Régression régularisée par contrainte de variabilité

Youness NAJI, Laurent LE BRUSQUET, Gilles FLEURY

École Supérieure d'Électricité – Service des Mesures
3 rue Joliot Curie, Plateau du Moulon, 91192 Gif-sur-Yvette, France
youness.naji@supelec.fr, laurent.lebrusquet@supelec.fr, gilles.fleury@supelec.fr

Résumé – La variabilité d'une base d'apprentissage est définie. Elle permet, pour les problèmes de régression, la construction d'a priori sur la solution cherchée : en imposant à la solution cherchée d'avoir une variabilité proche de celle de la base d'apprentissage, on augmente sa robustesse.
L'approche est illustrée sur un exemple de régression paramétrique (régression polynomiale) et sur un exemple de régression non-paramétrique (approximation par noyaux). Dans le premier exemple, le risque d'overfitting est contenu, ce qui permet d'explorer un vaste espace de solutions potentielles (régression polynomiale d'ordre élevé) alors que dans le deuxième exemple, l'approche donne une estimation de la largeur du noyau proche de la valeur optimale.

Abstract – The variability of a training set is defined. It is used for the construction of prior information in the context of regression problems: By forcing the variability of the searched solution to be close to the variability of the training set, the robustness is improved.
The method is applied to a parametric regression problem (polynomial regression) and to a non-parametric regression problem (kernel approximation). In the first problem, the risk of overfitting is avoided. This allows the definition of a large space for the searched solution (polynomial regression with a high order). In the second problem, the method gives an estimation of the kernel spread close to the optimal value.

1 Introduction

La démarche présentée vise à améliorer la robustesse des méthodes utilisées en régression ou en apprentissage [1] : soit un processus inconnu : $y \mapsto z = f^*(y)$ que l'on cherche à estimer à partir d'un ensemble de données (base d'apprentissage) $D_n = \{(y_i,\ z_i)_{i=1\cdots n}\}$ issu du processus f^* :

$$z_i = f^*(y_i) + \epsilon_i, \quad i = 1 \cdots n \tag{1}$$

où les ϵ_i sont supposés indépendants de f^*.
Les mesures indirectes s'inscrivent dans cette problématique : la relation permettant d'exprimer la grandeur d'intérêt z (mesure non accessible directement) à partir des observées y est identifiée à partir d'un ensemble d'expériences pour lesquelles on connaît à la fois des signaux d'intérêt et les observées correspondantes.

Le contexte des méthodes paramétrique et non-paramétrique de régression est étudié par la suite :

- Les méthodes paramétriques de régression consistent à modéliser $f^*(y)$ par une fonction paramétrée $f_{\boldsymbol{\theta}}(y)$ dont les paramètres $\boldsymbol{\theta}$ sont estimés à partir de la base d'apprentissage D_n :
$$\hat{\boldsymbol{\theta}}(D_n) = \arg\min_{\boldsymbol{\theta}\in\Theta} J\left(f_{\boldsymbol{\theta}},\ D_n\right) \tag{2}$$
J peut par exemple être le coût quadratique entre les z_i mesurés et leurs valeurs issues du modèle.
- Les méthodes non-paramétriques de régression approchent la fonction $f^*(y)$ par une fonction $f^{(D_n)}(y)$ calculée directement à partir de D_n.
Dans la suite, $f^{(D_n)}(y)$ dépend de paramètres de réglage $\boldsymbol{\theta}$ pouvant être optimisés :
$$\hat{\boldsymbol{\theta}}(D_n) = \arg\min_{\boldsymbol{\theta}\in\Theta} J\left(f_{\boldsymbol{\theta}}^{(D_n)},\ D_n\right) \tag{3}$$

Qu'elles soit paramétriques ou non, les méthodes de régression sont susceptibles de fournir des solutions manquant de robustesse (problème d'*overfitting*). C'est particulièrement vrai lorsque le nombre d'exemples dans la base d'apprentissage est faible [2]. Ce problème peut-être résolu de deux manières :

- se restreindre à des modélisations "douces" [3] (peu de paramètres). Cela peut priver l'espace des solutions potentielles de solutions intéressantes.
- régulariser le traitement pour en améliorer les performances [4]. En effet, si les solutions possibles sont susceptibles d'avoir de fortes oscillations, il faut contraindre leur régularité à être aussi élevée que celle de la base d'apprentissage. C'est dans cette voie que s'inscrit ce travail.

2 Approche proposée

2.1 Principe de régularisation

Étant donnée une base d'apprentissage, la meilleure exploitation de l'information qu'elle contient est celle qui conduit à une solution robuste (qui présente de bonnes qualités de généralisation). La définition de critères associés à la base d'apprentissage issue d'une prise d'observations a un objectif double :

1. introduire la régularisation nécessaire pour aboutir à une solution robuste,
2. prévoir la qualité de la solution apprise sur cette base.

Seul le premier point est ici abordé.
La régularisation proposée consiste à ne considérer que des modèles d'ajustement ($f_{\boldsymbol{\theta}}^{(D_n)}$ ou $f_{\boldsymbol{\theta}}$, notés indifféremment dans la suite $f_{\boldsymbol{\theta}}$) plus réguliers que les bases d'apprentissage ou de régularité équivalente.

La régularité d'une base d'apprentissage (resp. modèle) peut être perçue sous l'angle de sa variabilité : une base d'apprentissage (resp. modèle) est régulière si elle a une faible variabilité. Les problèmes d'optimisation (2) et (3) deviennent alors :

$$\begin{aligned}\hat{\boldsymbol{\theta}}(D_n) &= \arg\min_{\boldsymbol{\theta}\in\Theta} J_{ad}\left(f_{\boldsymbol{\theta}}, D_n\right) \\ &\quad \text{S. C. variabilité}(f_{\boldsymbol{\theta}}) \leq \text{variabilité}(D_n) \quad (4)\end{aligned}$$

où J_{ad} est un critère quantifiant l'adéquation entre les données observées et les données modélisées.
Les critères de variabilité d'une base d'apprentissage devront, dans la mesure du possible, ne prendre en compte que les variations dues au signal utile et donc être peu sensibles au bruit.

2.2 Critères de variabilité

Seul le cas de grandeurs y et z scalaires est dans la suite exposé. La variabilité d'une fonction $g : [a, b] \to \mathbb{R}$ a été définie en utilisant un critère lié à ses fluctuations :

$$V_{fc}(g) = \int_a^b \left(\frac{dg(y)}{dy}\right)^2 dy \quad (5)$$

Le critère de variabilité proposé pour les bases d'apprentissage tente d'approcher la variabilité définie précédemment :

$$V_{BA}(D_n) = \sum_{i=1}^{n-1} \frac{(z_{i+1} - z_i)^2}{(y_{i+1} - y_i)^2} \quad (6)$$

Le caractère quadratique de l'estimateur V_{BA} lui confère l'avantage d'avoir un biais et une variance indépendants du problème traité. Dans le cas de données d'apprentissage (y_i, z_i) régulièrement espacées (y_i issus d'un échantillonnage uniforme de $[a, b]$) et bruitées par un bruit blanc de variance σ_ϵ^2 (cf. équation (1)), un calcul d'espérance conduit à :

$$\begin{cases} \mathrm{E}_\epsilon\left[V_{BA}\right] = V_{fc}(f^*) + 2(n-1)\sigma_\epsilon^2 \quad [n \to \infty] \\ \text{var}(V_{BA}) \approx \sigma_\epsilon^4(12n^3 - 4n^2) \end{cases} \quad (7)$$

ce qui permet de corriger le biais de l'estimateur :

$$V_{BA}(D_n) = \sum_{i=1}^{n-1} \frac{(z_{i+1} - z_i)^2}{(y_{i+1} - y_i)^2} - 2(n-1)\sigma_\epsilon^2 \quad (8)$$

Une correction prenant en compte les valeurs des y_i peut également être obtenue dans le cas d'un échantillonnage quelconque de $[a, b]$.

2.3 Mise en œuvre

La caractérisation statistique de l'estimateur V_{BA}, associé à l'hypothèse de normalité des ϵ_i permet, via un maximum de vraisemblance, de modifier le problème d'optimisation (4), et ainsi d'obtenir un estimateur de $\boldsymbol{\theta}$ moins sensible aux variations de l'estimateur V_{BA}. On obtient :

$$\begin{aligned}\hat{\boldsymbol{\theta}}(D_n) &= \arg\min_{\boldsymbol{\theta}\in\Theta} \textstyle\sum_{i=1}^n (z_i - f_{\boldsymbol{\theta}}(y_i))^2 + \cdots \\ &\quad \cdots + \frac{\sigma_\epsilon^2}{\text{var}(V_{BA})}\left(V_{fc}(f_{\boldsymbol{\theta}})) - V_{BA}(D_n)\right)^2 \quad (9)\end{aligned}$$

Contraindre la variabilité de la solution cherchée à un a priori construit à partir de la base d'apprentissage est une approche générale permettant des extensions :

1. grandeurs y de dimension supérieure à 1,
2. critères de variabilité définis différemment, en particulier des critères liés à l'occupation spectrale,
3. critères de variabilité locale (dépendants de y) afin de restreindre localement le risque de sur-apprentissage (utiles si la fonction f^* est oscillante dans un domaine et plus régulière ailleurs).

3 Applications

Le critère de variabilité proposé a été utilisé pour l'approximation de la fonction $f^* : [0, 1] \to \mathbb{R}$ définie par :

$$z = f^*(y) = \frac{1}{2}\sin(4\pi y) + \frac{1}{2}\sin\left(\frac{20\pi y}{3}\right) \quad (10)$$

Les données d'apprentissage (y_i, z_i) ont été générées selon (1) avec une valeur de σ_ϵ conduisant à un RSB de 10dB.

3.1 Régression paramétrique

On considère ici la classe $\{f_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\mathbb{R}^{d+1}}$ des polynômes de degré d :

$$f_{\boldsymbol{\theta}}(y) = P_{\boldsymbol{\theta}}(y) = \sum_{k=0}^{d} \theta_k y^k \quad (11)$$

Le polynôme obtenu par la relation (9) a été comparé au polynôme d'ordre d' obtenu sans contrainte de variabilité. d' a été calculé avec la méthode MDL (Maximum Description Length [5]) :

$$d' = \arg\min_d \left[\sum_{i=1}^{n}\left(z_i - f_\theta^{(d)}(x_i)\right)^2 + (d+1)\frac{log(n)}{n}\right] \quad (12)$$

La figure 1 compare les deux estimateurs pour une trajectoire du bruit $\{\epsilon_i\}_{i=1\cdots n}$, ainsi que les erreurs de généralisation E_g :

$$E_g\left(f_{\hat{\boldsymbol{\theta}}}\right) = \int_0^1 \left(f_{\hat{\boldsymbol{\theta}}} - f^*(y)\right)^2 dy \quad (13)$$

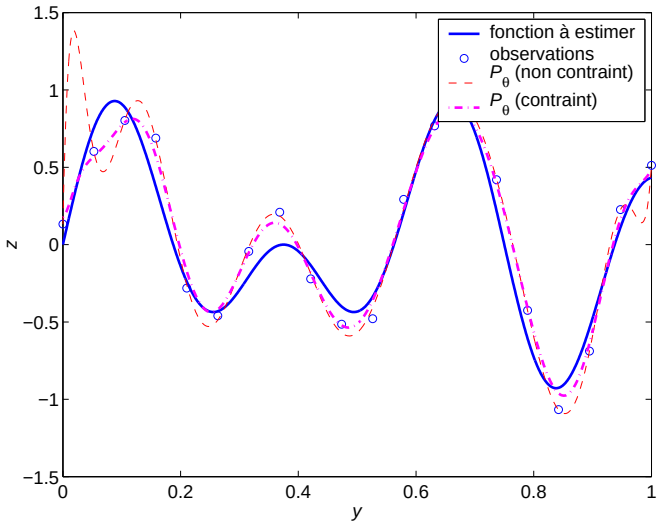

FIG. 1: Exemple de résultat (avec $n = 20$, $d = 16$). Pour le polynôme contraint : $E_g = 0,0126$. Pour le polynôme non contraint : $E_g = 0,0518$ ($d' = 13$ - estimation MDL).

La figure 2 et le tableau 1 donnent les résultats obtenus pour 1000 trajectoires de $\{\epsilon_i\}_{i=1\cdots n}$. L'estimateur régularisé est plus robuste que l'estimateur polynomial obtenu par optimisation du critère MDL.

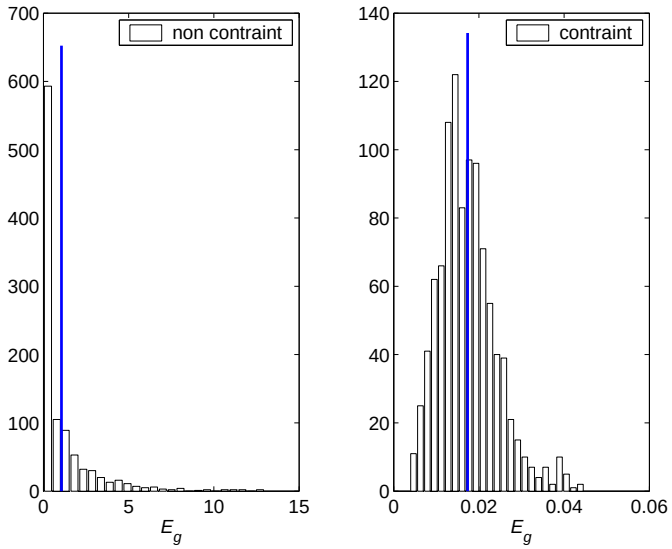

FIG. 2: Histogrammes des erreurs de généralisation dans le cas de la régression polynomiale (1000 simulations).

TAB. 1: Erreurs de généralisation moyennes.

	contraint	non contraint
E_g	0,0172	1,03

3.2 Régression par noyaux

L'estimateur non paramétrique de Nadaraya-Watson [6][7] a été utilisé pour approximer la fonction $f^*(y)$:

$$f_h^{NW(D_n)}(y) = \frac{\sum_{i=1}^n z_i K\left(\frac{y-y_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{y-y_i}{h}\right)} \qquad (14)$$

où le noyau $K(y)$ a été choisi gaussien : $K(y) = \exp\left(-\frac{y^2}{2}\right)$.
Le coefficient h réalise un compromis entre un estimateur robuste (h grand) et un estimateur non-biaisé (h petit). Il a été estimé en contraignant la variabilité de la fonction $f_h^{NW(D_n)}$ (équation (9)).
L'estimateur ainsi obtenu a été comparé à l'estimateur obtenu avec la valeur optimale de h :

$$h_{opt} = \arg\min_h E_g\left(f_h^{NW(D_n)}\right) \qquad (15)$$

La figure 3 montre sur l'exemple de la figure 1 que les deux critères (9) et (15) ont leur minimum pour des valeurs de h proches l'une de l'autre. Cela conduit naturellement à estimateur de Nadaraya-Watson proche de l'estimateur optimal, comme le montre la figure 4. Les courbes en pointillés sont les estimateurs que l'on aurait obtenus pour des valeurs arbitraires de h ($h_1 = 0,01$ et $h_2 = 0,05$).

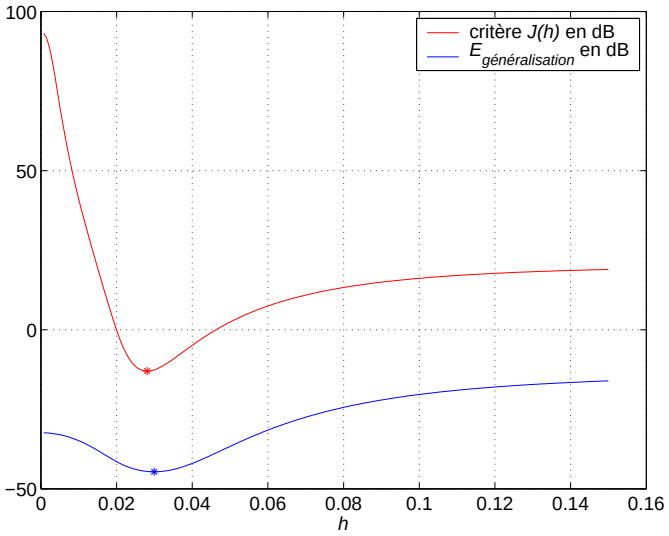

FIG. 3: Critères (9) et (15) pour les points (y_i, z_i) de la figure 1. $h_{est} = 0,0281$. $h_{opt} = 0,0299$.

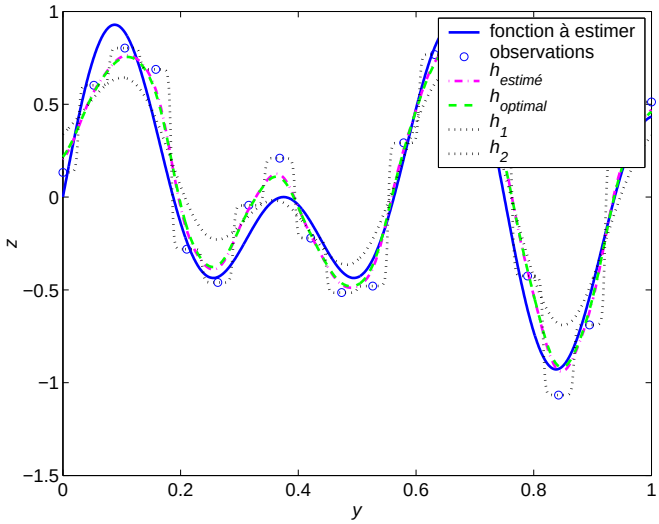

FIG. 4: Résulats avec les données de Fig. 1. Pour $h = h_{opt}$: $E_g = 0,0183$. Pour $h = h_{est}$: $E_g = 0,0190$. Pour $h = h_1$: $E_g = 0,0377$. Pour $h = h_2$: $E_g = 0,0366$.

La figure 5 et le tableau 2 donnent les erreurs de généralisation obtenues pour 1000 simulations avec les 4 valeurs de h. L'estimateur proposé a un comportement proche de l'estimateur obtenu avec le noyau de largeur optimale.

TAB. 2: Erreurs de généralisation moyennes.

	h_{otp}	h_{est}	h_1	h_2
E_g	0,0183	0,0190	0,0377	0,0366

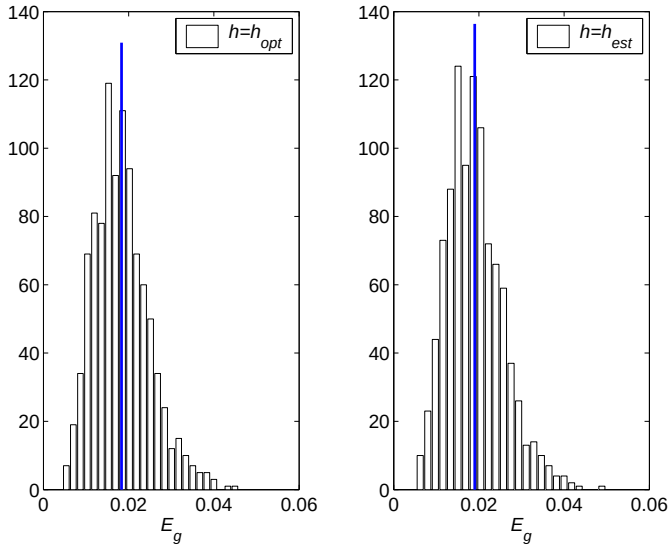

FIG. 5: Estimateur de Nadaraya-Watson : histogrammes des erreurs de généralisation pour $h = h_{opt}$ et $h = h_{est}$.

Ce résultat est confirmé par l'histogramme de l'erreur d'estimation de h_{opt} de la figure 6 et par le calcul de l'erreur moyenne :

$$\left\langle \frac{|h_{est} - h_{opt}|}{h_{opt}} \right\rangle = 9.5\% \qquad (16)$$

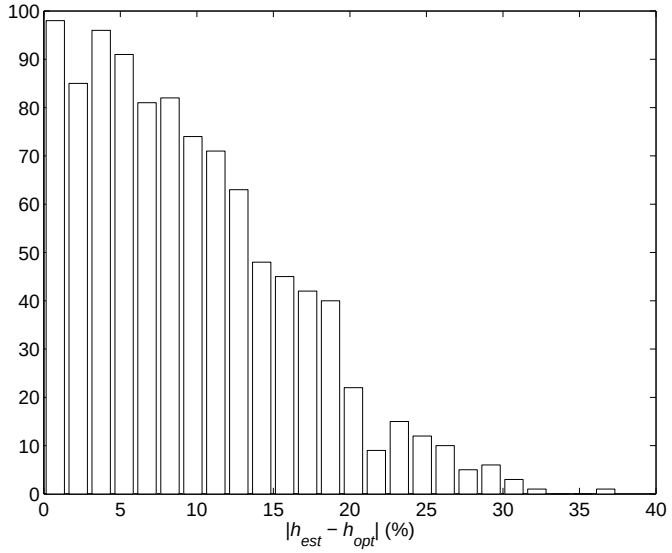

FIG. 6: Histogramme des erreurs sur l'estimation de h_{opt}

4 Conclusion

L'approche proposée dans cet article s'inscrit dans le cadre des méthodes de régression où l'on cherche une solution à partir d'un ensemble de données (base d'apprentissage). L'idée émise est de régulariser le terme d'adéquation aux données par un a priori calculé à partir des points de la base d'apprentissage.

On extrait ainsi de la base d'apprentissage des informations non exploitées par le critère d'adéquation aux données.

Dans cet article, la variabilité de la base d'apprentissage (information liée à sa régularité) est définie et sert de contrainte à la solution cherchée : en effet, asservir la variabilité de la solution cherchée à une valeur proche de celle de la base d'apprentissage permet d'obtenir une solution compatible avec la richesse de la base d'apprentissage, et ainsi limite le risque d'*overfitting*.

La démarche proposée, appliquée à des cas simples de régression scalaire paramétrique (approximation polynomiale) et non-paramétrique (approximation par noyaux), a abouti à des solutions robustes. Elle peut être étendue à d'autres méthodes de régression, y compris dans le cas non-scalaire.

Des travaux en cours visent à construire de nouveaux estimateurs de la variabilité, à la fois par réduction de la variance de l'estimateur actuel, et par définition de nouvelles grandeurs définies dans le domaine spectral.

Références

[1] V. Vapnik. *An overview of statistical learning theory*. IEEE Transactions on Neural Networks, Vol 10, No 5, sep-1999.

[2] C. Schaffer. *Overfitting Avoidance as Bias*. Machine Learning, Vol 10, No 2, pp 153-78, fev 1993.

[3] M. Bekara, A. K. Seghouane et G. Fleury. *A small sample model selection criterion based on the kullback symetric divergence*. IEEE Internationnal Conference on acoustic Speech and Signal processing, 2003.

[4] D. Schuurmans et F. Southey. *An adaptive Regularisation Criterion for Supervised learning*. Proceedings of 17^{th} Int. Conf. on Machine Learning, (ICML-2000), Stanford, CA, juin 2000.

[5] M .B. Vitanyi et M. Li. *Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity*. IEEE Trans. Information Theory , IT-46:2, pp 446-464, 2000.

[6] E. A. Nadaraya. *On estimating regression*. Theory of probability and application, Vol. 10, pp 186-190, 1964.

[7] G. S. Watson. *Smooth regression analysis*. Sanklya Series, A26, pp 359-372, 1964.