Approche Bayésienne pour l'estimation du paramètre de lissage dans les splines

S.AMROUN, S.ADJABI

Doctoriales de Recherche Opérationnelle, le 12 et 13 Décembre 2018

Résumé

Dans ce travail, nous présentons l'estimateur non paramétrique par la méthode des splines de la courbe de régression de la moyenne. Le paramètre de lissage qui intervient dans la forme de l'estimateur de la fonction régression sera estimé par l'approche bayésienne. Cette approche bayésienne sera comparée sur une fonction de régression cible connue et sur des données réelles avec la méthode Validation croisée généralisée (GCV) en utilisant le critère de l'erreur moyenne quadratique (ASE).

Introduction

On veut étudier la relation non paramétrique entre une variable explicative (indépendante) X et une variable réponse (dépendante) Y. Soit pour cela l'échantillon (x_i, y_i), i=1,…, n issu du couple (X,Y). Le modèle de régression non paramétrique s'écrit sous la forme :

$$y_i = m(x_i) + \epsilon_i$$

où $\varepsilon_i \rightarrow$ i.i.d $N(0,\sigma^2)$, m est une fonction inconnue qu'on cherche à estimer.

Estimateur splines: L'estimateur par splines de lissage de m est obtenu en minimisant le critère S(m) qui combine deux mesures: la mesure de fidélité des données et la mesure de lissage, S(m) est de la forme :

$$S(m) = \sum_{i=1}^{n}(y_i - m(x_i))^2 + \lambda \int_a^b (m''(x))^2\, dx$$

- $\sum_{i=1}^{n}(y_i - m(x_i))^2$, mesure la fidélité des données,
- $\int(\hat{m})^2 dx$, mesure de lissage.
- λ>0 est un paramètre de lissage, il contrôle le compromis entre le biais et la variance de $\widehat{m}$, il influe sur la qualité de l'estimateur $\widehat{m}$, le choix de λ est un problème important.

Des méthodes ont été proposées dans la littérature pour la sélection du paramètre λ, à savoir, CV, GCV,...

Le but de ce travail est d'introduire l'approche bayésienne comme méthode alternative à ces méthodes.

Le paramètre de lissage λ varie entre 0 et $+\infty$

$\lambda \rightarrow 0$, La pénalité de régularité disparaît et l'estimateur interpole les données (estimateur flexible).

$\lambda \rightarrow +\infty$, La pénalité de régularité domine et l'estimateur est forcé d'être constant.

L'estimateur spline de lissage optimale dépend linéairement des observations

$$\widehat{m}_\lambda = A_\lambda Y;$$

avec

$A_\lambda = (I + \lambda K)^{-1}$ est la matrice de lissage de l'estimateur spline. I est la matrice identité,

et K une matrice dépendant uniquement des $x_1, \dots, x_n$.

Voici quelques méthodes (dites classiques) pour la sélection du paramètre de lissage

Validation croisée : la fonction CV est donnée par

$$CV = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \widehat{m}_\lambda(x_i))^2}{(1-(A_\lambda)_{ii})^2}$$

Validation croisée généralisée : la fonction GCV est donnée par

$$GCV = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \widehat{m}_\lambda(x_i))^2}{(1-\frac{1}{n}tr(A_\lambda))^2}$$

Critère de Mallows (Cp) : le critère est donné par

$$C_p = \frac{1}{n}\{\|(A_\lambda - I)y\|^2 + 2\sigma^2\ tr(A_\lambda) + \sigma^2\}$$

Critère d'information d'Akaike AIC : le critère est donné par

$$\log\frac{\|(A_\lambda - I)y\|^2}{n} + \frac{2(tr(A_\lambda)+1)}{n - tr(A_\lambda) - 2} + 1$$

Dans toutes ces méthodes λ est choisi de telle manière qu'il minimise les fonctions CV, GCV, Cp et AIC.

On appelle estimation bayésienne du paramètre λ la moyenne de la loi a posteriori. Cette moyenne est notée $E(\lambda\,|y)$.

$$E(\lambda\,|y) = \int \lambda\,\pi(\lambda\,|y)\,d\lambda = \int \frac{\lambda\,\pi(y|\lambda)\pi(\lambda)}{\int \pi(y|\lambda)\,\pi(y)d\lambda}\,d\lambda.$$

Nous proposons les estimateurs bayésiens de λ et de la variance des erreurs σ^2. Supposons que les lois a priori des deux paramètres λ et σ^2 sont $G(\alpha,\beta)$ et $IG(a,b)$ respectivement.

La loi a posteriori de (λ, σ^2) est donnée par

$$\pi(\lambda,\sigma^2|y_1,\dots,y_n) \propto \pi(\lambda)\left(\frac{1}{\sigma^2}\right)^{\frac{n+2a}{2}+1}\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \widehat{m}(x_i))^2 + 2b\right)\right]$$

La loi a posteriori de $(\sigma^2/\lambda, y_1, \dots, y_n)$ est une inverse gamma et donnée par :

$$\pi(\sigma^2/\lambda, y_1, \dots, y_n) \sim IG\left(\frac{n+2a}{2}, \frac{1}{2}\sum_{i=1}^{n}(y_i - \widehat{m}(x_i))^2 + b\right)$$

La loi de λ sachant les données :

$$\pi(\lambda/y_1, \dots, y_n) \propto \pi(\lambda)\frac{1}{2}\left(\sum_{i=1}^{n}(y_i - \widehat{m}(x_i))^2 + b\right)^{-\frac{n+2a}{2}}$$

Résultats

Calcul des estimateurs

Pour le calcul des estimateurs des paramètres nous utilisons les méthodes MCMC et l'échantillonnage de Gibbs. Nous présentons une application sur des données simulées et deux cas réels.

Cas des données simulées :

$m = 4.26(\exp(*3.25x) - 4\exp(-6.5x) + 3\exp(-9.75x))$;

$\varepsilon \rightarrow N(0,\sigma^2)$, avec $\sigma^2 \in \{0.05^2, 0.08^2, 0.1^2\}$

Convergence de la méthode MCMC

Pour vérifier la convergence de la méthode MCMC pour l'estimation du paramètre de lissage et de la variance, nous utilisons les indicateurs SIF et BMSE. Selon les résultats exposés dans le tableau 1, les indicateurs (BMSE et SIF) montrent une bonne convergence de la méthode MCMC et nous avons obtenu une bonne performance des estimateurs. Nous remarquons aussi un taux d'acceptation raisonnable qui avoisine 0.50.

Table 1 Estimated parameter and their corresponding BMSE and SIF evaluation

n	Error model	Parameter	Estimation	BMSE	SIF	Acceptance rate
20	$N(0,0.05^2)$	σ^2	0.1182	0.0020	0.0003	
		λ	0.2063	0.0507	0.4795	0.42
20	$N(0,0.08^2)$	σ^2	0.1244	0.0024	0.0004	
		λ	0.2820	0.0650	1.0947	0.42
20	$N(0,0.1^2)$	σ^2	0.1246	0.0021	0.0003	
		λ	0.3089	0.0592	0.2420	0.44
50	$N(0,0.05^2)$	σ^2	0.0452	0.0012	0.0010	
		λ	0.0127	0.0086	24.1373	0.51
50	$N(0,0.08^2)$	σ^2	0.0479	0.0009	0.0009	
		λ	0.0129	0.0129	27.9162	0.46
50	$N(0,0.1^2)$	σ^2	0.0530	0.0013	0.0009	
		λ	0.0310	0.0090	25.6872	0.64

Etude comparative : Approche bayésienne et GCV

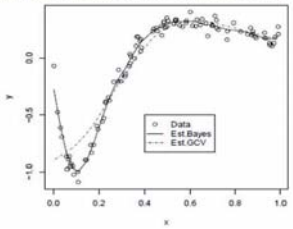
En utilisant le critère de l'erreur moyenne quadratique ASE, nous examinons et comparons les performances de l'approche bayésienne proposée et de la validation croisée généralisée. Les résultats présentés dans le tableau (2) montrent que, ASE diminue quand la taille de l'échantillon n augmente et cette constatation est vérifiée pour toutes les variances. Le ASE associé à l'approche bayésienne est toujours plus petit que celui associé à la méthode GCV. Ce qui montre que l'approche bayésienne est plus performante au sens de ce critère ASE que la méthode GCV.

Table 2 The average ASE for the Bayesian approach and generalised cross-validation method

	n	Approach Bayesian	Generalised cross validation
$\sigma^2 = 0.05^2$	20	**0.0124**	0.0173
	50	**0.0055**	0.0195
	100	**0.0025**	0.0165
	200	**0.0018**	0.0119
	500	**0.0018**	0.0119
$\sigma^2 = 0.1^2$	20	**0.0203**	0.0267
	50	**0.0116**	0.0263
	100	**0.0068**	0.0237
	200	**0.0061**	0.0226
	500	**0.0060**	0.0170
$\sigma^2 = 0.08^2$	20	**0.2518**	0.4209
	50	**0.0116**	0.0263
	100	**0.0068**	0.0237
	200	**0.0035**	0.0173
	500	**0.0020**	0.0034

Note: Bold values indicate the best results.

Figure 1 Bayesian approach and generalised cross-validation for the simulated data



Cas réels:

1) L'étude concerne une propriété physique du titane en fonction de la chaleur. X est la température et Y est une propriété physique

Table 3 RMSE for the Bayesian approach and generalised cross-validation method and the corresponding BMSE and SIF for the titanium heat data

	Parameter	Estimate	RMSE	BMSE	SIF	Acceptance rate
Bayesian approach	σ^2	0.04113		0.00022	0.00002	
	λ	0.57106	9.7464e-10	0.36618	0.06529	0.76
Generalised cross-validation	λ	0.99993	2.9732e-09			

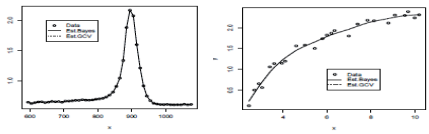
2) L'étude porte sur la sortie du courant continue en fonction de la vitesse du vent dans un moulin à vent. X est la vitesse du vent et Y est la sortie du courant.

Table 4 RMSE for the Bayesian approach and generalised cross-validation method and the corresponding BMSE and SIF for the windmill data

	Parameter	Estimate	RMSE	BMSE	SIF	Acceptance rate
Bayesian approach	σ^2	0.08634		0.00068	0.00006	
	λ	1.02108	0.0038	0.11615	0.01294	0.68
Generalised cross-validation	λ	0.9999	0.00715			

Pour étudier les performances des estimateurs dans les cas réels nous avons utilisé le critère de la racine de l'erreur moyenne quadratique (RMSE).

Les meilleures performances sont obtenues avec l'approche bayésienne au sens de RMSE en comparant avec la méthode de validation croisée généralisée. Notez que les indicateurs BMSE et SIF sont faibles, ce qui indique une bonne convergence de l'approche bayésienne en utilisant la méthode MCMC.



Bayesian approach and generalised cross-validation fitted the titanium heat and the windmill.

Conclusion

Nous avons présenté l'estimateur non paramétrique par la méthode des splines de la courbe de régression de la moyenne. Nous avons estimé Le paramètre de lissage par l'approche bayésienne.

Selon le critère de l'erreur moyenne quadratique (ASE), nous avons comparé cette approche bayésienne avec la méthode Validation croisée généralisée (GCV) sur une fonction de régression cible connue et sur des données réelles. Les résultats numériques montrent que l'approche bayésienne est plus performante que la méthode GCV .

Références

1. Craven, P. et G. Wahba (1979). Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math., 31(4)*, 377–403.
2. Eubank, R. (1999). Nonparametric regression and spline smoothing. *New York, 2nd ed, Marcel Dekker, Inc.*
3. Lee, T. (2002). Smoothing parameter selection for smoothing plines: a simulation study. *Computational Statistics and Data Analysis.*, 42:139–148.