3 ETUDE THEORIQUE DU LEAVE-ONE-OUT

Résumé

Dans le cas d'un modèle non-linéaire par rapport aux paramètres, il est possible d'obtenir une approximation de la solution des moindres carrés $\boldsymbol{\theta}_{LS}$ en effectuant un développement de Taylor au premier ordre de la sortie du modèle. On utilise également cette approche pour estimer la solution des moindres carrés $\boldsymbol{\theta}_{LS}^{(-i)}$ obtenue après avoir supprimé l'exemple i de la base d'apprentissage. En combinant ces deux expressions, on obtient une relation approchée entre $\boldsymbol{\theta}_{LS}$ et $\boldsymbol{\theta}_{LS}^{(-i)}$, expression qui est valable sous réserve que les deux développements de Taylor le soient, c'est-à-dire que la courbure du sous-espace des solutions soit suffisamment faible.

On utilise la relation précédente pour prédire l'effet du retrait d'un exemple de l'ensemble d'apprentissage. Ainsi, l'erreur de prédiction sur cet exemple est multipliée par un coefficient qui tend vers l'infini lorsque h_{ii} tend vers 1. *De même, l'intervalle de confiance sur la prédiction de l'exemple retiré de la base d'apprentissage devient infini lorsque h_{ii} tend vers* 1. *Ceci nous amène à interpréter la grandeur h_{ii} comme l'influence de l'exemple i sur l'estimation des paramètres du modèle : plus h_{ii} (qui est positif) est grand et tend vers* 1, *plus cette influence est grande.*

Afin de comparer l'effet du retrait d'un exemple, tel qu'il est prédit par ces formules, à celui obtenu par apprentissage, nous avons introduit une classification des modèles en modèles "propices au leave-one-out" ou non. Il s'agit en réalité de définir les modèles pour lesquels l'effet du retrait d'un exemple ne peut pas être raisonnablement estimé par apprentissage, et donc pour lesquels une comparaison avec les résultats des formules de linéarisation n'a pas de sens. Nous proposons une méthode géométrique qui permet de s'assurer - en partie seulement - qu'une solution est propice au leave-one-out.

Nous montrons enfin qu'un modèle est souvent non propice au leave-one-out à cause de la présence d'un exemple à forte influence sur les coefficients et dont le retrait fait converger l'apprentissage vers un autre minimum de la fonction de coût. Dans ce cas, la seule manière d'estimer l'effet du retrait de cet exemple sur le modèle est d'utiliser les formules fondées sur le développement de Taylor.

3.1 Introduction

Lorsqu'un modèle est linéaire par rapport à ses paramètres, on obtient facilement la solution des moindres carrés en résolvant le système d'équations canoniques. Ce n'est pas le cas pour les modèles non linéaires par rapport à leurs paramètres ; cependant, en considérant une zone suffisamment petite de l'espace des paramètres, on peut trouver une expression approchée de la solution des moindres carrés en réalisant un développement limité au premier ordre de la sortie du modèle. Après un bref rappel de la démonstration de ce résultat classique (voir par

exemple [Seber 89]), nous allons montrer comment il peut s'appliquer pour détecter les modèles surajustés et prédire l'effet du retrait d'un exemple de l'ensemble d'apprentissage.

3.2 Approximation locale de la solution des moindres carrés

Notons $\boldsymbol{\theta}_{LS}$ la solution des moindres carrés, c'est-à-dire le vecteur des paramètres qui minimise la fonction de coût quadratique :

$$J(\boldsymbol{\theta}) = {}^t[\boldsymbol{y_p} - \boldsymbol{f}(X, \boldsymbol{\theta})]\ [\boldsymbol{y_p} - \boldsymbol{f}(X, \boldsymbol{\theta})] \tag{3.1}$$

avec $\boldsymbol{f}(X, \boldsymbol{\theta}) = {}^t[f(\boldsymbol{x}^1, \boldsymbol{\theta}), ..., f(\boldsymbol{x}^N, \boldsymbol{\theta})]$, où X est la matrice ${}^t[\boldsymbol{x}^1, ..., \boldsymbol{x}^N]$ de dimensions (N, n).

Si le modèle est linéaire par rapport aux paramètres, c'est-à-dire si $\boldsymbol{f}(X, \boldsymbol{\theta}) = {}^t[{}^t\boldsymbol{x}^1\boldsymbol{\theta}, ..., {}^t\boldsymbol{x}^N\boldsymbol{\theta}]$, alors la solution des moindres carrés s'écrit :

$$\boldsymbol{\theta}_{LS} = ({}^tX\,X)^{-1}\,{}^tX\,\boldsymbol{y}_p \tag{3.2}$$

Si le modèle n'est pas linéaire par rapport aux paramètres, il n'existe pas d'expression similaire. Néanmoins, on peut obtenir une solution locale approchée de la solution des moindres carrés à partir d'un développement limité de $\boldsymbol{f}$ au voisinage d'un point $\boldsymbol{\theta}^*$ de l'espace des paramètres.

Ce développement limité permet d'obtenir une expression approchée de $\boldsymbol{f}(X, \boldsymbol{\theta})$ et donc de $J(\boldsymbol{\theta})$. La solution des moindres carrés $\boldsymbol{\theta}_{LS}$ est ensuite obtenue, comme dans le cas linéaire, en annulant le gradient de J, mais après n'avoir conservé que les termes du premier ordre en $(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$.

Pour obtenir un développement limité cohérent de $\frac{\partial J}{\partial \boldsymbol{\theta}}$ au premier ordre en $(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, il faut partir d'un développement limité au second ordre de $\boldsymbol{f}(X, \boldsymbol{\theta})$ au voisinage de $\boldsymbol{\theta}^*$:

$$\boldsymbol{f}(X, \boldsymbol{\theta}) \cong \boldsymbol{f}(X, \boldsymbol{\theta}^*) + Z\,(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + {}^t(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\,\boldsymbol{S}\,(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \tag{3.3}$$

Dans la formule précédente :

- Z désigne la matrice jacobienne du modèle, définie par $Z = {}^t[\boldsymbol{z}^1, ..., \boldsymbol{z}^N]$ où $\boldsymbol{z}^i = \left.\frac{\partial f(\boldsymbol{x}^i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*}$,
- $\boldsymbol{S} = \sum_{i=1}^{N} S(\boldsymbol{x}^i)\,\boldsymbol{e}_i$ est un tenseur d'ordre 3 dans lequel $S(\boldsymbol{x}^i)$ est une matrice de dimensions (q, q) définie par $S(\boldsymbol{x}^i) = \left(\left.\frac{\partial^2 f(\boldsymbol{x}^i, \boldsymbol{\theta})}{\partial\theta_j\,\partial\theta_k}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*}\right)_{\substack{j = 1, ..., q \\ k = 1, ..., q}}$ et $\boldsymbol{e}^i$ est le $i^{\text{ème}}$ vecteur de la base orthonormale de $\Re^N$.

En utilisant (3.3) dans l'expression du coût, on obtient, après dérivation et en négligeant les termes d'ordre supérieur à 1 en $(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$:

$$\frac{\partial J}{\partial \boldsymbol{\theta}} = \frac{\partial J}{\partial (\boldsymbol{\theta} - \boldsymbol{\theta}^*)} \cong -2\,{}^tZ\left(\boldsymbol{y}_p - \boldsymbol{f}(x, \boldsymbol{\theta}^*)\right) + \left\{2\,{}^tZZ - 2\sum_{i=1}^{N}\left(y_p^i - f(\boldsymbol{x}^i, \boldsymbol{\theta}^*)\right) S(\boldsymbol{x}^i)\right\}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \tag{3.4}$$

Or, la matrice de dimensions (q, q) située à l'intérieur des accolades n'est autre que le Hessien de la fonction de coût, définit par $H = \left(\frac{\partial^2 J(\boldsymbol{\theta})}{\partial \theta_j \, \partial \theta_k} \right)_{\substack{j = 1, \ldots, q \\ k = 1, \ldots, q}}$. Dans la pratique, la deuxième partie du Hessien peut être négligée, ce qui conduit à l'approximation suivante, dite de Levenberg-Marquardt :

$$H \cong 2 \, {}^t Z Z \tag{3.5}$$

Dans tout ce qui suit, on se place dans le cas où la matrice Z est de rang plein, c'est-à-dire dans un cas où il n'y a ni surajustement avec déficience du rang, ni redondance de coefficients.

Finalement, en annulant le gradient de J, on obtient une approximation de $\boldsymbol{\theta}_{LS}$ sous la forme :

$$\boldsymbol{\theta}_{LS} \cong \boldsymbol{\theta}^* + ({}^t Z \, Z)^{-1} \, {}^t Z \, [\boldsymbol{y}_p - \boldsymbol{f}(X, \boldsymbol{\theta}^*)] \tag{3.6}$$

Cette expression est donc valable sous deux hypothèses :

1. le développement limité au second ordre de la fonction de coût est valable, c'est-à-dire que les termes du second ordre du développement limité de $\frac{\partial J}{\partial \boldsymbol{\theta}}$ sont négligeables par rapport aux termes du premier ordre,
2. le Hessien de la fonction de coût peut être approché par $2 \, {}^t Z Z$, comme discuté dans [Bishop 95].

Dans le cas d'un modèle linéaire par rapport à ses paramètres, la relation (3.6) n'est pas une approximation, mais une égalité. Cette approche a été utilisée par plusieurs auteurs dans des buts différents, y compris pour estimer des intervalles de confiances sur les paramètres et sur les prédictions du modèle (voir par exemple [Seber 89]).

Si ce résultat (formule 3.6) est classique, il est important d'insister sur sa démonstration, et sur les deux hypothèses permettant d'y arriver. En effet, il est possible d'arriver au même résultat en partant d'un développement de Taylor de $\boldsymbol{f}$ au premier ordre, ce qui conduit à oublier le second terme du Hessien de J au lieu de le négliger (voir [Seber 89]).

3.3 Effet du retrait d'un exemple de l'ensemble d'apprentissage

Nous allons montrer comment l'approximation de la solution des moindres carrés peut être utilisée pour prédire l'effet du retrait d'un exemple de l'ensemble d'apprentissage.

Dans tout ce qui suit, toutes les grandeurs concernant les modèles dont l'apprentissage a été réalisé à l'aide de tous les exemples sauf le $i^{\text{ème}}$ seront munies d'un exposant $^{(-i)}$. Ainsi, $\boldsymbol{f}^{(-i)}(X, \boldsymbol{\theta})$ et $\boldsymbol{y_p}^{(-i)}$ sont des vecteurs de dimension N - 1, de même que $Z^{(-i)}$ est une matrice de dimensions (N - 1, q). Inversement, toutes les grandeurs sans exposant feront référence à des modèles ajustés sur les N exemples. Par ailleurs, la différence entre la sortie mesurée et la prédiction d'un modèle sera appelée "résidu" si l'exemple correspondant fait partie de la base d'apprentissage, et "erreur" dans le cas contraire.

3.3.1 Effet du retrait d'un exemple sur sa prédiction

Si l'on suppose que le retrait de l'exemple i de la base d'apprentissage ne provoque qu'une légère modification de la solution des moindres carrés, alors on peut, de même que pour la relation (3.6), établir une expression approchée de $\boldsymbol{\theta}_{LS}^{(-i)}$ au voisinage de $\boldsymbol{\theta}^*$:

$$\boldsymbol{\theta}_{LS}^{(-i)} \cong \boldsymbol{\theta}^* + \left({}^t Z^{(-i)}\, Z^{(-i)}\right)^{-1} {}^t Z^{(-i)}\, [\boldsymbol{y}_p^{(-i)} - \boldsymbol{f}^{(-i)}(X, \boldsymbol{\theta}^*)] \tag{3.7}$$

En combinant (3.6) et (3.7), on obtient le résultat suivant (voir par exemple [Antoniadis 92]) :

$$\boldsymbol{\theta}_{LS}^{(-i)} \cong \boldsymbol{\theta}_{LS} - \left({}^t Z\, Z\right)^{-1} \boldsymbol{z}^i \frac{R_i}{1 - h_{ii}} \tag{3.8}$$

dans lequel $\boldsymbol{z}^i$ est le vecteur dont les éléments constituent la $i^{ème}$ colonne de la matrice Z, et R_i est le résidu du $i^{ème}$ exemple : $R_i = y_{pi} - f(\boldsymbol{x}^i, \boldsymbol{\theta}_{LS}) = y_{pi} - f(\boldsymbol{x}^i, \boldsymbol{\theta}^*) - {}^t\boldsymbol{z}^i\, \boldsymbol{\theta}_{LS}$ et h_{ii} est la $i^{ème}$ composante de la projection, sur le sous-espace des solutions, du vecteur unité le long de l'axe i : $h_{ii} = {}^t\boldsymbol{z}^i\, ({}^t Z\, Z)^{-1}\, \boldsymbol{z}^i$. La démonstration de la relation (3.8) est donnée en Annexe 3.

Ceci nous permet d'estimer l'erreur $R_i^{(-i)}$ sur la prédiction du $i^{ème}$ exemple quand celui-ci est retiré de la base d'apprentissage : $R_i^{(-i)} = y_{p\,i} - f(\boldsymbol{x}^i, \boldsymbol{\theta}^*) - {}^t\boldsymbol{z}^i\, \boldsymbol{\theta}_{LS}^{(-i)}$. On a donc : $R_i^{(-i)} \cong R_i + {}^t\boldsymbol{z}^i \left(\boldsymbol{\theta}_{LS} - \boldsymbol{\theta}_{LS}^{(-i)}\right)$. En utilisant la relation (3.8), on obtient la même relation que dans le cas linéaire :

$$R_i^{(-i)} \cong \frac{R_i}{1 - h_{ii}} \tag{3.9}$$

De même, en utilisant les formules précédentes, on trouve une approximation de la fonction de coût quadratique :

$$J(\boldsymbol{\theta}_{LS}^{(-i)}) \cong J(\boldsymbol{\theta}_{LS}) - \frac{R_i^2}{1 - h_{ii}} \tag{3.10}$$

Une idée analogue a été proposée par [Larsen 96] et [Sorensen 96]. Cependant, l'utilisation que ces auteurs ont faite de leur développement limité n'était pas correcte : pour s'en rendre compte, il suffit de remarquer que leurs résultats, contrairement aux formules (3.8) à (3.10), ne sont pas exacts dans le cas d'un modèle linéaire.

Les figures 3.1.b et 3.1.c permettent d'apprécier la précision des formules (3.9) et (3.10), dans le cas du modèle considéré sur la figure 3.1.a. La figure 3.1.b représente l'erreur de prédiction pour chaque exemple extrait de l'ensemble d'apprentissage, en fonction de l'erreur estimée par la relation (3.9). De même, la figure 3.1.c représente les valeurs de la fonction de coût obtenues après retrait de chaque exemple, par apprentissage et à l'aide de la relation (3.10).

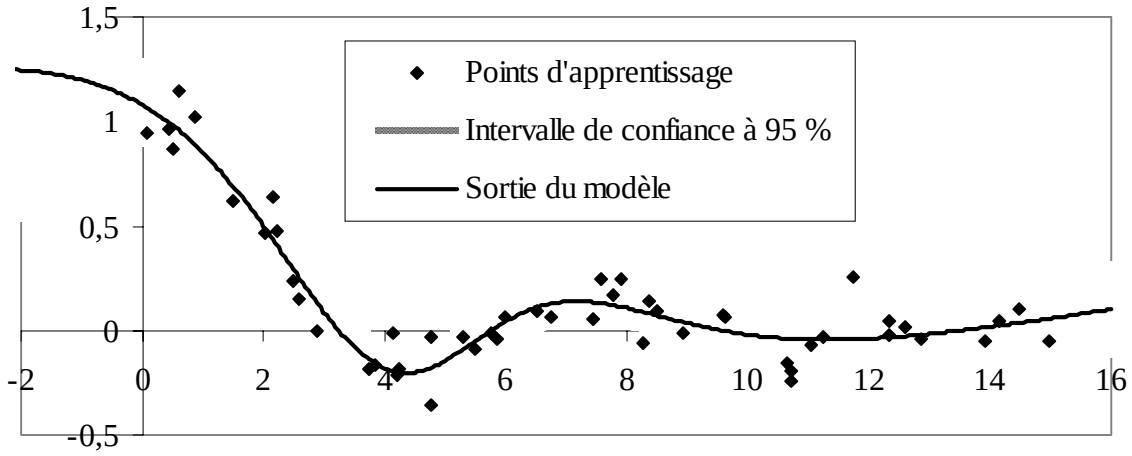

Figure 3.1.a : Ensemble d'apprentissage et modèle considéré

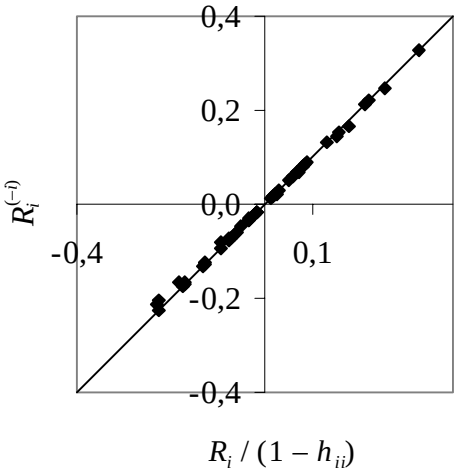

Figure 3.1.b : Erreurs sur exemple retiré, par apprentissage et formule (3.9)

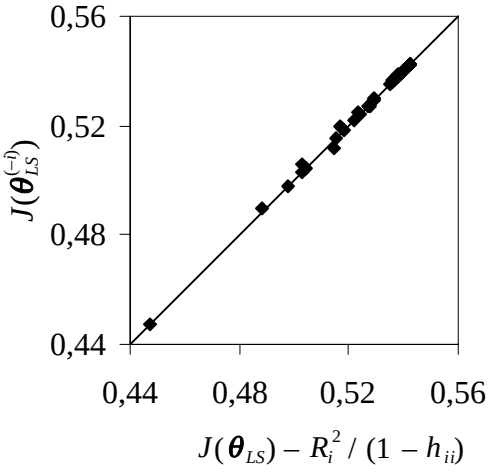

Figure 3.1.c : Coûts avec exemple retiré, par apprentissage et formule (3.10)

3.3.2 Effet du retrait d'un exemple sur l'intervalle de confiance de sa prédiction

Dans le chapitre 2, nous avons indiqué l'expression approchée d'un intervalle de confiance sur la sortie du modèle, dans l'hypothèse d'un bruit de mesure gaussien, et en supposant le modèle-hypothèse vrai. Rappelons qu'à un niveau de confiance 1 - α, l'intervalle de confiance approché pour $E(Y_p \mid \boldsymbol{x})$ est :

$$E\left(Y_p \middle| \boldsymbol{x}\right) \in f\left(\boldsymbol{x}, \boldsymbol{\theta}_{LS}\right) \pm t_\alpha^{N-q}\, s \sqrt{{}^t\boldsymbol{z}\left({}^tZZ\right)^{-1}\boldsymbol{z}}\ , \tag{3.11}$$

Pour l'exemple numéro i de la base d'apprentissage, l'intervalle de confiance précédent s'écrit donc :

$$E\left(Y_p \middle| \boldsymbol{x}^i\right) \in f\left(\boldsymbol{x}^i, \boldsymbol{\theta}_{LS}\right) \pm t_\alpha^{N-q}\, s \sqrt{h_{ii}} \tag{3.12}$$

Ainsi, d'après la propriété (2.2), la demi-largeur de l'intervalle de confiance sur les exemples d'apprentissage est inférieure à $t_{\alpha}^{N-q}\, s$.

Par ailleurs, l'expression (3.12) montre que l'approche analytique des intervalles de confiance fait intervenir la même grandeur (h_{ii}) que les grandeurs associées au retrait d'un exemple de la base d'apprentissage, ce qui est normal car les deux approches utilisent le même développement limité. C'est la raison pour laquelle nous avons choisi cette expression des intervalles de confiance. Nous verrons dans le chapitre 4 comment exploiter cette similitude dans le cadre de la sélection de modèles.

Dans le paragraphe précédent, nous avons montré que l'on peut estimer à la fois la valeur de la fonction de coût (3.10), après avoir retiré un exemple de la base d'apprentissage, et l'erreur de prédiction (3.9) sur cet exemple. De la même façon, il est possible d'estimer les intervalles de confiance sur cette prédiction.

Étant donné un vecteur d'entrée $\boldsymbol{x}^i$, l'intervalle de confiance approché pour l'espérance mathématique de $\boldsymbol{Y}_p$, avec un niveau de confiance 1 - α, est, pour le modèle obtenu après avoir supprimé l'exemple i de la base d'apprentissage :

$$E^{(-i)}\left(Y_p \left| \boldsymbol{x}^i \right.\right) \in f\left(\boldsymbol{x}^i, \boldsymbol{\theta}_{LS}^{(-i)}\right) \pm t_{\alpha}^{N-q-1}\, s^{(-i)} \sqrt{{}^t\boldsymbol{z}^i \left({}^tZ^{(-i)} Z^{(-i)} \right)^{-1} \boldsymbol{z}^i} \tag{3.13}$$

Notons $h_{ii}^{(-i)} = {}^t\boldsymbol{z}^i \left({}^tZ^{(-i)} Z^{(-i)} \right)^{-1} \boldsymbol{z}^i$. D'après le lemme d'inversion matricielle présenté en Annexe 3, on montre facilement que :

$$h_{ii}^{(-i)} \cong \frac{h_{ii}}{1 - h_{ii}}, \tag{3.14}$$

et que cette relation est une égalité dans le cas d'un modèle linéaire. En combinant (3.9), (3.13) et (3.14), on obtient l'expression suivante :

$$E^{(-i)}\left(Y_p \left| \boldsymbol{x}^i \right.\right) \in f\left(\boldsymbol{x}^i, \boldsymbol{\theta}_{LS}\right) - \frac{h_{ii}}{1 - h_{ii}} R_i \pm t_{\alpha}^{N-q-1}\, s^{(-i)} \sqrt{\frac{h_{ii}}{1 - h_{ii}}}\ . \tag{3.15}$$

Dans l'expression précédente, la seule inconnue reste l'estimation $s^{(-i)}$ de la performance de généralisation du modèle ajusté sans l'exemple i. Dans le cas où l'on utilise la valeur de J pour estimer cette performance, $s^{(-i)}$ se déduit de s par la formule (3.10). Dans les autres cas, nous supposerons que $s^{(-i)} \approx s$.

3.3.3 Interprétation des h_{ii}

En résumé, l'approximation de la solution des moindres carrés permet d'estimer l'effet du retrait d'un exemple sur toutes les grandeurs utilisées lors d'une modélisation non linéaire, notamment par réseaux de neurones, y compris sur les intervalles de confiance.

En examinant le récapitulatif du tableau 3.1, il apparaît que les $\{h_{ii}\}_{i\,=\,1,\,\ldots,\,N}$ jouent un rôle déterminant dans l'estimation de l'influence de chaque exemple sur le modèle.

Base d'apprentissage	Tous les exemples	Exemple *i* retiré
Solution des moindres carrés	$\boldsymbol{\theta}_{LS}$	$\boldsymbol{\theta}_{LS} - \left({}^t Z\, Z\right)^{-1} \boldsymbol{z}^i \frac{R_i}{1-h_{ii}}$
Coût	$J(\boldsymbol{\theta}_{LS})$	$J(\boldsymbol{\theta}_{LS}) - \frac{R_i^2}{1-h_{ii}}$
Prédiction de l'exemple *i*	$f\left(\boldsymbol{x}^i, \boldsymbol{\theta}_{LS}\right)$	$f\left(\boldsymbol{x}^i, \boldsymbol{\theta}_{LS}\right) - \frac{h_{ii}}{1-h_{ii}} R_i$
Résidu / erreur sur l'exemple *i*	R_i	$\frac{R_i}{1-h_{ii}}$
Intervalle de confiance sur la prédiction de l'exemple *i*	$\pm\, t_\alpha^{N-q}\, s \sqrt{h_{ii}}$	$\pm\, t_\alpha^{N-q-1}\, s \sqrt{\frac{h_{ii}}{1-h_{ii}}}$

Tableau 3.1 : Résumé de l'influence du retrait d'un exemple de la base d'apprentissage

D'un point de vue théorique, deux cas extrêmes sont à considérer :

- si l'axe *i* est orthogonal au sous-espace des solutions, défini par les vecteurs colonne de *Z*, tous ces vecteurs colonne ont leur $i^{\text{ème}}$ composante égale à zéro ; par conséquent $\boldsymbol{z}^i = 0$ et $h_{ii} = 0$. L'exemple *i* n'a pas d'influence sur le modèle, ce qui est confirmé par les relations (3.8) à (3.10). Rappelons que ce cas ne peut se produire que si le modèle ne possède pas de biais,
- si l'axe *i* se trouve dans le sous-espace des solutions, $h_{ii} = 1$ et $R_i = 0$. En d'autres termes, l'exemple *i* a été parfaitement appris, ce qui conduit à une indétermination dans les relations (3.8) à (3.10).

La grandeur h_{ii} apparaît ainsi comme une véritable mesure de l'influence de l'exemple *i* sur l'estimation des paramètres du modèle. Plus h_{ii} est proche de 1, plus son influence est grande : en effet, l'intervalle de confiance sur la prédiction de cet exemple, si on l'enlève de la base d'apprentissage, devient très grand, puisqu'il tend vers l'infini lorsque h_{ii} tend vers 1. Cela signifie que le modèle a utilisé certains degrés de liberté spécifiquement, de façon à s'ajuster au plus près à cet exemple (R_i très petit). Lorsque cet exemple est supprimé de la base d'apprentissage, le modèle "ne sait plus quoi faire" de ces degrés de libertés, ce qui se traduit, localement, par des intervalles de confiance très élevés.

L'indétermination provoquée par un exemple de forte influence, pour lequel h_{ii} tend vers 1 et son résidu R_i tend vers 0, peut être partiellement levée. En effet, on peut supposer - sauf cas pathologique - que le retrait de la base d'apprentissage d'un exemple à forte influence fait varier la valeur de sortie du modèle pour cet exemple. Cela signifie que le rapport $h_{ii} \frac{R_i}{1-h_{ii}}$ ne tend en général pas vers 0. On est donc sûr qu'il en va de même pour l'estimation de l'erreur de prédiction de cet exemple $R_i^{(-i)} \cong \frac{R_i}{1-h_{ii}}$. Cette remarque est très importante pour la suite car elle signifie que, pour un exemple à forte influence, le résidu R_i tend en général vers 0 **moins rapidement** que $1 - h_{ii}$.

On peut utiliser le tableau précédent de deux manières :

- ces formules peuvent servir d'outil pour valider les résultats obtenus par une procédure conventionnelle de leave-one-out (qui implique la réalisation de N apprentissages distincts si l'on dispose de N exemples) ; nous expliquerons ceci dans le paragraphe suivant,
- elles peuvent également éviter au concepteur de réaliser le leave-one-out conventionnel ; la sélection de modèles se fonde alors sur *l'estimation* de l'effet du retrait d'un exemple de la base d'apprentissage. Cette application sera détaillée dans les chapitres 4 et 5 de ce mémoire.

3.4 Validation des résultats de leave-one-out

Nous allons montrer que les résultats présentés dans le paragraphe précédent peuvent s'avérer très utiles pour vérifier la validité de l'estimation des performance obtenues en effectuant la procédure conventionnelle de leave-one-out.

Pour ce faire, rappelons que l'hypothèse implicite de la procédure de leave-one-out est que la suppression d'un exemple de la base d'apprentissage n'affecte pas de manière importante l'estimation des paramètres d'un modèle, c'est-à-dire que les solutions des moindres carrés $\boldsymbol{\theta}_{LS}^{(-i)}$ sont très proches de la solution $\boldsymbol{\theta}_{LS}$ obtenue sur l'ensemble des données disponibles ; par conséquent, tous les minima des fonctions de coût $J^{(-i)}$ - obtenus à l'issue des N apprentissages - devraient se trouver dans une petite région de l'espace des paramètres. En pratique, la validation de résultats en leave-one-out devrait obligatoirement passer par cette vérification.

Vérifier cette propriété en comparant entre elles les distances $\left\| \boldsymbol{\theta}_{LS} - \boldsymbol{\theta}_{LS}^{(-i)} \right\|_{i = 1, \ldots, N}$ nécessiterait la définition d'un seuil à partir duquel on considère que la solution $\boldsymbol{\theta}_{LS}^{(-i)}$ est trop éloignée de $\boldsymbol{\theta}_{LS}$. Une méthode plus satisfaisante consiste à vérifier qu'en remettant l'exemple i dans la base d'apprentissage, et en poursuivant l'apprentissage à partir de $\boldsymbol{\theta}_{LS}^{(-i)}$, celui-ci converge de nouveau vers $\boldsymbol{\theta}_{LS}$. Dans la suite de ce mémoire, ceci tiendra lieu de définition.

Définition :

> Un minimum $\boldsymbol{\theta}_{LS}$, de l'erreur quadratique moyenne J sur un ensemble de N observations, est dit ***propice au leave-one-out*** si et seulement si $\forall i \in [1, \ldots, N]$, la poursuite de l'apprentissage à partir de $\boldsymbol{\theta}_{LS}$, en retirant l'exemple i de la base d'apprentissage jusqu'à convergence vers un minimum $\boldsymbol{\theta}_{LS}^{(-i)}$, puis en réintégrant l'exemple i dans la base d'apprentissage, fait revenir celui-ci à $\boldsymbol{\theta}_{LS}$.

Cette définition est illustrée graphiquement sur les figures 3.2 et 3.3.

Notons dès à présent un point fondamental : il ne faut pas confondre cette définition avec les notions de stabilité introduites pour obtenir des bornes sur l'estimation des performances de généralisation (voir paragraphe 2.7). En effet, la propriété de stabilité concerne un algorithme d'apprentissage, indépendamment des données utilisées, et donc des différents minima de la fonction de coût.

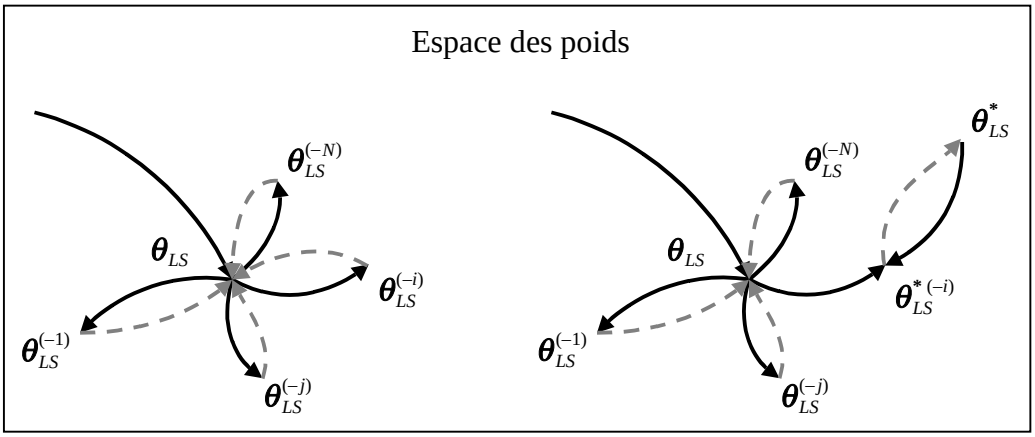

Figure 3.2 : Minima respectivement propice (à gauche) et non propice (à droite) au leave-one-out

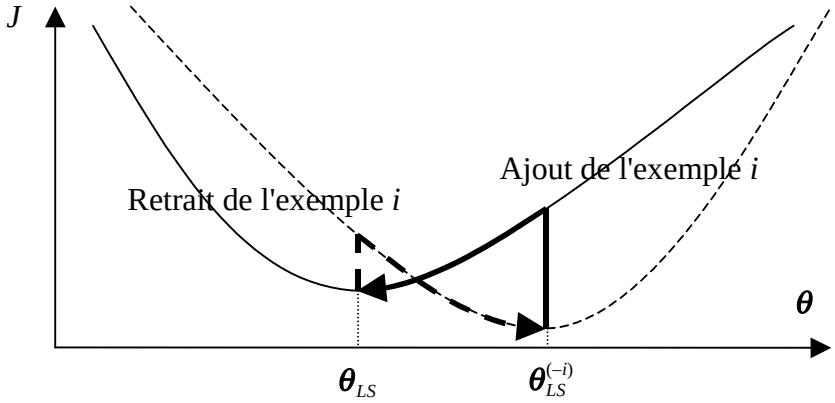

Figure 3.3 : Retrait et ajout de l'exemple i de la base d'apprentissage : cas d'un minimum propice au leave-one-out

Nous avons introduit cette définition pour déceler les minima pour lesquels il n'est pas raisonnable d'estimer des performances de leave-one-out par apprentissage. Néanmoins, ceci ne signifie pas que, pour un minimum propice au leave-one-out, l'erreur quadratique moyenne obtenue par leave-one-out est nécessairement une bonne estimation des performances de généralisation. Nous reviendrons sur ce point dans le paragraphe 4.4.2.

Nous allons montrer dans ce paragraphe :

- que tous les minima ne sont pas forcément propices au leave-one-out,
- que les inégalités (2.2) à (2.4), couplées aux formules (3.9) et (3.10), permettent en partie de s'assurer qu'un minimum est propice au leave-one-out,

3.4.1 Interprétation géométrique de l'estimation des performances en leave-one-out

Si l'on suppose que le modèle possède un terme constant, ce qui est généralement le cas, les propriétés (2.2) à (2.4) s'écrivent :

$$\frac{1}{1-h_{ii}} \geq \frac{N}{N-1} \geq 0 \tag{3.16}$$

En combinant (3.16) aux relations (3.9) et (3.10), on obtient les deux inégalités suivantes :

$$J(\boldsymbol{\theta}_{LS}^{(-i)}) \leq J(\boldsymbol{\theta}_{LS}), \tag{3.17}$$

$$\left(R_i^{(-i)}\right)^2 \geq \frac{N}{N-1}\left(J(\boldsymbol{\theta}_{LS}) - J(\boldsymbol{\theta}_{LS}^{(-i)})\right). \tag{3.18}$$

Ces deux conditions sont illustrées graphiquement sur la figure 3.4, qui représente $\left(R_i^{(-i)}\right)^2$ en fonction de $J(\boldsymbol{\theta}_{LS}^{(-i)})$. Sur un tel graphe, à partir d'un modèle dont l'apprentissage a été effectué avec l'ensemble de la base d'apprentissage, les points $\{M_i\}_{i = 1, ..., N}$, représentant les N modèles obtenus après suppression d'un exemple, devraient tous se situer dans le secteur angulaire représenté par la zone grisée.

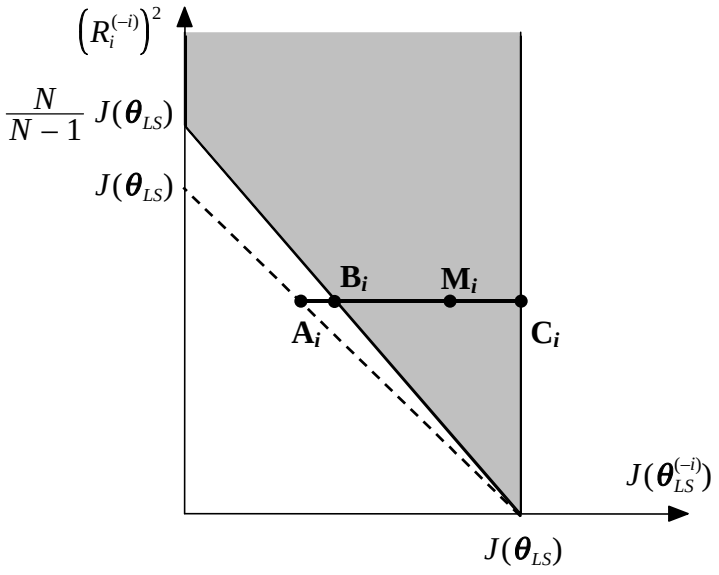

Figure 3.4 : Localisation théorique des performances en leave-one-out

Le graphique de la figure 3.4 permet également d'interpréter graphiquement les $\{h_{ii}\}_{i = 1, ..., N}$. On démontre en effet grâce à (3.9) et (3.10) la relation géométrique suivante :

$$\frac{\left|A_iM_i\right|}{\left|A_iC_i\right|} = \frac{1}{N} + \frac{\left|B_iM_i\right|}{\left|A_iC_i\right|} \cong h_{ii}. \tag{3.19}$$

En d'autres termes, h_{ii} est le rapport suivant lequel le point M_i partage le segment horizontal $[A_i\ C_i]$. Si le point M_i est à la limite gauche du secteur angulaire, c'est-à-dire confondu avec B_i, alors $h_{ii} = \frac{1}{N}$.

Nous allons montrer que la présence de tous les points obtenus après suppression d'un exemple à l'intérieur d'un tel secteur angulaire est une **condition nécessaire** pour qu'un minimum soit propice au leave-one-out. En effet, en raisonnant par l'absurde, deux cas se présentent (cf. figure 3.5) :

1. supposons qu'il existe un exemple i_0 dont les performances d'apprentissage sont situées à droite du domaine : alors l'apprentissage correspondant n'a pas convergé vers le bon minimum. En effet, si l'on remettait i_0 dans la base d'apprentissage, la minimisation de la fonction de coût conduirait forcément à une solution pour laquelle la fonction de coût serait supérieure à $J(\boldsymbol{\theta}_{LS}^{(-i_0)})$.

2. supposons qu'il existe un exemple i_1 dont les performances d'apprentissages sont situées à gauche du domaine. Cela signifie qu'il existe un autre minimum $\boldsymbol{\theta}^*_{LS}$ tel que $J(\boldsymbol{\theta}^*_{LS}) < J(\boldsymbol{\theta}_{LS})$. En effet, considérons le modèle $f(\boldsymbol{x}, \boldsymbol{\theta}^*_{LS}) = f(\boldsymbol{x}, \boldsymbol{\theta}_{LS}^{(-i_1)}) + \frac{R_{i_1}^{(-i_1)}}{N}$. On montre facilement que la fonction de coût correspondante est majorée par $J(\boldsymbol{\theta}_{LS}^{(-i_1)}) + (N-1)\left(\frac{R_{i_1}^{(-i_1)}}{N}\right)^2 + \left((1-\frac{1}{N})\, R_{i_1}^{(-i_1)}\right)^2$. Le minimum $\boldsymbol{\theta}^*_{LS}$ vérifie donc :

$$J(\boldsymbol{\theta}_{LS}^{(-i_1)}) \le J(\boldsymbol{\theta}^*_{LS}) \le J(\boldsymbol{\theta}_{LS}^{(-i_1)}) + \frac{N-1}{N}\left(R_{i_1}^{(-i_1)}\right)^2 < J(\boldsymbol{\theta}_{LS}) \,. \tag{3.20}$$

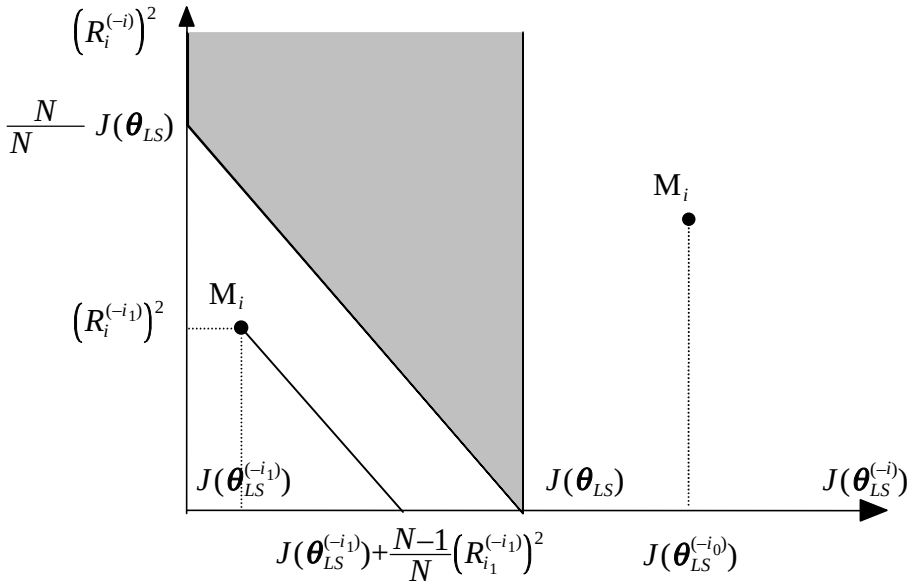

Figure 3.5 : Exemples correspondant à des performances situées en dehors du secteur angulaire

Cette interprétation géométrique montre qu'il est toujours possible, pour la meilleure solution trouvée $\boldsymbol{\theta}_{LS}$ (au sens du minimum de la fonction de coût), de placer toutes les performances obtenues après retrait d'un exemple dans le secteur angulaire de $J(\boldsymbol{\theta}_{LS})$. Il suffit pour cela de respecter les deux règles suivantes lors de l'application du leave-one-out :

Règle n°1 : démarrer les N apprentissages à partir du meilleur minimum atteint sur l'ensemble des exemples d'apprentissage, en supposant qu'il s'agit du minimum global,

Règle n°2 : vérifier - graphiquement - que les performances des N modèles obtenus se situent bien à l'intérieur d'un secteur angulaire tel que celui défini précédemment. Dans le cas contraire, c'est-à-dire si un point correspondant à un exemple i_1 se situe à gauche du secteur angulaire, redémarrer un apprentissage sur les N exemples à

partir des poids $\boldsymbol{\theta}_{LS}^{(-i_1)}$. Nous venons en effet de démontrer que cet apprentissage convergera vers un autre minimum, situé plus bas que le précédent.

Dans [Moody 94], les auteurs affirment que la première de ces deux règles est suffisante pour s'assurer de la convergence des *N* apprentissages vers le "même" minimum. Ceci est inexact car - même en respectant également la deuxième règle - rien n'assure que l'apprentissage converge de nouveau vers $\boldsymbol{\theta}_{LS}$ en remettant chaque exemple dans la base d'apprentissage.

Remarques :

- nous avons fait l'hypothèse que le modèle possède un biais. Dans le cas contraire, la seule différence réside dans la pente de la droite délimitant - à gauche - le secteur angulaire : celle-ci est égale à 1 au lieu de $\frac{N}{N}$; tout le reste du raisonnement est identique.
- la démonstration précédente suppose qu'avec les algorithmes d'apprentissage utilisés, la fonction de coût est une fonction monotone décroissante du nombre d'itérations. Ceci est le cas des algorithmes utilisés dans le cadre de ce travail (Quasi-Newton et Levenberg-Marquardt), mais exclut des méthodes du type recuit simulé.

Cette interprétation géométrique est donc une condition **nécessaire** à l'estimation de la performance de généralisation d'un modèle par la méthode du leave-one-out. Si les *N* apprentissages d'un leave-one-out n'ont pas tous convergé vers des solutions situées à l'intérieur d'un tel secteur angulaire, l'estimation correspondante des performances n'a aucun sens. Cette condition n'est cependant pas **suffisante** pour assurer que l'on considère bien des solutions ayant convergé vers le même minimum. En effet, les secteurs angulaires correspondant à deux minima distincts $\boldsymbol{\theta}_{LS}$ et $\boldsymbol{\theta}^*_{LS}$ ont toujours une intersection non vide.

3.4.2 Limite de l'approche : cas du retrait d'un exemple avec forte influence

Nous venons de montrer que, même en appliquant les deux règles précédentes de manière à ce que toutes les performances après retrait d'un exemple soient situées à l'intérieur d'un secteur angulaire, la seule façon de s'assurer qu'un minimum est propice au leave-one-out est, à partir de chaque solution $\boldsymbol{\theta}_{LS}^{(-i)}$, de remettre l'exemple *i* dans la base et de vérifier que la poursuite de l'apprentissage converge de nouveau vers $\boldsymbol{\theta}_{LS}$. Trois cas peuvent alors se présenter :

1. tous les apprentissages reviennent à $\boldsymbol{\theta}_{LS}$, auquel cas le minimum est propice au leave-one-out,
2. un apprentissage converge vers un minimum $\boldsymbol{\theta}^*_{LS}$ situé plus bas que $\boldsymbol{\theta}_{LS}$; ce dernier n'est donc pas propice au leave-one-out. Ce cas est le même que lorsque les performances correspondant au retrait d'un exemple se situaient à gauche du secteur angulaire, sauf que ceci ne peut pas se détecter graphiquement. De même qu'au paragraphe 3.4.1, il est possible de recommencer les apprentissages avec retrait d'un exemple à partir de ce nouveau minimum,

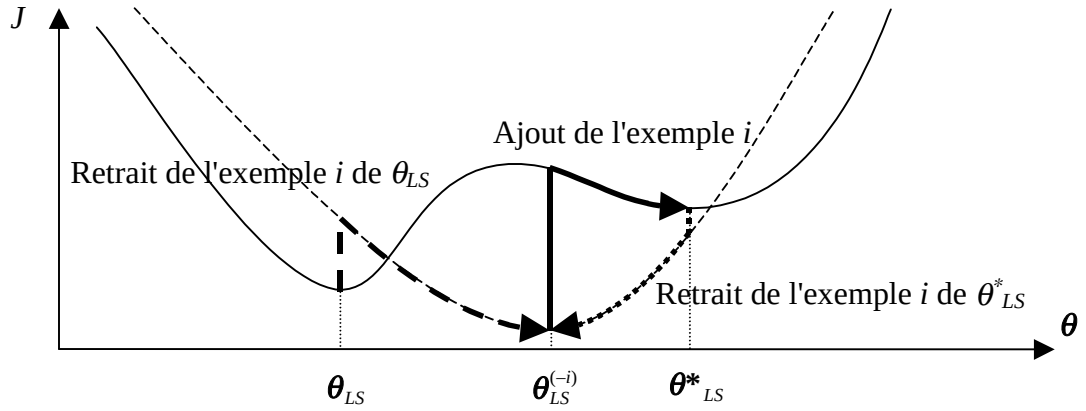

Figure 3.6 : Retrait et ajout de l'exemple i de la base d'apprentissage : cas d'un minimum non propice au leave-one-out

3. un apprentissage converge vers un minimum $\boldsymbol{\theta}^*_{LS}$ situé plus haut que $\boldsymbol{\theta}_{LS}$. Cela signifie que la solution $\boldsymbol{\theta}_{LS}$ n'existe que lorsque l'exemple en question se trouve dans la base d'apprentissage (voir figure 3.6). On ne peut donc pas connaître l'effet du retrait de l'exemple *i* sur la solution $\boldsymbol{\theta}_{LS}$ mais uniquement sur $\boldsymbol{\theta}^*_{LS}$: $\boldsymbol{\theta}^{(-i)}_{LS}$ est en réalité $\boldsymbol{\theta}^{*\,(-i)}_{LS}$.Cette solution n'est donc pas propice au leave-one-out.

Intuitivement, on peut s'attendre à rencontrer ce dernier cas plus particulièrement lors du retrait d'un exemple dont l'influence sur les poids du modèle est grande. Illustrons ceci par un exemple : celui du modèle représenté sur la figure 3.7.a. L'exemple *i* est celui dont l'influence sur les poids du modèle est la plus grande, en l'occurrence $h_{ii} = 0.944$.

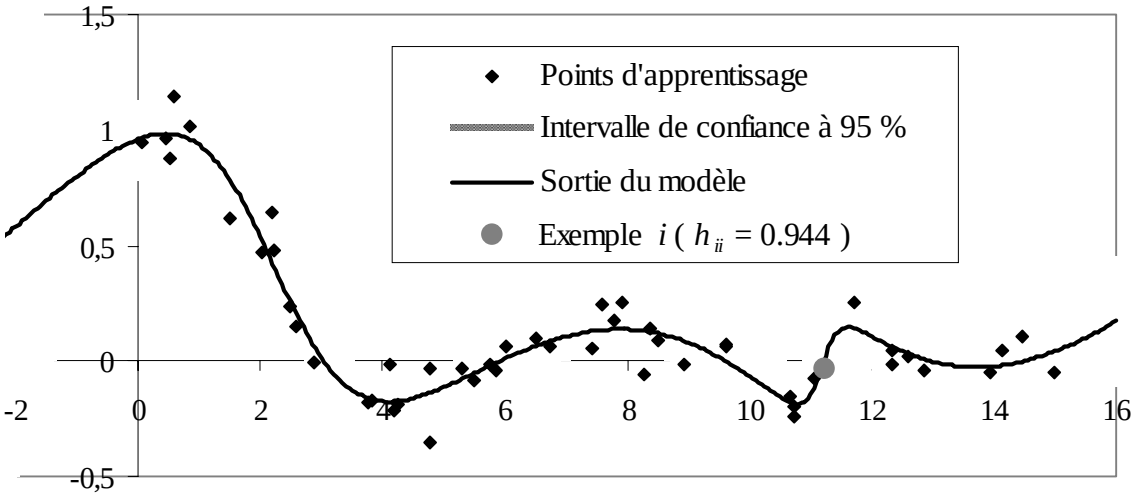

Figure 3.7.a : Exemple de modèle avec grande influence de certains points

Nous allons étudier en détail le retrait de cet exemple, suivant que ce retrait est calculé à partir des formules de linéarisation ou qu'il est effectué par la poursuite de l'apprentissage, en respectant les règles définies au paragraphe précédent.

Examinons, dans le tableau 3.2, les erreurs de prédiction sur cet exemple, obtenues respectivement par apprentissage et par utilisation de la formule (3.9) : l'erreur calculée est

environ trois fois plus grande que l'erreur constatée après apprentissage. Cette différence se reflète également sur la valeur de la fonction de coût : le coût obtenu par apprentissage est significativement plus élevé que celui prédit par la formule (3.10).

	Poursuite de l'apprentissage	Formule de linéarisation
$\left(R_i^{(-i)}\right)^2$	-0.277	-0.770
$J(\boldsymbol{\theta}_{LS}^{(-i)})$	0.358	0.333

Tableau 3.2 : Comparaison entre poursuite de l'apprentissage et utilisation des formules de linéarisation pour l'exemple i sorti

Cette différence sur les performances du modèle après retrait de l'exemple *i* est naturellement visible lorsque l'on considère (figure 3.7.b) les modèles obtenus respectivement :

1. par poursuite de l'apprentissage,
2. par linéarisation de la sortie du modèle au voisinage de $\boldsymbol{\theta}_{LS}$, c'est-à-dire :

$$f(\boldsymbol{x}, \boldsymbol{\theta}_{LS}^{(-i)}) \cong f(\boldsymbol{x}, \boldsymbol{\theta}_{LS}) + \left.\frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}_{LS}} \left(\boldsymbol{\theta}_{LS}^{(-i)} - \boldsymbol{\theta}_{LS}\right), \tag{3.21}$$

expression dans laquelle la différence $\boldsymbol{\theta}_{LS}^{(-i)} - \boldsymbol{\theta}_{LS}$ est calculée par la formule (3.8).

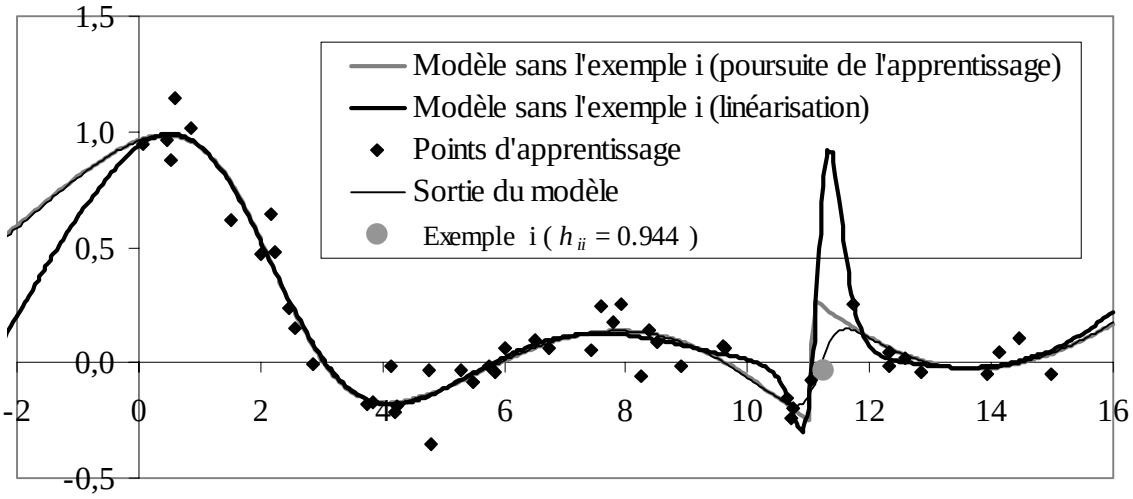

Figure 3.7.b : Effet du retrait d'un exemple à forte influence sur les modèles obtenus respectivement par apprentissage et par linéarisation

La figure 3.7.b permet de constater que la poursuite de l'apprentissage, après avoir éliminé l'exemple *i*, ne s'est traduite que par un ajustement local du modèle. En dehors de l'intervalle des entrées [10 ; 12], le modèle n'a pratiquement pas été modifié. En revanche, pour la solution obtenue par linéarisation, les modifications se situent certes au voisinage de l'exemple *i*, mais également en dehors de la plage des entrées définie par l'ensemble des points

d'apprentissage : ceci nous permet d'appréhender réellement ce qui se passe lorsqu'on "relâche" l'influence de cet exemple sur l'estimation de l'ensemble des poids du modèle.

Certes, cette différence est sans doute en partie due à une distance trop grande entre $\boldsymbol{\theta}_{LS}$ et $\boldsymbol{\theta}_{LS}^{(-i)}$, rendant imprécise l'utilisation du développement limité (3.21). Nous allons voir que ceci ne constitue pas l'explication principale des différences observées sur la figure 3.7.b.

En effet, considérons, sur la figure 3.7.c, la localisation des performances des modèles après retrait d'un exemple : celles-ci se situent effectivement dans un secteur angulaire tel que celui de la figure 3.4, ce qui est normal compte tenu du fait que nous avons procédé en respectant les règles définies au paragraphe 3.3.1.

Sur le graphique 3.7.c, nous avons en outre représenté l'effet du retour dans la base d'apprentissage de l'exemple qui en avait été retiré : ceci est illustré graphiquement par un trait horizontal qui part du point correspondant à l'exemple et dont la longueur est égale à $\left| J(\boldsymbol{\theta}_{LS}^{*}) - J(\boldsymbol{\theta}_{LS}^{(-i)}) \right|$. Ainsi, la valeur de la fonction de coût, pour le minimum atteint après retour d'un exemple dans la base d'apprentissage, se lit en abaissant la parallèle à l'axe des ordonnées passant par l'extrémité droite du trait. Pour des raisons de lisibilité du graphique, ce trait n'est représenté que dans le cas où le minimum atteint lors du retour diffère du minimum correspondant au secteur angulaire.

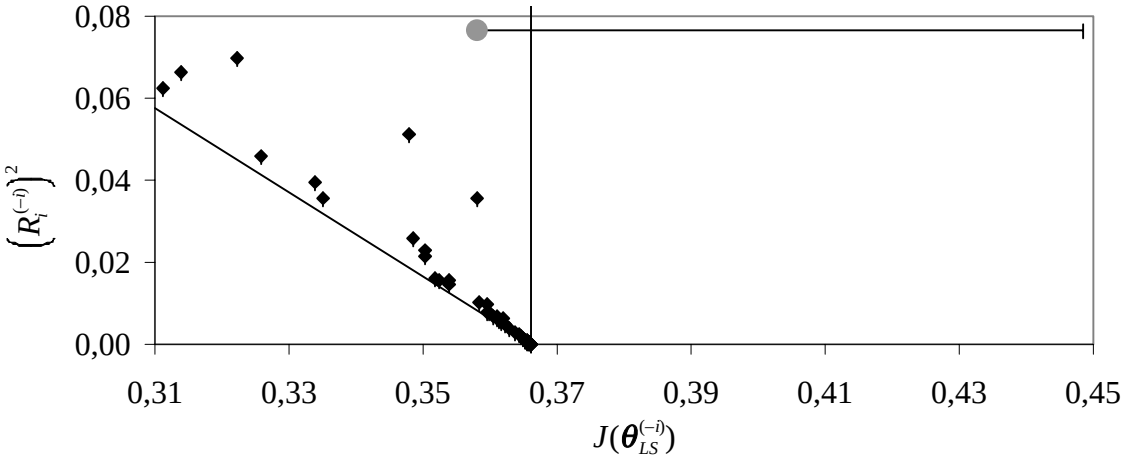

Figure 3.7.c : Secteur angulaire et localisation des performances des N modèles après apprentissage

Dans le cas du modèle précédent, il apparaît qu'il n'y a qu'un seul point dont le retour dans la base d'apprentissage provoque une convergence vers un autre minimum situé plus haut que le précédent. Cela signifie que cette solution n'est pas propice au leave-one-out. Or il s'avère que ce point correspond justement à l'exemple *i*, dont l'influence sur l'estimation des coefficients du modèle est maximale et pour lequel nous avions mentionné les différences dans le tableau 3.2.

Il apparaît ainsi que les différences entre les erreurs de prédiction d'un exemple à forte influence, lorsque celui-ci n'est pas dans la base d'apprentissage, ne suffisent pas remettre en cause la précision des formules de linéarisation (3.9) et (3.10). En effet, si le retrait d'un tel

exemple fait converger la minimisation du coût quadratique vers un autre minimum (local), les deux approches ne peuvent pas être comparées l'une à l'autre.

Cet exemple nous a permis de mettre en évidence l'existence de minima non propices au leave-one-out, ce qui se traduit par des différences entre les modèles obtenus après suppression d'un exemple, suivant que l'on supprime cette influence à partir des formules de prédiction ou par poursuite de l'apprentissage, même en validant ce dernier par le critère du secteur angulaire. Notre expérience montre que ceci se produit très souvent lors du retrait d'exemples qui ont une forte influence sur l'estimation des coefficients du modèle.

3.5 Conclusion

Nous avons montré qu'un développement de Taylor du premier ordre au voisinage de la solution des moindres carrés permet d'estimer l'effet du retrait d'un exemple à la fois sur la sortie du modèle (et donc sur son erreur de prédiction donnée par la formule (3.9)) et sur l'intervalle de confiance sur la sortie du modèle (3.15).

Ces estimations sont toutes fondées sur la grandeur h_{ii}, c'est-à-dire sur le terme diagonal de la matrice de projection orthogonale sur le sous-espace des solutions, qui n'est défini que lorsque le rang de la matrice jacobienne Z est égal au nombre de paramètres ajustables du modèle. Nous avons en outre interprété h_{ii} comme une mesure de l'influence de l'exemple i sur l'estimation des paramètres du modèle.

Afin de pouvoir comparer les résultats obtenus par apprentissage et par utilisation des formules fondées sur le développement de Taylor, nous avons montré qu'en dépit du respect d'une procédure adéquate, certains minima de la fonction de coût quadratique n'étaient pas propices au leave-one-out. Il s'agit des minima qui ne sont pas conservés par le retrait d'un exemple à forte influence. Pour ces minima, l'utilisation des formules de linéarisation semble être le seul moyen d'estimer l'effet du retrait d'un tel exemple sur les paramètres du modèle.

Nous verrons dans le chapitre 4 comment tout ceci doit être pris en considération lors de la sélection de modèles neuronaux.