Choix d'un noyau pour la régression à vecteurs de support par analyse structurelle; application à la régression multivariable.

Emmanuel Vazquez, Eric Walter

Laboratoire des Signaux et Systèmes,
CNRS – Supélec – Univ. Paris-Sud
91192 Gif-sur-Yvette, France
vazquez@lss.supelec.fr, walter@lss.supelec.fr

Résumé – Le choix du noyau pour une application spécifique est une étape importante dans les méthodes à noyaux reproduisants (SVM ou SVR, processus gaussiens, RBF, ou splines). Nous souhaitons montrer que la théorie du krigeage, issue de la géostatistique, forme un cadre utile pour aider l'utilisateur dans ce choix. Grâce à ce point de vue, nous montrons comment étendre les SVR au cas multivariable lorsque les sorties sont corrélées.

Abstract – The choice of a kernel for a specific application is an important step in reproducing kernel methods (such as SVM or SVR, Gaussian processes, RBF, or splines). We would like to show that the theory of Kriging, developped in Geostatistics, is useful to assist the user in this choice. This viewpoint makes it easy to extend SVR to multi-output regression with correlated outputs.

1 Introduction

L'intérêt des méthodes à vecteurs de support (SVM et SVR, pour *support vector machines* et *support vector regression*) est souvent justifié par le fait qu'elles s'appuient sur la théorie des espaces de Hilbert à noyaux reproduisants ; or le choix d'un noyau pour une application spécifique est très rarement argumenté. La plupart du temps, les noyaux sont choisis a priori, de type radial gaussien par exemple, indépendamment de la nature des données observées, et le seul problème habituellement considéré consiste à adapter la « largeur » de la gaussienne. Dans le cadre de la théorie de la prédiction linéaire des processus aléatoires, le noyau est interprété comme une fonction de covariance. Nous nous inspirons de la géostatistique et des statistiques spatiales, où la prédiction linéaire, appelée *krigeage* (*Kriging* en anglais), est précédée d'une phase dite d'*analyse structurelle*, qui consiste à choisir une fonction de covariance (le noyau reproduisant) adaptée aux données observées. Les géostatistitiens ont une longue expérience du traitement de données expérimentales et ont largement exploré ce domaine [1]. Notons pour l'anecdote que le krigeage, développé par le mathématicien français Georges Matheron dans les années 60 à partir des travaux de Krige, a été réintroduit dans les années 90 sous le nom de *théorie des processus gaussiens* (voir [8] par exemple). Nous verrons aussi qu'à partir de ce point de vue il est très facile d'étendre les méthodes à vecteurs de support au cas de la prédiction de processus à *plusieurs sorties corrélées*.

2 La covariance comme noyau reproduisant

La modélisation de type boîte noire cherche à prédire un phénomène ou la sortie d'un système à partir d'un nombre fini d'observations du vecteur de ses entrées, ou facteurs, $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \cdots, n$ et de ses sorties $\mathbf{f}(\mathbf{x}_i) \in \mathbb{R}^q$. D'un point de vue mathématique, il s'agit d'un problème d'approximation ou d'interpolation. Pour le moment, fixons $q = 1$, ce qui correspond à une sortie scalaire. Soit $F(\mathbf{x})$, un processus aléatoire gaussien de moyenne nulle, fonction du vecteur de facteurs $\mathbf{x} \in \mathbb{R}^d$. $F(\mathbf{x})$ prend ses valeurs dans un espace de variables aléatoires sur un espace probabilisé, muni du produit scalaire $(X, Y) = \mathsf{E}[XY]$. Dans ce cas simple, le krigeage cherche la meilleure prédiction linéaire de $F(\mathbf{x})$, c'est-à-dire une combinaison linéaire

$$\hat{F}(\mathbf{x}) = \sum_{i=1}^{n} \lambda_{i,\mathbf{x}} F(\mathbf{x}_i) \tag{1}$$

minimisant l'erreur quadratique moyenne $||\hat{F}(\mathbf{x}) - F(\mathbf{x})||$, où la norme est déduite du produit scalaire. Dans ce contexte, on utilise la fonction $k(\mathbf{x}, \mathbf{y})$, covariance de $F(\mathbf{x})$ et $F(\mathbf{y})$, qui décrit comment deux valeurs de la sortie du processus à modéliser sont corrélées, en fonction des valeurs $\mathbf{x}$ et $\mathbf{y}$ prises par le vecteur des facteurs. La quantité $k(\mathbf{x}, \mathbf{y})$ peut être vue comme un produit scalaire dans l'espace de Hilbert généré par $F(\mathbf{x})$.

Le lien entre covariance et noyau reproduisant est bien connu aujourd'hui notamment dans le cadre de la théorie des splines [6, 7], mais cela concerne toutes les méthodes

fondées sur les noyaux reproduisants comme les réseaux de fonctions de base radiales (RBF) et les SVM. Il existe en effet une isométrie bijective entre un espace de Hilbert à noyau reproduisant et l'espace de Hilbert engendré par $F(\mathbf{x})$, les deux espaces étant munis du produit scalaire déduit de $k(\mathbf{x}, \mathbf{y})$.

3 Ce qu'apporte le krigeage

3.1 Adaptation du noyau aux données

La covariance permet de choisir une distance naturelle sur l'espace des observations. Dans le cas d'une fonction de covariance radiale $k(\mathbf{x}, \mathbf{y}) = r(||\mathbf{x} - \mathbf{y}||)$, deux valeurs observées de la sortie, $f(\mathbf{x}_i)$ et $f(\mathbf{x}_j)$, modélisées par les variables aléatoires $F(\mathbf{x}_i)$ et $F(\mathbf{x}_j)$, seront proches si la distance dans l'espace des facteurs $||\mathbf{x}_i - \mathbf{x}_j||$ est petite. Dans certaines applications où le processus est un phénomène physique, il est possible d'accéder à sa covariance théorique. L'analyse spectrale peut aussi apporter beaucoup de renseignements (rappelons que pour un processus stationnaire, la densité spectrale est la transformée de Fourier de la covariance). Dans les nombreux cas où la covariance n'est pas connue, il reste la possibilité de l'estimer à partir des données, en intégrant éventuellement les informations partielles dont on dispose. C'est l'objet de l'*analyse structurelle*, qui consiste à choisir une covariance adaptée aux données. Les ouvrages de référence de géostatistique comme [1] accordent beaucoup d'importance à cette étape de modélisation.

L'hypothèse la plus courante est de considérer le phénomène étudié comme stationnaire. Sa covariance est alors invariante par translation. De plus, si la structure de corrélation ne dépend que de la distance euclidienne dans l'espace des facteurs, le phénomène sera isotrope. Lorsque ce n'est pas le cas, l'espace des facteurs peut être modifié par une transformation linéaire pour tenter de se rapprocher du cas isotrope. Bien sûr, ces choix ne sont pas arbitraires et nous insistons sur le fait que l'analyse des données combinée à l'exploitation de l'information a priori sur le processus doit permettre un choix rationnel des hypothèses sur le noyau.

Un choix important en pratique est celui de la régularité. La régularité de la covariance à l'origine (par exemple, sa continuité, sa dérivabilité, etc) fixe la régularité du processus aléatoire $F(\mathbf{x})$. Ainsi, si la covariance admet des dérivées jusqu'à l'ordre deux, le processus aléatoire sera différentiable (en moyenne quadratique). L'utilisation d'un noyau gaussien entraîne que le processus aléatoire est indéfiniment dérivable et doit donc être réservée aux phénomènes physiques admettant des dérivées à tout ordre. Cette analyse est formalisée dans [5], où Stein montre dans un cadre asymptotique qu'il est important d'estimer avec précision la régularité de la covariance. Il plaide pour l'utilisation du *noyau de Matérn*,

$$r(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}h}{\rho} \right)^{\nu} \mathcal{K}_\nu \left(\frac{2\nu^{1/2}h}{\rho} \right) , \qquad (2)$$

qui est un modèle isotrope admissible (pour la condition de positivité de la covariance) pour toute dimension d de l'espace des facteurs. $\mathcal{K}_\nu(.)$ désigne la fonction de Bessel modifiée de seconde espèce. Ce modèle comporte trois paramètres réels positifs, σ^2, ρ et ν, servant respectivement à ajuster la variance du processus, sa distance caractéristique de corrélation (correspondant à la largeur du noyau) et sa régularité de façon indépendante. La figure 1 représente cette fonction pour quelques valeurs des paramètres.

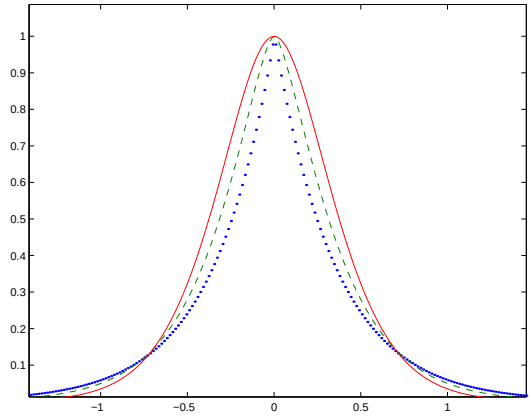

FIG. 1 – Covariance de Matérn avec $\sigma^2 = 1$, $\rho = 0.5$ et trois régularités différentes : $\nu = 3$ en trait plein, $\nu = 1$ en trait discontinu, et $\nu = 0.5$ en pointillés.

Enfin mentionnons le *variogramme expérimental*, qui est l'outil privilégié des géostatistitiens pour effectuer l'analyse structurelle. Le variogramme expérimental est l'estimée de la fonction

$$\gamma(\mathbf{h}) = \gamma(\mathbf{x} - \mathbf{y}) = \frac{1}{2}\mathsf{var}(F(\mathbf{x}) - F(\mathbf{y})) . \qquad (3)$$

Le variogramme est en général une fonction croissante avec $||\mathbf{h}||$ et s'annule à l'origine. Si le processus est stationnaire, notons que cette fonction filtre la moyenne de $F(\mathbf{x})$ et que $\gamma(\mathbf{h}) = k(\mathbf{0}, \mathbf{0}) - k(\mathbf{0}, \mathbf{h})$. L'utilisation du variogramme permet de mieux modéliser les données observées. Par exemple, une discontinuité du variogramme à l'origine traduit la présence d'un bruit blanc qui peut provenir d'une erreur de mesure sur la sortie. Estimer le variogramme dans différentes directions permet de déceler des anisotropies et de choisir la structure du noyau en conséquence. Toutefois, ceci n'est possible que si l'on dispose d'un nombre de données suffisant.

3.2 Méthode probabiliste de choix des paramètres du noyau

Le choix d'un noyau comprend certes le choix d'une structure, mais aussi l'estimation des paramètres θ_i, $i = 1, \cdots, p$ de cette structure à partir des données observées. Le variogramme peut être utilisé à cette fin puisqu'il permet aussi d'estimer la covariance. D'autres alternatives existent, et dans un cadre probabiliste l'approche du maximum de vraisemblance est naturelle. La log-vraisemblance des données s'écrit très simplement, pour un modèle à moyenne nulle,

$$l(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det\mathbf{K}(\boldsymbol{\theta}) - \frac{1}{2}\mathbf{f}_{\text{obs}}^{\mathsf{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{f}_{\text{obs}} , \qquad (4)$$

où $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_p]^\mathsf{T}$ est le vecteur des paramètres du noyau, $\mathbf{f}_{\text{obs}} = [f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n)]^\mathsf{T}$ le vecteur des données expérimentales et $\mathbf{K}(\boldsymbol{\theta})$ la matrice de covariance du vecteur $[F(\mathbf{x}_1), \cdots, F(\mathbf{x}_n)]^\mathsf{T}$.

Parmi les avantages du maximum de vraisemblance, on peut noter que cette méthode fournit un cadre pour évaluer la précision de l'estimation des paramètres grâce aux bornes de Cramér-Rao et à la notion d'information de Fischer.

3.3 Des noyaux généralisés

Au lieu de considérer la covariance d'un processus aléatoire $F(\mathbf{x})$, on peut s'intéresser aux accroissements de ce processus. Par exemple, le variogramme défini par (3) peut être utilisé comme covariance généralisée (la covariance généralisée d'un processus $F(\mathbf{x})$ est seulement *conditionnellement* positive). Ceci conduit à une extension du krigeage dite *krigeage intrinsèque* [2]. Notons dans ce cas que la moyenne de $F(\mathbf{x})$ est filtrée par les différences, ce qui permet très facilement de modéliser un processus de moyenne inconnue.

Parmi les covariances généralisées, les covariances dites polynomiales permettent de retrouver le cas des splines de type « plaques minces » [3].

4 Un cas particulier : SVR à plusieurs sorties corrélées

En considérant un noyau comme une covariance, il s'avère très facile d'étendre une SVR au cas multivariable ($q > 1$) où les composantes du vecteur de sortie sont corrélées. (En l'absence de corrélation, on peut se contenter d'appliquer à chaque composante une SVR séparée ce qui permet de se ramener au cas $q = 1$.) Dans le cadre du krigeage, chaque composante du vecteur des sorties du phénomène étudié est modélisée par un processus aléatoire $F_\alpha(\mathbf{x})$, $\alpha \in \Pi = \{1, \cdots, l\}$. Pour effectuer la prédiction, il est alors nécessaire d'utiliser les covariances $k_{\alpha,\alpha}(\mathbf{x}, \mathbf{y})$ et inter-covariances $k_{\alpha,\beta}(\mathbf{x}, \mathbf{y})$ de chaque processus. L'utilisation de plusieurs noyaux ne pose cependant pas de difficulté. Un simple changement de notations permet en effet de se ramener au cas mono-variable. Si l'on note $F(\alpha, \mathbf{x})$ le processus aléatoire, et $k\left[(\alpha, \mathbf{x}), (\beta, \mathbf{y})\right]$ la covariance de $F(\alpha, \mathbf{x})$ et $F(\beta, \mathbf{y})$, cela revient à considérer un facteur supplémentaire, dont la valeur permet d'indicer la sortie considérée. Il n'y a donc qu'un seul noyau, dépendant d'un paramètre supplémentaire.

Transposé aux SVR, ceci revient à construire une fonction

$$f_{\mathsf{sv}} : \begin{array}{ccc} \Pi \times \mathbb{R}^d & \rightarrow & \mathbb{R} \\ \alpha, \mathbf{x} & \mapsto & (w, \psi(\alpha, \mathbf{x}))_{\mathcal{F}} + b \end{array} \tag{5}$$

minimisant le coût

$$\frac{1}{2}||w||^2_{\mathcal{F}} + C \sum_{i,\alpha} [f_{\mathsf{sv}}(\alpha, \mathbf{x}_i) - f(\alpha, \mathbf{x}_i)]_\varepsilon \tag{6}$$

où $f(\alpha, \mathbf{x}_i)$ représente les données observées sur les différentes sorties α. Les notations sont celles qui apparaissent de manière classique dans la littérature consacrée aux SVM (voir [4] par exemple). Le paramètre $b \in \mathbb{R}$ peut parfois être retiré. Le coût comporte deux termes, dont le poids relatif est spécifié par la constante $C \in \mathbb{R}_+$ à choisir, le premier étant un terme de *régularisation*, le second un terme d'*attache aux données*. Ainsi $[.]_\varepsilon$ désigne la fonction de perte *ε-insensible* qui fait la spécificité des SVM et s'écrit

$$[x]_\varepsilon = \begin{cases} 0 \text{ si } |x| \leq \varepsilon\,, \\ |x - \varepsilon| \text{ sinon}\ . \end{cases} \tag{7}$$

Les fonctions $\psi(\alpha, \mathbf{x}) \in \mathcal{F}$ et $w \in \mathcal{F}$, où $\mathcal{F}$ est l'*espace des caractéristiques* selon la terminologie SVM, s'écrivent en fonction du noyau reproduisant, c'est-à-dire

$$\psi(\alpha, \mathbf{x}) = k\left[(\alpha, \mathbf{x}), (.,.)\right] \tag{8}$$

et

$$w = \sum_{\beta,i} w_{\beta,i} k[(\beta, \mathbf{x}_i), (.,.)]\,. \tag{9}$$

Le produit scalaire dans $\mathcal{F}$, noté $(.,.)_{\mathcal{F}}$ ci-dessus, se déduit du noyau $k\left[(.,.), (.,.)\right]$ et par conséquent

$$(w, \psi(\alpha, \mathbf{x}))_{\mathcal{F}} = \sum_{\beta,i} w_{\beta,i} k[(\beta, \mathbf{x}_i), (\alpha, \mathbf{x})]\,. \tag{10}$$

Notons que les algorithmes d'optimisation dans le cas multivariable sont les mêmes que dans le cas monovariable, puisque les méthodes sont formellement identiques. Cette extension au cas multivariable est donc très simple en pratique et permet facilement d'améliorer la prédiction en profitant de la corrélation entre les sorties.

Les modèles d'inter-covariances entre les sorties dépendent naturellement du problème. Un modèle simple est par exemple celui où deux processus $F_1(\mathbf{x})$ et $F_2(\mathbf{x})$ admettent la même covariance radiale $k(\mathbf{x}, \mathbf{y}) = r(||\mathbf{x} - \mathbf{y}||) = r(h)$ et où il existe une constante γ telle que $-1 < \gamma < 1$ et que les inter-covariances s'écrivent $\mathsf{cov}(F_1(\mathbf{x}), F_2(\mathbf{y})) = \mathsf{cov}(F_2(\mathbf{x}), F_1(\mathbf{y})) = \gamma r(h)$. Ce type de modèle s'appelle *modèle proportionnel* [1]. Comme précédemment nous insistons sur le fait que c'est l'analyse du système qui doit guider l'utilisateur vers un modèle donné et que la physique du processus permet souvent d'accéder aux propriétés des covariances et inter-covariances.

Un cas particulier qui peut servir d'illustration de la méthode est celui où l'on désire imposer des *conditions sur le gradient* (ou la dérivée) de la fonction f_{sv}. Ce type de contrainte peut sembler difficile à prendre en compte mais se traite en fait très naturellement. Pour simplifier la présentation, considérons le cas où $x \in \mathbb{R}$. La dérivée de $F(x)$ est définie par

$$F'(x) = \lim_{h \to 0} \frac{F(x + h) - F(x)}{h}\,. \tag{11}$$

(La limite est définie en moyenne quadratique.) Si $F(x)$ admet une covariance radiale $r(h)$, $h \in \mathbb{R}_+$, deux fois différentiable, la structure de covariance du modèle multivariable s'écrit :

$$\begin{aligned} &\mathsf{cov}[F(x_1), F'(x_2)] \\ &\qquad = \lim_{h \to 0} \frac{r(|x_1 - x_2 - h|) - r(|x_1 - x_2|)}{h} \\ &\qquad = -\operatorname{sgn}(x_1 - x_2)\, r'(|x_1 - x_2|)\,, \end{aligned} \tag{12}$$

$$\mathsf{cov}[F'(x_1), F(x_2)] = \operatorname{sgn}(x_1 - x_2)\, r'(|x_1 - x_2|)\,, \tag{13}$$

$$\mathsf{cov}[F'(x_1), F'(x_2)] = -r''(|x_1 - x_2|)\,. \tag{14}$$

La structure de covariance ci-dessus permet donc d'utiliser des données sur la dérivée du processus à prédire, pour imposer des conditions aux limites par exemple. Cette application est illustrée sur la figure 2. Notons que cette application nécessite d'imposer la valeur de la dérivée. Ceci implique que (6) doit faire intervenir des constantes C et ε différentes pour chaque sortie α et que dans le cas de la sortie correspondant à la dérivée, C doit être grand et ε petit. Tenir compte de cette exigence ne pose pas de problème en pratique dans les algorithmes d'optimisation.

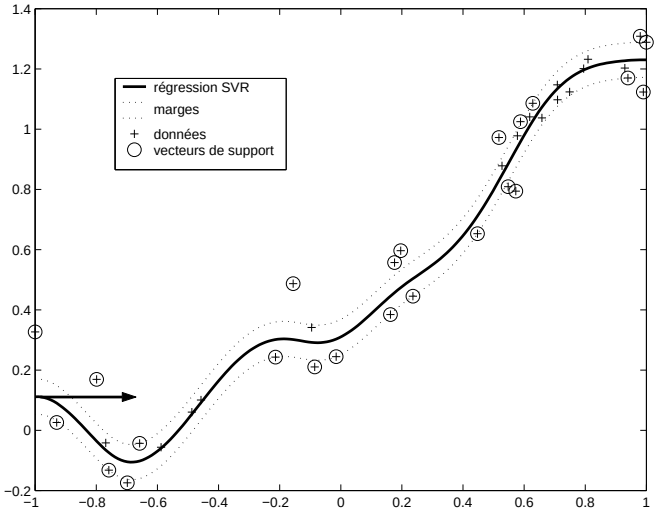

FIG. 2 – SVR avec données sur la dérivée. On impose une tangente horizontale en $x = -1$, matérialisée par la flèche.

5 Conclusions et perspectives

La méthodologie de l'analyse structurelle permet de choisir un noyau adapté aux données observées et à la physique du phénomène étudié et ainsi d'obtenir par une démarche rationnelle des modèles boîte noire satisfaisants pour l'utilisateur. Nous espérons avoir contribué à montrer que la géostatistique peut être appliquée dans le domaine de la modélisation en ingénierie avec des résultats intéressants. La présentation du krigeage et de la géostatistique faite dans cet article reste cependant très partielle. Par exemple, nous n'avons pas parlé du krigeage non-linéaire (ou krigeage d'indicatrices), qui nous semble très intéressant à étudier dans le cadre de la classification puisque dans ce cas, les données, souvent binaires, ne sont pas correctement modélisables par des processus gaussiens. Nous n'avons pas non plus parlé des résultats sur les équations différentielles stochastiques pour la modélisation des systèmes dynamiques. Il reste également des questions ouvertes, comme par exemple celle du choix de modèles de covariances ou de noyaux dans des espaces de facteurs de dimension élevée.

Références

[1] J.-P. Chiles and P. Delfiner. *Geostatistics : Modeling Spatial Uncertainty.* Wiley Series in Probability and Statistics. Wiley Interscience, New York, 1999.

[2] G. Matheron. The intrinsic random functions, and their applications. *Adv. Appl. Prob.*, 5 :439–468, 1973.

[3] G. Matheron. Splines and Kriging : their formal equivalence. In D.F. Merriam, editor, *Down-to-Earth Statistics : Solutions Looking for Geological Problems*, pages 77–95. Syracuse univ. of geology contributions edition, 1981.

[4] B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, Cambridge, 2002.

[5] M.L. Stein. *Interpolation of Spatial Data : Some Theory for Kriging.* Springer Verlag, New York, 1999.

[6] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* SIAM, Philadelphia, 1990.

[7] G. Wahba and G.S. Kimeldorf. Spline functions and stochastic processes. *Sankhyä : the Indian Journal of Statistics : Series A*, 32(2) :173–180, 1970.

[8] C.K.I. Williams and C.E. Rasmussen. Gaussian processes for regression. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, volume 8. MIT Press, 1995.