Proceedings of 2012 4th International Conference on Machine Learning and Computing
IPCSIT vol. 25 (2012) © (2012) IACSIT Press, Singapore

Determining the Optimal Bandwidth Based on Multi-criterion Fusion

Hai-Li Liang [1+], Xian-Min Shen [2], Chun-Fang Ding [1] and Hai-Ying Li [1]

[1] Department of Information Science and Technology, Xingtai University, Xingtai 054001, China
[2] Department of Physics, Xingtai University, Xingtai 054001, China

Abstract. It is very important for the probability density estimation criteria to determine the optimal bandwidth. There are three typical bandwidth selection methods: the bootstrap method, the least-squares cross-validation (LSCV) method and the biased cross-validation (BCV) method. From the perspective of data analysis, producing a stable or robust solution is a desired property of a bandwidth selection method. However, the issue of robustness is often overlooked in real world applications. In this paper, we propose to improve the robustness of bandwidth selection method by using multiple bandwidth evaluation criteria. Based on this idea, a multi-criterion fusion-based bandwidth selection (MCF-BS) method is developed with the goal of improving the estimation performance. We carry out some numerical simulations on four univariate artificial datasets: Uniform dataset, Normal dataset, Exponential dataset and Rayleigh dataset. The finally comparative results show that our strategies are well-performed and the designed MCF-BS can obtain the best estimation accuracy than the existing bandwidth selection methods.

Keywords: probability density estimation, bandwidth selection, bootstrap, least-squares cross-validation, biased cross-validation, robustness

1. Introduction

Probability density estimation (short of "PDE") [1, 2] is a very important and necessary technique in many theoretical studies and practical applications of probability and statistic. PDE aims to explore the underlying probability density function $p(x)$ from the observed dataset $X=\{x_1, x2, \cdots, x_N\}$, N is the size of dataset X, by using some data-interpolation methods, e.g. Parzen window method [3, 4]. The estimated density function can be expressed according to Parzen window method as follows:

$$\hat{p}(x)=\frac{1}{Nh}\sum_{i=1}^{N}\exp\left[-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2\right], \tag{1}$$

where, we note that all our studies are based on the univariate data, N is the number of samples belonging to dataset X, h is the determined bandwidth.

The purpose of PDE is to make the estimated density $\hat{p}(x)$ near the true density $p(x)$ as soon as possible. That is to say, the error between $\hat{p}(x)$ and $p(x)$ should reach minimum. There are a lot of error criteria [5] to evaluate the estimated performance. In this study, the Mean Integrated Squared Error (MISE) [6] and Integrated Squared Error (ISE) [7] are selected as the researching pools. Their expressions are shown in the following equations (2) and (3):

$$\mathrm{MISE}(h)=\mathrm{E}\left\{\int\left[\hat{p}(x)-p(x)\right]^2 dx\right\}, \tag{2}$$

$$\mathrm{ISE}(h)=\int\left[\hat{p}(x)-p(x)\right]^2 dx\,. \tag{3}$$

[+] Corresponding author. Tel.: +86 15930241029.
E-mail address: haily.liang@gmail.com.

By using different bandwidth selection methods to solve the error criteria (2) or (3), we can obtain the different optimization expressions of bandwidth h. The mostly used bandwidth-solving methods are the bootstrap method [8], the least-squares cross-validation (LSCV) method [9] and the biased cross-validation

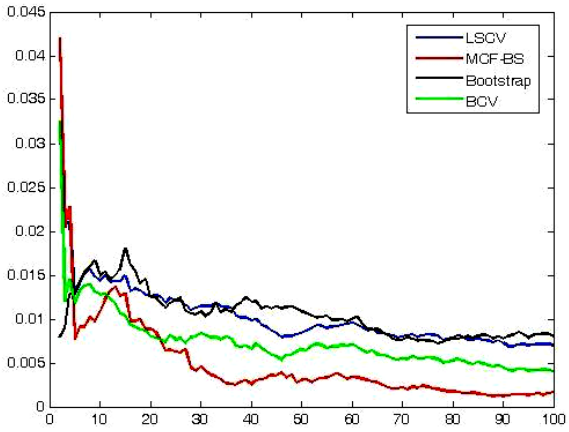

Fig. 1: The comparison on Uniform distribution.

(BCV) method [10]. No matter which method is used to determine the optimal bandwidth, the parallel between these bandwidth-selection methods is that the brute-force or exhaustive search strategies should be used to find the optimal parameters, e.g. a general quasi-Newton method in S-PLUS function *nlmin*. In our previous work [12], five particle swarm optimization (PSO) [11] algorithms are applied to solve the optimal bandwidth. In this study, we only use the standard PSO to find the optimal bandwidths for the sake of computational complexity. From the perspective of data analysis, producing a stable or robust solution is a desired property of a bandwidth selection method. However, the issue of robustness is often overlooked in real world applications when single bandwidth evaluation criterion is used. In this paper, we propose to improve the robustness of bandwidth selection method by using multiple bandwidth evaluation criteria. Based on this idea, a multi-criterion fusion-based bandwidth selection (MCF-BS) method is developed with the goal of improving the estimation performance.

In order to validate the feasibility and effectiveness of our proposed strategies, four univariate artificial datasets are generated randomly: Uniform dataset, Normal dataset, Exponential dataset and Rayleigh dataset. Then, we test the estimated performances of four bandwidth selection methods: bootstrap, LSCV, BCV, and MCF-BS. In order to test the estimation performances of four different bandwidth selectors, four different types of univariate artificial datasets are generated randomly: Uniform dataset, Normal dataset, Exponential dataset and Rayleigh dataset. The experimental results show that MCF-BS can obtain the best estimation performance. Because the mechanism of multi-criterion fusion can guarantee that the bandwidth selection algorithm selects a stable and robust bandwidth for the probability density estimation application.

2. Multi-Criterion Fusion-Based Bandwidth Selection (MCF-BS)

In There are many different bandwidth-solving methods which can be used as the bandwidth selectors. In this paper, three commonly used bandwidth selectors are introduced: bootstrap method [8], least-squares cross-validation (LSCV) method [9] and biased cross-validation (BCV) method [10].

The bootstrap method proposed by Taylor [8] finds the optimal bandwidth by solving the following error criterion:

$$\underset{\text{bootstrap}}{\text{MISE}}(h) = \text{E}\left\{\int\left[\hat{p}(x) - \hat{p}^*(x)\right]^2 dx\right\}, \tag{4}$$

where, $\hat{p}(x)$ is the estimated density (1) based on the given dataset X, $\hat{p}^*(x)$ is the bootstrap density which is estimated by using the re-sampling dataset from the density $\hat{p}(x)$. Taylor proved that when the bootstrap method uses the Gaussian kernel function to estimate the density, the process of re-sampling dataset is not necessary. So, the error criterion function (4) can be derived as the following form:

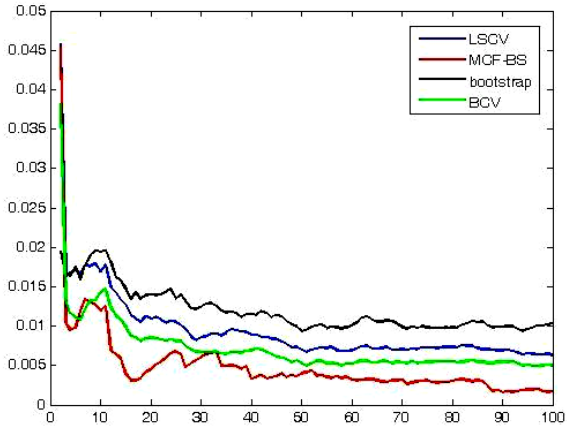

Fig. 2: The comparison on Normal distribution.

$$\underset{\text{bootstrap}}{\text{MISE}}(h)=\frac{1}{2\sqrt{\pi}Nh}+\frac{1}{2\sqrt{\pi}N^2h}\times\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left\{\frac{N-1}{\sqrt{2}N}\exp\left[-\frac{1}{8}\left(\frac{x_i-x_j}{h}\right)^2\right]+\exp\left[-\frac{1}{4}\left(\frac{x_i-x_j}{h}\right)^2\right]-\frac{2\sqrt{2}}{\sqrt{3}}\exp\left[-\frac{1}{6}\left(\frac{x_i-x_j}{h}\right)^2\right]\right\}. \tag{5}$$

LSCV uses one more direct method to obtain an optimal bandwidth by computing the following error criterion:

$$\underset{\text{LSCV}}{\text{ISE}}(h)=\int[\hat{p}(x)-p(x)]^2\,dx=\int[\hat{p}(x)]^2\,dx-2\int\hat{p}(x)p(x)dx+\int[p(x)]^2\,dx\,. \tag{6}$$

From the equation (6), we can find that the term is not relevant with the band-parameter h. So, the optimal bandwidth parameter can be obtained by minimizing the following error criterion (7):

$$\begin{aligned}\underset{\text{LSCV}}{\text{ISE}}^{*}(h)&=\int[\hat{p}(x)]^2\,dx-2\int\hat{p}(x)p(x)dx\\&=\frac{1}{2\sqrt{\pi}Nh}+\frac{1}{2\sqrt{\pi}N^2h}\times\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left\{\exp\left[-\frac{1}{4}\left(\frac{x_i-x_j}{h}\right)^2\right]-\frac{2}{\sqrt{2}}\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h}\right)^2\right]\right\}.\end{aligned} \tag{7}$$

Because $\text{MISE}(h)=\int Bias^2(\hat{p}(x))dx+\int Var|\hat{p}(x)|dx$, where $Bias[\hat{p}(x)]=\text{E}[\hat{p}(x)]-p(x)$ and $Var[\hat{p}(x)]=\text{E}\left[p^2(x)\right]-\text{E}^2[\hat{p}(x)]$, we can derive the following equation (8) by using (1) to substitute the corresponding $\hat{p}(x)$ in *Bias* and *Var*:

$$\text{MISE}(h)=\frac{1}{Nh}\int\left[\text{K}(x)\right]^2dx+\frac{1}{4}h^4\int x^2\text{K}(x)dx\int[p''(x)]^2\,dx\,, \tag{8}$$

where, $K(x)$ is the Gaussian kernel function, $\int[p''(x)]^2\,dx$ can not be computed because the true density $p(x)$ is unknown. $\int[p''(x)]^2\,dx=\int[\hat{p}''(x)]^2\,dx-\frac{1}{Nh^5}\int\left[\text{K}''(x)\right]^2dx$ is always used to estimate the unknown part in BCV, So, the error criterion of BCV is as follows:

$$\underset{\text{BCV}}{\text{MISE}}(h)=\frac{1}{2\sqrt{\pi}Nh}+\frac{1}{4N(N-1)h}\times\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left\{\left[\left(\frac{x_i-x_j}{h}\right)^4-6\left(\frac{x_i-x_j}{h}\right)^2+3\right]\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i-x_j}{h}\right)^2\right]\right\}. \tag{9}$$

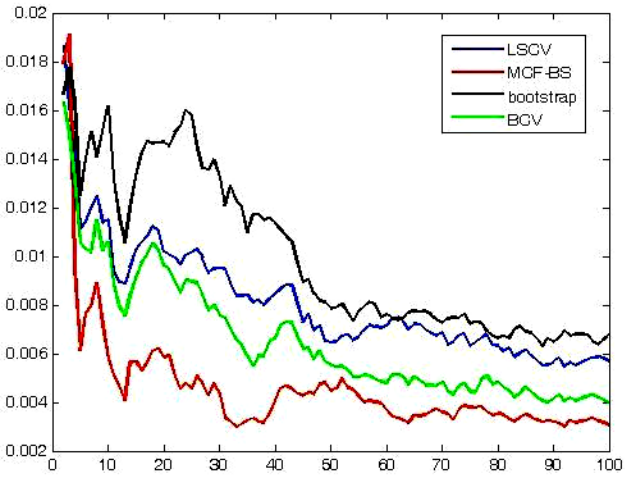

Fig. 3: The comparison on Exponential distribution.

In our study, we want to use the score-based multi-criterion fusion [13] to integrate three bandwidth selection criterion mentioned above. In score-based multi-criterion fusion, each basis criterion first produces a corresponding score; a score combination algorithm is then employed to aggregate the multiple scores into one consensus score. In score aggregating, it is essential to ensure that the scores produced by different basis criteria are comparable. Thus, score normalization should be done before score combination is performed. In this study, the scores produced by each basis criterion are normalized to the range of [0, 1]. Assume c_i is the score produced by basis criterion i, the score normalization is performed as follows:

$$, \ c_i' = \frac{c_i - c_{\min}}{c_{\max} - c_{\min}} \tag{10}$$

where, the criterion $c_i \in \left\{ \underset{\text{bootstrap}}{\text{MISE}}(h), \underset{\text{LSCV}}{\text{ISE}^*}(h), \underset{\text{BCV}}{\text{MISE}}(h) \right\}$, $c_{\max} = \max\left\{ \underset{\text{bootstrap}}{\text{MISE}}(h), \underset{\text{LSCV}}{\text{ISE}^*}(h), \underset{\text{BCV}}{\text{MISE}}(h) \right\}$, and $c_{\min} = \min\left\{ \underset{\text{bootstrap}}{\text{MISE}}(h), \underset{\text{LSCV}}{\text{ISE}^*}(h), \underset{\text{BCV}}{\text{MISE}}(h) \right\}$.

For all the basis criteria, it is assumed that the larger the score, the better the feature. A simple yet effective score combination method is to take the average of the normalized scores:

$$\underset{\text{MCF-BS}}{\text{MISE}}(h) = \frac{1}{3}\sum_{i=1}^{3} c_i' . \tag{11}$$

3. The Experiments and Results

In standard PSO algorithms, the number of particles in the initial population is 100 and the maximal iteration is 100. Our experiments are arranged as follows: For each bandwidth selector (Bootstrap, LSCV, BCV or MCF-BS), we use the standard PSO algorithm to search the optimal bandwidth based on four different types of univariate artificial datasets. Every type of dataset is generated 100 times randomly. The average results based on these 100 datasets is summarized for some distribution. The mean squared error is used to evaluate the estimation performance. The detailed comparative results are listed in Fig.1-Fig.4.

From the experimental results, we can get the following three observations: (1) With the increase of iteration, the MSE decreases gradually. And, when the optimal bandwidth is searched, MSE becomes steadily; (2) The estimating performances of bootstrap are worst among all the competitive bandwidth selection algorithms. From the pictures we can see that the curves corresponding bootstrap are located on top of the other curves; (3) MCF-BS obtains the best estimation performances. Because the mechanism of multi-criterion fusion can guarantee that the bandwidth selection algorithm selects a stable and robust bandwidth for the probability density estimation application.

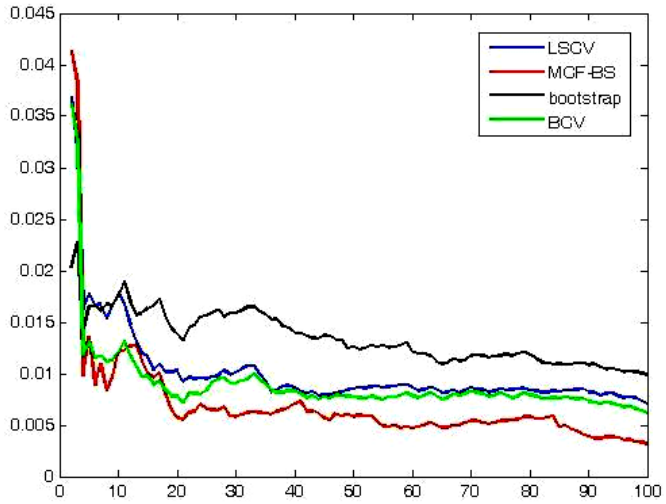

Fig. 4: The comparison on Rayleigh distribution.

4. Conclusions

In this paper, we propose to improve the robustness of bandwidth selection method by using multiple bandwidth evaluation criteria. Based on this idea, a multi-criterion fusion-based bandwidth selection (MCF-BS) method is developed. The finally comparative results show that our strategies are well-performed and the designed MCF-BS can obtain the best estimation accuracy among the existing selection methods.

5. References

[1] M.P. Wand, M.C. Jones. *Kernel Smoothing*. Chapman and Hall, 1995.

[2] D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, Inc, 1992.

[3] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 1962, **33** (3): 1065-1076.

[4] M.G. Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, 2001, **2**: 299-312.

[5] M.C. Jones, J.S. Marron, and S.J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 1996, **91** (433): 401-407.

[6] J.S. Marron, and M.P. Wand. Exact mean integrated squared error. *The Annals of Statistics*, 1992, **20** (2): 712-736.

[7] C.R. Heathcote. The integrated squared error estimation of parameters. *Biometrika*, 1977, **64** (2): 255-264.

[8] C.C. Taylor. Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrik*, 1989, **76** (4): 705-712.

[9] A.W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 1984, **71** (2): 353-360.

[10] D.W. Scott, and G.R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 1987, **82** (400): 1131-1146.

[11] J. Kennedy, R. Eberhart. Particle swarm optimization. *Proceedings of the 1995 International Conference on Neural Network, Perth, Australia*, 1995, pp. 1941-1948.

[12] H.L. Liang, X.M. Shen. Applying particle swarm optimization to determine the bandwidth parameter in probability density estimation. *Proceedings of the 2011 International Conference on Machine Learning and Cybernetics*, 2011, pp. 1362-1367.