Bandwidth Selection for Nadaraya-Watson Kernel Estimator Using Cross-Validation Based on Different Penalty Functions

Yumin Zhang(✉)

Management Department, Hebei Finance University, Baoding 071051, Hebei, China
zhangyumin071051@126.com

Abstract. The traditional cross-validation usually selects an over-smoothing bandwidth for kernel regression. The penalty function based cross-validation (e.g., generalized cross-validation (CV_{GCV}), the Shibata's model selector (CV_S), the Akaike's information criterion (CV_{AIC}) and the Akaike's finite prediction error (CV_{FPE})) are introduced to relieve the problem of selecting over-smoothing bandwidth parameter by the traditional cross-validation for kernel regression problems. In this paper, we investigate the influence of these four different penalty functions on the cross-validation based bandwidth selection in the framework of a typical kernel regression method, i.e., the Nadaraya-Watson kernel estimator (NWKE). Firstly, we discuss the mathematical properties of these four penalty functions. Then, experiments are given to compare the performance of aforementioned cross-validation methods. Finally, we give guidelines for the selection of different penalty functions in practical applications.

Keywords: Cross-validation · Kernel regression · Nadaraya-Watson kernel estimator · Penalty function

1 Introduction

The kernel regression [1] is a non-parametric technique to construct the conditional expectation of a given random variable in statistics and is one of the most commonly used regression analysis methods. Its objective is to find a non-linear mapping between the input variable X and the output Y [2]. The conditional expectation of input X with respect to the output Y can be written as Eq. (1):

$$E(Y|X) = g(X), \tag{1}$$

where $g(X)$ is an unknown regression function that needs to be estimated [3]. Alternatively, we can rewrite Eq. (1) by the functional relation as follows:

$$Y = g(X) + \varepsilon,\ E(\varepsilon) = 0. \tag{2}$$

X. Wang et al. (Eds.): ICMLC 2014, CCIS 481, pp. 88–96, 2014.
DOI: 10.1007/978-3-662-45652-1_10

Let $S = \{(x_i, y_i) \,|x_i \in R, y_i \in R, i = 1, 2, \cdots, N\}$ be the given dataset, where x_i and y_i are N observations of random variables X and Y. According to Eq. (3), the kernel regression finds the estimation of function g [4]:

$$\tilde{g}(x) = \sum_{i=1}^{N} w_i y_i, \tag{3}$$

where $\tilde{g}(x)$ is the estimated regression function. w_i is a function with respect to the input x and denotes the weight of the output y_i ($w_i > 0$ and $\sum_{i=1}^{N} w_i = 1$). The target of kernel regression is to find an optimal weight set $\{w_1, w_2, \cdots, w_N\}$ such that the estimated regression function $\tilde{g}(x)$ can approximate the true regression function $g(x)$ [5]. There are three commonly used methods to determine the output weight w_i: the Nadaraya-Watson kernel estimator [6], the Priestley-Chao kernel estimator [7] and the Gasser-Müller kernel estimator [8]. These three estimators are all implemented based on the Parzen window method [9].

It is well known that [10–12] the selection of smoothing parameter or bandwidth h is very important for the regression estimation performance of kernel regression methods. There are many sophisticated methods to find the optimal bandwidth, e.g., the cross-validation [13], penalizing functions [14], the plug-in [14] and the bootstrap methods [15]. None of them constantly outperforms the others. Therefore, we focus on studying the cross-validation based bandwidth selection which uses the leave-one-out strategy to determine the optimal bandwidth. In [14], some penalty functions are introduced to cross-validation to design an alternative calculation paradigm for cross-validation. In this paper, we introduce four different penalty functions based on cross-validation methods, i.e., the generalized cross-validation ($\mathrm{CV_{GCV}}$) [16], the Shibata's model selector ($\mathrm{CV_S}$) [17], the Akaike's information criterion ($\mathrm{CV_{AIC}}$) [17] and the Akaike's finite prediction error ($\mathrm{CV_{FPE}}$) [17]. We investigate the influence of these four different penalty functions on the cross-validation based bandwidth selection in the framework of the Nadaraya-Watson kernel estimator (NWKE). Firstly, mathematical properties of different penalty functions are discussed. Then, experiments are given to compare the performance of aforementioned cross-validation methods and we give guidelines for the selections of different penalty functions for future practical applications.

2 NWKE

NWKE [6] uses the following Eq. (4) to compute the output weight $w_i, i = 1, 2, \cdots, N$ for y_i:

$$w_i = \frac{1}{Nh} \frac{K\left(\frac{x - x_i}{h}\right)}{\tilde{p}(x)}, \tag{4}$$

where, $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$ is the Gaussian kernel function and $\tilde{p}(x)$ is the estimated p.d.f. of random variable X based on the given observations

$x_1, x_2, \cdots, x_N$. According to the Parzen window method [9], we can get the expression of $\tilde{p}(x)$ as the following Eq. (5):

$$\tilde{p}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right), \tag{5}$$

where, h is the bandwidth parameter.

By bringing Eqs. (4) and (5) into Eq. (3), we can get the regression function Eq. (6) with $\mathrm{NW}_{\mathrm{KRE}}$:

$$\tilde{g}_{\mathrm{NWKE}}(x) = \frac{\sum_{i=1}^{N} \left[K\left(\frac{x - x_i}{h}\right) y_i\right]}{\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)}. \tag{6}$$

3 Cross-Validation Bandwidth Choice

NWKE is a more natural method for the data usage in the regression analysis. The regression performance of NWKE in Eq. (6) mainly depends on the selection of bandwidth parameter h. Cross-validation [13] is one of available bandwidth selection schemes, which uses the following formulas to determine the optimal bandwidth for NWKE:

$$h_{opt} = \arg\min_{h \in \mathrm{H}} \left(CV(h)\right), \tag{7}$$

$$CV(h) = \sum_{i=1}^{N} \left\{y_i - \tilde{g}_{\mathrm{NWKE}-i}(x_i)\right\}^2, \tag{8}$$

where, H is the domain of discourse of bandwidth h, $\tilde{g}_{\mathrm{NWKE}-i}(x)$ is NWKE which is obtained without using the i-th instance (x_i, y_i).

The penalty function based cross-validation calculates the optimal bandwidth according to the following Eqs. (9), (10) and (11):

$$h'_{opt} = \arg\min_{h \in \mathrm{H}} \left(CV'(h)\right), \tag{9}$$

$$CV'(h) = \sum_{i=1}^{N} \left\{y_i - \tilde{g}_{\mathrm{NWKE}-i}(x_i)\right\}^2 \pi\left(W(x_i)\right), \tag{10}$$

$$W(x) = \frac{K(0)}{\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)}, \tag{11}$$

where, $\pi(u)$ is the penalty function. In this paper, we select four different penalty function based cross-validations as follows. Figure 1 gives the curve presentation of these four penalty functions.

- generalized cross-validation-$\mathrm{CV}_{\mathrm{GCV}}$:

$$\pi(u) = (1 - u)^{-2}; \tag{12}$$

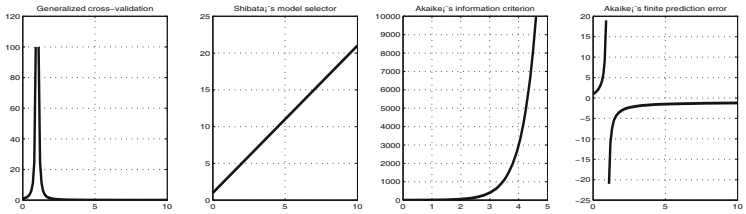

Fig. 1. 4 different penalty functions

- Shibata's model selector-$\mathrm{CV_S}$:

$$\pi(u) = 1 + 2u; \tag{13}$$

- Akaike's information criterion-$\mathrm{CV_{AIC}}$:

$$\pi(u) = \exp(2u); \tag{14}$$

- Akaike's finite prediction error-$\mathrm{CV_{FPE}}$:

$$\pi(u) = \tfrac{1+u}{1-u}. \tag{15}$$

4 The Experimental Comparison Among Different Cross-Validations

In this section, we will conduct some experiments to compare the regression performances of NWKE with different bandwidth selection schemes, i.e., traditional cross-validation (CV), $\mathrm{CV_{GCV}}$, $\mathrm{CV_S}$, $\mathrm{CV_{AIC}}$ and $\mathrm{CV_{FPE}}$. We compare the regression accuracy and method stability of above-mentioned five methods, and the regression accuracy and method stability are calculated as the following Eqs. (16) and (17) respectively:

- The fitting accuracy is measured by mean squared error between the real output y_i and the predicted output $\tilde{y}_i$:

$$mse = \tfrac{1}{N} \sum_{i=1}^{N} (\tilde{y}_i - y_i)^2. \tag{16}$$

- The stability is measured by the standard derivation of Q mses, i.e.,

$$std = \sqrt{\tfrac{1}{Q} \sum_{i=1}^{Q} (mse_i - \mu)^2}, \mu = \tfrac{1}{Q} \sum_{i=1}^{Q} mse_i. \tag{17}$$

For the preparation of experimental datasets, we select six testing functions:

$$\begin{aligned} &y_1 = 1 - x + \exp\left[-200\left(x - 0.5\right)^2\right] + \varepsilon, \\ &x \sim \mathrm{U}\left(0, 1\right), \varepsilon \sim \mathrm{N}\left(0, 0.1\right); \end{aligned} \tag{18}$$

$$\begin{aligned} &y_2 = x + \frac{4\exp\left(-2x^2\right)}{\sqrt{2\pi}} + \varepsilon, \\ &x \sim \mathrm{U}\left(-3, 3\right), \varepsilon \sim \mathrm{N}\left(0, 1\right); \end{aligned} \tag{19}$$

$$\begin{aligned} &y_3 = \sin\left[2\pi\left(1 - x\right)^2\right] + x + 0.2\varepsilon, \\ &x \sim \mathrm{U}\left(0, 1\right), \varepsilon \sim \mathrm{N}\left(0, 1\right); \end{aligned} \tag{20}$$

$$\begin{aligned} &y_4 = x + 2\sin\left(1.5x\right) + \varepsilon, \\ &x \sim \mathrm{U}\left(0, 10\right), \varepsilon \sim \mathrm{N}\left(0, 1\right); \end{aligned} \tag{21}$$

$$\begin{aligned} &y_5 = \left[\sin\left(2\pi x^3\right)\right]^3 + 0.2\varepsilon, \\ &x \sim \mathrm{U}\left(0, 1\right), \varepsilon \sim \mathrm{N}\left(0, 1\right); \end{aligned} \tag{22}$$

$$\begin{aligned} &y_6 = \sin\left(3\pi x\right) + 0.2\varepsilon, \\ &x \sim \mathrm{U}\left(0, 1\right), \varepsilon \sim \mathrm{N}\left(0, 1\right). \end{aligned} \tag{23}$$

We firstly give the descriptions concerning the relationship between the different CVs and bandwidths. Then, we compare the regression performances of NWKE with different optimal bandwidths h_{opt}. The comparative results on six testing functions are respectively summarized in Figures 2–7. For each testing function in this experiment, 200 data points are randomly generated. The minimal CVs and corresponding optimal bandwidths h_{opt} are listed in Table 1. We find the optimal bandwidth found by CV is smaller in comparison with other four CVs with different penalty functions. In other words, traditional CV in our comparisons is easier to obtain a rough bandwidth. Then, we give a more detailed comparison among aforementioned five methods (i.e., CV, $\mathrm{CV_{GCV}}$, $\mathrm{CV_S}$, $\mathrm{CV_{AIC}}$ and $\mathrm{CV_{FPE}}$) in terms of regression accuracy and stability. For each testing function, the final *mse* and *std* are respectively the average and standard derivation of Q=10 repetitions. In every run, there are 200 data points which are generated randomly. We compare the performance of NWKE with different optimal bandwidths solved by five CVs respectively. The comparative results are summarized in Table 2. Through observing the experimental results, we can get the following conclusions:

- The traditional CV obtains the better regression accuracy and stability. It tells us that CV can select a more rough bandwidth for the smaller input (e.g., $x \in [0, 1]$, $x \in [-3, 3]$ or $x \in [0, 10]$ in the employed testing functions).
- $\mathrm{CV_S}$ achieves a better regression accuracy compared with $\mathrm{CV_{GCV}}$, $\mathrm{CV_{AIC}}$ and $\mathrm{CV_{FPE}}$. However, its stability is worse on the testing functions y_1, y_2, y_4 and y_5.
- $\mathrm{CV_{AIC}}$ and $\mathrm{CV_{FPE}}$ obtain the comparative regression accuracy, but the stability of $\mathrm{CV_{AIC}}$ is better than $\mathrm{CV_{FPE}}$.

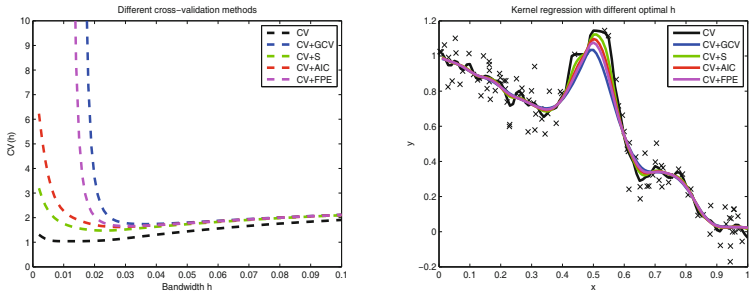

Fig. 2. The kernel regression on $y_1 = 1 - x + \exp\left[-200\,(x-0.5)^2\right]$ dataset

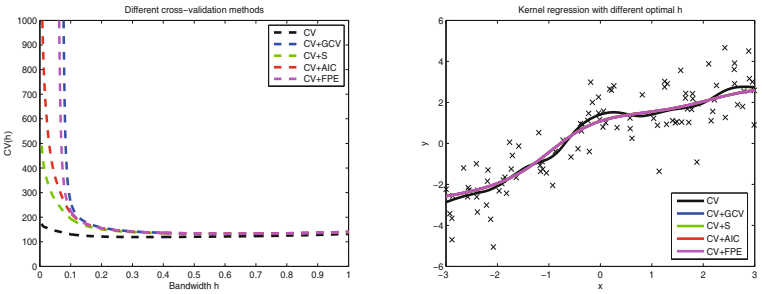

Fig. 3. The kernel regression on $y_2 = x + \frac{4\exp\left(-2x^2\right)}{\sqrt{2\pi}}$ dataset

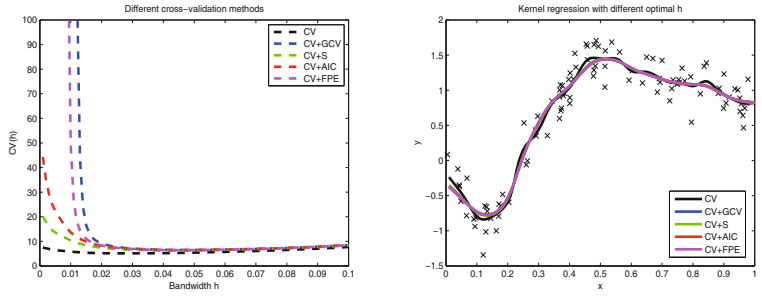

Fig. 4. The kernel regression on $y_3 = \sin\left[2\pi\,(1-x)^2\right] + x$ dataset

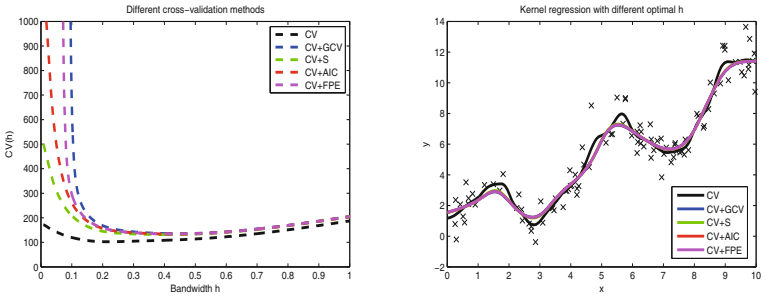

Fig. 5. The kernel regression on $y_4 = x + 2\sin(1.5x)$ dataset

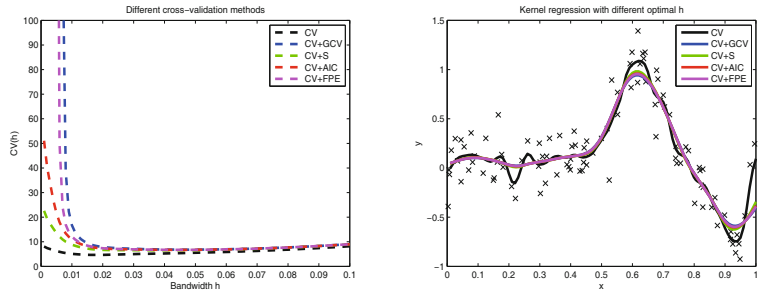

Fig. 6. The kernel regression on $y_5 = \left[\sin\left(2\pi x^3\right)\right]^3$ dataset

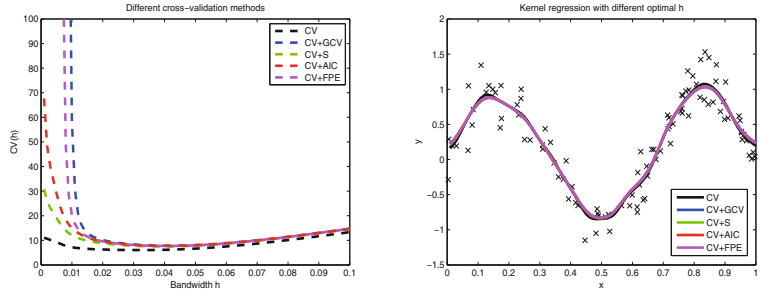

Fig. 7. The kernel regression on $y_6 = \sin(3\pi x)$ dataset

Table 1. The minimal CV and corresponding optimal bandwidth h_{opt}

CV	y_1		y_2		y_3		y_4		y_5		y_6	
methods	CV	h_{opt}	CV	h_{opt}	CV	h_{opt}	CV	h_{opt}	CV	h_{opt}	CV	h_{opt}
CV	1.042	0.012	119.364	0.330	5.115	0.029	102.400	0.218	4.685	0.021	6.224	0.029
CV+GCV	1.738	0.036	133.378	0.641	6.511	0.047	135.363	0.444	6.539	0.032	8.021	0.042
CV+S	1.474	0.023	132.633	0.620	6.313	0.043	131.399	0.414	6.176	0.029	7.749	0.038
CV+AIC	1.622	0.028	133.119	0.634	6.437	0.045	133.892	0.433	6.399	0.031	7.920	0.040
CV+FPE	1.656	0.031	133.126	0.634	6.443	0.046	133.994	0.434	6.413	0.031	7.928	0.041

Table 2. The regression performance of Nadaraya-Watson kernel estimator based on different CV methods

CV	**Regression accuracy**						**Method stability**					
methods	y_1	y_2	y_3	y_4	y_5	y_6	y_1	y_2	y_3	y_4	y_5	y_6
CV	0.0074	0.8217	0.0307	0.8549	0.0323	0.0276	0.0030	0.1898	0.0126	0.1705	0.0102	0.0113
CV+GCV	0.0097	0.9959	0.0489	1.1222	0.0488	0.0432	0.0039	0.1564	0.0064	0.2021	0.0165	0.0118
CV+S	0.0087	0.9693	0.0437	1.0434	0.0401	0.0342	0.0037	0.1551	0.0068	0.1930	0.0163	0.0110
CV+AIC	0.0093	0.9853	0.0461	1.0902	0.0448	0.0380	0.0037	0.1549	0.0072	0.1901	0.0168	0.0110
CV+FPE	0.0093	0.9866	0.0468	1.0966	0.0458	0.0398	0.0036	0.2873	0.0098	0.2678	0.0115	0.0131

5 Conclusions

In this paper, we investigate the influence of four different penalty functions on the cross-validation bandwidth selection in the framework of Nadaraya-Watson kernel regression estimator. The derived conclusions from our experiments give guidelines for the selection of different penalty functions for future applications.

References

1. Harta, J.D., Wehrlya, T.E.: Kernel Regression Estimation Using Repeated Measurements Data. Journal of the American Statistical Association **81**(396), 1080–1088 (1986)
2. Hart, J.D.: Kernel Regression Estimation With Time Series Errors. Journal of the Royal Statistical Society, Series B: Methodological **53**(1), 173–187 (1991)
3. Herrmanna, E.: Local Bandwidth Choice in Kernel Regression Estimation. Journal of Computational and Graphical Statistics **6**(1), 35–54 (1997)
4. Dabo-Nianga, S., Rhomarib, N.: Kernel Regression Estimation in a Banach Space. Journal of Statistical Planning and Inference **139**(4), 1421–1434 (2009)
5. Girarda, S., Guilloub, A., Stupflerc, G.: Frontier Estimation with Kernel Regression on High Order Moments. Journal of Multivariate Analysis **116**, 172–189 (2013)
6. Watson, G.S.: Smooth Regression Analysis. Sankhy: The Indian Journal of Statistics, Series A (1961–2002) **26**(4), 359–372 (1964)
7. Priestley, M.B., Chao, M.T.: Non-Parametric Function Fitting. Journal of the Royal Statistical Society. Series B (Methodological) **34**(3), 385–392 (1972)

8. Gasser, T., Müller, H.G.: Kernel estimation of regression functions. In: Gasser, T., Rosenblatt, M. (eds.) Smoothing Techniques for Curve Estimation. Lecture Notes in Mathematics, vol. 757, pp. 23–68. Springer, Heidelberg (1979)
9. Parzen, E.: On Estimation of a Probability Density Function and Mode. Annals of Mathematical Statistics **33**(3), 1065–1076 (1962)
10. Scott, D.W.: Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley and Sons, Inc. (1992)
11. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall (1995)
12. Wang, X.Z., He, Y.L., Wang, D.D.: Non-Naive Bayesian Classifiers for Classification Problems with Continuous Attributes. IEEE Transactions on Cybernetics (2013), doi:10.1109/TCYB.2013.2245891
13. Leung, D.H.Y.: Cross-Validation in Nonparametric Regression With Outliers. The Annals of Statistics **33**(5), 2291–2310 (2005)
14. Härdle, W.: Applied Nonparametric Regression. Cambridge University Press (1994)
15. Hardle, W., Marron, J.S.: Bootstrap Simultaneous Error Bars for Nonparametric Regression. The Annals of Statistics **19**(2), 778–796 (1991)
16. Golub, G.H., Heath, M., Wahba, G.: Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter **21**(2), 215–223 (1979)
17. Wechsler, H., Duric, Z., Li, F.Y., et al.: Motion estimation using statistical learning theory. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(4), 466–478 (2004)