⊠연구논문

비모수적 회귀함수 추정에 대한 Family Approach

정성석
전북대학교 통계학과

The Family Approach to Nonparametric Estimation of the Regression Function

Sung-Suk Chung
Dept. of Statistics, Chonbuk National University

Abstract

The smoothing parameter or bandwidth is crucial to performance of the kernel based regression estimator. So the choice of a "optimal" smoothing parameter produce a single curve estimate. If a single estimate is replaced by a family of estimates, it become easy that we understand what varies with choice of the smoothing parameter. This paper suggests the threshold of the maximum bandwidth and the number of the family members in the regression context.

1. Introduction

A regression function describes a general relationship between an explanatory variable and a response variable. To estimate the regression function nonparametrically, kernel-based smoothers are often used because of their simplicity and implementation. There are a great deal of theoretical research on kernel-based smoothers. See, for example, the books of Silverman(1986), Eubank(1988), Müller (1988), Härdle(1990), Scott(1992), Wand and Jones(1995) and Fan and Gijbels(1996).

Suppose that we have observations (X_i, Y_i); $i=1, \cdots, n$ from a population having a density $f(x, y)$. Let $f_X(x)$ be the marginal density of X. Denote the regression function by $m(x)=E(Y|X=x)$ and the constant variance by

$\sigma^2 = \mathrm{var}(Y|X=x)$. Here, $m(x)$ is assumed to be a smooth but unknown function. This relationship is expressed as follows

$$Y_i = m(X_i) + \varepsilon_i, \qquad i=1,\cdots,n \tag{1}$$

where ε_i's are independent random variables with the expectation 0 and the variance σ^2. Among the kernel-based methods for estimating $m(x)$, the local linear regression estimator(Stone, 1977, Cleveland, 1979, Müller, 1987, Fan, 1992, Fan and Gijbel, 1996) is often used because it shows good performance. It is based on moving locally weighted averaging and can be expressed as follows

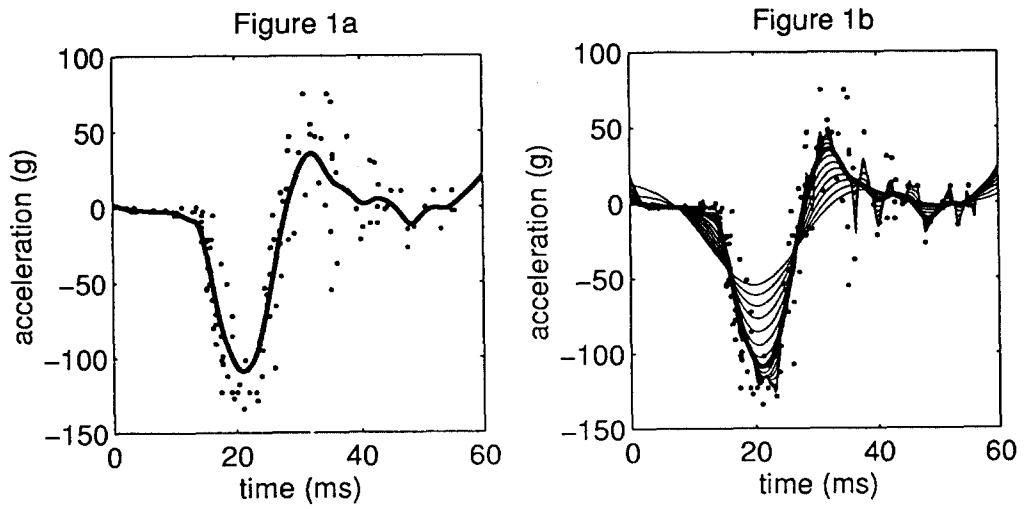
$$\hat{m}_{LL}(x;h) = n^{-1}\sum_{i=1}^{n} w(X_i, x; h)\, Y_i \tag{2}$$

where $w(X_i, x; h) = \dfrac{\{\hat{s}_2(x;h) - \hat{s}_1(x;h)(X_i - x)\}\, K_h(X_i - x)}{\hat{s}_2(x;h)\,\hat{s}_0(x;h) - \hat{s}_1(x;h)^2}$ and

$\hat{s}_r(x;h) = n^{-1}\sum_{i=1}^{n}(X_i - x)^r K_h(X_i - x)$, $r = 0, 1, 2$. The smoothing parameter or bandwidth, h, is crucial to the performance of the local linear regression estimators(Sivermann,1986). A choice of the smoothing parameter gives a single regression estimate. Therefore the goal of smoothing should be a single regression estimate. This idea comes naturally from parametric statistics, where all important information is summarized in a few parameters.

If a single regression estimate is replaced by a family of estimates, indexed by the smoothing parameter, smoothing becomes a more powerful graphical device. This reveals structure in the data more quickly and easily than is possible from any single estimate, because more information is summarized in a plot, as shown in <Figure 1>. And it easily adapts to situations where there is useful information in the data at several different level of the smoothing parameter.

<Figure 1> shows the local linear regression estimates using the Gaussian kernel, for the so-called motorcycle data set shown in Härdle(1990). The raw data $\{(X_i, Y_1)\}_{i=1}^{n}$, $n=133$, are shown as dots. Units are g (earth-acceleration) for Y and ms (milliseconds after impact in a simulated experiment) for X. Figure 1a shows a local linear estimate. The bandwidth used here, is h_{RSW} proposed by Ruppert, Sheather and Wand(1995) that provides a good compromise for "best" choice of global bandwidth for these data. Figure 1b shows a family of estimates centered at the estimate in Figure 1a.



< Figure 1 > Kernel regression estimates for Motorcycle data, overlaid with scatterplot of the raw data. Figure 1a shows a single estimate using Ruppert, Sheather and Wand bandwidth. Figure 1b shows a family of 15 estimates centered at the estimate in Figure 1a.

In this paper, we suggest the threshold in the choice of the maximum bandwidth of the family. In section 2, we describes the family approach of the nonparametric curve estimation. In section 3, we carry out the simulation study to determine the threshold of the maximum bandwidth of the family.

2. The Family Approach

In this section we consider the family approach of the nonparametric curve estimation. Minnotte and Scott(1993) and Marron and Chung(1997) introduced the family approach for the density estimation. Marron and Chung(1997) discussed also the regression estimation. The important points in the family approach are to determine the number of estimates in a family and the range of the bandwidths. We outline here the result of Marron and Chung(1997).

They recommended the family of the local linear regression estimates that used Ruppert, Sheather and Wand "Direct Plug in" bandwidth, h_{RSW}, as a central estimate, because it often gives an effective choice of global smoothing parameter. Figure 1a is the estimate using $h_{RSW} = 1.589$. They suggested using 15 estimates that made it easy to visually connect the estimates. We also use 15 members in

Figure 1b.

To determine the range of the bandwidths, we should first choose an extreme value of the bandwidth. Let $\hat{m}_h(x)$ be the local linear estimates using the smoothing parameter, h, and $\hat{m}_\infty(x)$ be the limit of the smooth $\hat{m}_h(x)$ as $h \to \infty$. Then $\hat{m}_\infty(x)$ is the least squares fit line. As the maximum bandwidth of the family, $h_{\max}$, Marron and Chung(1997) suggested

$$h'_{\max} = inf\{h > h_{RSW} : |\hat{m}_h(x^*) - \hat{m}_\infty(x^*)| \le c_1 |\hat{m}_{h_{RSW}}(x^*) - \hat{m}_\infty(x^*)|\} \quad (3)$$

$$h_{\max} = \max[h'_{\max}, c_2 h_{RSW}] \quad (4)$$

where x^* was a location that maximized $|\hat{m}_{h_{RSW}}(x) - \hat{m}_\infty(x)|$, i. e.

$$|\hat{m}_{h_{RSW}}(x^*) - \hat{m}_\infty(x^*)| = \max_x |\hat{m}_{h_{RSW}}(x) - \hat{m}_\infty(x)|.$$

Using this $h_{\max}$, the minimum bandwidth of the family could define in the same manner of the density estimation as follows

$$h_{\min} = h_{RSW}/(h_{\max}/h_{RSW}) = h_{RSW}^2/h_{\max}. \quad (5)$$

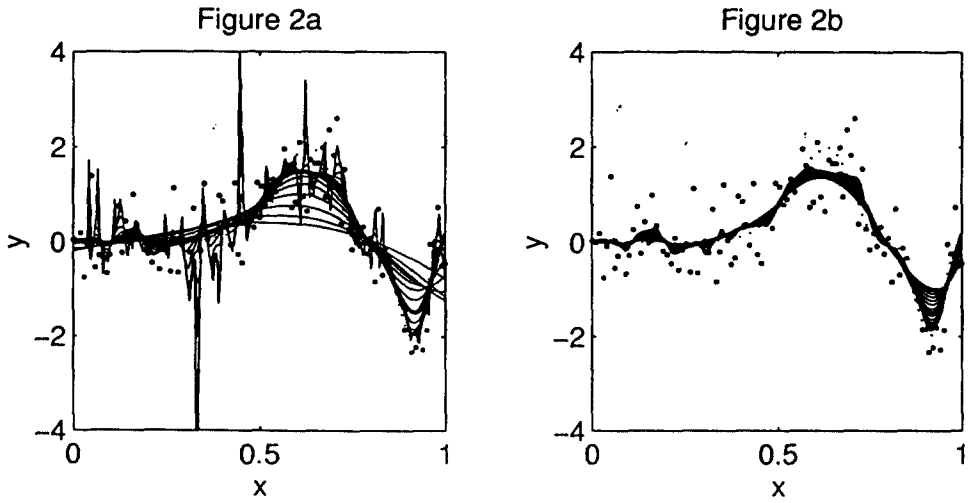
The bandwidths of the family is spaced as follows

$$\{h_{\min}(h_{\max}/h_{\min})^{(i-1)/14} : i = 1, \cdots, 15\}. \quad (6)$$

Since the bandwidth is considered as scale parameter, the logarithmic scale in (6) is better than the linear scale. For example, see <Figure 3> of Marron and Chung (1997).

In (3) and (4), c_1 and c_2 are personal choices. Marron and Chung(1997) recommended 0.3 and 3, respectively. Note that Figure 1a is used about $5h_{RSW}$ as $h_{\max}$. In the density estimation, they suggested using 0.6 and 3 as c_1 and c_2, respectively, after consideration of a number of examples. The idea of the density estimation were simply adapted to the regression estimation. While the density estimation is considered as one sided context, the regression estimation is two sided problem. So, the threshold of 0.6 is naturally replaced by 0.3. But in case of

the random design, the above method often gives the large value of $h_{\max}$ and causes $h_{\min}$ to have the very small value. The small value of $h_{\min}$ shows the bad performance of the estimate, as shown <Figure 2>. Figure 2a is the family of estimates using $h_{\max} = 2h_{RSW}$ and Figure 2b is that using $h_{\max} = 9h_{RSW}$ under the uniform random designs. Figure 2a shows that there is a little difference among the family members. This indicates $h_{\max} = 2h_{RSW}$ is relative small. On the other hand, Figure 2b shows that the estimate indexed by $h_{\min}$ is very noisy. This indicates $h_{\max} = 9h_{RSW}$ is too large so that the upper bound of $h_{\max}$ in (4) is needed



< Figure 2 > families of kernel regression estimates for 100 simulated data points from the true regression function $m(x) = 2\sin^3(2\pi x^3)$. Figure 2a is the family using $h_{\max} = 2h_{RSW}$. Figure 2b is that using $h_{\max} = 9h_{RSW}$. $h_{RSW} = 0.03466$. The dotted line is the true regression function.

3. Simulation Study

In this section, we carry out Monte Carlo simulation to determine the threshold of c_2 in (4). We consider three testing regression functions as follows

(m1) $m(x) = 2exp\{-(x-0.2)^2/0.4^2\} + 3\exp\{-(x-0.8)^2/0.05^2\}$.

(m2) $m(x)=2\sin(4\pi x)$.

(m3) $$m(x)=\begin{cases}1 & \text{if } x\in(0,0.25)\\ \cos(4\pi(x-1/4)) & \text{if } x\in(0.25,0.75)\\ 1 & \text{if } x\in(0.75,1)\end{cases}.$$

And we take the uniform random design and the equally spaced fixed design. The sample size is taken as $n=100$ and 500 replications are performed. We take $\varepsilon_i \sim N(0,\sigma^2)$ independent of X with $\sigma=0.5$ from randn function of MATLAB. We use the Gaussian kernel as the kernel function and the direct plug-in bandwidth selector, h_{RSW}, of Ruppert, Sheather and Wand(1995) as the bandwidth. To reduce computational effort, we use the binning approach suggested by Fan and Marron(1994). This is useful because the data only need to be binned once. So binned computation has the advantage of requiring only $O(N)$ kernel evaluation, this allows very fast computation of $\hat{m}(x;h)$ over the grid points. Here N denotes the number of grid points. We use $N=401$ recommended by Fan and Marron(1994). As the method for obtaining grid counts that has good properties, we use "linear binning"(Hall and Wand, 1993).

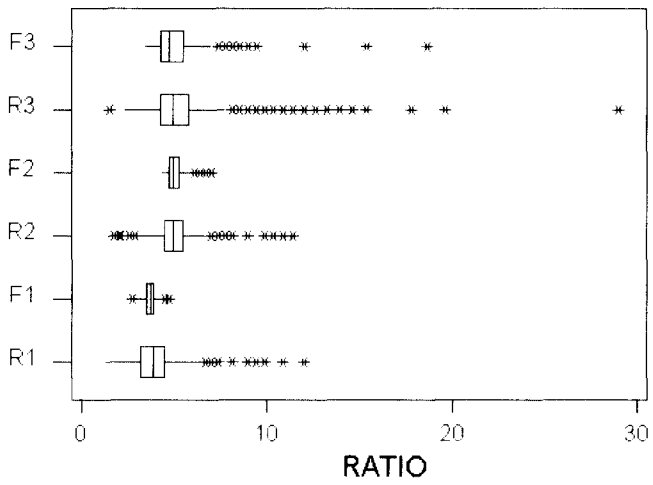
<Table 1> shows the median, 25th percentile(Q1) and 75th percentile(Q3) of ratio = $h'_{\max}/h_{RSW}$ and the numbers of ratios exceeding 7.0 in 500 replications. And <Figure 3> is box plot of ratios under three models. <Table 1> and <Figure 3> show that ratios under the fixed design is smaller and more stable than those under the random design. The numbers of ratios exceeding 7.0 under the random design are more than those under the fixed design. In case that ratio is large value, family approach gives bad appearance. This is because too small value of the smoothing parameter make the denominator of the weight function in (2) have negligible value. So this suggest that it is better the upper bound of $h_{\max}$ is limited. After consideration of a number of examples, we suggest using 7.0 as the upper bound of ratio as follows

$$h'_{\max}=inf\{h>h_{RSW}:|\hat{m}_h(x^*)-\hat{m}_\infty(x^*)|\le 0.3|\hat{m}_{h_{RSW}}(x^*)-\hat{m}_\infty(x^*)|\}.$$

$$h_{\max}=\min\{7h_{RSW},\ \max[h'_{\max},3h_{RSW}]\}.$$

< Table 1 > three quartiles and the numbers of ratios exceeding 7

	random design				fixed design			
	Q1	median	Q3	# of ratio's >7	Q1	median	Q3	# of ratio's>7
m1	3.225	3.920	4.538	12	3.556	3.734	3.920	0
m2	4.538	5.003	5.516	22	4.765	5.003	5.253	6
m3	4.322	5.003	5.792	69	4.322	4.765	5.516	29

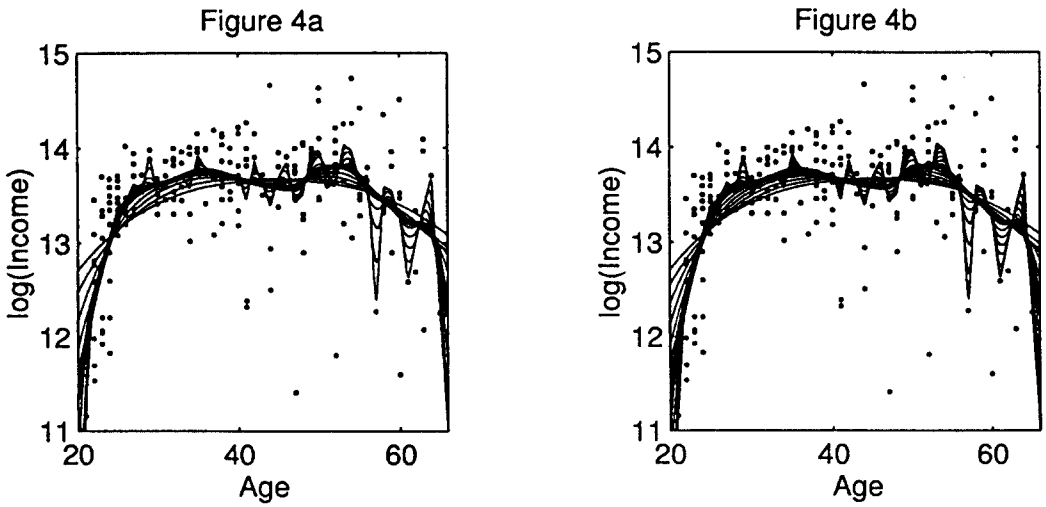


< Figure 3 > Box plot of the ratios. R1, R2 and R3 are the ratios under the random design of the model m1 m2 and m3, respectively. F1, F2 and F3 are the ratios under the fixed design of the model m1 m2 and m3, respectively.

Marron and Chung(1997) used 15 estimates in a family. 15 estimates in a family is adequate in the density context. But in regression context, it seems that 15 members are too many to visually distinguish the estimates. <Figure 4> shows the families of the estimates using 11 and 15 members for Canadian Earning Power Data, overlaid with scatterplot of the raw data. We suggest using 11 members after consideration of a number of examples like as <Figure 4>. Then the bandwidths of the family is spaced as follows

$$\{ h_{\min}(h_{\max}/h_{\min})^{(i-1)/10} : \ i = 1, \cdots, 11 \},$$

where $h_{\min}$ is same as in (5).



< Figure 4 > Kernel regression estimates for Canadian Earning Power Data, overlaid with scatterplot of the raw data. Family members are 11 for Figure 4a and 15 for Figure 4b.

4. Discussions

In section 3, we suggest the upper and lower bound of bandwiths, and the number of family members. When we restrict the bound of the bandwidths, each member of the family is smooth enough to visually connect adjacent members. So, the family approach becomes the powerful graphic tool. In the density estimation, there are a large of data in high density area so that the family approach could catch the finer peak of true density function. But in regression context, the family approach does not work as the density context because the region having the finer peak of true regression function does not accord the region of high density in X-space. But the family approach could have the same effect as the good location adaptive smoothing.

References

[1] Cleveland, W.(1979), "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, Vol. 74, pp. 829-836.

[2] Eubank, R.(1988), *Spline smoothing and nonparametric regression*, Marcel Dekker, New York.

[3] Fan, J.(1992), "Design-adaptive nonparametric regression," *Journal of the American Statistical Association*, Vol. 87, pp. 998-1004.

[4] Fan, J. and Gijbels, I.(1996), *Local polynomial modelling and its application*, Chapman and Hall, London.

[5] Fan, J. and Marron, J.S.(1994), "Fast implementations of nonparametric curve estimators, *Journal of Computational & Graphical Statistics*," Vol 3, pp. 35-56.

[6] Hall, P. and Wand, M.P.(1994), "On the accuracy of binned kernel density estimators," Submitted for publication.

[7] Härdle, W.(1990), *Applied nonparametric regression*, Cambridge University Press.

[8] Marron, J.S. and Chung, S.S.(1997), "Presentation of smoothers: the family approach," to appear *The American Statistician*

[9] Minnotte, M.C. and Scott, D.W.(1993), "The mode tree: a tool for visualization of nonparametric density features," *Journal of Computational & Graphical Statistics*, Vol 2, pp. 51-86.

[10] Müller, H.-G.(1987), "Weighted local regression and kernel methods for nonparametric curve fitting," *Journal of the American Statistical Association*, Vol. 82, pp. 231-238.

[11] Müller, H.-G.(1988), *Nonparametric Regression Analysis of Longitudinal Data*, Springer-Verlag, Berlin.

[12] Ruppert, D. Sheather, S.J. and Wand, M.P.(1995), "An effective bandwidth selector for local least squares regression," *Journal of the American Statistical Association*, Vol. 90, pp. 1257-1270.

[13] Scott, D.W.(1992), *Multivariate density estimation: theory, practice and visualization*, Wiley, New York.

[14] Silverman, B.W.(1986), *Density estimation for statistics and data analysis*, London: Chapman and Hall.

[15] Stone, C.J.(1977), "Consistent nonparametric regression," *The Annals of Statistics*, Vol. 5, pp. 595-620.

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