Rencontre EMSE/3MI - CEA/LETI

Splines et krigeage

L. Carraro – 12 juin 07

Ecole des mines de Saint-Etienne

carraro@emse.fr
www.emse.fr/~carraro

PLAN

- Approximation et interpolation
- Splines cubiques en 1D
- Krigeage
- Des noyaux
- Eléments de comparaison

Approximation/interpolation

Estimation non paramétrique :

données = couples (x_i, y_i) pour $1 \leq i \leq n$

modèle :

$$y_i = g(x_i) + \varepsilon_i, \ \text{où} \ \varepsilon_i \text{ est un bruit de variance } \sigma^2$$

estimation de g ?

$$-2 \log L(f) \propto (1/\sigma^2) \sum_{i=1}^{n} \left(y_i - g(x_i) \right)^2$$

Problème mal posé :

- Pénaliser
- Informations a priori

Régularisation

Fonction de coût pénalisée :

critère à minimiser :

$$C(g) = \boldsymbol{L}(\boldsymbol{Y}, g(\boldsymbol{X})) + \lambda \ J(g)$$

- J est une fonctionnelle mesurant la régularité de g
- λ est un paramètre de lissage
- $$\boldsymbol{L}(\boldsymbol{Y}, g(\boldsymbol{X})) = \sum_{i=1}^{n} \left(y_i - g(x_i) \right)^2$$

Plus généralement,

- $$\boldsymbol{L}(\boldsymbol{Y}, g(\boldsymbol{X})) = -2 \ \log L(g), \ ou \ \boldsymbol{L}(\boldsymbol{Y}, g(\boldsymbol{X})) = \sum_{i=1}^{n} D(y_i, g(x_i))$$

D mesure une distance entre y_i et $g(x_i)$

Cas limites

- $\lambda \to +\infty$ $$\underset{J(g)=0}{Min} \sum_{i=1}^{n} D(y_i, g(x_i))$$

Souvent, {g, J(g) = 0} est de dim finie

⇒ régression linéaire

- $\lambda \to 0$ $$\underset{g(x_i)=y_i}{Min} J(g)$$

⇒ interpolation

Splines cubiques

$$g : [0,1] \rightarrow \mathbb{R}$$

$$C(g) = \sum_{i=1}^{n} \left(y_i - g(x_i)\right)^2 + \lambda \int_0^1 g''(x)^2 \, dx$$

Solution de la forme (calcul des variations) :

$$g_\lambda(x) = \beta_\lambda(0) + \beta_\lambda(1)\, x + \sum_{i=1}^{n} \alpha_\lambda(i)\, K(x - x_i)$$

K est la fonction de Green de l'opérateur $\frac{d^4}{dx^4}$

$$K(h) = \frac{h_+^3}{6}$$

Solution complète

$$g_\lambda(x) = \beta_\lambda(0)+\beta_\lambda(1)x + \sum_{i=1}^{n} \alpha_\lambda(i) K(x-x_i)$$

Espace nul Noyau

En écriture matricielle :

$$g(x) = f(x)\beta + k(x)\alpha, \text{ et } f(x)=(1 \quad x), \beta=\begin{pmatrix}\beta(0)\\ \beta(1)\end{pmatrix} \text{ etc...}$$

$$\boldsymbol{G} = (g(x_1) \ldots g(x_n))' = \boldsymbol{F}\beta + \boldsymbol{K}\alpha$$

$$\text{et } \int_0^1 K''(\boldsymbol{x}-\boldsymbol{x}_i)K''(\boldsymbol{x}-\boldsymbol{x}_j)\,d\boldsymbol{x} = K(\boldsymbol{x}_i-\boldsymbol{x}_j) \text{ (K fonction de Green)}$$

$$C(g) = C(\alpha,\beta) = (\boldsymbol{Y}-\boldsymbol{F}\beta-\boldsymbol{K}\alpha)'(\boldsymbol{Y}-\boldsymbol{F}\beta-\boldsymbol{K}\alpha)+\lambda\alpha'\boldsymbol{K}\alpha$$

Un peu de calcul...

$$\beta \text{ fixé, } argmin\, C(\alpha, \beta) \text{ donne } \alpha(\beta) = \left(K + \lambda\, Id\right)^{-1}\left(\boldsymbol{Y} - \boldsymbol{F}\beta\right)$$

D'où l'expression (Spline) :

$$g_\lambda(x) = f(x)\hat{\beta} + k(x)\left(K + \lambda\, Id\right)^{-1}\left(\boldsymbol{Y} - \boldsymbol{F}\hat{\beta}\right)$$

$$\hat{\beta} = \left\{\boldsymbol{F}'\left[Id - K\left(K + \lambda\, Id\right)^{-1}\right]\boldsymbol{F}\right\}^{-1}\boldsymbol{F}'\left[Id - K\left(K + \lambda\, Id\right)^{-1}\right]\boldsymbol{Y}$$

Cas limites :

$$\lambda \to 0,\ g(x) = f(x)\hat{\beta} + k(x)K^{-1}\left(\boldsymbol{Y} - \boldsymbol{F}\hat{\beta}\right)$$

$$\hat{\beta} = \left\{\boldsymbol{F}'K^{-1}\boldsymbol{F}\right\}^{-1}\boldsymbol{F}'K^{-1}\boldsymbol{Y}$$

$$\lambda \to +\infty,\ g(x) = f(x)\hat{\beta}$$

$$\hat{\beta} = \left\{\boldsymbol{F}'\boldsymbol{F}\right\}^{-1}\boldsymbol{F}'\boldsymbol{Y}$$

Remarques

- Résultats matriciels généraux
  - Rôle clé du noyau K et de la formule :

$$\langle K(.-x_i) | f \rangle = f(x_i)$$

$$soit \langle K(.-x) | f \rangle = f(x)$$

Espaces de Hilbert à noyau reproduisant (RKHS)

- Critères pénalisés généraux : solutions de la forme :

$$g(x) = f(x)\beta + k(x)\alpha$$

  - Cas de coûts non quadratiques :

    système non linéaire en α et β.

Principes du krigeage

- ## Pénalisation probabiliste

On donne un poids à toutes les fonctions, selon leur régularité.

Utilisation de processus gaussiens ➡ poids déterminé par :

- Espérance : $E(g(x)) = f(x)\beta$
- Covariance : $cov(g(x), g(y)) = K(x, y)$

NB :

souvent, $K(x, y) = K(x-y)$ et $\sigma_g^2 = K(0)$

Principes du krigeage

- Modèle probabiliste des observations

$$y_i = g(x_i) + \varepsilon_i, \text{ où } \varepsilon_i \text{ est un bruit de variance } \sigma_\varepsilon^2$$

- Estimation de g en connaissant les y_i

Détermination de la loi conditionnelle du processus $(g(x))_x$ sachant **y**

Théorème (Duchon, Matheron, Wahba...)

$E[g(x)/Y]$ est donnée par l'expression (Spline)
avec $\lambda = \sigma_\varepsilon^2 / \sigma_g^2$

Remarques

- Pas de preuve « satisfaisante »
- Le noyau K est le même !
- Outil pivot = RKHS
  - Cas des splines cubiques :

    Le processus associé au noyau K est la primitive du mouvement brownien
  - Cas général :

    Notions de semi-RKHS et de processus intrinsèques

Propriétés du noyau K

- K est défini positif (ou conditionnellement d.p.) :

$$\forall a = (a_x)_x \in \mathbb{R}^{(X)}, \sum_x \sum_y a_x \, a_y \, K(x, y) \geq 0$$
$$et \;=\; 0 \;\; ssi \;\; \sum_x a_x \, K_x \;\equiv\; 0$$

- K détermine :
  - les fonctions de base stockées dans k(x)
  - vision spline : la pénalisation
  - vision krigeage : la régularité du processus

Propriétés du noyau K

- Pénalisations du type :

$$\|g\|_{\mathrm{H}}^{2} = \int_{\mathbb{R}^d} [P(x, D)\, g](x)^2 \, dx$$

P(x,.) est un polynôme en d variables (à coefficients non constants en général)

D est l'opérateur de différentiation

P(.,D) est un opérateur différentiel

- Le noyau K est la fonction de Green de l'opérateur :

  P(.,D)* P(.,D)

Splines

- Splines = vision fonctionnelle
  - propriétés de convergence
  - terme $f(x)\beta$ souvent implicite, et négligé
  - noyau découlant d'une norme
  - noyau non paramétré car norme « standard »
  - erreur commise de forme simple (norme)
  - lien avec krigeage via les RKHS

Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

Krigeage

- Krigeage = vision probabiliste
  - pas de propriétés de convergence
  - terme f(x)β explicite, mais souvent négligé
  - noyau explicite et paramétré
    - grand intérêt pratique
    - difficultés d'estimation
    - mélange de modèles (approche bayésienne)
  - erreur commise sous forme de loi de proba
  - lien avec splines via les RKHS

Bibliographie

Aronszajn N. (1950), Theory of reproducing kernels, Trans. AMS, 68, p. 337-404

Amato U., Antoniadis A., Pensky M. (2006), Wavelet kernel penalized estimation for non-equispaced design regression, Statistics and Computing, 16, 1, 37-56, 2006

Carraro L., Bay X., Roustant O. (2007), Introduction aux RKHS, GDR MASCOT NUM, www.lsp.ups-tlse.fr/Fp/Gamboa/GDR/rencontres_mars2007.html.

Carraro L., Corre B., Helbert C., Roustant O., Josserand S. (2007), Optimal designs for the propagation of uncertainty in computer experiments, to appear in Chemometrics and Intelligent Laboratory Systems.

Duchon J. (1980), Fonctions splines homogènes à plusieurs variables, Thèse d'Etat, Université de Grenoble.

Duchon J. (2007), Modèles intrinsèques et probabilités conditionnelles, en préparation.

Feuillard V. (2007), Analyse d'une base de données pour la calibration d'un code de calcul, Thèse de doctorat, Univ. Paris VI.

Girosi F., Jones M., Poggio T. (1995), Regularization theory and neural network architectures, Neural Computation, 7, p. 219-269.

Bibliographie

Lu F., Keles S., Lin Y., Wright S.J., Wahba G. (2006), Kernel Regularization and Dimension Reduction, Tech. Report n°1119, U. of Wisconsin.

Vazquez E. (2005), Modélisation comportementale de systèmes non linéaires multivariables par méthodes à noyaux et applications, thèse de doctorat de l'U d'Orsay.

Wahba G. (1990), Spline models for observational data, SIAM.

Wahba G. (2000), An Introduction to Model Building with Reproducing Kernel Hilbert Spaces, Tech. Report n°1020, U. of Wisconsin.

Yue R., Hickernell F.J. (1999), Robust designs for fitting linear models with misspecification, Stat. Sinica, 9, p. 1053-1069.

Ecole Nationale Supérieure des Mines
SAINT-ETIENNE