un PMC se ramène toujours à la minimisation d'une fonction de coût, les plus classiques étant celles des moindres carrés simples ou généralisés. Le troisième paragraphe présente les propriétés de la fonctions de coût des moindre carrés généralisés. Le quatrième paragraphe présente le cas particulier de la classification. Le cinquième paragraphe présente la régression non-

linéaire par PMC que nous abordons à partir de l'approche Bayesienne. De cette manière, il devient possible d'unifier la présentation de deux problèmes: la régression et la détermination des matrices de variance-covariance. Le sixième paragraphe aborde la résolution des problèmes inverses. Nous présentons dans ce paragraphe les problèmes spécifiques liés à l'inversion et les approches neuronales qui permettent de répondre à ces problèmes "mal posés" (approximation de fonctions densité, introduction de connaissance physique à priori). Le dernier paragraphe est une bibliographie commentée de différentes applications réalisées en géophysique à l'aide de ces méthodes.

2 Les Perceptrons Multi-Couches (PMC)

2.1 définitions

Un réseau de neurones est un ensemble de processeurs élémentaires, les neurones qui sont largement connectés les uns aux autres et qui sont capables d'échanger des informations au moyen des connexions qui les relient. Les connexions sont directionnelles et à chacune d'elle est associé un réel appelé poids de la connexion. L'information est ainsi transmise de manière unidirectionnelle du neurone j vers le neurone i, affectée du coefficient pondérateur w_{ij}. Un neurone calcule son état à partir d'informations venues de l'extérieur, ou bien il détermine son entrée à partir des neurones auxquels il est connecté et calcule son état comme une transformation souvent non linéaire de son entrée. Il transmet à son tour son état vers d'autres neurones ou vers l'environnement extérieur.

Un neurone est donc défini par trois caractéristiques: son état, ses connexions avec d'autres neurones et sa fonction de transfert. Nous utiliserons dans la suite, les notations suivantes:

- O: l'ensemble des états possibles des neurones.
- o_i: l'état du neurone i, où $o_i \in O$.
- f_i: la fonction de transfert associée au neurone i.
- s_i: l'entrée du neurone i, $s_i \in \mathbf{R}$
- w_{ij}: le poids de la connexion du neurone j vers le neurone i; $w_{ij} \in \mathbf{W}$ l'ensemble des poids du PMC.

Ainsi le neurone i recevant les informations de n_i neurones effectue l'opération suivante:

$$o_i = f(s_i) \text{ avec } s_i = \sum_{j=1}^{n_i} w_{ij} o_j \quad (1)$$

Les fonctions de transfert les plus souvent utilisées sont la fonction identité, la fonction sigmoïde, la fonction exponentielle et plus rarement les fonctions ondelettes [62].

- **la fonction identité**: un neurone dont la fonction de transfert est la fonction identité appelé neurone linéaire. Pour un tel neurone, l'état est calculé à l'aide de l'équation suivante:

$$o_i = s_i = \sum_j w_{ij} o_j \quad (2)$$

- **la fonction sigmoïde**: est la plus utilisée car elle introduit de la non-linéarité, mais c'est aussi une fonction continue, différentiable. Une fonction sigmoïde peut être définie par l'une des deux formes suivantes :

$$f(s) = A \frac{e^{Ks} - 1}{e^{Ks} + 1} = A \, tanh \, (\frac{K}{2} s) \quad (3)$$

ou bien

$$f(s) = \frac{A}{e^{-Ks} + 1} \quad (4)$$

contrairement à la fonction identité, ces fonctions sont bornées, elles tendent vers A quand $s \to +\infty$ et tendent respectivement vers -A ou 0 quand $s \to -\infty$. Le paramètre A régule la valeur de saturation, le paramètre K sert à réguler la pente de la courbe en tout point hors saturation.

- **la fonction exponentielle**: Elle est souvent utilisée au niveau de la couche de sortie afin d'assurer des valeurs de sorties positives et non bornées.

L'utilisation des fonctions de transfert non-linéaires permet l'obtention de modèles statistiques non-linéaires. Les **Perceptrons Multi-Couches (PMC)** [53] [28] [42] [5] sont des réseaux de neurones pour lesquels les neurones sont organisés en couches successives, les connections sont toujours dirigées des couches inférieures vers les couches supérieures et les neurones d'une même couche ne sont pas interconnectés. Un neurone ne peut donc transmettre son état qu'à un neurone situé dans une couche

postérieure à la sienne. Choisir l'architecture d'un PMC consiste à fixer le nombre de couches, le nombre de cellules par couche, la nature des différentes connexions entre les neurones et la nature des neurones sur chaque couche.

La première couche du réseau est la couche d'entrée, on suppose qu'elle contient p neurones, la dernière couche du réseau est sa couche de sortie, on suppose qu'elle contient q neurones. Les états des neurones de la première couche seront fixés par le problème traité à travers un vecteur $\mathbf{x} = (x_1, x_2, \ldots x_p)$. Les états de la première couche étant fixés, le réseau va pouvoir calculer les états de ses autres neurones en appliquant l'équation (1) de proche en proche d'une couche vers les couches supérieures. Cette partie du calcul est appellée propagation avant en opposition au calcul effectué par rétropropagation qui sera présenté à la section (2.2). Ainsi, la propagation avant permet de récupérer sur la couche de sortie un vecteur $\mathbf{y} = (y_1, y_2, \ldots y_q)$. C'est pourquoi un PMC défini une fonction de $\mathcal{R}^p$ dans $\mathcal{R}^q$.

Pour un PMC d'architecture fixé, la fonction définie par le réseau dépend des valeurs des poids $\mathbf{W}$ de ses différentes connexions. Une architecture génère donc une famille de fonctions :

$$\begin{array}{rcl} \mathcal{R}^p & \longrightarrow & \mathcal{R}^q \\ \mathbf{x} & \longmapsto & \mathbf{y} = \mathbf{F}(\mathbf{W}, \mathbf{x}) \end{array} \qquad (5)$$

Tenant compte de l'équation (1) et du fait que les fonctions de transfert des différents neurones sont indéfiniment dérivables (sigmoïde, linéaire, ...), cette famille est formée de fonctions non-linéaires et indéfiniment dérivables.

L'ensemble des familles définies par l'ensemble des architectures possibles est très adapté à la recherche de fonctions de régression non-linéaires. Plusieurs résultats théoriques ont été établis concernant leurs capacités d'approximateurs universels de fonctions [18] [23] [29] [41] [60]. Le résultat fondamental de ces travaux est que toute fonction continue de $\mathcal{R}^p$ dans $\mathcal{R}^q$ peut être approximée, d'une manière uniforme sur un compact de $\mathcal{R}^p$, par une fonction définie par un PMC à une seule couche cachée et ceci avec une précision ϵ choisie à l'avance. Cependant, plusieurs études ont montré l'intérêt de considérer des PMC à deux couches cachés [32] [48]. Ces résultats théoriques ne permettent pas d'avoir une idée précise du nombre de neurones cachés nécessaires pour approximer une fonction donnée. La détermination de l'architecture optimale se fait en utilisant les théories statistiques concernant le choix d'un modèle. Un grand nombre de résultats existent qui appliquent les résultats généraux au cas particulier des architectures PMC [28].

2.2 Apprentissage

Dans la suite de l'article nous utiliserons les PMC pour modéliser à partir des données empiriques acquises en géophysique. Nous supposerons que les appareils de mesure utilisés permettent de collecter un ensemble de couples d'observations $\mathcal{D} = \left\{(\mathbf{x}_i^{obs}, \mathbf{y}_i^{obs}), i = 1 \cdots N^{obs}\right\}$. Comme nous l'avons mentionné dans l'introduction, ces couples d'observations sont souvent des représentations bruitées ou ambigües d'une réalité sous-jacente. Tenant compte des bruits et de la manière dont sont générées les données, nous supposons que $(\mathbf{x}^{obs}, \mathbf{y}^{obs})$ sont des réalisations de variables aléatoires $\mathbf{X}$ et $\mathbf{Y}$ de fonction densité $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}/\mathbf{x})p(\mathbf{x})$.

Pour chacun des problèmes évoqués (classification, régression, estimation de fonction densité, ...) il existe une méthodologie particulière à mettre en oeuvre. Chacune aboutit à la minimisation d'une fonction de coût empirique $R_{emp}(\mathbf{W})$, fonction qui dépend des paramètres $\mathbf{W}$ du PMC et de l'ensemble $\mathcal{D}$. L'apprentissage d'un PMC consiste à minimiser cette fonction coût par descente de gradient. L'algorithme le plus connu est celui de la rétropropagation du gradient [42], [5], [28], il permet de calculer de manière récursive le gradient de $R_{emp}(\mathbf{W})$ par rapport à l'ensemble des paramètres $(\mathbf{W})$. Comme mentionné au paragraphe (2.1) o_i représente l'état du neurone i, s_i son entrée, f_i sa fonction de transfert et w_{ij} le poids synaptique du neurone j vers le neurone i: $o_i = f_i(s_i)$, $s_i = \sum_j w_{ij} o_j$. Le calcul du gradient se décompose de la façon suivante:

$$\frac{\partial R_{emp}}{\partial w_{ij}} = \frac{\partial R_{emp}}{\partial s_i} \frac{\partial s_i}{\partial w_{ij}} \tag{6}$$

Si l'on note δ_i les dérivées partielles de R_{emp} par rapport à l'entrée du neurone i, le gradient par rapport aux paramètres $\mathbf{W}$ s'écrit alors:

$$\frac{\partial R_{emp}}{\partial w_{ij}} = \delta_i o_j \tag{7}$$

les quantités $\delta_i = \frac{\partial R_{emp}}{\partial s_i}$ se calculent recursivement par rétropropagation à partir de la couche de sortie :

- **Si l'indice i caractérise un neurone de la couche de sortie**, nous avons alors:

$$\delta_i = f_i'(s_i) \frac{\partial R_{emp}}{\partial o_i} \tag{8}$$

  Ceci suppose que la fonction de coût R_{emp} est définie explicitement en fonction des états des cellules de sorties o_i.

- **Si i est l'indice d'un neurone caché**, en notant par k l'indice des neurones qui prennent leur information du neurone i, nous avons :

$$\delta_i = f_i'(s_i) \sum_k \delta_k w_{ki} \tag{9}$$

Pour une architecture de PMC donnée, seule la connaissance de la dérivées de la fonction coût par rapport aux états des neurones de sortie ($\frac{\partial R_{emp}}{\partial o}$) intervient pour initialiser le calcul du gradient. L'algorithme de rétropropagation du gradient est donc général et peut s'appliquer à n'importe quelle fonction coût (R_{emp}) dont on sait calculer les dérivées partielles par rapport aux états des cellules de sortie. Cet algorithme permet de calculer deux types de dérivées :

- Les dérivées par rapport aux poids synaptiques du réseau w_{ij}.
- Les dérivées par rapports aux entrées des neurones et en particulier par rapport aux variables présentées sur la couche d'entrée du réseau (vecteur $\mathbf{x} = (x_1, x_2, \ldots, x_p)$). Ainsi, dans ce dernier cas, pour toute variable d'entré x_i on a :

$$\frac{\partial \mathbf{F}(\mathbf{W}, \mathbf{x})}{\partial x_i} = \sum_k \delta_k w_{ki} \tag{10}$$

Ainsi, l'algorithme de rétropropagation du gradient permet, d'une manière très simple, de calculer les dérivées partielles de la fonction $\mathbf{F}(\mathbf{W}, \mathbf{x})$ qui est générée par le PMC. Cette simplicité de mise en oeuvre est particulièrement importante pour l'apprentissage des PMC comportant un grand nombre de poids. L'inversion par modèle adjoint qui sera présentée au paragraphe (6.3.1) doit son efficacité à cette propriété, il est tout à fait réaliste d'envisager son utilisation pour traiter d'applications réelles de grande taille.

De cette rapide présentation de l'apprentissage, il ressort qu'une information prépondérente est introduite par l'intermédiaire de la fonction de coût. La plus connue et la plus classiquement utilisée en statistique est la fonction des moindres carrés dont nous présentons les principales propriétés dans le paragraphe suivant.

3 La fonction de coût des moindre carrés

3.1 Présentation générale

Comme nous venons de le dire l'apprentissage d'un PMC revient à la minimisation d'une fonction de coût, la plus classique étant la fonction de coût des moindres carrés généralisés dont l'expression est la suivante:

$$R(\mathbf{W}) = \int\int \Big(\mathbf{y} - \mathbf{F}(\mathbf{W},\mathbf{x})\Big)^T \Sigma^{-1}(\mathbf{x}) \Big(\mathbf{y} - \mathbf{F}(\mathbf{W},\mathbf{x})\Big) p(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} \tag{11}$$

où $\Sigma^{-1}(\mathbf{x})$ est une matrice définie positive qui dépend en général de $\mathbf{x}$, dans l'expression des moindre carrés simple cette matrice est remplacée par la matrice identité. On peut démontrer que la minimisation de cette fonction coût implique la minimisation de :

$$\int \Big(\mathbf{E}(\mathbf{Y}/\mathbf{x}) - \mathbf{F}(\mathbf{W},\mathbf{x})\Big)^T \Sigma^{-1}(\mathbf{x}) \Big(\mathbf{E}(\mathbf{Y}/\mathbf{x}) - \mathbf{F}(\mathbf{W},\mathbf{x})\Big) p(\mathbf{x})\, d\mathbf{x} \tag{12}$$

Ce qui montre qu'un bon minimum de (11) donne une bonne approximation de :

$$\mathbf{E}(\mathbf{Y}/\mathbf{x}) = \int \mathbf{y} p(\mathbf{y}/\mathbf{x})\, d\mathbf{y} \tag{13}$$

et que les sorties d'un PMC sont telles que :

$$\mathbf{F}(\mathbf{W},\mathbf{x}) \approx \mathbf{E}(\mathbf{Y}/\mathbf{x}) \tag{14}$$

où $\mathbf{E}(\mathbf{Y}/\mathbf{x})$ est la moyenne conditionnelle des observations $\mathbf{y}^{obs}$.

Dans certains cas, où la matrice $\Sigma(\mathbf{x}) = \Sigma$ est constante et ne dépend plus de $\mathbf{x}$, il est possible de donner des informations sur la précision de cette approximation. Si l'on note $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q)$ la matrice orthonormale des vecteur propres de la matrice Σ et σ_j^2 la valeur propre associée à $\mathbf{u}_j$, la matrice Σ^{-1} peut alors se décomposer sous la forme:

$$\Sigma^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^T \quad \text{où} \quad \mathbf{D}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_q^2} \end{bmatrix} \tag{15}$$

la fonction de coût (11) s'écrit:

$$R(\mathbf{W}) = \iint \Big(\mathbf{U}^T\mathbf{y} - \mathbf{U}^T\mathbf{F}(\mathbf{W},\mathbf{x})\Big)^T \mathbf{D}^{-1}\Big(\mathbf{U}^T y - \mathbf{U}^T\mathbf{F}(\mathbf{W},\mathbf{x})\Big) p(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}$$
$$= \sum_{j=1}^{q} \iint \Big(\frac{\mathbf{u}_j^T\mathbf{y} - \mathbf{u}_j^T\mathbf{F}(\mathbf{W},\mathbf{x})}{\sigma_j}\Big)^2 p(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} \qquad (16)$$

Si l'on note $\mathbf{Y} = \mathbf{U}^T\mathbf{y} = (Y_1, \cdots, Y_p)$ les coordonnées de $\mathbf{y}$ dans les axes principaux, on obtient:

$$R(\mathbf{W}) = \sum_{j=1}^{q} \iint \Big(\frac{Y_j - \mathbf{u}_j^T\mathbf{F}(\mathbf{W},\mathbf{x})}{\sigma_j}\Big)^2 p(\mathbf{x},\mathbf{U}\mathbf{Y})\, d\mathbf{x}\, d\mathbf{Y} \qquad (17)$$

Du résultat présenté en (12) on déduit:

$$R(\mathbf{W}) = \sum_{j=1}^{q} \int \Big(\frac{\mathbf{E}(Y_j/\mathbf{x}) - \mathbf{u}_j^T\mathbf{F}(\mathbf{W},\mathbf{x})}{\sigma_j}\Big)^2 p(\mathbf{x})\, d\mathbf{x} + \text{constant} \qquad (18)$$

ou de manière équivalente:

$$R(\mathbf{W}) = \sum_{j=1}^{q} \int \Big(\frac{\mathbf{u}_j^T\big[\mathbf{E}(\mathbf{Y}/\mathbf{x}) - \mathbf{F}(\mathbf{W},\mathbf{x})\big]}{\sigma_j}\Big)^2 p(\mathbf{x})\, d\mathbf{x} + \text{constant} \qquad (19)$$

Cette équation montre bien que la précision de l'approximation dépend de $p(\mathbf{x})$ qui représente la concentration locale des données, elle est d'autant meilleure que $p(\mathbf{x})$ est grand. De plus, pour un $\mathbf{x}$ donné la précision de l'approximation sur l'axe principal $\mathbf{u}_j$ dépend de la valeur propre σ_j^2 qui lui est associée: plus cette valeur est petite meilleure est la précision. Dans la pratique, il est presque impossible d'avoir accès à la fonction de coût R, on utilise la fonction de coût empirique définie à partir de l'ensemble d'apprentissage $\mathcal{D} = \big\{(\mathbf{x}_i^{obs}, \mathbf{y}_i^{obs}), i = 1 \cdots N^{obs}\big\}$ par:

$$R_{emp}(\mathbf{W}) = \sum_{i=1}^{N^{obs}} R_i \qquad (20)$$

Avec

$$R_i = \Big(\mathbf{y}_i^{obs} - \mathbf{F}(\mathbf{W},\mathbf{x}_i^{obs})\Big)^T \Sigma^{-1}(\mathbf{x}_i^{obs})\Big(\mathbf{y}_i^{obs} - \mathbf{F}(\mathbf{W},\mathbf{x}_i^{obs})\Big) \qquad (21)$$

La fonction de coût des moindre carrés

Cette fonction de coût constitue une approximation discrète de la fonction de coût théorique R et suppose implicitement :

$$p(\mathbf{x}) = \sum_{i=1}^{N^{obs}} \frac{1}{n} Dirac_{x_i^{obs}} \tag{22}$$

Où $Dirac_x$ représente la distribution de Dirac au point $\mathbf{x}$. Si nous notons par $\mathbf{f} = (f_1, f_2, \ldots, f_q)$ la suite des q fonctions de transfert des neurones de la couche de sortie, par $\mathbf{s} = (s_1, s_2, \ldots, s_q)$ les entrées des neurones de sortie et par $\mathbf{o} = (o_1, o_2, \ldots, o_q)$ les états des neurones de sortie associés à une observation $\mathbf{x}^{obs}$, l'initialisation de l'algorithme de la rétropropagation du gradient pour un neurone de sortie j au moment de l'apprentissage d'un vecteur $\mathbf{x}^{obs}$ est alors :

$$\frac{\partial R}{\partial s_j} = -2\,\Sigma^{-1}\left(\mathbf{x}^{obs}\right)\Big)(\mathbf{y}^{obs} - o) f^{'}(s_j) \tag{23}$$

Dans cette formule R correspond à un R_i particulier défini par l'expression (21). Dans la pratique, on choisit le plus souvent la métrique de Mahalanobis, ce qui revient à remplacer Σ^{-1} par l'inverse de la matrice de variance-covariance de la variable aléatoire conditionnelle $\mathbf{Y}/\mathbf{x}$ (notée dans la suite $C_y^{-1}(\mathbf{x})$). Cette métrique permet de prendre en compte la dispersion des données pendant le processus de minimisation. Dans le cas où les composantes de $\mathbf{Y}/\mathbf{x}$ sont indépendantes et ont le même écart-type σ, on a alors $\mathbf{C}_y(x) = \sigma^2 I$ et R_{emp} n'est autre que la fonction de coût des moindre carrés simples. Si la loi de $\mathbf{Y}/\mathbf{x}$ dépend de $\mathbf{x}$, l'utilisation de l'erreur quadratique simple mène à des imprécisions. En effet, cette fonction de coût examine toutes les régions d'une manière identique et affecte aux différentes sorties du réseau la même pondération sans tenir compte de leur variabilité respective.

Les propriétés que nous venons de rappeler s'appliquent à toutes les méthodes qui minimisent des expressions aux moindres carré. La classification peut faire partie de ces méthodes, il suffit d'introduire un codage spécifique pour représenter les classes. Les spécificités liées aux problèmes de la classification permettent d'obtenir des propriétés supplémentaires.

Le domaine de la classification est certainement celui dans lequel les PMC sont les plus populaires et où leurs performances sont le plus reconnues. Les PMC fournissent dans de nombreux cas des classifieurs de très bonne qualité. Le but de cet article étant de traiter de la régression,

nous présentons dans ce qui suit uniquement les propriétés des classifieurs qui permettront de proposer des méthodes utiles pour résoudre des problèmes inverses. Les méthodes utilisant la classification pour la résolution de problèmes inverses sont présentée au paragraphe (6.2.2). Nous renvoyons à [43], [47], [5], [28] pour une présentation exhaustive des différentes approches de la classification à l'aide des PMC.

4 Perceptron et Classification

La classification par PMC suppose d'introduire un codage qui permet de représenter les différentes classes. Dans le cas d'un problème de classification à q classes où l'on désigne par $\mathbf{x}$ les observations ($\mathbf{x} \in \mathcal{R}^p$), la classe associée à $\mathbf{x}$ sera représentée par un vecteur $\mathbf{y} = (y_1, y_2, \ldots y_q)$ dont les composantes ne peuvent prendre que deux valeurs (ou code) a ou b. le vecteur qui code la classe k ($Classe_k$) a toutes ses composantes égales à b exceptée la $k$ième qui prend la valeur a. Ce codage permet d'utiliser comme classifieur un PMC dont la couche de sortie comporte q neurones, chaque neurone représentant une des classes. L'apprentissage s'effectue de la même manière que pour la régression en minimisant la fonction de coût des moindres carrés simples ($\Sigma = \mathbf{I}$). Les poids $\mathbf{W}$ du PMC sont estimés à partir d'un ensemble d'apprentissage $\mathcal{D}$ où chaque observation $\mathbf{x}$ est associée au vecteur $\mathbf{y}$ qui code sa classe. Si l'ensemble d'apprentissage est statistiquement représentatif de la population étudiée et si l'architecture est bien adaptée à la complexité du problème de classification sous jacent, les propriété présentées en (13) s'appliquent donc et le PMC approxime la moyenne de la variable aléatoire conditionnelle $\mathbf{Y}/\mathbf{x}$. Si l'on calcule l'espérance conditionnelle du vecteur de sortie en utilisant le codage vectoriel présentée plus haut, on obtient pour la k^{ime} composante de ce vecteur:

$$E(\mathbf{Y}/\mathbf{x})_k = a \times p(Classe_k/\mathbf{x}) + b \times \big(1 - p(Classe_k/\mathbf{x})\big) \qquad (24)$$

En fin d'apprentissage, l'état du neurone k de la couche de sortie du PMC est donc une approximation de (24). Si l'on utilise le codage particulier dans lequel $a = 1$ et $b = 0$, l'état du $k$ième neurone de sortie approxime $p(Classe_k/\mathbf{x})$ qui correspond à la probabilité a posteriori pour l'observation $\mathbf{x}$ d'appartenir à $Classe_k$. Le résultat de la section (3) nous permet de voir que la précision de cette approximation, dans une région donnée, dépend de la densité de probabilité $p(\mathbf{x})$: plus la densité est grande, plus l'approximation est précise.

La densité $p(\mathbf{x})$ est une fonction qui décrit la répartition de l'ensemble des individus dans l'espace des données, elle n'est pas accessible dans la pratique sous sa forme théorique. La détermination des paramètres $\mathbf{W}$ du PMC se fait par apprentissage en minimisant l'erreur aux moindres carrés déterminée sur la base d'apprentissage $\mathcal{D}$. L'équation (22) montre que, dans ce cas, $p(\mathbf{x})$ décrit une répartition apparente qui est celle des individus de la base d'apprentissage. La constitution de la base d'apprentissage permet d'influencer la précision de la fonction de classification que l'on cherche à obtenir. Si l'on choisit de construire la base d'apprentissage en respectant la distribution naturelle des données, on favorise la reconnaissance des classes les plus probables au détriment des classes rares. Si le but recherché est de reconnaître le mieux possible toutes les classes, le modélisateur peut influencer la précision de l'approximation en choisissant une distribution qui fait disparaître l'inégalité des répartitions des différentes classes. Une méthode simple consiste à utiliser un échantillon dans lequel chaque classe est également représentée. Si l'on ne dispose pas d'assez de données pour procéder à cet équilibre, on peut soit utiliser une fonction de coût utilisant des moindre carrés pondérés, soit avoir recours à des techniques de génération de données. Dans le premier cas, les sorties correspondantes aux classes les moins représentées sont pondérées par des coefficients plus importants. Dans le second cas, si l'on a une connaissance de la variabilité des mesures, il est possible de dupliquer les mesures existantes en ajoutant un bruit [5]. En fin d' apprentissage, le réseau classifie un individu donné $\mathbf{x}$ en lui attribuant la classe k_0 telle que:

$$k_0 = \arg\max_k(o_k) \tag{25}$$

Où o_k représente l'état de la $k$ième cellule de sortie du PMC lorsqu'on lui présente en entrée $\mathbf{x}$. Cette régle de décision qui utilise les probabilités estimées par le réseau correspond à la règle de décision de Bayes. Elle minimise, si le réseau est un bon estimateur des probabilités a posteriori, le risque d'erreur de classification [22].
Enfin, il est à noter que dans le cas de la classification, il est possible d'utiliser comme fonction de coût la distance de Kullback-Leibler plus adaptée pour l'estimation des probabilités [43].

Nous nous plaçons maintenant dans un cadre plus général, qui est le plus souvent celui de la géophysique et dans lequel les données utilisées sont entachées d'erreur. Il est nécessaire dans ce cas d'utiliser le formalisme probabiliste et de formuler d'une manière différente le problème de l'apprentissage par PMC.

5 Perceptron et Régression non linéaire

5.1 Formulation probabiliste

La régression non linéaire est une des méthodes classiques de la statistique largement utilisée dans le traitement des données. Son but principal est d'aider à la détermination d'une fonction univoque permettant de relier deux variables distinctes pour lesquels on fait l'hypothèse qu'il existe une relation de dépendance fonctionnelle. Si l'on se place dans le cadre concret de la géophysique, la première variable $\mathbf{x}$ représente un vecteur de paramètres physiques (par exemple la température, la vitesse du vent ...) et la seconde $\mathbf{y}$ contient les observations effectuées par rapport à cette variable (par exemple l'émissivité, la rugosité d'une surface..). Il est alors justifié de supposer que la relation est univoque et qu'il existe un modèle théorique idéal $\mathbf{G}$ qui permet d'inférer $\mathbf{y}$ à partir de la connaissance de $\mathbf{x}$:

$$\mathbf{y} = \mathbf{G}(\mathbf{x}) \tag{26}$$

Dans cette expression $\mathbf{G}$ représente la fonction théorique sous-jacente que l'on cherche à estimer à l'aide d'un PMC, cette fonction est appelée modèle direct . Dans la réalité, la dépendance fonctionnelle n'est pas analytiquement évidente, mais on dispose à la fois d'un certain nombre d'observations empiriques et de connaissances physiques sur la nature de la relation recherchée. Les données disponibles se présentent sous la forme de couples d'observations $(\mathbf{x}^{obs}, \mathbf{y}^{obs})$ qui proviennent d'appareils de mesure différents, ou pour les paramètres physiques ($\mathbf{x}^{obs}$ il s'agit parfois de valeurs obtenues à partir de modèles numériques. Ces données présentent donc une variabilité autour de "vraies" valeurs inconnues. La variabilité peut provenir de plusieurs causes qui sont liées à la sensibilité de l'appareil, à des variables non prises en compte dans la modélisation, ou bien encore au manque de précision des modèles numériques représentant la physique étudiée. Dans tous les cas il importe de prendre en compte ces incertitudes. Une manière classique de formaliser le problème est de faire l'hypothèse que chaque observation ($\mathbf{x}^{obs}$, $\mathbf{y}^{obs}$) se décompose en une somme de deux termes le premier étant déterministe et le second stochastique. Dans une telle approche, les différentes incertitudes relatives au phénomène physique sont prises en compte par l'intermédiaire de la partie stochastique de l'expression. Nous noterons $\mathbf{x}^{vrai}$ et $\mathbf{y}^{vrai}$ les valeurs des composantes déterministes, une donnée particulière $\mathbf{x}^{obs}$ est alors égale à :

$$\mathbf{x}^{obs} = \mathbf{x}^{vrai} + \boldsymbol{\epsilon} \tag{27}$$

De la même manière l'observation correspondante $\mathbf{y}^{obs}$ s'écrit :

$$\mathbf{y}^{obs} = \mathbf{y}^{vrai} + \boldsymbol{\eta} \tag{28}$$

dans ces expressions $\boldsymbol{\epsilon}$ et $\boldsymbol{\eta}$ représentent les parties stochastiques de ces observations.

Le but de la régression non linéaire est alors de trouver une fonction qui approche au mieux la relation déterministe :

$$\mathbf{y}^{vrai} = \mathbf{G}(\mathbf{x}^{vrai}) \tag{29}$$

Dans la suite nous cherchons à approximer cette relation à l'aide d'un PMC, c'est à dire à l'aide d'une fonction $\mathbf{F}(\mathbf{W}, \mathbf{x})$. Cette fonction permet de relier de manière univoque $\mathbf{x}$ à l'observation $\mathbf{y}$, elle représente la solution du "problème direct" qui consiste à modéliser le comportement de l'appareil de mesures. Dans cette terminologie, l'expression "problème direct" est opposé à "problème inverse". On qualifie d'inverse un problème où l'on cherche à inférer les paramètres physiques à partir des mesures observées; il s'agit d'un problème plus complexe puisque la relation recherchée peut être multivaluée. Les méthodes adaptées à la résolution de problèmes inverses font appel à des techniques spécifiques; dans le paragraphe 6 nous exposons celles utilisant les PMC. Nous ne traitons pas dans la suite de l'article du cas général pour lequel il existe des incertitudes sur les observations $\mathbf{x}^{obs}$. Ce problème peut être abordé à l'aide d'algorithmes neuronaux spécialisés utilisant des PMC qui permettent de filtrer le bruit des paramètres d'entrée et de détecter les mesures aberrantes [59] [2]. Nous supposerons donc dans la suite qu'il n'y a pas d'incertitudes sur les entrées ($\boldsymbol{\epsilon} = 0$), et que pour chaque élément de l'ensemble d'apprentissage $\mathcal{D}$ nous avons:

$$\mathbf{x}_i^{obs} = \mathbf{x}_i^{vrai} \tag{30}$$

Nous supposerons par contre que chaque observation $\mathbf{y}_i^{obs}$ est perturbée par un bruit additif $\boldsymbol{\eta}_i$ de moyenne nulle:

$$\mathbf{y}_i^{obs} = \mathbf{y}_i^{vrai} + \boldsymbol{\eta}_i \tag{31}$$

et que les differents bruits $\boldsymbol{\eta}_i$ sont indépendants. Sous ces hypothèses, quelle que soit l'observation $\mathbf{y}$ on a:

$$\mathbf{y}^{vrai} = \mathbf{E}(\mathbf{Y}/\mathbf{x}^{obs}) \tag{32}$$

Nous détaillerons la méthode permettant, dans les conditions que nous venons de présenter, d'estimer le modèle direct $\mathbf{G}$ par un réseau PMC ayant des neurones linéaires sur sa couche de sortie. Le choix d'un tel modèle se justifie par sa propriété d'approximateur universel de fonctions (paragraphe 2.1). La prise en compte durant l'apprentissage, des différents bruits de mesure peut se faire en utilisant l'approche bayesienne qui estime $\mathbf{W}$ en maximisant la probabilité $p(\mathbf{W}/\mathcal{D})$. La formule de Bayes permet d'écrire:

$$p(\mathbf{W}/\mathcal{D}) = \frac{p(\mathcal{D}/\mathbf{W})p(\mathbf{W})}{p(\mathcal{D})} \tag{33}$$

La maximisation de l'équation (33) revient à la minimisation de :

$$\begin{aligned} -2\ln\Big(p(\mathbf{W}/\mathcal{D})\Big) \quad &= \quad -2\ln\Big(p(\mathcal{D}/\mathbf{W})\Big) \\ & \qquad -2\ln p(\mathbf{W}) + \text{ constante} \end{aligned} \tag{34}$$

Dans cette expression:

- la probabilité des données $p(\mathcal{D})$ est constante par rapport à $\mathbf{W}$ et n'intervient pas dans le processus de minimisation.
- Le terme $\ln p(\mathbf{W})$ apparaît comme une contrainte supplémentaire sur la distribution des poids, il joue le rôle d'un terme de régularisation. Si l'on suppose que la distribution à priori des poids suit une loi normale de moyenne nulle et de matrice de variance-covariance $\boldsymbol{\sigma}\mathbf{I}$, $\ln p(\mathbf{W})$ apparaît comme un terme de régularisation de type "weight decay" [5]. Si la distribution à priori des poids suit une loi uniforme le terme $\ln p(\mathbf{W})$ devient une constante et peut être supprimé de la modélisation. Dans la suite, pour simplifier la présentation, nous choisissons cette hypothèse et supprimons le terme $\ln p(\mathbf{W})$.

En tenant compte de l'indépendance des mesures on obtient:

$$\ln\Big(p(\mathcal{D}/\mathbf{W})\Big) = \sum_{i=1}^{N^{obs}} \ln\Big(p\big((\mathbf{x}_i^{obs}, \mathbf{y}_i^{obs})/\mathbf{W}\big)\Big) \tag{35}$$

qui peut se décomposer:

$$\ln\Big(p(\mathcal{D}/\mathbf{W})\Big) = \sum_{i=1}^{N^{obs}} \ln\Big(p(\mathbf{y}_i^{obs}/\mathbf{x}_i^{obs},\mathbf{W})\Big) + \sum_{i=1}^{N^{obs}} \ln\Big(p(\mathbf{x}_i^{obs}/\mathbf{W})\Big) \quad (36)$$

Le premier terme de l'égalité (36) représente la probabilité que les observations $\mathbf{y}_i^{obs}$ découlent de $\mathbf{x}_i^{obs}$ quand la probabilité conditionnelle est génerée par le PMC ($\mathbf{F}(\mathbf{W},\mathbf{x})$). Par hypothèse $\mathbf{x}_i^{obs}$ est connu sans erreur et le second terme de l'équation (36) ne dépend pas de $\mathbf{W}$.

La poursuite du calcul nécessite de faire des hypothèses supplémentaires sur la nature du bruit additif qui représente la partie stochastique de l'observation $\mathbf{y}_i^{obs}$. Dans la suite de ce paragraphe, nous faisons l'hypothèse que le bruit $\boldsymbol{\eta}$ dépend de $\mathbf{x}$ et suit une loi gaussienne de moyenne nulle et de matrice de variance-covariance $\mathbf{C}_y(\mathbf{x})$, la relation (28) permet d'exprimer la densité conditionnelle du bruit sous la forme:

$$p(\eta/\mathbf{x}) = \frac{1}{(2\pi)^{\frac{q}{2}} det(\mathbf{C}_y(\mathbf{x}))^{\frac{1}{2}}} exp\Big(-\frac{1}{2}\Big(\mathbf{y}-\mathbf{y}^{vrai}\Big)^T \mathbf{C}_y^{-1}(\mathbf{x})\Big(\mathbf{y}-\mathbf{y}^{vrai}\Big)\Big) \quad (37)$$

Si l'on suppose que le PMC utilisé est acceptable, c'est à dire que les sorties du réseaux approximent "correctement" l'espérance conditionnelle:

$$\mathbf{y}^{vrai} = E(\mathbf{Y}/\mathbf{x}_i^{obs}) \approx \mathbf{F}(\mathbf{W},\mathbf{x}_i^{obs}) \quad (38)$$

On peut alors simplifier l'expression de $\ln\Big(p(\mathbf{y}_i^{obs}/\mathbf{x}_i^{obs},\mathbf{W})\Big)$ en remplaçant $\mathbf{y}^{vrai}$ par $\mathbf{F}(\mathbf{W},\mathbf{x})$, approximation calculée par le PMC .

$$-2\ln\Big(p(\mathbf{y}_i^{obs}/\mathbf{x}_i^{obs},\mathbf{W})\Big) = \Big(\mathbf{y}_i-\mathbf{F}(\mathbf{W},\mathbf{x}_i)\Big)^T \mathbf{C}_y^{-1}(\mathbf{x}_i)\Big(\mathbf{y}_i-\mathbf{F}(\mathbf{W},\mathbf{x}_i)\Big) - \ln\Big(\det\big(\mathbf{C}_y^{-1}(\mathbf{x}_i)\big)\Big) + q\ln 2\pi \quad (39)$$

L'équation (36) permet de redéfinir la fonction de coût:

$$R_{emp}(\mathbf{W}) = \sum_{i=1}^{N^{obs}} R_i(\mathbf{W}) + \text{ constante} \quad (40)$$

avec

$$R_i(\mathbf{W}) = \left(\mathbf{y}_i^{obs} - \mathbf{F}(\mathbf{W}, \mathbf{x}_i^{obs})\right)^T \mathbf{C}_y^{-1}(\mathbf{x}_i^{obs}) \left(\mathbf{y}_i^{obs} - \mathbf{F}(\mathbf{W}, \mathbf{x}_i^{obs})\right) - \ln\left(\det\left(\mathbf{C}_y^{-1}(\mathbf{x}_i^{obs})\right)\right) \quad (41)$$

Le premier terme de la partie droite de R_i est tout simplement le coût des moindres carrés généralisés (21). L'équation (41) apparaît comme une généralisation de l'équation (21), le second terme du second membre de cette équation permet de tenir compte de la variabilité du bruit en sortie. Si nous disposons d'une estimation acceptable de $\mathbf{C}_y(\mathbf{x})$ alors ce second terme est une constante par rapport à $\mathbf{W}$ et il peut être supprimé de la fonction coût. Par contre, si $\mathbf{C}_y(\mathbf{x})$ est inconnu et doit être estimé durant l'apprentissage, il est alors important de tenir compte de ce terme.

5.2 Détermination des paramètres de la loi normale Y/x

La résolution générale du problème de régression consiste à estimer, pour tout $\mathbf{x}$, la moyenne conditionnelle $\mathbf{E}(\mathbf{Y}/\mathbf{x})$ ainsi que les coefficients de la matrice de variance-covariance $\mathbf{C}_y(\mathbf{x})$. Il faut donc définir un PMC adapté à cette tâche avec une configuration de poids $\mathbf{W}$ qui minimise l'expression (40). Le réseau PMC doit avoir autant de neurones de sorties que de paramètres à estimer. Ces paramètres sont: les q composantes de $\mathbf{E}(\mathbf{Y}/\mathbf{x})$, les q variances des composantes de $\mathbf{Y}/\mathbf{x}$ ainsi que les $\frac{q(q-1)}{2}$ coefficients de corrélations.

Une méthodologie générale permettant d'estimer l'inverse de la matrice de variance covariance a été présentée dans [61], l'idée de base est d'estimer les coefficients de la décomposition de Cholesky :

$$\mathbf{C}_y^{-1}(\mathbf{x}) = \mathbf{A}^T(\mathbf{x})\mathbf{A}(\mathbf{x}) \quad (42)$$

où $\mathbf{A}(\mathbf{x}) = [a_{ij}(\mathbf{x}); i < j)]$ est une matrice triangulaire admettant des eléments positifs sur la première diagonale. Nous présentons par la suite l'algorithme d'apprentissage dans un cas simplifié où la matrice de variance-covariance est diagonale (coefficients de corrélations nuls). Le k^{me} élément de la diagonale étant $\sigma_k^2(\mathbf{x})$. Cette version simplifiée de l'algorithme a été présentée dans [40] et [5]
Pour une matrice de variance-covariance diagonale l'équation (41) devient:

$$R(\mathbf{W}) = \sum_{k=1}^{q} \frac{\left(\mathbf{y}_k - \mathbf{F}_k(\mathbf{W}, \mathbf{x})\right)^2}{\sigma_k^2(\mathbf{x})} + \sum_{k=1}^{q} \ln\left(\sigma_k^2(\mathbf{x})\right) + constante \qquad (43)$$

où pour simplifier l'écriture de l'équation nous avons omis d'écrire l'indice i caractérisant un exemple particulier de la base d'apprentissage pour une observation $\mathbf{x}^{obs}$ quelconque.
L'architecture du réseau possède deux types différents de neurones de sortie qui seront notés M et D, les états de ces neurones représentent les différentes valeurs à estimer:

- Les sorties de type M estiment les q valeurs moyennes distinctes $\mathbf{E}(\mathbf{Y}/\mathbf{x}_i^{obs})$, dans ce cas on utilise des neurones linéaires.
- Les sorties de type D estiment les q variances $\sigma_k^2(\mathbf{x})$; comme ces valeurs doivent être positives, les neurones du groupe D utilisent des fonctions de transfert exponentielles.

Le nombre de neurones de la couche d'entrée du réseau est égal à la dimension du vecteur $\mathbf{x}$. Le nombre de couches et de cellules cachées dépendra de la complexité du problème à résoudre. La couche de sortie admet $2q$ cellules dont les états constituent deux ensembles différents:

$$O^M = \{o_k^M(\mathbf{W}, \mathbf{x}^{obs}) \ , \ k = 1 \cdots q\}$$
$$O^D = \{ \ o_k^D(\mathbf{W}, \mathbf{x}^{obs}) \ , \ k = 1 \cdots q\}$$

Le vecteur O^M correspond à $\mathbf{F}_k(\mathbf{W}, \mathbf{x})$ de l'équation (43). Afin de noter les entrées des cellules de sorties nous utiliserons des notations similaires en remplaçant o par s et O par S pour les deux groupes de cellules $S = \{S^M, S^D\}$:

- les neurones de type M ont des fonctions de transferts linéaires on a donc $o_k^M = s_k^M$.
- les neurones de type D ont des fonctions de transfert exponentielles on a donc : $o_k^D = e^{s_k^D}$.

Afin de simplifier les formules utilisées pendant l'apprentissage nous introduisons quelques simplifications et quelques notations intermédiaires. Nous omettrons l'indice i, $\mathbf{x}^{obs}$ représentera donc une observation particulière et R la fonction coût partielle R_i qui lui est associée. Pour chacune des observations $\mathbf{x}^{obs}$ nous noterons $\mathbf{\Delta} = [\Delta_k^M] = (\mathbf{y}^{obs} - \mathbf{O}^M)$, $k = 1 \cdots q$ l'erreur attachée à cette observation.
La fonction de coût qui représente l'erreur attachée à cette observation s'écrit:

$$R(\mathbf{W}) = \sum_{k=1}^{q} \frac{(\Delta_k^M)^2}{e^{s_k^D}} + \sum_{k=1}^{q} s_k^D \tag{44}$$

L'apprentissage consiste à déterminer les poids $\mathbf{W}$ qui minimisent (44), l'initialisation de l'algorithme de rétro-propagation du gradient demande le calcul de $\frac{\partial R}{\partial \mathbf{S}}$ pour chaque neurone de sortie :

$$\text{pour } k \in \text{ M} : \quad \frac{\partial R}{\partial s_k^M} = -2\frac{\Delta_k^M}{e^{s_k^D}} \tag{45}$$

$$\text{pour } k \in \text{ D} : \quad \frac{\partial R}{\partial s_k^D} = -\frac{(\Delta_k^M)^2}{e^{s_k^D}} + 1 \tag{46}$$

5.2.1 L'algorithme

Nous présentons dans ce paragraphe l'algorithme permettant de minimiser la fonction (44). Cet algorithme comporte trois phases. Les deux premières phases (1 et 2) sont exécutées à la suite d'une manière itérative, une phase est itérée un certain nombre de fois avant de passer à la suivante (phase 1 itérée, phase 2 itérée,..). La phase 3 est exécutée à la suite de cette initialisation pour raffiner la solution obtenue:

- **Phase 1** : Le but de cette phase est de proposer une bonne estimation de la valeur moyenne $E(\mathbf{Y}/\mathbf{x})$ avant de passer aux deux phases 2 et 3, cette phase ne concerne que les sorties de type M.

  Au cours de la phase 1, on suppose que les q écart-types $\sigma_k(\mathbf{x})$ ne dépendent pas des poids $\mathbf{W}$ et restent constants. Seul le premier terme de l'équation (44) est minimisé. La minimisation porte alors sur le coût lié à l'erreur quadratique généralisée. L'algorithme de rétropropagation du gradient s'applique uniquement aux neurones qui permettent d'estimer les q moyennes (neurones de type M). Leurs gradients de sortie sont initialisés par l'équation (45) au début de chaque itération.

  Lors du premier passage de la phase 1 nous supposons que les q écart-types σ_i sont constants et égaux à 1. Dans ce cas R_{emp} se réduit à l'erreur quadratique simple. Au cours des itérations qui suivent les écart-types supposés constants ont pour valeur celles estimées durant la phase 2.

  Dans cette phase, l'apprentissage calcule une approximation de $\mathbf{E}(\mathbf{Y}/\mathbf{x}^{obs})$, la précision obtenue peut être analysée en utilisant les

résultats du paragraphe (5.1). En fin de phase on obtient les sorties des neurones de type M:

$$\big(\mathbf{F}_1(\mathbf{W}^*,\mathbf{x}),\ldots,\mathbf{F}_q(\mathbf{W}^*,\mathbf{x})\big) = (o_1^M, o_2^M, \ldots, o_q^M) \qquad (47)$$

- **Phase 2**: Le but de cette phase est de donner une première estimation des écart-types $\sigma_k(\mathbf{x})$. Les q états des neurones de sortie du type M sont figés, ils conservent les valeurs $\mathbf{F}_j(\mathbf{W}^*, \mathbf{x}^{obs})$ calculées précédemment. On ne prend en compte que les erreurs qui apparaissent sur les sorties de type D, la minimisation ne porte que sur le second terme de $\mathbf{R}(\mathbf{W})$ (44). Les q neurones de sorties de type D estiment les variances:

$$\big(\mathbf{F}_{q+1}(\mathbf{W}^*,\mathbf{x}),\ldots,\mathbf{F}_{2q}(\mathbf{W}^*,\mathbf{x})\big) = (o_{q+1}^D, o_{q+2}^D, \ldots, o_{2q}^D) \qquad (48)$$

  les gradients de l'algorithme de rétropropagation sont initialisés au début de chaque itération par l'équation (46). Au cours de la première itération de cette phase on utilise la matrice $\mathbf{W}^*$ calculée en fin de phase précédente.

- **Phase 3** : Durant cette phase, la fonction de coût (44) est minimisée dans sa globalité. On suppose que les $2q$ sorties du réseau sont variables, chaque itération de la rétropropagation du gradient est initialisée par l'équation (45) pour les sorties de type M et par l'équation (46) pour les sorties de type D.

6 Perceptron et Modèle Inverse

6.1 Position du problème

Nous introduisons dans ce paragraphe des formalismes adaptés à la résolution de problèmes inverses à l'aide de PMC. Sous sa forme générale le problème se formule de la manière suivante: connaissant les observations $\mathbf{y}$, quelles sont les valeurs des paramètres, représentés par $\mathbf{x}$, qui les ont inférées. On recherche donc maintenant une fonction $\mathbf{F}$ telle que:

$$\begin{array}{rcl} \mathcal{R}^q & \longrightarrow & \mathcal{R}^p \\ \mathbf{y} & \longmapsto & \mathbf{x} = \mathbf{F}^{-1}(\mathbf{W},\mathbf{y}) \end{array} \qquad (49)$$

En géophysique, on cherche à inférer des paramètres physiques à partir d'observations fortement bruitées issues de capteurs, il est donc impossible de ne pas tenir compte des incertitudes. Nous introduisons de la manière la plus complète possible le traitement de ces incertitudes,

nous abordons en particulier un problème fondamental en géophysique qui est celui de la cohérence des champs de mesures retrouvées.
D'une manière générale, si la relation qui lie les observations $\mathbf{y}$ aux paramètres physiques est univoque, le problème inverse est un problème de régression. Toutes les méthodes présentées dans les premiers paragraphes peuvent être utilisées et vont permettre une résolution acceptable du problème (5,49). L'inversion consiste alors à associer à chaque mesure observée la moyenne de $E(\mathbf{X}/\mathbf{y})$, quand la correspondance est bien univoque cette méthode donne de bons résultats [24] [25] [26] [51] [49]. Cependant la relation inverse recherchée n'est pas toujours univoque, il s'agit souvent d'un problème mal posé: la solution recherchée peut ne pas être unique. L'ambiguïté du problème inverse peut être intrinsèque, elle peut également provenir d'une analyse incomplète du problème. En effet, la présence de variables non prises en compte par la modélisation peut mener à des valeurs d'observations identiques $\mathbf{y}$ en réponse à des valeurs de paramètres $\mathbf{x}$ distinctes. D'autre part, le fait que les données soient entachées d'erreurs de mesure peut accentuer l'apparition d'ambiguïtés.

D'une manière générale, une simple régression aux moindres carrés qui conduit à l'estimation de la moyenne $E(\mathbf{X}/\mathbf{y})$ peut se révéler être une solution inacceptable si l'ensemble des valeurs possibles de $\mathbf{x}$ qui peuvent inférer une observation $\mathbf{y}$ ne constitue pas un ensemble convexe. La moyenne $E(\mathbf{X}/\mathbf{y})$ calculée par le PMC, peut se trouver à l'extérieur de l'ensemble des solutions acceptables et peut ne correspondre à aucune solution physique plausible [31].

La résolution d'un problème inverse s'effectue donc de manière différente selon la nature du problème considéré multivalué ou non. Nous détaillons maintenant les approches possibles de résolution d'un problème inverse par PMC dans le cas d'un problème multivalué. Dans ce cas aussi deux approches sont à considérer selon que l'on cherche une résolution exhaustive ou non.

- Si l'on veut déterminer l'ensemble des solutions avec si possible leur probabilité d'apparition (voir 5.2), il faut alors calculer la loi de la variable aléatoire $\mathbf{X}/\mathbf{y}$ ($\mathbf{X}$ conditionné par l'observation $\mathbf{y}$).
- Dans le cas où l'on recherche, parmi toutes les valeurs ambiguës possibles, une solution particulière possédant certaines particularités physiques, des méthodes moins lourdes peuvent être utilisées. On cherche alors à proposer une des solutions qui a pu inférer l'observation $\mathbf{y}$.

Nous exposons maintenant les méthodologies utilisant des PMC qui permettent d'aborder les problèmes que nous venons d'évoquer. Une

présentation probabiliste du problème de l'inversion permettra de discuter d'une manière générale les différentes résolutions possibles. Le formalisme probabiliste permet, comme dans les paragraphes précédents, d'intégrer facilement les modèles PMC dans les différentes résolutions possibles.

6.2 Détermination de la loi conditionnelle de X/y

6.2.1 Modèle de mélanges

Dans le paragraphe (5) nous avons présenté une méthode permettant d'estimer la fonction densité de la variable aléatoire conditionnelle $\mathbf{X/y}$, cette méthode n'est pas générale puisqu'elle fait l'hypothèse que celle-ci suit une loi normale. Souvent la variable aléatoire conditionnelle n'est pas unimodale et l'hypothèse gaussienne n'est pas adaptée, il faut alors avoir recours à une approche plus générale. L'estimation de la fonction densité peut se faire en modélisant celle-ci par une mixture de $\mathbf{K}$ fonctions normales [4]. On pose dans ce cas :

$$p(\mathbf{x/y}) = \sum_{k=1}^{K} \alpha_k(\mathbf{y}) \mathbf{f}_k(\mathbf{x/y}) \tag{50}$$

où $\mathbf{f}_k$ est une fonction densité normale de moyenne $\mu_k(\mathbf{y})$ et d'écart-type $\sigma_k(\mathbf{y})$, dans cette expression les coefficients du mélange $\alpha_k(\mathbf{y})$ vérifient la relation : $\sum_{k=1}^{K} \alpha_k(\mathbf{y}) = 1$. L'utilisation d'un simple écart-type $(\sigma_k(\mathbf{y}))$ pour caractériser $\mathbf{f}_k$ introduit implicitement que l'on considère des distributions sphériques, donc des matrices de variance-covariance de la forme $\sigma_k I$ où I est la matrice identité de dimension p. Une propriété importante des mélanges de Gaussienne est qu'ils permettent d'approximer toute fonction densité continue avec la précision désirée, à la condition que le mélange contienne un nombre suffisant de gaussiennes et que les paramètres de ces gaussiennes soient bien estimés. Le nombre K de gaussiennes utilisées dépend de la complexité du problème à résoudre. La méthode utilisée pour la détermination des moyennes et des écart-types de ces gaussiennes à l'aide de PMC est présentée dans la suite du paragraphe.

Effectuer l'inversion revient donc à estimer la fonction densité de la variable aléatoire conditionnelle $\mathbf{X/y}$ (sous la forme présentée par l'expression 50), soit à déterminer les $(p+2)K$ quantités : $\mu_k(\mathbf{y})$, $\alpha_k(\mathbf{y})$ et $\sigma_k(\mathbf{y})$. Cette estimation peut se faire par un PMC dont les $(p+2)K$ sorties estiment chacune l'une des quantités recherchées, la couche

d'entrée du réseau étant égale à la dimension de $\mathbf{y}$. Comme pour les PMC classiques, le nombre de couches et de neurones cachés dépend de la difficulté du problème traité. On distingue pour ce PMC trois types de neurones sur la couche de sortie:

- Le groupe M estime les moyennes $\mu_k(\mathbf{y})$.
- Le groupe D estime les écart types $\sigma_k(\mathbf{y})$.
- le groupe P estime les pondérations $\alpha_k(\mathbf{y})$.

Par la suite, comme au paragraphe (5.2) les états des neurones de sorties seront notés par l'ensemble $O = \{O^M, O^D, O^P\}$ qui est défini de la manière suivante:

$$O^M = \{o_j^M(\mathbf{W}, \mathbf{y}^{obs}) \ , \ j = 1 \cdots Kp\}$$
$$O^D = \{ o_j^D(\mathbf{W}, \mathbf{y}^{obs}) \ , \ j = 1 \cdots K\}$$
$$O^P = \{ o_j^P(\mathbf{W}, \mathbf{y}^{obs}) \ , \ j = 1 \cdots K\}$$

Pour les entrées de ces neurones on utilisera des notations similaires en remplaçant o par s et O par S pour les trois groupes de cellules

$$S = \{S^M, S^D, S^P\} \tag{51}$$

Les neurones de type M ont des fonctions de transferts linéaires ($o_j^M = s_j^M$) et les neurones de type D et P ont des fonctions de transfert exponentielles ($o_j^D = e^{s_k^D}$ et $o_j^P = e^{s_k^P}$). D'autre part, les pondérations α_j représentent une répartition de probabilités, elles seront calculées à partir de o_j^P par des fonctions dites "softmax":

$$\alpha_j(\mathbf{y}) = \frac{o_j^P(\mathbf{y})}{\sum_{k=1}^{\mathbf{K}} o_k^P(\mathbf{y})} \tag{52}$$

L'estimation de ces différents paramètres se fait par apprentissage des poids du PMC en maximisant la vraisemblance, ou en minimisant:

$$R(\mathbf{W}) = -\ln p(\mathcal{D}/\mathbf{W}) \tag{53}$$

En tenant compte de l'équation (50), nous obtenons :

$$R(\mathbf{W}) = -\sum_{i=1}^{\mathbf{N}^{obs}} \ln \Big(\sum_{k=1}^{K} \alpha_k(\mathbf{y}_i^{obs}) \mathbf{f}_k(\mathbf{x}_i^{obs}/\mathbf{y}_i^{obs}) \Big) \tag{54}$$

avec

$$\mathbf{f}_k(\mathbf{x}/\mathbf{y}) = \frac{1}{(2\pi)^{\frac{p}{2}}\sigma_k(\mathbf{y})^p} exp\Big(- \frac{\left\|\mathbf{x} - \mu_k^2(\mathbf{y})\right\|^2}{2\sigma_k^2(\mathbf{y})}\Big) \qquad (55)$$

On peut utiliser, pour minimiser $R(\mathbf{W})$ la rétropropagation du gradient initialisée à l'aide des dérivées partielles par rapport aux entrées des neurones de sortie du PMC: $\frac{\partial R}{\partial S^M}$, $\frac{\partial R}{\partial S^P}$ et $\frac{\partial R}{\partial S^D}$. L'ensemble des formules de dérivation est présentée dans [5]. Des procédures en mathlab sont disponibles sur le cite web de l'Université de Aston à l'adresse: http://www.ncrg.aston.ac.uk. La minimisation du log de vraisemblance peut également être obtenue en utilisant l'algorithme E-M (Estimation-Maximisation) qui est plus adapté au traitement des variables cachées. Une description complète de cet algorithme et de sa mise en oeuvre est développée dans [58] [21].

En fin d'apprentissage, si la fonction densité de la variable aléatoire $\mathbf{Y}/\mathbf{x}$ est bien approximée, l'étude de la fonction $\mathbf{X}/\mathbf{y}$ permet de déterminer les valeurs les plus probables de $\mathbf{x}$ qui correspondent à ses maxima locaux. À une observation $\mathbf{y}$ on peut donc associer différentes valeurs possibles des paramètres géophysiques $\mathbf{x}$ avec leur probabilité d'apparition. La valeur moyenne $E(\mathbf{X}/\mathbf{y})$ est égale à $\sum_{k=1}^{K} \alpha_k(\mathbf{y})\mu_k(\mathbf{y})$ qui est une somme pondérée des différents $\mu_k(\mathbf{y})$.

Dans le domaine des réseaux de neurones, les PMC approximateurs de fonction densité sont également connus sous le nom de réseaux multi-expert [30]. On considère alors que le PMC représente un réseau de K experts dans lequel le kième expert estime la moyenne $\mu_k(\mathbf{y})$. Un superviseur, représenté par les sorties des neurones de type P, permet d'attribuer une importance relative aux sorties des réseaux experts. Les PMC multi-expert ont été proposés pour modéliser des problèmes de régression ou de prévision de séries temporelles lorsque celles-ci changent de régime ou de comportement dans les différentes régions de l'espace des entrées [58] [38]. Dans les applications traitées chaque expert se spécialise automatiquement dans une région de l'espace et peut être étudié séparément par la suite. Ces mêmes réseaux multi experts sont également utilisés dans le cas de la classification en échangeant les densités gaussiennes par des distributions de Bernouilli [28].

La principale difficulté de la mise en oeuvre des réseaux approximateurs de densité provient du nombre et de la complexité des calculs à effectuer pour effectuer la minimisation. Nous présentons dans le paragraphe suivant une méthode d'approximation de fonctions densité alternative plus robuste, mais dont l'utilisation est restreinte à des variables dont la dimension n'est pas trop élevée.

6.2.2 Approximation par histogramme

L'approximation de la fonction densité de la variable aléatoire $E(\mathbf{X}/\mathbf{y})$ peut se faire à l'aide d'un PMC utilisé dans son mode classifieur (§ 4). Cette approche basée sur une discrétisation des valeurs de la variable x est générale, cependant sa mise en oeuvre qui nécessite un ensemble de données qui croit très rapidement avec la dimension de x ne peut être envisagée que si celle-ci est petite. Nous détaillons la méthodologie pour $p = 1$, la généralisation à des dimensions supérieure est directe. L'idée générale de la méthode consiste à transformer le problème d'inversion en un problème de classification. L'ensemble des valeurs possibles de la variable $\mathbf{x}$ est discrétisée en N intervalles ($p = 1$), on associe alors à chaque paramètre physique $\mathbf{x}$ l'intervalle auquel il appartient. Par la suite, chaque intervalle est considéré comme une classe à laquelle on associe son indicatrice. Un intervalle I_j correspondant à la classe j sera codé par le vecteur $\mathbf{d}$ de dimension N dont la j^{me} composante prend la valeur 1 et dont les autres composantes sont nulles.

On utilise pour l'approximation de la fonction densité un PMC qui possède q neurones sur la couche d'entrée et N sur celle de sortie, ces derniers neurones correspondent aux N intervalles de la discrétisation. Le PMC réalise maintenant une fonction de $\mathcal{R}^q \longrightarrow \mathcal{R}^N$.

La détermination des poids du réseau s'effectue en prenant comme ensemble d'apprentissage $\mathcal{D}' = \left\{(\mathbf{y}_i^{obs}, \mathbf{d}_i), i = 1 \cdots N^{obs}\right\}$ où $\mathbf{d}_i$ est obtenu en transformant les paramètres physiques $\mathbf{x}_i^{obs}$ par l'intermédiaire du codage en classe d'intervalles. L'utilisation du réseau sous forme de classifieur permet d'approximer les probabilités d'appartenance a posteriori aux différentes classes qui sont ici les N intervalles I_j (voir paragraphe 4). le PMC calcule, pour chaque $\mathbf{y}$, un vecteur de $\mathcal{R}^N$ constitué par les états des N neurones de la couche de sortie. L'état du neurone j de la couche de sortie approxime la probabilité conditionnelle d'appartenance à la classe I_j. Il s'agit de la probabilité que $\mathbf{x}$, inverse de $\mathbf{y}$, soit dans l'intervalle considéré I_j. L'ensemble des N valeurs proposées en sortie du réseau représente donc une approximation de l'histogramme de la variable aléatoire $\mathbf{X}/\mathbf{y}$, cette approximation est effectuée sans faire aucune hypothèse paramétrique sur la distribution. L'interprétation de l'histogramme permet d'obtenir les différentes solutions du problème inverse. Si on représente chaque intervalle I_j par son centre x_j, la détermination par interpolation des différents pics de la courbe dont les abscisses sont les valeurs x_j et les ordonnées les états des neurones de sortie associés permet de calculer les valeurs les plus probables de $\mathbf{x}$ et les probabilités qui leurs sont associées. L'ensemble des différents pics rangés en ordre décroissant de probabilité constitue, pour une observation $\mathbf{y}$ don-

née, un ensemble de solutions au problème inverse. Des applications à l'inversion des données diffusiométriques ont montré la robustesse de l'approche [35] [34] [45].

Le fait de proposer plusieurs solutions possibles avec des probabilités associées pose le problème du levé d'ambiguïtés. Une inversion est rarement recherchée pour une mesure isolée et seule l'inversion d'un ensemble de mesures dont les paramètres physiques doivent présenter une cohérence globale peut permettre d'accéder à la solution recherchée. Comme dans le cas de la régression, nous proposons au paragraphe suivant une approche probabiliste de l'inversion qui va permettre d'introduire cette notion de cohérence de champs des paramètres retrouvés. L'utilisation de ce formalisme, le plus souvent commun aux méthodes d'inversion classiques [52], permet d'aborder d'une manière méthodologique le problème de l'inversion. Nous introduisons par la suite les méthodologies les plus connues et les plus utilisées par les spécialistes des réseaux de neurones,

6.3 Une formulation probabiliste de l'inversion

Une formulation complète du problème inverse sous sa forme la plus générale recherche à inverser globalement un ensemble de mesures. L'idée sous jacente est que les différents inverses recherchés sont corrélés et qu'une inversion globale permet de prendre en compte un nombre beaucoup plus important d'informations. Nous introduisons donc deux ensembles $\mathcal{Y}$ et $\mathcal{X}$ qui représentent les deux ensembles qui nous intéressent: $\mathcal{Y} = \{\mathbf{y}^i, i = 1 \ldots N^{mes}\}$ représente l'ensemble des mesures observées et $\mathcal{X} = \{\mathbf{x}^i, i = 1 \ldots N^{mes}\}$ l'ensemble des paramètres physiques associés que l'on cherche à déterminer. Le problème est donc de déduire les valeurs de $\mathcal{X}$ connaissant son champ d'observations $\mathcal{Y}$. Comme tout problème inverse il peut se ramener à déterminer les valeurs de $\mathcal{X}$ qui maximisent la vraisemblance (ou qui minimisent l'opposé de son logarithme):

$$p(\mathcal{X}/\mathcal{Y}) = \frac{p(\mathcal{Y}/\mathcal{X})p(\mathcal{X})}{p(\mathcal{Y})} \tag{56}$$

Pour un ensemble d'observations $\mathcal{Y}$ donné, la probabilité p($\mathcal{Y}$) est constante, on peut alors écrire :

$$p(\mathcal{X}/\mathcal{Y}) \propto p(\mathcal{Y}/\mathcal{X})p(\mathcal{X}) \tag{57}$$

Dans cette équation $p(\mathcal{Y}/\mathcal{X})$ représente la vraisemblance des mesures observées et $p(\mathcal{X})$ représente la probabilité à priori de l'ensemble des paramètres physiques $\mathcal{X}$ associés. La probabilité $p(\mathcal{X})$ représente la connaissance à priori que nous avons sur la solution recherchée. La présence de ce terme dans la fonction coût force la recherche vers une solution physiquement plausible. En général, nous pouvons supposer qu'une mesure observée particulière $\mathbf{y}^i$ ne dépend que du paramètre physique $\mathbf{x}^i$ en ce point de mesure. Sous cette condition de dépendance locale, l'équation (57) devient :

$$p(\mathcal{X}/\mathcal{Y}) \propto \prod_i p(\mathbf{y}^i/\mathbf{x}^i) p(\mathcal{X}) \tag{58}$$

L'indice i parcourt l'ensemble des observations à inverser. L'équation (58) permet de retrouver l'ensemble des paramètres $\mathcal{X}$ en utilisant les probabilités conditionnelles $p(\mathbf{y}/\mathbf{x})$. Cette probabilité correspond à la densité de répartition des mesures pour un paramètre physique donné, elle est connue sous le nom de relation directe locale. Son utilisation dans la formulation probabiliste (58), en appliquant à nouveau la formule de Bayes permet d'obtenir une seconde formulation de l'inversion:

$$p(\mathbf{y}^i/\mathbf{x}^i) = \frac{p(\mathbf{x}^i/\mathbf{y}^i) p(\mathbf{y}^i)}{p(\mathbf{x}^i)} \tag{59}$$

L'équation (58) devient:

$$p(\mathcal{X}/\mathcal{Y}) \propto \prod_i \frac{p(\mathbf{x}^i/\mathbf{y}^i) p(\mathbf{y}^i)}{p(\mathbf{x}^i)} p(\mathcal{X}) \tag{60}$$

L'équation (60) permet de retrouver le champ des paramètres $\mathcal{X}$ en utilisant les probabilités conditionnelles $p(\mathbf{x}/\mathbf{y})$. Cette probabilité correspond à la densité de répartition des différentes valeurs possibles des paramètres physiques connaissant l'observation, elle est connue sous le nom de relation inverse locale. Les équations (58) et (60) permettent de proposer deux méthodes d'inversion: une inversion par modèle direct local et une inversion par modèle inverse local.

Cette modélisation d'un problème inverse en utilisant cette double formulation probabiliste a été introduite par Cornford et all pour résoudre la difficile inversion des signaux diffusiométriques [16]. Nous présentons cette application plus en détails dans le paragraphe 7 et donnons toutes les références qui permettent de comprendre comment

un même problème peut être abordé de manières différentes. Cette approche reste toujours valable et peut être utilisée d'une manière générale pour la résolution de problèmes inverses. Nous présentons en détails dans ce qui suit les différentes méthodes à mettre en oeuvre dans les modélisations concernées. Afin de faciliter, en présence d'un problème inverse particulier, le choix de la méthode à utiliser, nous discutons des différentes hypothèses sous-jacentes qui conditionnent l'efficacité des différentes méthodes possibles.

6.3.1 inversion par modèle direct local

Cette méthode consiste à maximiser la vraisemblance et pour cela à utiliser son expression proposée à l'équation (58). Pour pouvoir effectuer la maximisation il est nécessaire de donner une forme analytique aux quantités manipulées et donc de faire des hypothèses sur les distributions que l'on utilise. On peut par exemple admettre que la densité $p(\mathbf{y}/\mathbf{x})$ qui apparaît dans cette formule est unimodale. On peut également supposer que le phénomène étudié se décompose sous la forme d'une somme d'un phénomène déterministe auquel s'ajoute un bruit (28). Il faut noter que ces deux hypothèses sont le plus souvent acceptables dans le cas d'un problème direct, par exemple pour un capteur donné on peut supposer qu'à une mesure déterministe s'ajoutent différents bruits qui suivent dans leur ensemble une loi normale.

Sous de telles conditions, l'utilisation de l'ensemble d'apprentissage $\mathcal{D} = \left\{(\mathbf{x}_i^{obs}, \mathbf{y}_i^{obs}), i = 1 \cdots N^{obs}\right\}$ permet d'estimer la moyenne et les coefficients de l'inverse de la matrice de variance-covariance de $p(\mathbf{y}/\mathbf{x})$ à l'aide d'un PMC (voir 5.2). Si les poids $\mathbf{W}$ du PMC qui résultent de l'apprentissage sont figés, le PMC détermine une fonction qui estime la moyenne $F(\mathbf{x}) \simeq E(\mathbf{Y}/\mathbf{x})$ et l'inverse de la matrice de variance-covariance $\mathbf{C}_x^{-1}$. Maximiser l'équation (58) par rapport aux paramètres $\mathcal{X}$ revient à minimiser:

$$\begin{aligned} -2\ln p(\mathcal{X}/\mathcal{Y}) &= \sum_i \left(\mathbf{y}^i - F(\mathbf{x}^i)\right)^T \mathbf{C}_{x^i}^{-1} \left(\mathbf{y}^i - F(\mathbf{x}^i)\right) \\ &\quad -2\ln p(\mathcal{X}) - \sum_i \ln \mid \det \mathbf{C}_{x^i}^{-1} \mid + \textit{constante} \quad (61) \end{aligned}$$

Dans cette expression $F(\mathbf{x}^i)$ et $\mathbf{C}_x^{-1}$ représentent les états des neurones de sortie du PMC, ils dépendent de $\mathbf{x}$ et interviennent dans la minimisation de (61). Cependant pour simplifier les calculs, on fait l'hypothèse supplémentaire que les valeurs de variance et de covariance de $\mathbf{C}_x^{-1}$ varient très lentement en fonction de $\mathbf{x}$. Cette hypothèse permet de négliger

les dérivées de $\mathbf{C}_x^{-1}$ dans le calcul du gradient de la fonction coût (61) par rapport à $\mathbf{x}$. Ce gradient ne dépend alors que des dérivées $\frac{\partial F(\mathbf{x})}{\partial \mathbf{x}}$ et $\frac{\partial Lnp(\mathcal{X})}{\partial \mathbf{x}}$. Etant donné que la fonction F est un PMC, il est facile de calculer par rétropropagation ses dérivées $\frac{\partial F(\mathbf{x})}{\partial \mathbf{x}}$ par rapport aux variables d'entrée (paragraphe 2.2). Par contre $\frac{\partial Lnp(\mathcal{X})}{\partial \mathbf{x}}$ dépend de l'expression de la probabilité a priori $p(\mathcal{X})$. Le premier terme du second membre de l'équation (61) mesure l'écart entre les mesures $\mathbf{y}^i$ et les sorties du modèle direct $F(\mathbf{x}^i)$ relativement aux paramètres physiques $\mathbf{x}^i$. Si nous admettons que le modèle direct décrit bien la physique de la mesure, alors la minimisation de ce terme permet de déterminer les paramètres physiques $\mathbf{x}^i$ afin qu'ils reproduisent "au mieux" les mesures $\mathbf{y}^i$. La présence du terme $\ln p(\mathcal{X})$ dans l'expression à minimiser permet de maintenir une certaine cohérence à l'ensemble des paramètres physiques $\mathcal{X}$ qu'il faut retrouver. Citons par exemple le cas des mesures satellitaires pour lesquelles les valeurs des paramètres physiques retrouvés possèdent une cohérence spatiale. La minimisation de (61) par méthode du gradient nécessite la détermination d'une solution initiale. Le choix d'une solution initiale $\mathcal{X}_0$ contraint fortement la solution trouvée, une "bonne" solution de départ permet de trouver un "bon" ensemble de solutions $\mathcal{X}$ au problème inverse. Le choix de $\mathcal{X}_0$ peut être fait en faisant intervenir des connaissances extérieures, nous discuterons ce problème au paragraphe (6.3.3).

6.3.2 Inversion par modèle inverse local

Cette deuxième méthode consiste à maximiser la vraisemblance à partir de l'équation (60) en utilisant la relation inverse locale. Dans cette équation le terme $p(\mathbf{y}^i)$ est constant puisqu'il correspond à la mesure observée en un point particulier. Le problème revient donc à minimiser l'expression:

$$-\ln p(\mathcal{X}/\mathcal{Y}) = -\sum_i \ln p(\mathbf{x}^i/\mathbf{y}^i) + \sum_i \ln p(\mathbf{x}^i) - \ln p(\mathcal{X}) + constante \tag{62}$$

Comme dans l'inversion par modèle direct local, la maximisation de l'équation demande de faire des hypothèses sur les différentes distributions qui apparaissent. Ces distributions sont d'une part $p(\mathbf{x}^i)$ qui représente la densité de probabilité des paramètres physiques au point d'observation i, et d'autre part $p(\mathbf{x}/\mathbf{y})$ qui représente celle des paramètres physiques connaissant les observations. Dans les situations où l'on peut

supposer que les lois considérées suivent une loi normale, la méthodologie utilisée pour l'inversion par modèle direct local peut être appliquée. On remplace $p(\mathbf{x}^i)$ par son expression analytique. Un PMC dont les poids sont estimés à partir de

$$\mathcal{D}' = \left\{(\mathbf{y}_i^{obs}, \mathbf{x}_i^{obs}), i = 1 \cdots N^{obs}\right\} \tag{63}$$

permet d'effectuer la régression entre les variables explicatives $\mathbf{y}$ et les variables expliquées $\mathbf{x}$, ce PMC estime la moyenne $F^{-1}(\mathbf{y}) \simeq E(\mathbf{X}/\mathbf{y})$ et l'inverse de la matrice de variance-covariance $\mathbf{C}_y^{-1}$ de la distribution $p(\mathbf{x}/\mathbf{y})$.

Dans le cas où aucune hypothèse ne peut être faite sur la distribution $p(\mathbf{x}/\mathbf{y})$, celle ci peut être modélisée par un PMC en utilisant la méthode de mixture de gaussiennes présentée dans le sous-paragraphe (6.2.1). La relation liant $\mathbf{y}$ à $\mathbf{x}$ étant souvent multivoque ce second cas arrive beaucoup plus fréquemment que si l'on utilise l'inversion par inverse direct local qui fait apparaître la distribution $p(\mathbf{y}/\mathbf{x})$.

La minimisation de (62) peut se faire en utilisant une méthode de gradient. Dans ce cas, le gradient de cette dernière fonction de coût par rapport à $\mathbf{x}$ se calcule d'une manière simple puisque les différents moments et paramètres de $p(\mathbf{x}/\mathbf{y})$ qui sont estimés par le réseau dépendent de $\mathbf{y}$ et non de $\mathbf{x}$. Comme il s'agit de méthode de gradient et étant donné que l'on peut avoir à utiliser des distributions multimodales, la solution de départ choisie pour initialiser la méthode de gradient va jouer un rôle primordial. Nous discutons dans la suite de ce paragraphe les différentes solutions possibles en fonction des hypothèses supplémentaires que l'on peut faire concernant le problème physique.

D'autre part, il importe de noter que l'estimation de la fonction de densité conditionnelle se fait par apprentissage à partir de données qui peuvent pour certaines d'entre elles provenir de modèles numériques. C'est le cas en télédétection, domaine dans lequel on effectue des collocations entre les mesures satellitaires et les sorties des modèles numériques de prévision. Les prévisions des modèles numériques correspondent à des valeurs moyennes qui filtrent les variabilités réelles locales. Avec un tel ensemble de données, la solution calculée par la méthode d'inversion par modèle inverse local ne peut proposer qu'une solution lisse présentant des caractéristiques identiques au modèle de prévision numérique utilisé. L'utilisation de l'inversion à l'aide du modèle direct local (paragraphe 6.3.1) peut permettre une recherche plus fine de la solution. D'une certaine manière, la solution proposée à l'aide du modèle inverse local ($\mathcal{X}_0$) peut servir d'initialisation à une minimisation dans l'espace des observations $\mathbf{y}$, ce qui permet de reconstituer la variabilité locale.

6.3.3 Choix de $p(\mathcal{X})$ et de la solution initiale

Comme nous l'avons mentionné plus haut $p(\mathcal{X})$ représente la probabilité à priori de l'ensemble des paramètres physiques $\mathcal{X}$ aux points de mesures. Le choix du champs initial de ces paramètres va dépendre des différentes hypothèses qu'il est possible de faire sur $p(\mathcal{X})$. Si l'on suppose que le champs est formé de paramètres indépendants ($p(\mathcal{X}) = \prod_i p(\mathbf{x}^i)$), le second membre de l'expression (62) se simplifie et devient:

$$-\ln p(\mathcal{X}/\mathcal{Y}) = -\sum_i \ln p(\mathbf{x}^i/\mathbf{y}^i) + constante \tag{64}$$

La solution qui minimise cette expression est celle qui pour chaque $\mathbf{y}^i$ associe le paramètre géophysique $\mathbf{x}^i$ qui maximise $p(\mathbf{x}^i/\mathbf{y}^i)$. Sous l'hypothèse d'indépendance des paramètres de $\mathcal{X}$, le problème inverse est plus ou moins facile selon que la distribution $p(\mathbf{x}/\mathbf{y})$ est unimodale ou non.

- Si nous supposons que la fonction densité $p(\mathbf{x}/\mathbf{y})$ est unimodale et suit une loi normale. Le champs inverse recherché qui maximise la vraisemblance est obtenu en associant à chaque mesure $\mathbf{y}^i$ le paramètre physique $F^{-1}(\mathbf{y}) \simeq E(\mathbf{X}/\mathbf{y})$ et si cela est nécessaire l'inverse de la matrice de variance-covariance $\mathbf{C}_y^{-1}$ de la distribution $p(\mathbf{x}/\mathbf{y})$ proposé par le modèle inverse local. Cette méthode simplifiée est souvent utilisée dans la pratique pour résoudre l'inversion des données satellitaires. Son inconvénient majeur est de faire implicitement deux hypothèses fortes qui consistent à supposer que localement la solution du problème inverse est unique et que les paramètres physiques $\mathbf{x}^i$ sont indépendants et ne dépendent pas d'un contexte local.
- Si en plus de l'indépendance des paramètres physiques nous supposons que la densité $p(\mathbf{x}/\mathbf{y})$ n'est pas unimodale, il est alors possible de l'estimer par un réseau de mixture de gaussiennes (section 6.2.1). Dans ce cas , le choix d'une solution initiale de l'équation (64) est plus délicat à cause du caractère non univoque de la relation entre $\mathbf{y}$ et $\mathbf{x}$. La mise au point d'une "bonne" solution initiale $\mathcal{X}_0$ peut se faire en commençant par déterminer la localisation des principaux modes de la fonction densité conditionnelle $p(\mathbf{x}/\mathbf{y})$ et en utilisant des méthodes adaptées au problème particulier afin de choisir un ensemble (ou champs) $\mathcal{X}$ parmi ces modalités locales [44], [16], [17].

En général l'hypothèse d'indépendance des paramètres physiques est trop limitative et il est important de prendre en compte les dépendances

spatiales ou temporelles. Le terme $p(\mathcal{X})$ qui intervient dans les expressions (61,62) doit être pris en compte. Pour pouvoir effectuer le calcul il faut donc donner une forme fonctionnelle à l'expression, la plus classique étant de supposer qu'il s'agit d'une distribution gaussienne avec une matrice de variance-covariance représentant les dépendances spatiales ou temporelles. Une autre manière consiste à considérer directement le terme $\ln p(\mathcal{X})$ qui pourra représenter des contraintes permettant de favoriser certaines solutions physiquement plus plausibles. Dans tous les cas $p(\mathcal{X})$ apparaît comme un terme régularisateur de l'expression à minimiser.

Si l'on dispose d'une solution initiale $\mathcal{X}$ et que l'on suppose que la distribution statistique de l'ensemble des paramètres géophysiques $\mathcal{X}$ est normale de moyenne $\mathcal{X}_0$ et que l'on dispose en plus d'informations sur la matrice de variance- covariance Σ des erreurs spatiales ou temporelles du modèle. la minimisation de (61) peut se réduire, si l'on suppose connue la matrice C_x^{-1}, à minimiser:

$$\sum_i \left(\mathbf{y}^i - F(x^i)\right)^T C_{x^i}^{-1} \left(\mathbf{y}^i - F(x^i)\right) + \left(\mathcal{X} - \mathcal{X}_0\right)^T \Sigma^{-1} \left(\mathcal{X} - \mathcal{X}_0\right) \quad (65)$$

Cette modélisation permet de faire intervenir, par le biais de $\mathcal{X}_0$, une information qui contraint la solution recherchée. La solution initiale $\mathcal{X}_0$ peut être celle obtenue par inversion directe en supposant l'indépendance des paramètres physiques. Les contraintes introduites au niveau du champs peuvent également contenir des informations sur la physique du phénomène étudié. Elles peuvent, par exemple, provenir d'un modèle physique de bonne qualité dont on sait qu'il calcule des solutions moyennes acceptables. On fait dans ce cas l'hypothèse que la solution réelle, donc le champs inverse $\mathcal{X}$ proposé, ne s'écarte pas trop de cette solution moyenne. La minimisation de (65) peut se faire par une méthode de gradient qui calcule les dérivées partielles par rapport aux paramètres physiques $\mathbf{x}^i$, elle détermine ainsi le champs $\mathcal{X}$ des paramètres physiques retrouvés.

Nous présentons dans le dernier paragraphe une approche, dérivée de celle-ci, trés utilisée dans le domaine des réseaux de neurones qui permet de proposer une solution au difficile problème de l'inversion des données corrélées.

6.3.4 Prise en compte du voisinage dans le cadre du problème inverse local

Comme nous venons de le signaler l'ensemble des paramètres géophysiques $\mathcal{X} = \left\{\mathbf{x}^i, i = 1 \cdots N^{mes}\right\}$ à retrouver correspond le plus souvent à des

phénomènes corrélés dans le temps ou dans l'espace (cas des mesures satellitaires). Ces paramètres ne sont pas statistiquement indépendants et chaque paramètre géophysique dépend d'un voisinage local (temporel ou spatial). Il est souvent possible de faire l'hypothèse selon laquelle:

$$p(\mathbf{x}^i/\mathcal{X}) = p(\mathbf{x}^i/\mathcal{V}(x^i)) \tag{66}$$

où $\mathcal{V}(\mathbf{x}^i)$ représente un sous ensemble de paramètres géophysiques dans un contexte temporel ou spatial du point $\mathbf{x}^i$. Cette hypothèse est souvent réaliste si l'on se place dans un contexte géophysique (mesures satellitaires, ...). Accepter cette hypothèse revient à supposer que le champs des paramètres géophysique est un champs Markovien. Si l'on introduit le contexte local des observations $\mathcal{V}(\mathbf{y}^i)$ correspondant à l'ensemble des paramètres géophysiques de $\mathcal{V}(\mathbf{x}^i)$, une décomposition de la probabilité $p(\mathcal{X}/\mathcal{Y})$ à l'aide des différents voisinages en parcourant le champs $\mathcal{X}$ selon un trajet prédéfini (balayage du passé vers le futur pour un champs temporel, de gauche à droite et de haut en bas pour un champs spatial) permet d'introduire la pseudo vraisemblance:

$$\prod_{i}^{N^{mes}} p(\mathbf{x}^i/\mathcal{V}(\mathbf{y}^i)) \tag{67}$$

Cette pseudo vraisemblance est en général différente de $p(\mathcal{X}/\mathcal{Y})$. On peut vérifier que $p(\mathcal{X}/\mathcal{Y})$ est égal à la pseudo-vraisemblance dans le cas où l'information contenue dans $\mathcal{V}(\mathbf{y}^i)$ contient toute l'information contenue dans $\mathcal{V}(\mathbf{x}^i)$. Cette hypothèse n'est vraiment acceptable que dans les régions de faible ambiguïté.

L'expression (67) permet d'avoir une approximation simplifiée de $p(\mathcal{X}/\mathcal{Y})$ où les variables géophysiques sont décorrélées relativement au contexte utilisé. En pratique la maximisation de (67) donne de bons résultats. La modélisation de $p(\mathbf{x}^i/\mathcal{V}(y^i))$ peut se faire par un PMC en estimant directement cette fonction densité (section 6.2.1) ou bien en procédant à une approximation par histogramme (§ 6.2.2). Le choix des PMC a l'avantage de fournir une famille de fonctions multidimensionnelles flexibles. La prise en compte du voisinage se fait par l'intermédiaire de la couche d'entrée du PMC en ajoutant le nombre de neurones nécessaires à la prise en compte des mesures du contexte. Les deux méthodes d'estimation de la densité sont très générales, elles ne dépendent pas de la nature des observations traitées, et s'appliquent sans modification à $\mathcal{V}(\mathbf{y}^i)$. Généralement $\mathcal{V}(\mathbf{y}^i)$ est déterminée de manière empirique, un compromis doit être fait quant à la taille de $\mathcal{V}(\mathbf{y}^i)$, celle-ci doit être assez grande pour tenir compte du contexte de dépendance et suffisamment

petit pour que la taille du PMC reste raisonnable. Une application de cette méthode est détaillée dans [45], [34], [53].

7 Applications à la Géophysique

Dans ce paragraphe nous avons choisi de présenter quelques problèmes géophysiques dont la résolution utilise les techniques que nous venons de présenter. Notre but est de montrer que la régression et la résolution de problèmes inverses par PMC permettent d'élaborer des méthodes opérationnelles. Dans le cas des méthodes neuronales le passage de la théorie vers la pratique est parfois délicat; le traitement des données nécessite un grand nombre de pré traitements et la validation un grand nombre d'opérations et de vérifications que nous n'avons pas abordés dans le souci de simplifier la présentation formelle des théories. Les manuscrits de thèses nous ont paru être les documents les plus adaptés pour dispenser les informations nécessaires à cet usage. Pour chacune des thèses que nous avons choisies, nous présentons le problème géophysique étudié, les problèmes théoriques sous jacents, nous décrivons brièvement les modèles mis en œuvre pour le résoudre. Nous renvoyons également aux articles parus sur le sujet. Pour chaque problème évoqué nous indiquons en référence le paragraphe de l'article qui traite les points théoriques concernés.

"Contribution à l'étude des diffusiomètres NSCAT et ERS-2 par modélisation neuronale. Influence de la hauteur des vagues sur le signal diffusiométrique."
Les satellites d'observation de la terre permettent d'obtenir une couverture presque complète de la terre. L'utilisation de diffusiomètres embarqués à bord de satellites permet d'obtenir en tout point de l'océan plusieurs mesures de la rugosité de la mer qui sont vus sous différents angles d'incidences. L'analyse de ces signaux permet de faire le lien avec le vent de surface et en particulier de déterminer en tout point de l'océan sa vitesse et sa direction. La thèse présente une étude des mesures de rétrodiffusion des radars diffusiométriques spatiaux NSCAT et ERS-2, le but principal étant de comprendre les mécanismes complexes de l'interaction des ondes électromagnétiques radar avec la surface marine. La mise au point de modèles empiriques par régression non-linéaire en utilisant la gigantesque quantité de données disponibles permet une meilleure compréhension de la relation existant entre les signaux rétrodiffusés et le vecteur vent. La méthode employée est donc la régression simple par PMC (3). Les données utilisées étant fortement

bruitées une partie de la thèse présente des modèles neuronaux qui estiment les erreurs dues à la modélisation incomplète et à la mesure (5.2). L'étude des propriétés physiques des fonctions neuronales obtenues mettent en évidence les capacités des MLP pour retrouver l'espérance mathématique et donc la relation "vraie" sous jacente. Un grand nombre de comparaisons avec des algorithmes proposés par l'IFREMER ou la NASA permettent de conclure sur la très bonne qualité des fonctions proposées.
Références: [55] [56] [36] [35] [34].

"Architectures Neuronales pour l'Approximation des Fonctions de Transfert: application à la télédétection."
L'inversion des données diffusiométriques est un problème complexe, c'est un problème multivalué. Une méthodologie neuronales complète a été proposée pour effectuer cette inversion. Elle consiste en un système modulaire de réseaux de neurones qui permet d'inverser directement d'une manière séquentielle la vitesse, puis la direction du vent. Un premier PMC permet d'inverser la vitesse du vent à l'aide d'une inversion directe avec prise en compte du contexte local (§6.3.4). (6.3.4); un second PMC, qui prend en compte la première inversion et le contexte local, retrouve la direction du vent en approximant la densité conditionnelle par la méthode des histogrammes (6.2.2). La modélisation permet d'obtenir les différentes directions possibles (au nombre de 4). Cette méthode a été appliquée aux signaux du diffusiomètre de ERS2, les résultats sont excellents puisque le module du vent est obtenu avec une RMS de l'ordre de 1.2 m/s et la direction est retrouvée avec une précision de 20 degrés dans 76% des cas pour le premier ambigu et dans 96 % des cas si l'on considère l'ensemble des quatre directions possibles. Un premier levé d'ambiguïtés qui utilise les probabilités calculées par l'inverse neuronal a permis de montrer la validité physique des champs retrouvés. La méthode est implémentée en tant que prototype dans le centre opérationnel de Prévimar.
Références: [34], [53] [45].

"Approche réseaux de neurones pour la classification d'émission structurées de type sifflement".
Il s'agit d'une des premières recherches qui montre l'applicabilité des méthodes neuronales à des problèmes de télédétection satellitaire. Le travail réalisé traite de la reconnaissance et de la détermination des paramètres pertinents de deux types d'émissions naturelles "Extrêmement Basse Fréquence" qui sont observées à partir du satellite AU-

REOL3 : les sifflements électroniques et les sifflements protoniques. La reconnaissance est effectuée à l'aide d'une classification qui utilise une architecture faisant intervenir un contexte temporel. Ce contexte est introduit en utilisant un modèle neuronal spécifique : le modèle TDNN (TimeDelay Neural Network (**??**)). Le manuscrit de thèse comporte une présentation des réseaux TDNN (Time Delay Neural Network) ainsi qu'une discussion sur l'implémentation matérielle.
Références: [37].

"Approche connexionniste pour le pilotage temps réel du récepteur numérique WAVES/TNR embarqué sur la sonde spatiale WIND."
Les sondes spatiales, telle que WIND, permettent d'opérer des mesures dans l'espace et de les retransmettre au sol pour être analysées. Un accroissemnt des capacités d'acquisitions à bord induit une augmentation du volume des informations qui met en difficulté le système de retransmission dont les capacités sont limitées. Le but de la thèse est de proposer un système neuronal de sélection de l'information permettant de ne retransmettre que les mesures utiles à l'étude des phénomènes visés. Ce système a été intégré au récepteur numérique TNR de l'expérience WAVES embarquée sur la sonde spatiale WIND. TNR est chargé d'étudier le bruit thermique par une analyse spectrale à haute résolution du plasma du vent solaire. Le bruit thermique produit une raie, appelée raie plasma, dont la position en fréquence et la forme caractérisent certains paramètres physiques du milieu étudié. TNR réalise une analyse spectrale en haute résolution en partitionnant sa bande passante totale en 3 bandes (A,C,E) plus étroites produisant un "agrandissement" du signal qui permet donc une analyse spectrale plus fine. Etant donné que la largeur de la bande passante de la raie plasma est inférieure à la largeur des bandes de TNR, il faut s'assurer que le raie est bien visible dans la bande d'analyse car seule le spectre de la bande d'analyse courante est transmis au sol. Le système qui a été réalisé permet de maximiser le nombre de spectres transmis au sol contenant une raie plasma en détectant la bande qui la contient et en localisant finement sa position dans la bande. La modélisation utilise un MLP en mode classifieur qui permet d'estimer les probabilités a posteriori d'appartenance de la fréquence de la raie plasma à l'une des 32 classes de fréquences qui constituent les canaux de mesures d'une bande (§4). Le système a été mis au point sur des données simulées, la sonde spatiale WIND, lancée le 1er novembre 1994, a embarqué le récepteur numérique WAVES/TNR utilisant le prototype. Après 6 mois de vol, le réseau a été mis à jour sur les données

réelles mesurées et ses nouveaux paramètres ont été téléchargés. Le mode le plus sûr du récepteur consiste à balayer tout l'espace des fréquences en commutant cycliquement les bandes de TNR (ACEACE...). Ce mode garantit donc une mesure régulière de la raie mais une résolution temporelle faible. Les résultats obtenus avec le système neuronal permettent de gagner directement un facteur 3. Les capacités de localisation fine de la position de la raie dans une bande permettraient théoriquement de pousser ce gain à un facteur maximal de 12 pour une résolution temporelle de 325 msec.
Références; [44].

"Approche neuronale de l'inversion, application à la tomographie acoustique océanique".
La Tomographie Acoustique Océanique (TAO) est une méthode de mesure des paramètres acoustiques de l'océan qui permet de retrouver certains de ces paramètres physiques. Elle repose sur la capacité de l'océan à guider les signaux sonores sur des longues distances. Dans l'eau de mer, le son se propage vite (environ à $1500 m/s$ soit $4,5$ fois plus vite que dans l'air). Or, le temps mis par un signal sonore à se propager dépend de la célérité du son qui est elle même reliée à la température, la salinité et la pression. A partir des mesures des temps de propagation, on peut par un processus d'inversion estimer les paramètres physiques internes de l'océan. Il s'agit d'un problème d'inversion bien posé, ne présentant pas d'ambiguïtés, il est traité à l'aide d'une méthode directe locale (6.3.1). Celle ci fait l'hypothèse implicite de l'indépendance des variables géophysiques (utilisation des moindres carrés simples) et n'introduit pas de probabilité à priori sur le champs à retrouver. Le manque d'observations réelles permettant de relier les paramètres physiques étudiés aux temps d'arrivés amène à résoudre le problème par simulation en utilisant les modèles acoustiques existants. La physique liée à la propagation des ondes étant bien connue, la précision des modèles acoustiques est largement supérieure à celle qui serait obtenue en utilisant les données d'observations. Les temps de propagation d'un signal sonore entre deux paires sources-récepteurs. (propagation à trajets multiples) sont calculés à partir d'un modèle acoustique. Ces temps d'arrivée sont inversés pour retrouver le champ de célérité du milieu, ainsi que ses caractéristiques physiques (température, salinité, densité,...). Au moment d'inverser les temps d'arrivées provenant d'observations réelles, on initialise le PMC en déterminant un premier inverse par linéarisation autour d'un état de référence du milieu. Une validation de l'approche est faite en inversant les données d'une expérience de tomographie (GASTOM'90) menée en Atlantique

Nord-Est. Les expériences présentées prouvent que l'utilisation de modèles permet d'obtenir de très bons résultats en présence d'observations réelles.
Références: [51], [49], [50].

" La modélisation du transfert radiatif à des fins climatiques: une nouvelle approche fondée sur les réseaux de neurones artificiels."
Ce travail a permis de mettre au point un modèle de transfert radiatif aux GO (grandes longueur d'onde), NeuroFlux, dont les paramétrisations utilisent le PMC. Le calcul par NeuroFlux des flux montants et descendants de la surface jusqu'au sommet de l'atmosphère est effectué par une batterie de 39 PMC. Un premier PMC calcule les flux associés à un ciel clair, les autres PMC sont chacun spécialisés dans le calcul des flux, soit montants, soit descendants, en présence d'un seul nuage opaque dans une couche précise du modèle. Les flux finaux sont obtenus par combinaison linéaire des quantités estimées par les 39 PMC. Suivant une approximation courante, les paramètres de la combinaison linéaire sont calculés à partir des caractéristiques de la nébulosité. Chacun des PMC effectue donc un calcul simple, il relève des méthodes présentées au paragraphe (3). L'originalité de la méthode provient du découpage du calcul complet en 39 sous-tâches effectuant des calculs plus simples. Une information physique complémentaire est introduite dans les équations qui permettent le calcul des flux totaux. La paramétrisation de l'équation du transfert radiatif à l'aide de PMC permet des gains en rapidité importants. Par exemple, NeuroFlux est 16,5 fois plus rapide que le code de transfert radiatif opérationnel dans le modèle de circulation générale du Centre Européen pour les Prévisions Météorologiques à Moyen Terme (CEPMMT). Le mémoire de thèse contient trois types de validation qui toutes prouvent l'efficacité de la méthode neuronale: des comparaisons de code à code à partir de situation atmosphériques provenant de radiosondages et de mesure satellitaire, des études de sensibilié à la variation d'une seule variable géophysique, l'application de NeuroFlux dans le cadre d'une simulation climatique d'un modèle de circulation générale. Une extension de l'approche au calcul des flux de transfert radiatif verticaux aux GO à partir du radiomètre multi-fréquence TOVS (TIROS-N Operational ertical Sounder) est aussi développée.
Références: [10] [11] [12] [8] [9][7].

"Utilisation des réseaux de neurones pour l'inversion des observations spatiales et la détermination des concentrations de

constituants minoritaires dans la troposphère."
Le travail de recherche s'inscrit dans le cadre de la préparation de la mission de l'instrument IASI (Infrared Atmospheric Sounding Interferometer) qui sera lancée en l'an 2003. Il s'agit de développer un algorithme qui permet de restituer, à partir des luminances mesurées, les concentrations d'ozone (O3), de méthane (CH4) et de monoxyde de carbone (CO). Il s'agit d'un problème d'inversion dont la résolution est abordée à partir de simulations. La base de données doit réunir les grandeurs définissant l'état de l'atmosphère (profils de température et de constituants chimiques) et les mesures spectrales associées. Les mesures spectrales proviennent d'un code de transfert radiatif qui prend en entrée des profils de température et de constituants chimiques issus d'un modèle de chimie-transport. Le travail réalisé a permis de mettre au point les bases méthodologiques de l'inversion, le but final étant d'obtenir une méthode opérationnelle d'inversion. La phase de faisabilité étant achevée, le travail aborde maintenant la phase opérationnelle. Une application sur des données mesurées par l'instrument IMG (Interferometric Monitor for Greenhouse Gases) et une inter comparaison avec différents algorithmes valident l'approche. La méthode d'inversion choisie est celle présentée au paragraphe (6.3.2), elle fait l'hypothèse implicite de l'indépendance des variables géophysiques et n'introduit pas de probabilité a priori sur le champs à retrouver.
Références: [25] [26] [14] [13].

"Système de reconnaissance automatique de signatures électromagnétiques en radioastronomie basse fréquence."
Il s'agit d'un système automatique pour la détection, l'estimation et la classification de signaux radiosplanétaires en bande kilométrique. Il s'agit donc, d'un problème de classification en basses fréquences radio astronomiques dans le plan temps fréquences. Ce problème est compliqué, car il s'agit d'identifier plusieurs sources émettant au même moment et reçu par un même capteur. Les caractéristiques et le type d'émissions à détecter sont : les types IIIsolaires, HOM (Jupiter Hectometric radiation), BKOM (Jupiter Broadbandkilometric radiation), NKOM (Jupiter Narrow band kilometric radiation) et autres signaux. Ces signaux se distinguent par leur localisation dans la bande de fréquences, par leurs formes et présentent des ambiguïtés. La solution proposée procède en deux étapes. La première étape classifie ces signaux en utilisant un réseau TDNN (Time Delay Neural Network) qui permettent la prise en compte d'un contexte temporel [57],l'ensemble des réponses du TDNN détermine une sorte de squelétisation du spectrogramme. La seconde

étape implémente un certain nombre de lois physiques locales sous forme d'automates cellulaires et fait évoluer la classification trouvée afin de recouvrir le signal autour de la "squelétisation" trouvée. Ce mémoire contient aussi une comparaison entre TDNN et le réseau à filtre mémoire récursive de type gamma.
Références: [20], [19].

Restitution des contenus en eau liquide atmosphérique à partir des mesures radiométriques des capteurs spatiaux"
L'objectif de cette thèse est la mise au point d'une méthode de restitution des contenus en eau liquide atmosphérique (eau liquide nuageuse et précipitations en phase liquide) en utilisant des mesures radiométriques provenant de capteurs spatiaux (SSM/I, TMI, AMSR). Il s'agit d'un problème d'inversion dont la résolution est abordée à partir d'algorithmes neuronaux développés sur une base de donnée simulée. Pour la réalisation de cette base, l'étude et la modélisation du problème physique direct est fondamentale, elle a nécessité le développement d'un modèle de transfert radiatif performant de manière à créer une base de données contenant les grandeurs atmosphériques que l'on souhaite restituer et les mesures spectrales associées. La méthode d'inversion mise en œuvre est une méthode neuronale basée sur l'utilisation d'une architecture multi-expert [30], composée d'un réseau contrôleur non linéaire et de plusieurs experts (également non linéaires). L'espace d'entrée est divisé, par le réseau contrôleur, en sous parties correspondant à des régimes différents (ciel faiblement ou fortement nuageux par exemple) Chaque expert est ainsi dédié à une partie de l'espace d'entrée, c'est à dire qu'il apprend un modèle inverse local à partir des données de ce sous espace ; il peut ainsi extraire des caractéristiques qui ne peuvent être généralisées à l'ensemble des données. La validation est ensuite effectuée en appliquant les algorithmes obtenus à des données mesurées. Concernant l'eau liquide nuageuse, restituée à partir du radiomètre SSM/I, une inter comparaison avec différents algorithmes valident l'approche. L'amélioration porte pour l'essentiel sur l'élargissement du domaine de validité de l'inversion réalisée. Les taux de pluie restitués à partir du radiomètre TMI par l'algorithme neuronal sont validés à partir des données fournies par le radar également embarqué sur le satellite TRMM. Une nette amélioration est apporté par rapport à l'algorithme standard avec des performances globales meilleures et la suppression des points aberrants.
Références: [38].

8 Conclusion

Ce document présente dans un cadre général, celui de la régression non linéaire multiple, la régression par Perceptron Multicouches. Les Perceptrons MultiCouches constituent une famille particulière très riche de fonctions, dont la principale caractéristique est de permettre une grande souplesse de modélisation. Ces fonctions sont particulièrement adaptées quand il s'agit de traiter des données présentant des incertitudes. Nous avons développé en détails les principaux modèles permettant de déterminer selon les besoins de l'application la valeur moyenne calculée par la régression quand les incertitudes sont prises en compte au niveau des variables à expliquer. Une généralisation complète du formaliseme probabiliste aux cas où les variables explicatives présentent des incertitudes peut être développé. Il est alors possible de définir de nouveaux modèles et algorithmes d'apprentissage pour les Perceptrons MultiCouches qui corrigent les variables explicatives "au mieux" de manière à retrouver la relation "vraie" liant variables explicatives et variables expliquées [2]. La famille de fonctions générée par les architectures PMC est suffisamment vaste pour permettre de proposer des fonctions de régression précises dans les cas où la relation sous jacente est univoque. L'introduction d'hypothèses probabilistes sur les variables à régresser et l'utilisation du maximum de vraisemblance permet de définir différentes fonctions de coût adaptées à la résolution de différents types de problèmes. Il devient alors possible de proposer des solutions pour résoudre des problèmes plus complexes comme celui de la détermination des matrices de variance-covariance conditionnelles. Le formalisme probabiliste permet de traiter le problème de la régression, mais aussi celui plus général de la détermination de fonctions densité, la densité recherchée étant approximée en tant que mélange de lois normales. Dans la seconde partie de l'article nous avons présenté un formalisme probabiliste permettant d'aborder de différentes manière la résolution de problèmes inverses. Nous avons montré que tous les résultats relatifs à la régression simple et à l'estimation de fonctions densité s'intégraient naturellement dans l'élaboration de méthodes inverses. Tous ces résultats ont été généralisés et présentés dans le cadre plus vaste de l'inversion de champs d'observations (spatiaux ou temporelles). On peut alors utiliser les corrélations existantes qui apparaissent au niveau des données et des paramètres physiques en les introduisant explicitement dans la résolution. La discussion et les applications présentées qui suivent l'exposé des résultats théoriques permettent de mieux de mieux appréhender l'utilisation des PMC pour traiter des problèmes inverses et montrent l'opérationalité et l'étendue de leur domaine d'applications.

References

[1] Anderson T.W. (1958). *"An introduction to multivariate statistical analysis"*, John Wiley & Sons, Inc, New York 374p

[2] Badran F.,Stephan Y., Metoui N., and Thiria S. (1999) A general formulation of non-linear least square regression using multi-layered perceptrons"., Submited to IEEE transaction on neural networks

[3] Battiti, R. (1992) First- and second-order methods for learning: Between steepest descent and Newton's Method, Neural Computation, Vol. **4**, pp. 141–166.

[4] Bishop, C. (1994) Mixture density networks , technical report NCRG4288, Aston University, Birmingham.

[5] Bishop, C. (1995) *"Neural Networks for Pattern recognition"*, Calendon Press - Oxford.

[6] Bricaud, A., Babin, M., Morel, A. and Claustre, H. (1995) Variability in the chlorophyll-specific absorbtion coefficients of natural phytoplantankon : Analysis and parameterization, Journal of geophysical Research, vol. 100, No. C7, pp. 13,321-13,332.

[7] Chéruy F, F. Chevallier, N. A. Scott, et A. Chédin, (1996) : *Une méthode utilisant les techniques neuronales pour le calcul rapide de la distribution verticale du bilan radiatif thermique terrestre. C. R. Acad. Sci. Paris, t. 322, S. IIb, p. 665-672.*

[8] Chevallier F, F. Chéruy, N. A. Scott, et A. Chédin, (1998) : *A neural network approach for a fast and accurate computation of longwave radiative budget. J. Appl. Meteor., 37:11, 1385-1397.*

[9] Chevallier F. (1998): " La modélisation du transfert radiatif à des fins climatiques: une nouvelle approche fondée sur les réseaux de neurones artificiels" Thèse de doctorat de l'Université de Paris 7.

[10] Chevallier F., A. Chédin, F. Chéruy, J.-J. Morcrette, (2000): *TIGR-like atmospheric profile databases for accurate radiative flux computation. Q. J. R. Meteor. Soc, 126, 777-785.*

[11] Chevallier F, F. Chéruy, R. Armante, N. A. Scott, C. Stubenrauch, et N. A. Scott, (2000): *Retrieving the clear sky vertical longwave radiative budget from TOVS: comparison of a neural network-based retrieval and a method using geophysical parameters. J. Appl. Meteor, sous presse.*

[12] Chevallier F, J.-J Morcrette, F. Chéruy, et N. A. Scott, (2000): *Use of a neural network-based longwave radiative transfer scheme in the ECMWF atmospheric model. Q. J. R. Meteor. Soc., 126, 761-776.*

[13] Clerbaux, C., P. Chazette, J. Hadji-Lazaro, G. Megie, J.-F. Muller, and S. A. Clough. (1998). Remote sensing of CO, CH4, and O3 using a spaceborne nadir-viewing interferometer J. Geophys. Res 103 (D15), 18,999-19,013

[14] Clerbaux, C., J. Hadji-Lazaro, S. Payan, C. Camy-Peyret, and G. Megie (1999) Retrieval of CO columns from IMG/ADEOS spectra , IEEE trans. Geosci. Remote Sensing, 37 (3)

[15] Cornford, D., Ramage, G., Nabney, I. (1999), A scatterometerneural network neural model with input noise, Neurocomputing (30) 1-4 pp13-21

[16] Cornford, D and Nabney, I T and Bishop, C M (1999): Neural Network-Based Wind Vector Retrieval from Satellite Scatterometer Datal Neural Comp. Appli., 8, pp 206-217.

[17] Cornford, D and Nabney, I T and Williams, C K I (1999): Adding Constrained Discontinuities to Gaussian Process Models of Wind Fields Accepted in Advances in Neural Information Processing Systems, 11, editor, Kearns, M S and Solla, S A and Cohn, C A, MIT Press

[18] Cybenko, G. (1989), Approximation by superposition of a sigmoidal function, Math. Control Signal Systems, **2**, pp. 303–314.

[19] De Lassus H., Lecacheux A., Thiria S., Badran F. (1996): "Neural Network Clusters and cellular automata for the Detection and Classification of overlapping Transient Signals on Radio Astronomy Spectrograms from Spacecraft. In International Symposium on Time-Frequency and time-Scale Analysis, pp 253-256, Paris France, june 1996. IEEE Signal Processing Society.

[20] De Lassus H. (1999): "Système de reconnaissance automatique de signatures électromagnétiques en radioastronomie basse fréquence." Thèse de doctorat du Conservatoire National des Arts et Métiers.'

[21] Dempster A.P., Laird N.M. and Rubin D.B. (1977), maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, B 39 (1), pp 1-38.

[22] Duda R.O., and Hart P.E. (1973), *Pattern Classification and Scene Analysis* New York John Wiley.

[23] Funahashi, K.I. (1989), On the approximate realization of continuous mapping by neural networks, Neural Networks, Vol. **2**, pp. 185–192.

[24] Gross L., Thiria S., Frouin R., Mitchell B.G. (2000) Artificial neural networks for modeling transfer function between marine reflectance and phytoplankton pigment concentration Journal of. Geophys. Res. Vol 105,no.C2, pp3483-3949, february 15

[25] Hadji-Lazaro J. (1999): "Utilisation des réseaux de neurones pour l'inversion d'observations spatiales et la détermination des concentrations de constituants minoritaires dans la troposphère." Thèse de doctorat de l'Université de Paris 7.

[26] Hadji-Lazaro, J., C. Clerbaux, and S. Thiria. (1999), An inversion algorithm usingneural networks to retrieve atmospheric CO total columns from high-resolution nadir radiances, J. Geophys. Res Vol 104 NO 19 pp 23841-23854.

[27] Hardel, W. (1990), *Applied Nonparametric regression*. Cambridge university press.

[28] Haykin, S.(1996) *Neural Networks a comprehensive foundation*, Prentice Hall.

[29] Hornik, K., Stinchcomb, M. and White, H. (1989) Multi-Layered feedforward networks are universal approximators, Neural Networks, **2**, pp. 359–366.

[30] Jordan, M.I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm.Neural Computation, **16**: pp 181-214.

[31] Jordan, M. and rumelhard, D. (1992) Forward models: Supervised learning with a distal teacher, Cognitive Science, **16**, pp. 307–354.

[32] Lippman, R.P. (1987) An introduction to computing with neural nets, ASSP, Magazine, pp. 4–22, April 87.

[33] MacKay, D. (1992a) Bayesian interpolation, Neural Computation, **4**, pp. 415–447.

[34] Mejia C. (1992): "Architectures Neuronales pour l'Approximation des Fonctions de Transfert: application à la télédétection." Thèse de doctorat de l'Université d'Orsay.

[35] Mejia C , S. Thiria, F. Badran, N. Tran and M. Crepon. (1998) Determination of the Geophysical Model Function of ERS1 Scatterometeer by the use of Neural Networks., Journal of geophysical Research, vol. 103, pp. 12853-12868

[36] Mejia, C., Badran, F., Bentamy, A., Crepon, M., Thiria, S. and Tran, N. (1999) Determination of the geophysical model function of NSCAT and its corresponding variance by the use of the neural

networks, Journal of geophysical Research, vol. 104, No. C5, pp. 11,539-11,556.

[37] Miniere X. (1994): "Approche réseaux de neurones pour la classification d'émission structurées de type sifflement" Thèse de doctorat de l'Université de Paris 7.

[38] Moreau E. (2000): "Restitution de paramètres atmosphériques par radiométrie hyperfréquence spatiale. Utilisation de méthodes neuronales." Thèse de doctorat de l'Université de Paris 7.'

[39] Nabney, I.T., Cornford, D.,Williams, C.K.I. (1999) Structured neural network modelling of multi-valued functions for wind vector retrieval from satellite scatterometer measurements, Neurocomputing (30) 1-4 pp 3-11.

[40] Nix, D. and Weigend, A. (1995) Learning local error bars for nonlinear regression, in Advanced in neural Information Processing Systems, G.Tesauro & Al. Eds, MIT Press, Cambridge, pp. 489–496.

[41] Poggio, T. and Girosi, F. (1990) Networks for approximation and learning, Proceedings of the IEEE, Vol. **78**, pp. 1481–1497.

[42] Reed R.D. and Marks R.J. (1998) "Neural Smithing. Supervised Learning in Feedforward Neural Networks" A Bradford Book MIT Press.

[43] Richard M.D., and Lippman R.P. (1991) Neural network classifiers estimate Bayesian a-posteriori probabilities, Neuralcomputation 3(4) pp 461-483.

[44] Richaume P. (1996): "Approche connexionniste pour le pilotage temps réel du récepteur numérique WAVES/TNR embarqué sur la sonde spatiale WIND." Thèse de doctorat du Conservatoire national des Arts et métiers.

[45] Richaume, P.,Mejia, C.,Thiria, S., Tran, N., Crepon, M., Roquet, H., Badran, F.(2000) Neural Network Wind retrieval from ERS1 Scatterometer Data. Journal of Geophysical Research, vol 105 , NO. C4, pages 8737-8751, april, 15.

[46] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Parallel Distributed Processing, Vol. **1**, MIT Press, Cambridge.

[47] Schumann J. (1996) "Pattern Classification: a unified view of statistical and neural approaches", Wiley-Interscience.

[48] Sontag, E.D. (1992) Feedback stabilization using two-hidden-layer nets, IEEE Transactions on Neural Networks, **3(6)**, pp. 981–990.

[49] Stephan Y. (1996): "Approche neuronale de l'inversion. application à la Tomographie acoustique océanique" Thèse de doctorat du Conservatoire National des Arts et Métiers.

[50] Stéphan, Y., Thiria, S. and Badran, F. (1996) Application of multilayered neural networks to ocean tomography inversions,Inverse Problems in Engineering, Vol. **1**, pp. 181–304.

[51] Stéphan, Y., Démoulin, X. and Sarzeaud, O. (1998) Neural direct approaches for Geoacoustic inversion, Journal of Computational Acoustics, Vol.**6**, No 1-2, 151-166.

[52] Tarantola, A. (1987) "Inverse Problem Theory", Elseiver Science Publisher, Amsterdam, 613 p.

[53] Thiria, S., Mejia, C., Badran F. and Crépon, M. A neural approach for modeling nonlinear transfer function: Application for wind retrieval from spaceborn scatterometer data, Journal of Geophysical Research, 98, C12, 22,827–22,841 (1993).

[54] Thodberg, H. (1996) A review of Bayesian neural networks with an application to near infrared spectroscopy, IEEE Transactions on Neural Networks, Vol. **7**, no 1, pp. 56–72.

[55] Tran N., Thiria S., Crepon M., Badran F. and Freilich M. (2000): 'Validation of the QSCAT NRCS on the advanced neural network NSCAT GMF and estimation of neural network QSCAT GMF' IGARSS'2000, Honolulu Hawai, July 24-28, 2000.

[56] Tran N. (1999): "Contribution à l'étude des diffusiomètres NSCAT et ERS-2 par modélisation neuronale. Influence de la hauteur des vagues sur le signal diffusiométrique." Thèse de doctorat de lÕUniversité de Paris VI.

[57] Waibel A., Hanazawa T. Hinton G., Shikano K. and Lang K.J.,(1989). Phoneme recognition using time-delay neural networks. IEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-37, pp. 328-339.

[58] Weigend, A.S., Mangeas, M. and Srivastava, A. N. (1995). Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting. International Journal of Neural Systems. **6**: 373-399.

[59] Weigend, A.S., Zimmermann, H.G.,Neuneier, R. (1996). Clearning, Neural Networks in Financial Engineering, Proceedings of the Third International Conference on Neural Networks in the Capital Markets, NNCM-95, pp. 511-522

[60] White, H. (1990). Connectionnist Nonparametric Regression: Multi-Layer Feedforward Networks Can Learn Arbitrary Mappings. Neural Networks **3**, 535–549.

[61] Williams, P.M. (1996). Using Neural Networks to Model Conditional Multivariate Densities. Neural Computation**8**, 843–854.

[62] Zhang J., Walter G.G., Miao Y., Wayne Lee N.,(1995). Wavelet Neural Networks For Function Learning. IEE Trans on signal Processing 43 (6) 1485-1497.