Journal of the Faculty of Environmental Science and Technology,
Okayama University
Vol.19, No.1, pp.1-7, March 2014

影響関数による近似の数値評価

包　紅梅[1], 笛田　薫[2]

Numerical Evaluation of the Approximation by an Influence Function

Hong Mei Bao, Kaoru Fueda

For evaluating statistical models one of the most effective criteria is cross-validation. But it requires a large amount of computation. Various alternative schemes are considered to reduce its computation. Modified generalized information criterion is one of those alternative schemes. In this criterion an influence function is used to estimate the parameters of the models. By the numerical simulation we studied the effect of an influence function.

Surveying data of the lake depth are used as the sample data. We estimate the shape of lake bottom as spline surface. The estimated parameters and the estimated depths obtained by two criteria are compared and the effect of an information function is analysed.

Key words: *influence function, information criteria, CV, mGIC, B-spline*

1 Introduction

We have tried to determine the optimal model which estimate the shape of lake bottom by using spline surface. We use the model of the bivariate B-spline approximation with a penalized term. There are many factors or parameters to be determined.

To evaluate those models Leave-one-out cross-validation (LOOCV) test is quite useful, but it requires large amount of computation. Modified GIC (mGIC) can reduce it by using an influence function. The influence function is related to the first term of a Taylor expansion and it can estimate the value of the parameters. But we cannot avoid the errors. We study about these errors by the numerical simulation.

2 Statistical Model

For the nonlinear statistical modeling we use the maximum penalized likelihood methods [5],[6],[7]. We have $n = 278$ observations $\{(z_\alpha, x_\alpha); \alpha = 1, \ldots, n\}$, where z_α is the response variables generated from the unknown true distribution $G(z|x)$ having a probability density of $g(z|x)$ and x_α is the vectors of explanatory variables. We estimate w, which is a vector consisting of the unknown parameters and determines the model $z = u(x|w)$. Let $f(z_\alpha|x\alpha;\theta)$

[1] 岡山大学環境学研究科
[2] 岡山大学環境生命科学研究科

be a specified parametric model, where θ is a vector of unknown parameters included in the model. The regression model with Gaussian noise is denoted as

$$z_\alpha = u(x_\alpha|w) + \varepsilon_\alpha, \varepsilon_\alpha \sim N(0,\sigma^2), \alpha = 1,\dots,n \tag{1}$$

$$f(z_\alpha|x_\alpha;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\{z_\alpha - u(x_\alpha;w)\}^2}{2\sigma^2}\right] \tag{2}$$

where $\theta = (w', \sigma^2)'$. The parameter will be determined by the maximization of the penalized log-likelihood function expressed as

$$\ell_\lambda(\theta) = \sum_{\alpha=1}^{n} \log f(z_\alpha|x_\alpha;\theta) - \frac{n}{2}\lambda H(w). \tag{3}$$

As the regularized term or penalized terms $H(w)$ with an m-dimensional parameter vector w, we use

$$H(w) = \iint \left\{ \left(\frac{\partial^2 u}{\partial x^2}\right)^2 + \left(\frac{\partial^2 u}{\partial y^2}\right)^2 \right\} dxdy, \tag{4}$$

for the three dimensional approximation [10]. $H(w)$ can be represented in the quadratic form by the $m \times m$ nonnegative matrix K as follows

$$H(w) = w'Kw.$$

3 Samples and Conditions

We use the data of the lake depth measured from the boat by using GPS, echo sounder, clinometer and azimuth meter. We have the data measured along the wake of the boat. The total number of data used in this simulation is 278. The total number of knots of B-spline along the x and y axis is 20 respectively.

$$u(x,y) = \sum_{i=1}^{20}\sum_{j=1}^{20} w_{ij} M_i(x) N_j(y), \tag{5}$$

where $M_i(x), N_j(y)$ are the spline functions with order four along the x and y axis respectively and w_{ij} are the coefficients of products of the spline functions.

The total number of estimated coefficients of the spline function is $(20-4)\times(20-4) = 256$. And the variance is also to be estimated. The result of surface estimation is shown in Fig. 1. The curved line is the estimated surface, cross points are the location of samples and vertical lines show the size of residuals.

In this paper we set $\lambda\hat{\sigma}^2 = 10^{-5}$ as the experiment.

4 Numerical simulation

4.1 Information criteria

LOOCV is calculated as below

$$\mathrm{CV} \;=\; -2\sum_{\alpha=1}^{n} \log(f(\boldsymbol{x}_\alpha, \hat{\boldsymbol{\theta}}^{(-\alpha)})) = \sum_{\alpha=1}^{n} \left\{ \log(2\pi\hat{\sigma}^{(-\alpha)^2}) + \frac{(z_\alpha - \hat{u}^{(-\alpha)})^2}{\hat{\sigma}^{(-\alpha)^2}} \right\}. \tag{6}$$

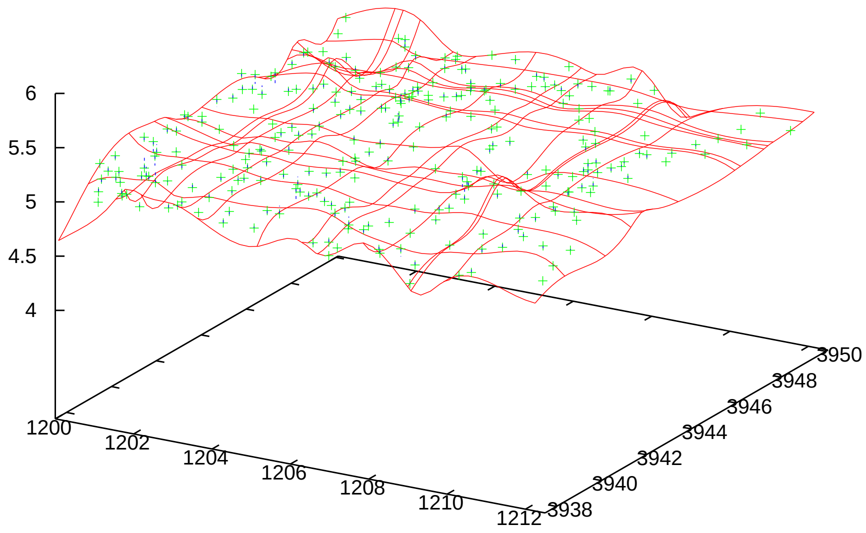

Figure 1: estimated surface

And we adopt the next approximation for CV

$$\boldsymbol{T}(\hat{G}^{(-\alpha)}) \;\approx\; \boldsymbol{T}(G) + \frac{1}{n-1}\sum_{i\neq\alpha}^{n}\boldsymbol{T}^{(1)}(z_i;G) \approx \boldsymbol{T}(\hat{G}) - \frac{1}{n}\boldsymbol{T}^{(1)}(z_\alpha;\hat{G}). \tag{7}$$

In the equation(6) of CV we replace the $\hat{\boldsymbol{\theta}}^{(-\alpha)}$ with $\tilde{\boldsymbol{\theta}}_\alpha = \hat{\boldsymbol{\theta}} - \frac{1}{n}\boldsymbol{T}^{(1)}(z_\alpha;\hat{G})$ and its scheme is called as modified GIC (mGIC)[11]. The calculation is shown as below

$$\mathrm{mGIC} = -2\sum_{\alpha=1}^{n}\log(f(\boldsymbol{x}_\alpha, \hat{\boldsymbol{\theta}} - \frac{1}{n}\boldsymbol{T}^{(1)}(z_\alpha;\hat{G}))) = \sum_{\alpha=1}^{n}\left\{\log(2\pi\tilde{\sigma}_\alpha^2) + \frac{(z_\alpha - \tilde{u}_\alpha)^2}{\tilde{\sigma}_\alpha^2}\right\}, \tag{8}$$

where $\tilde{\boldsymbol{\theta}}_\alpha = (\tilde{\boldsymbol{w}}'_\alpha, \tilde{\sigma}_\alpha^2), \tilde{u}_\alpha = u(\boldsymbol{x}_{\boldsymbol{\alpha}}|\tilde{\boldsymbol{w}}'_\alpha)$.

4.2 Numerical result

The numerical result shows that the first terms (variance) of (6),(8) are almost same and the second terms (depth) of them are quite different.

Table 1: Comparison of CV and mGIC

	CV	mGIC
variance term	-1299.783	-1298.879
depth term	1227.449	544.995

About the variance term, the maximum difference is 0.024155 and the average is 0.003258. These are quite small. But about the depth term, there are large differences.

By the influence function we estimate not the depth but the parameter. We calculate 278 samples and every sample has 257 parameters. So the total number of parameters which we estimated is 71446. The ratios of the parameters estimated by CV and estimated by mGIC are calculated and those results are shown in Table 2. The average of the ratios is 0.99999 and the variance is 5.2410×10^{-6}.

Table 2: ratio of the estimated parameters (mGIC / CV)

ratio	total number	percentage
0.995 - 1.005	70416	98.55%
0.990 - 1.010	71010	99.38%
0.975 - 1.025	71342	99.85%
0.950 - 1.050	71419	99.96%

The total number of parameters which have the differences larger than 5.0% is only 27. The smallest ratio is 0.87720 and the largest is 1.10598. The estimation of the parameter is quite accurate. But the estimation of the depth using these parameters cause large differences in some samples. Consequently the value of the information criterion has the large difference.

4.3 Difference between smaples

About the depth term the half of 278 samples have the difference smaller than 0.71 and the sum of those differences is only 30.91 . But the rest of 278 samples have the large differences and the sum of them is 651.56 . The samples which have the large differences are distributed near the boundary area. Those locations are shown in Fig.2. The maximum difference is 43.06839 and the average is 2.45486. The ratio of the parameters of the sample which has the largest difference are shown in Table 3.

Table 3: ratio of parameters of the sample which has the largest difference

	average	maximum	minimum
all 257 parameters	0.99958		
valid 16 spline coefficients	0.99342	1.03259	0.88941
other 241 parameters	1.00002	1.00428	0.99701

4.4 Difference between parameters

On the other hand the parameters which have large errors are shown in Table 4.

The function number in Table 4 means the number of spline function in (5). The support of the spline functions related with the coefficients in Table 4 are shown in Fig. 3.

Those areas are distributed southeast mainly.

5 Conclusion

We can predict the values of spline coefficients of LOOCV using the influence function of order one. But the difference of the information criteria is quite large. In this paper we show the

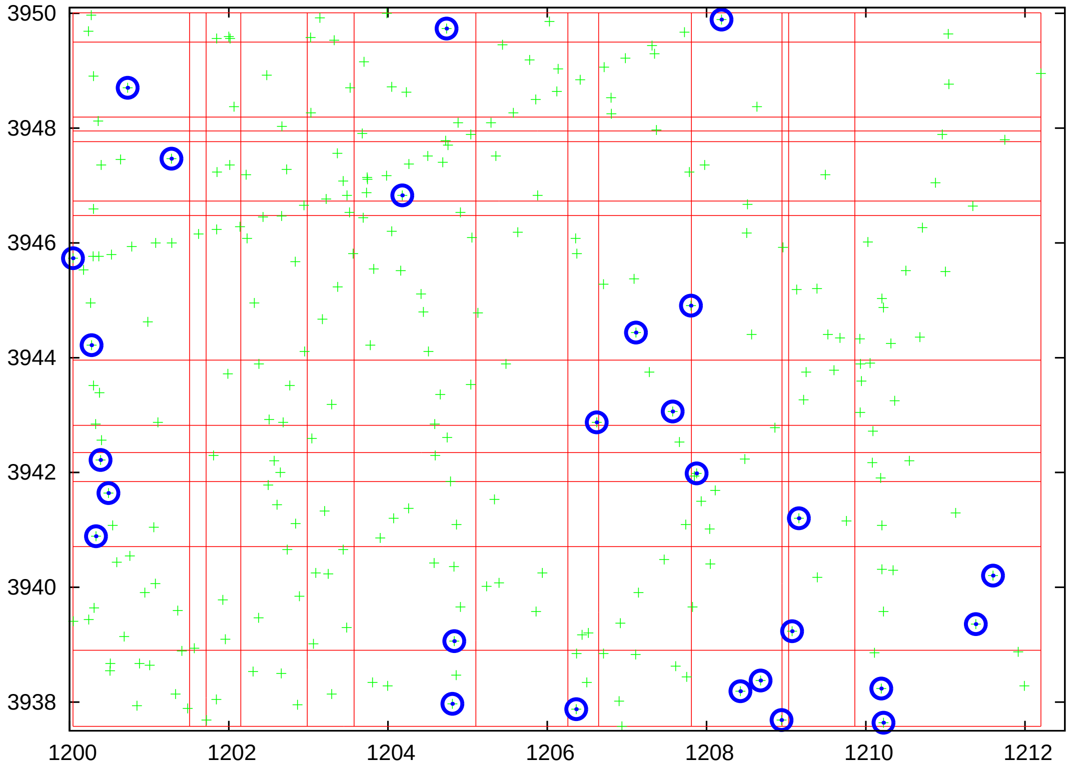

Figure 2: locations of samples with large errors

Table 4: Parameters with large variance

function number(i)	function number(j)	minimum ratio	maximum ratio	variance	standard deviation
3	15	0.87720	1.01827	0.0000735	0.0085736
16	12	0.99835	1.10571	0.0000514	0.0071676
1	14	0.88941	1.00357	0.0000499	0.0070617
8	11	0.98520	1.09219	0.0000455	0.0067451
4	16	0.97364	1.10598	0.0000445	0.0066704
4	15	0.98694	1.08300	0.0000363	0.0060287
3	14	0.95697	1.04481	0.0000330	0.0057406
15	12	0.99693	1.08498	0.0000324	0.0056886
2	14	0.97906	1.07677	0.0000296	0.0054376
3	15	0.99109	1.07392	0.0000288	0.0053688

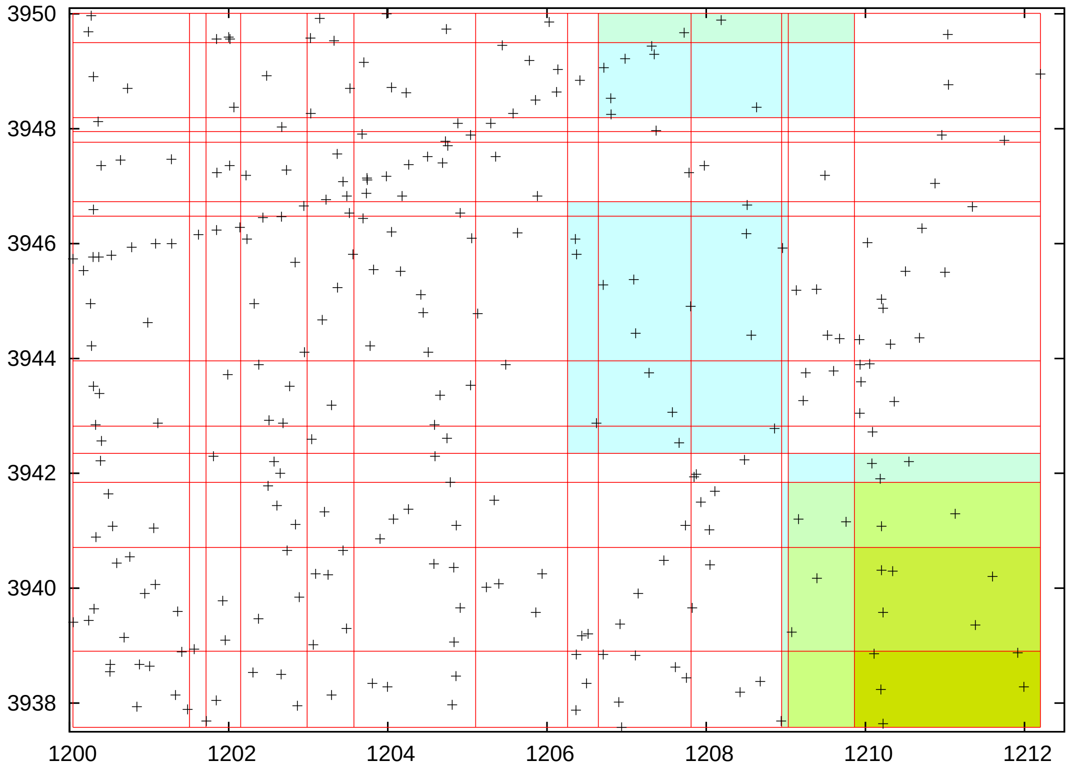

Figure 3: area

result of computation and the distribution of errors of samples or parameters. It will be our future work to clarify the reason of these errors.

References

[1] Cox, M.G., ”The numerical evaluation of B-splines”, *J. Inst. Math. Appl.*, **10**(1972), pp.134-149.

[2] Cox, M.G., ”An algorythm for spline interpolation”, *J. Inst. Math. Appl.*, **15**(1975), pp.95-108.

[3] de Boor, C., ”On calculation with B-splines”, *J. Approx. Theory*, **6**(1972), pp.50-62.

[4] Schoenberg, I. J. and Whitney, A., ”On Pólya frequency functions III”, *Trans. Amer. Math. Soc*, Vol. **74**(1953). pp. 246-259, pp. 246-259

[5] Good, I. J. and Gaskins, R.A., ”Non parametric roughness penalties for probability densities”, *Biometrika*, Vol. **58**(1971). pp. 255-277

[6] Good, I. J. and Gaskins, R.A., ”Density estimation and bump hunting by the penalized likelihood method exemplified by scattering and meteorite data”, *Journal of American Standard Association*, Vol. **75**(1980). pp. 42-56

[7] Green, P. J., Silverman, B. W., ”Nonparametric Regression and Generalized Linear Models”, Chapman and Hall, London(1994).

[8] Konishi, S., Kitagawa, G. (1996). "Generalised information criteria in model selection.", *Biometrika*, **83**, 875—890.

[9] Konishi, S., Kitagawa, G. (2008). "Information Criteria and Statistical Modeling", Springer Science+Business Media, LLC.

[10] Umeyama, S., (1996). "Discontinuity extraction in regularization using robust statistics", *Technical report of IEICE.*,PRU95-217 (1996). pp. 9-16

[11] Ueki, M. and Fueda, K., "Optimal Tuning Parameter Estimaiton In Maximum Penalized Likelihood Method" *Annals of the Institute of Statistical Mathematics*, **62** (2010), 413-438.

[12] Stone,M., "Cross-validatory choice and assessment of statistical predictions (with discussion)", *Journal of the Royal Statistical Society*, Series B, 36 (1974), 111—147.

[13] von Mises, R., "On the asymptotic distribution of differentiable statistical function.", *Annals of Mathematical Statistics*,**18**(1947),309-348.

[14] Withers, C. S.,"Expansions for the distribution and quantiles of a regular functional of the empirical distribution with applications to nonparamteric confidence intervals.",*Annals of Mathematical Statistics*(1983),**11**(2),(1983),577-587.