The Korean Communications
in Statistics Vol. 5, No. 2, 1998
pp. 277-285

Smoothing Parameter Selection Using Multifold Cross-Validation in Smoothing Spline Regression[1)]

Changkon Hong[2)], Choongrak Kim[3)] and Misuk Yoon[4)]

Abstract

The smoothing parameter λ in smoothing spline regression is usually selected by minimizing cross-validation (CV) or generalized cross-validation (GCV). But, simple CV or GCV is poor candidate for estimating prediction error. We defined MGCV (Multifold Generalized Cross-validation) as a criterion for selecting smoothing parameter in smoothing spline regression. This is a version of cross-validation using leave- k-out method. Some numerical results comparing MGCV and GCV are done.

1. Introduction

Consider a nonparametric regression model

$$y_j = \mu(t_j) + e_j \qquad j = 1, \cdots, n \tag{1.1}$$

where $a \le t_1 < \cdots < t_n \le b$ and the errors e_j are zero mean, uncorrelated random variables with common variance σ^2. Assume that μ is smooth in the sense that it belongs to the m-th order Sobolev space W_2^m of functions on $[a,b]$ defined as

$$W_2^m = \{f\text{: function on } [a,b] \mid f^{(i)} \text{ is absolutely continuous}, i = 0, \cdots, m-1, \; f^{(m)} \in L^2[a,b]\}$$

Discussions of smoothing splines and their statistical applications may be found in Wegman and Wright (1983), Silverman (1985), Eubank (1988), and Wahba (1990).

One of many possible estimators of μ in (1.1) is the minimizer over $g \in W_2^m[a,b]$

1) This research was supported by Non Directed Research Fund, Korea Research Foundation, 1996
2) Associate Professor, Dept. of Statistics, Pusan National University, Pusan 609-735
3) Associate Professor, Dept. of Statistics, Pusan National University, Pusan 609-735
4) Dept. of Statistics, Pusan National University, Pusan 609-735

of

$$\frac{1}{n}\sum_{j=1}^{n}(y_j-g(t_j))^2+\lambda\int_a^b\{g^{(m)}(t)\}dt, \quad \lambda>0. \tag{1.2}$$

The parameter λ in (1.2) is called smoothing parameter, and the choice of λ is usually accomplished by minimizing cross-validation (CV) or generalized cross-validation (GCV). Typically, the minimizer $\hat{\lambda}$ of CV or GCV is found by numerical methods and this choice of λ is used for fitting the data.

There are various model selection methods in smoothing spline, such as the Akaike information criterion (AIC) (Akaike 1974; Shibata 1981), the C_p (Mallow 1973), the jackknife, and the bootstrap (Efron 1983,1986). Especially one of the most useful methods in selection problem is cross-validation. A lot of work have been done on this, for example Stone (1974), Bowman (1984), and Härdle and Marron (1985). The cross-validation idea is simply splitting the data into two parts, using one part to fit a model and then using the other part to measure prediction ability. However the version of CV is unsatisfactory in several respects. Efron (1986) shows that the simple CV is poor candidate for estimating the prediction error and suggested that some version of bootstrap would be better off. When selecting the correct model is concerned, it is well known that the model selection by CV criterion is apt to overfit. The idea of multifold cross-validation (MCV) first appeared in Geisser (1975) where instead of deleting one observation as in simple CV, $k>1$ observations deleted. Some recent developments under linear regression models can be found in Burman (1989), Zhang (1993), and Shao (1993).

GCV criterion introduced by Wahba (1977) is most widely used in the selection of smoothing parameter λ. This is due to various optimality properties studied by Craven and Wahba (1979), Speckman (1982), Cox (1984), and Li (1986). See Eubank (1988) for detailed discussions.

In this paper, we introduce a version of multifold generalized cross-validation (MGCV) criterion. In Section 2, CV and GCV are defined, and MCV and MGCV in smoothing spline are suggested. Some numerical results comparing MGCV to GCV are done in Section 3, and concluding remarks are in Section 4.

2. Multifold Generalized Cross-Validation

2.1 Cross-validation and generalized cross-validation

The choice of smoothing parameter λ in smoothing spline is usually accomplished by minimizing CV or GCV.

The estimate of λ based on the CV criterion is the minimizer of

$$CV(\lambda)=\frac{1}{n}\sum_{j=1}^{n}\{y_j-\hat{\mu}_{(j)}(t_j)\}^2,$$

where $\hat{\mu}_{(j)}(t_j)$ is the spline estimate of μ at t_j, when the j-th observation has been deleted from the data. By the delete - one lemma (Craven and Wahba 1979),

$$\hat{\mu}_{(j)}(t_j)=(\hat{\mu}(t_j)-h_{jj}y_j)/(1-h_{jj})$$

the CV criterion can be expressed as

$$CV(\lambda)=\frac{1}{n}\sum_{j=1}^{n}\{y_j-\hat{\mu}(t_j)\}^2/(1-h_{jj})^2, \tag{2.1}$$

where h_{jj} is the j-th diagonal element of the hat matrix H(λ). (see Eubank (1988) for details.)

The GCV criterion replaces the individual leverage values $h_{11}(\lambda),\cdots,h_{nn}(\lambda)$ in (2.1) by their averages $n^{-1}\sum_{j=1}^{n}h_{jj}(\lambda)$, i.e., GCV criterion is

$$GCV(\lambda)=\frac{1}{n}\sum_{j=1}^{n}\{y_j-\hat{\mu}(t_j)\}^2/(1-n^{-1}\sum_{j=1}^{n}h_{jj})^2. \tag{2.2}$$

2.2 MCV and MGCV

Let K = $\{i_1,\cdots,i_k\}$ be a size k of index set. Then the delete -k multifold cross-validation in smoothing spline can be defined as

$$MCV(\lambda)=\frac{1}{\binom{n}{k}}\sum_{K}(y_K-\mu_{(K)}(t_K))'(y_K-\mu_{(K)}(t_K)), \tag{2.3}$$

where $\sum_{K}$ denotes $\binom{n}{k}$ numbers of all possible summation, $y_K=(y_{i1},\cdots,y_{ik})'$ and $\mu_{(K)}(t_K)=(\mu_{(K)}(t_{i1}),\cdots,\mu_{(K)}(t_{ik}))'$.

By the delete - k lemma (Kim 1996),

$$\hat{\mu}_{(K)}(t_K)=y_K-(I-H_K)^{-1}r_K,$$

where H_K is $k\times k$ submatrix of $H(\lambda)$ and $r_K=y_K-\hat{\mu}(t_K)$ is residual vector, we can express MCV(λ) in (2.3) as

$$MCV(\lambda)=\frac{1}{\binom{n}{k}}\sum_{K}r_K'(I-H_K)^{-2}r_K. \tag{2.4}$$

Therefore, the MGCV is obtained by replacing $(I-H_K)^{-2}$ by its average matrix $A=\frac{1}{\binom{n}{k}}\sum_K (I-H_K)^{-2}$, that is

$$MGCV(\lambda)=\frac{1}{\binom{n}{k}}\sum_K r_K' A\, r_K. \tag{2.5}$$

Remark : Natural and intuitive extension of GCV is $\frac{1}{\binom{n}{k}}\sum_K r_K' B\, r_K$, where $B=[\frac{1}{\binom{n}{k}}\sum_K (I-H_K)]^{-2}$. Note that this reduces to GCV when $k=1$. However, by simulation study, we found that our definition of MGCV shows better results than this.

3. Simulation study

Consider a cubic smoothing spline method, i.e., m = 2 in (1.2). The smoothing parameter λ governs the trade - off between the goodness - of - fit and the smoothness. As an extreme case, if $\lambda=\infty$, then the corresponding model becomes simple linear regression.

So far we have seen many criteria to estimate λ. For a certain criterion to be good, the estimate of λ based on that criterion should be close to the optimal λ in some sense. For example, if we generate random numbers from a simple linear regression, then good criterion gives $\hat{\lambda}=\infty$. In the case of GCV which is most widely used to estimate λ, the proportion of $\hat{\lambda}=\infty$ based on the random numbers from a simple linear regression is about 58 %. (Wahba 1990, Jeong 1996). In this section we will see the performance of MGCV numerically, and compare it with that of GCV.

Consider

$$y_j=\beta_0+\beta_1 x_j+e_j, \qquad j=1,\cdots,n$$

For simplicity, we let $\beta_0=\beta_1=1$, $x_j=(j-1)/(n-1)$, and $e_j\sim N(0,\sigma^2)$. For n = 20, 50, and 100, we generate random numbers from $N(0,0.1^2)$ and compute $\hat{\lambda}$ based on MGCV for k=2, 3, and replicate 1,000 times. Table 1 shows the proportion of $\hat{\lambda}=\infty$ out of 1,000. We see that the proportion decreases as n becomes large, and the effect of k is almost negligible.

When GCV is used, the proportions are .512, .557, .581 for $n=20, 50, 100$, respectively. Hence, we might say that MGCV is more efficient than GOV when n is small.

Table 1. Proportion of $\hat{\lambda}=\infty$ out of 1000 replications based on MGCV.

k \ n	20	50	100
2	.661	.622	.582
3	.662	.621	.589

Next, consider

$$y_j=\beta_0+\beta_1 x_j+\beta_2 x_j \exp(-2x_j)+e_j, \quad j=1,\cdots,n$$

Again, let $\beta_0=\beta_1=1$, $x_j=(j-1)/(n-1)$, and $e_j \sim N(0, 0.1^2)$.

For β_2 = .2 and .5, we compute $\hat{\lambda}_{GCV}$ and $\hat{\lambda}_{MGCV}$ based on GCV and MGCV, respectively. Let $\hat{\lambda}_{opt}$ be the minimizer of the average squared error (ASE),

$$ASE=\frac{1}{n}\sum_{j=1}^{n}(\hat{\mu}(x_j)-\mu(x_j))^2$$

where μ is the true function. We wish to see how close $\hat{\lambda}$ based on GCV or MGCV, to $\hat{\lambda}_{opt}$.

Let

P1 = proportion of $|\hat{\lambda}_{opt}-\hat{\lambda}_{GCV}| < 0.1$

P2 = proportion of $|\hat{\lambda}_{opt}-\hat{\lambda}_{GCV}| < 0.2$

Q1 = proportion of $|\hat{\lambda}_{opt}-\hat{\lambda}_{MGCV}| < 0.1$

Q2 = proportion of $|\hat{\lambda}_{opt}-\hat{\lambda}_{MGCV}| < 0.2$

Table 2 contains the results for k =2, 3, and 4. We see that $\hat{\lambda}_{MGCV}$ is always better than $\hat{\lambda}_{GCV}$. P1 and P2 increases as n becomes larger, however, Q1 and Q2 decreases as n becomes larger. Also, the effect of k is almost negligible. To see other aspects of GCV and MGCV. let

P3 = proportion of $|\hat{\lambda}_{opt} - \hat{\lambda}_{GCV}| > 0.3$

P4 = proportion of $|\hat{\lambda}_{opt} - \hat{\lambda}_{GCV}| > 1.0$

Q3 = proportion of $|\hat{\lambda}_{opt} - \hat{\lambda}_{MGCV}| > 0.3$

Q4 = proportion of $|\hat{\lambda}_{opt} - \hat{\lambda}_{MGCV}| > 1.0$.

Table 2. Performance of $\hat{\lambda}_{GCV}$ and $\hat{\lambda}_{MGCV}$.

P1 = $|\hat{\lambda}_{opt} - \hat{\lambda}_{GCV}| < 0.1$, P2 = $|\hat{\lambda}_{opt} - \hat{\lambda}_{GCV}| < 0.2$,

Q1 = $|\hat{\lambda}_{opt} - \hat{\lambda}_{MGCV}| < 0.1$, Q2 = $|\hat{\lambda}_{opt} - \hat{\lambda}_{MGCV}| < 0.2$,

β_2		n =20	n=50	n=100
.2	P1	.670	.754	.789
	Q1	.929 (k=2)	.853 (k=2)	.827 (k=2)
		.952 (k=3)	.865 (k=3)	
		.966 (k=4)		
	P2	.680	.766	.804
	Q2	.942 (k=2)	.859 (k=2)	.841 (k=2)
		.960 (k=3)	.872 (k=3)	
		.974 (k=4)		
.5	P1	.613	.723	.746
	Q1	.876 (k=2)	.820 (k=2)	.811 (k=2)
		.894 (k=3)	.818 (k=3)	
		.909 (k=4)		
	P2	.648	.748	.784
	Q2	.903 (k=2)	.851 (k=2)	.833 (k=2)
		.921 (k=3)	.845 (k=3)	
		.937 (k=4)		

Table 3 contains the result for k = 2, 3, and 4. Again, $\hat{\lambda}_{MGCV}$ is much better than $\hat{\lambda}_{GCV}$, especially when n is small.

Table 3. Performance of $\hat{\lambda}_{GCV}$ and $\hat{\lambda}_{MGCV}$.

$P3 = |\hat{\lambda}_{opt} - \hat{\lambda}_{GCV}| > 0.3$, $P4 = |\hat{\lambda}_{opt} - \hat{\lambda}_{GCV}| > 1.0$,

$Q3 = |\hat{\lambda}_{opt} - \hat{\lambda}_{MGCV}| > 0.3$, $Q4 = |\hat{\lambda}_{opt} - \hat{\lambda}_{MGCV}| > 1.0$,

β_2		n =20	n=50	n=100
.2	P3	.308	.221	.181
	Q3	.049 (k=2)	.128 (k=2)	.146 (k=2)
		.031 (k=3)	.116 (k=3)	
		.021 (k=4)		
	P4	.236	.151	.109
	Q4	.014 (k=2)	.066 (k=2)	.078 (k=2)
		.006 (k=3)	.057 (k=3)	
		.000 (k=4)		
.5	P3	.322	.228	.119
	Q3	.073 (k=2)	.126 (k=2)	.152 (k=2)
		.053 (k=3)	.131 (k=3)	
		.040 (k=4)		
	P4	.236	.159	.116
	Q4	.014 (k=2)	.065 (k=2)	.075 (k=2)
		.005 (k=3)	.060 (k=3)	
		.000 (k=4)		

Conclusively, MGCV is much more efficient and stable than GCV when n is small (less than 50, say).

4. Concluding remarks

In this thesis, we consider the problem of the choice of λ in smoothing spline. The selection of λ is usually accomplished by minimizing cross-validation (CV) or generalized cross-validation (GCV). However, the simple CV is poor candidate for estimating the prediction error and tend to overfit when the correct model is concerned. So, we suggest MCV and MGCV which is an extension of CV and GCV in smoothing spline. From the results of simulation study, MGCV is more efficient and stable than GCV in the sense of choosing λ close to the optimal value.

One disadvantage of MGCV is computation time when n and k are large, however, k-fold MGCV based on the idea of the k-fold MCV by Burman (1989) would be a good alternative. Also, the choice of k which seems to be still an open problem can be a good future research area.

References

[1] Akaike, H. (1974). A new look at statistical model identification, *IEEE Transactions on Automatic Control*, 19, 716-723.

[2] Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, 71, 353-60

[3] Burman, P. (1989). A comparative study of ordinary cross-validation, u-fold cross-validation and the repeated learning-testing methods, *Biometrika* 76, 503-514.

[4] Cox, D. D. (1984). Multivariate smoothing spline functions, *Journal of SIAM* 21, 789-813.

[5] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline function : Estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematik*, 31, 377-403.

[6] Efron, B. (1983). Estimating the error rate of a prediction rule : Improvement on cross-validation, *Journal of the American Statistical Association*, 78, 316-331.

[7] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*. 81, 461-470.

[8] Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression,* Marcel Dekker, New York.

[9] Geisser, S. (1975). The predictive sample reuse method with application, *Journal of the American Statistical Association* , 70, 320-328.

[10] Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation, *The Annals of Statistics*, 13, 1465-81.

[11] Jeong, M. (1996). A study on model checking in nonparametric regression, Ph. D. Thesis, Department of Statistics Pusan National University.

[12] Kim, C. (1996). Cook's distance in spline smoothing, *Statistics and Probability Letters,* 31, 139-144.

[13] Li, K. C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing, *The Annals of Statistics*, 14, 1101-1112.

[14] Mallows, C. L. (1973). Some comment on C_p, {\em Technometrics}, 15, 661-675.

[15] Shao, J. (1993). Linear model selection by cross-validation, *Journal of the American Statistical Association*, 88, 486-494.

[16] Shibata, R. (1981). On optimal selection of regression variables, *Biometrika*, 68, 45-54.

[17] Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion), *Journal of the Royal Statistical Society*, ser. B., 47, 1-52.

[18] Speckman, P. (1982). Efficient nonparametric regression with cross-validated smoothing splines, *Manuscript.*

[19] Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction (with discussion), *Journal of the Royal Statistical Society*, ser. B., 36, 114-147.

[20] Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy, *Journal of SIAM*, 14, 651-667.

[21] Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia.

[22] Wegman, E.J. and Wright, I.W. (1983). Splines in statistics, *Journal of the American Statistical Association*, 78, 351-365.

[23] Zhang, P. (1993). Model selection via multifold cross-validation, *The Annals of Statistics*, 21. 299-313.