The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 413-423

# Robust Cross Validation Score[1)]

**Dongryeon Park[2)]**

## Abstract

Consider the problem of estimating the underlying regression function from a set of noisy data which is contaminated by a long tailed error distribution. There exist several robust smoothing techniques and these are turned out to be very useful to reduce the influence of outlying observations. However, no matter what kind of robust smoother we use, we should choose the smoothing parameter and relatively less attention has been made for the robust bandwidth selection method. In this paper, we adopt the idea of robust location parameter estimation technique and propose the robust cross validation score functions.

*Keywords* : Cross validation, Local regression, Location parameter estimators, Robust regression

## 1. Introduction

Consider the problem of estimating the underlying regression function from a set of noisy data. Suppose we are given $n$ pairs of random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$, where $X_i$ is the independent variable and $Y_i$ is the corresponding dependent variable. We assume that there exist a smooth function $m$ such that

$$Y_i = m(X_i) + \epsilon_i, \qquad i = 1, \cdots, n \tag{1.1}$$

where $\epsilon_i$ is an error term representing random errors in the observations. The aim is the estimation of the unknown function $m$. This problem is called the nonparametric regression problem and there now exist several approaches to this problem. Some of the more popular are kernel estimators, local polynomial regression estimators, smoothing spline estimators and orthogonal series estimators. For a more detailed discussion of these estimators, see Gasser

1) This research work has been supported by the Hanshin University Special Research Grant in 2005.
2) Associate professor, Department of Statistics, Hanshin University, 411 Yangsan-dong, Osan, Kyunggi-do, Korea. E-mail: drpark@hs.ac.kr

and Müller (1979), Fan and Gijbels (1996), and Wand and Jones (1995).

Each of these methods has its own particular strengths and weakness and the local polynomial regression estimators are generally accepted as one of the best methods. Fan (1992, 1993), Fan and Gijbels (1992) and Ruppert and Wand (1994) give a detailed picture of the advantages of local polynomial fitting. Here we briefly introduce the basic concept of local polynomial regression estimators. Suppose that the $(p+1)^{th}$ derivative of $m(x)$ at the point $x_0$ exist. A Taylor expansion gives, for $x$ in a neighborhood of $x_0$,

$$m(x) \approx m(x_0) + m'(x_0)(x-x_0) + \frac{m''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x-x_0)^p. \quad (1.2)$$

This relationship suggests that we can approximate the unknown regression function $m(x)$ locally by a polynomial of order $p$ and this polynomial is fitted locally by a weighted least squares regression. Denote by $\hat{\beta}_j$, $j = 0, 1, \cdots, p$ the solution to the least squares problem

$$\min_{\beta} \sum_{i=1}^{n} \left( Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right)^2 K_h(X_i - x_0) \quad (1.3)$$

where $\beta = (\beta_0, \cdots, \beta_p)^T$, $K_h(\cdot) = K(\cdot/h)/h$ with $K$ a kernel function assigning weights to each data point, and $h$ is a bandwidth controlling the size of local neighborhood. Then (1.2) suggests that an estimator for $m(x_0)$ is

$$\hat{m}(x_0) = \hat{\beta}_0. \quad (1.4)$$

There are some important issues for using local polynomial regression estimators. First of all, we need to choose the order of polynomial, $p$. For a given bandwidth $h$, a large value of $p$ would reduce a bias, but cause a large variance. Fan and Gijbels (1995) show that there is a general pattern of increasing variability, and recommend that the use of the lowest odd order, i.e. $p = 1$, or occasionally $p = 3$.

The more important issue is the choice of the bandwidth $h$. A too large bandwidth results in over-smoothed estimates, causing a large bias, while a too small bandwidth results in noisy estimates, causing a large variance. Thus the choice of bandwidth is the tradeoff between the bias and the variance, and the good bandwidth selector is inevitable for the good performance of local polynomial regression estimator. A lot of research has been done for the bandwidth selection problem (Bowman, 1984; Fan and Gijbels, 1995; Jone, Marron, and Sheather, 1996; Rice, 1984). The idea of cross-validation (CV) is probably the most popular bandwidth selection rule. The CV score function is defined as

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_{-i}(X_i))^2 \tag{1.5}$$

where $\hat{m}_{-i}$ is the so-called "leave-one-out" version of $\hat{m}$. That is $\hat{m}_{-i}$ is constructed with $n-1$ data points by leaving out the data point $(X_i, Y_i)$. The cross-validation selection rule chooses the bandwidth $h$ to minimize the CV score function.

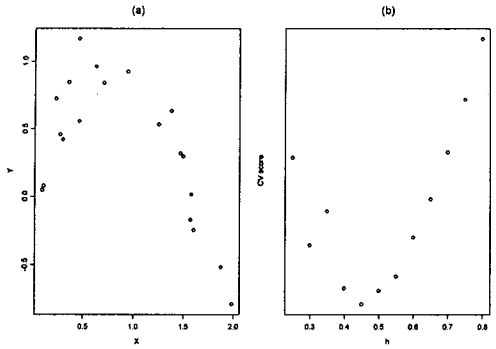
However, there are some criticism against CV. Jones, Marron, and Sheather (1996) call the cross-validation bandwidth selection rule as the first generation method and argue that it is much inferior to the second generation methods such as the plug-in rules. One of the major problem of CV is that it has unacceptably large variance. However, according to Loader (1999a), large variability is not a problem of CV itself. Rather, it is a symptom of the difficulty of bandwidth selection and problem of resolving uncertainty in the data. She also argues that plug-in methods reflect the difficulty by over-smoothing difficult problem and have less ability to resolve uncertainty, so she recommends to use cross-validation method.

In this paper, we assume that the error terms in model (1.1) have the long-tailed distribution, so we might have a data set which contains a few outliers. Local polynomial regression estimators are based on $L_2$ norm, so they can be highly influenced by outliers in the response variable. In this situation it is preferable to have an estimation method which is more resistant for extreme observations. Lowess (Cleveland, 1979), $L_1$ local regression estimators (Wang and Scott, 1994), and local-WMD estimators (Park, 2004) were introduced for this purpose.

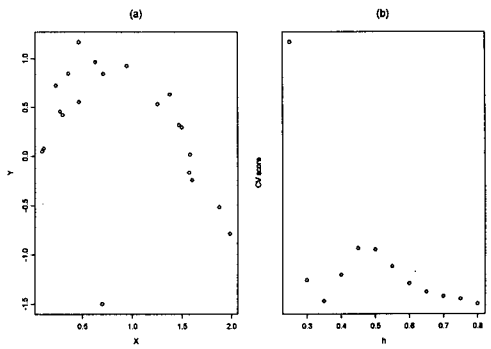
Extreme observations would have huge influence not only to local polynomial regression estimators but also to cross-validation score function. Figure 1 shows that the scatter plot of two variables and their cross-validation score values for various bandwidths. In this figure, the CV plot has a global minimum at $h = 0.45$. Now we include one outlier in the data set and make the CV plot again to see the effect of outlier to the CV score. Figure 2 shows the totally different shape of the CV plot. It is very difficult to find the global minimum. What happens in Figure 2 is that the data set has a very large negative outlier at $X = 0.7$ and whatever bandwidth is chosen, this point produces a very large squared error and this error is so large in comparison to the errors from all other points that the CV values are all about the same. In this case the CV function is essentially worthless for the purpose of choosing $h$.

This example gives us the reason why we need a robust version of the CV score function when we have the outliers, but only a few researches have been done for the robust version of the cross validation score function. Cantoni and Ronchetti (2001) proposed the robust versions of cross-validation and $C_p$ for smoothing splines. Wang and Scott (1994) proposed $L_1$ version of CV score function for $L_1$ local regression estimators. The aim of this paper is to propose several robust version of the CV score functions for local polynomial regression

estimators. They are based on robust location parameter estimators.




<Fig. 1> Scatter plot and CV plot for original data set




<Fig. 2> Scatter plot and CV plot for modified data set

The paper is organized as follows. In Section 2, we propose several robust version of CV score functions. In Section 3, we compare the performance of CV score functions by a simulation study. Section 4 presents some conclusions.

## 2. Robust Cross Validation Score Function

The cross validation selection rule using CV score function of (1.5) requires a heavy computation. To derive a simplification of CV score function, we need to define more terms in local polynomial regression. Since the local polynomial regression estimate solves a least squares problem, $\hat{m}(x)$ is a linear estimate. That is, for each $x$, $\hat{m}(x)$ can be written as

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i. \tag{2.1}$$

The hat matrix is the $n \times n$ matrix $S$ with $i^{th}$ row $(w_1(X_i), \cdots, w_n(X_i))^T$, which maps the data to the fitted values:

$$\begin{pmatrix} \hat{m}(X_1) \\ \vdots \\ \hat{m}(X_n) \end{pmatrix} = S \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \tag{2.2}$$

The trace of the hat matrix $S$, $tr(S) = \sum_{i=1}^{n} S_{ii}$ is called the degrees of freedom of a local fit, which provide a generalization of the number of parameters of a parametric model. Now

we can define a simplified version of CV score function. The following theorem is proved in Loader (1999b).

**Theorem** [Loader, 1999b] Let $S_{ii}$ be the $i^{th}$ diagonal element of the hat matrix $S$. If $S_{ii} < 1$, the leave-one-out cross validation estimate $\hat{m}_{-i}(x)$ is

$$\hat{m}_{-i}(X_i) = \frac{\hat{m}(X_i) - S_{ii} Y_i}{1 - S_{ii}}. \tag{2.3}$$

Using the theorem the CV score function can be written as

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \hat{m}(X_i))^2}{(1 - S_{ii})^2}. \tag{2.4}$$

The estimator $\hat{m}(X_i)$ can be either classical local estimator or robust local estimator, but we assume that it is robust local estimator throughout this paper. The idea of defining the robust version of CV score function is to treat the values of the cross validation score function as a realization of a random variable. That is, we treat $(Y_i - \hat{m}(X_i))^2/(1 - S_{ii})^2$ as a random variable. Then the CV score function of (2.4) is just the mean of the random variables, which is very vulnerable statistic to outliers. What is needed here is a robust location parameter estimates for the realization of these random variables.

The median is the simple and classical robust location parameter estimator. Modern research on robust methods offers even better performance if we can accept more complicated estimators of location (see Hoaglin, Mosteller, and Tukey 1983; Rousseeuw and Leroy 1987). Among others, we briefly introduce the M-estimators and the least trimmed squares (LTS) estimator of location.

The mean of the random variables $Z_1, \cdots, Z_n$ is related with the least squares estimator which is

$$\min_{\hat{\theta}} \sum_{i=1}^{n} (Z_i - \hat{\theta})^2. \tag{2.5}$$

This estimator has a poor performance in the presence of contamination. Huber has lowered the sensitivity of the least squares objective function by replacing the squares in (2.5) by a suitable function $\rho$. This leads to location M-estimators, defined by

$$\min_{\hat{\theta}} \sum_{i=1}^{n} \rho(Z_i - \hat{\theta}) \tag{2.6}$$

which satisfy the necessary condition

$$\sum_{i=1}^{n} \psi(Z_i - \hat{\theta}) = 0 \tag{2.7}$$

where $\psi$ is the derivative of $\rho$. In general, the M-estimator of location must take account of the scale of the sample in order to be location and scale equivariant, so we choose an auxiliary estimator of scale $S_n$ and use it with a constant $c$ to rescale the $Z_i - \hat{\theta}$. The constant $c$ is known as the tuning constant. Huber himself used the functions

$$\psi(x) = \begin{cases} -k & , x \leq -k \\ x & , -k \leq x \leq k \\ k & , x \geq k \end{cases} \tag{2.8}$$

and the corresponding estimator is called Huber estimator. The influence curve of Huber estimator is constant for all observations beyond a certain point. An M-estimator can be made more resistant by having the $\psi$ function, and hence the influence curve, return to 0. Tukey biweight estimator has the following redescending $\psi$ function

$$\psi(x) = \begin{cases} (1-x^2)^2 & , |x| \leq 1 \\ 0 & , |x| > 1 \end{cases} \tag{2.9}$$

Let us now look at the LTS estimator. In order to determine the LTS location estimate we have to consider the following $n-h+1$ subsamples:

$$\{Z_{(1)}, \cdots, Z_{(h)}\}, \{Z_{(2)}, \cdots, Z_{(h+1)}\}, \cdots, \{Z_{(n-h+1)}, \cdots, Z_{(n)}\} \tag{2.10}$$

where $Z_{(1)}, \cdots, Z_{(n)}$ represent the order statistics and $h = [n/2] + 1$. Each of these subsamples contains $h$ observations and for each subsample, we calculate the mean

$$\bar{Z}^{(1)} = \frac{1}{h} \sum_{i=1}^{h} Z_{(i)}, \ \cdots, \ \bar{Z}^{(n-h+1)} = \frac{1}{h} \sum_{i=n-h+1}^{n} Z_{(i)} \tag{2.11}$$

and the corresponding sum of squares

$$SQ^{(1)} = \sum_{i=1}^{h}(Z_{(i)} - \bar{Z}^{(1)})^2, \cdots, SQ^{(n-h+1)} = \sum_{i=n-h+1}^{n}(Z_{(i)} - \bar{Z}^{(n-h+1)})^2 \quad (2.12)$$

The LTS estimate then corresponds to the mean $\bar{Z}^{(j)}$ with the smallest associated sum of squares $SQ^{(j)}$.

Now we define the robust version of CV score functions using these robust location parameter estimators. We put $(Y_i - \hat{m}(X_i))^2/(1-S_{ii})^2$ of (2.4) as a random variable $Z_i$, $i = 1, \cdots, n$. Then we can apply any robust location parameter estimation methods to random variables $Z_1, \cdots, Z_n$ and this procedure leads to the robust version of CV score function. We here consider the median, Huber estimator, Tukey bisquares estimator, and the LTS location estimator as the robust location parameter estimation methods. The robust CV score function induced by these robust estimators are denoted by MCV, HCV, BCV, and LCV, respectively.

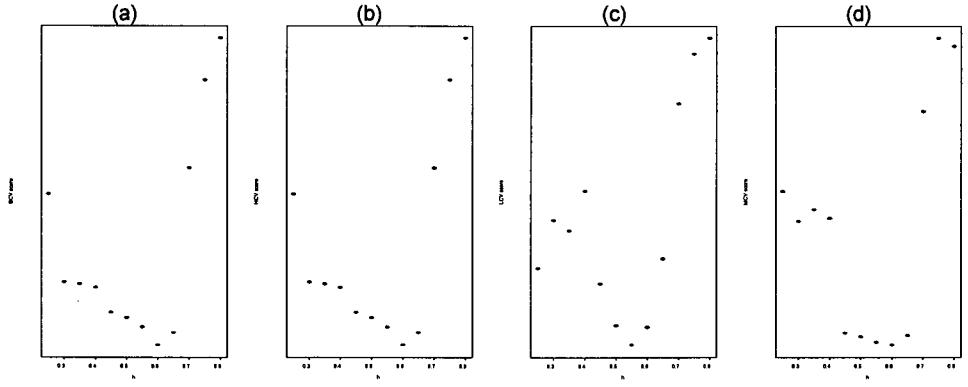
We revisit the example data set of Figure 2 and apply the proposed robust CV score functions. Figure 3 shows the corresponding CV plots. We have much improved CV plots for BCV, HCV, and LCV. To investigate the properties of these robust CV score functions thoroughly, we need to derive the asymptotic behavior of these estimators, but these are left as a further research topic. In Section 3, we compare the empirical properties of these estimators by a simulation.

## 3. Simulation Study

For the simulation study the model was taken to be

$$Y_i = sin(2\pi(1-X_i)^2) + \epsilon_i,\ i = 1, \cdots, n. \quad (3.1)$$

We performed a comparison of the existing methods with the robust methods proposed in this paper. As the existing methods, we considered the standard CV score function of




<Fig. 3> CV plots for proposed methods. (a) BCV (b) HCV (c) LCV (d) MCV

(1.5) and the absolute CV (ACV) function (Wang and Scott, 1994) which is defined as

$$ACV(h) = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{m}_{-i}(X_i)|. \tag{3.2}$$

The tuning constant $c$ used here was 5 for BCV, and 1.45 for HCV. The auxiliary scale estimator for both BCV and HCV was taken to be the median absolute deviation (MAD) which is defined as

$$MAD = med_i\{|X_i - med_j\{X_j\}|\}. \tag{3.3}$$

Errors were generated from several symmetric distributions; standard normal distribution, t-distribution, Cauchy distribution, and a contaminated normal distribution. We denote $CN(\alpha;\sigma)$ the contaminated normal distribution whose distribution function is given by

$$F(x) = (1-\alpha)\Phi(x) + \alpha\Phi\left(\frac{x}{\sigma}\right). \tag{3.4}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. A list of the error distribution used is given in Table 1, together with their corresponding tail index $\tau(F)$ (Hoaglin, Mosteller, and Tukey, 1983) which is defined by

$$\tau(F) = \frac{F^{-1}(0.99) - F^{-1}(0.5)}{F^{-1}(0.75) - F^{-1}(0.5)} \Big/ \frac{\Phi^{-1}(0.99) - \Phi^{-1}(0.5)}{\Phi^{-1}(0.75) - \Phi^{-1}(0.5)}. \tag{3.5}$$

The tail index expresses how the extreme portion of the distribution spreads out relative to the width of the center. Lighter tailed distributions such as the uniform distribution have index values less than 1.0 and heavier tailed distributions have index values greater than 1.0.

<Table 1> Error distributions used in simulation and their tail index

| $F(x)$ | $\tau(F)$ | $F(x)$ | $\tau(F)$ |
|---|---|---|---|
| $N(0,1)$ | 1.00 | $CN(0.3,5)$ | 2.84 |
| $t(4)$ | 1.46 | $CN(0.1,10)$ | 4.93 |
| $t(3)$ | 1.72 | $CN(0.2,10)$ | 5.57 |
| $CN(0.1,5)$ | 2.50 | $CN(0.3,10)$ | 5.34 |
| $CN(0.2,5)$ | 2.88 | $Cauchy(0,0.1)$ | 9.22 |

We considered the random design case for the independent variable and took $X \sim Unif(0,1)$. For each error distribution, we generated the random sample of size $n = 50$ and then chose the bandwidth by each CV function. We used the S-plus function $location.m()$ for both HCV and BCV, and $location.lts()$ for LCV. Using the bandwidth, we finally got the estimate of true regression function by lowess, the robust local linear

regression estimator proposed by Cleveland (1979). The performance of each CV score function was measured by the Monte Carlo MISE (Mean Integrated Squared Error) of the $\hat{m}(x)$ which was computed as the average of

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{m}(x_i)-m(x_i))^2 \tag{3.6}$$

over 5,000 Monte Carlo simulation samples. For computing $\hat{m}(x)$, we used the function *locfit.robust*( ) of the *locfit* library in the S-plus.

The simulation results are listed in Table 2. The standard errors of the simulations were estimated at most 2 % for all cases, so these are not the significant factors for interpreting the results.

For the normal error distribution, which is the uncontaminated case, the standard CV score function produces the best result. However, for all of the contaminated cases, HCV shows the superior performance over other methods. Especially, it is worthy of note that at the heavier tail distribution like $CN(.2,10)$ and $Cauchy(0,.1)$, HCV shows better performance than BCV which is based on redescending $\psi$ function.

The Huber estimator is designed to sacrifice efficiency at heavier tail distribution for higher efficiency near the normal distribution, so it is known that the redescending estimators are better than the Huber estimator at heavier tail distribution for the general location parameter estimation problem (Hoaglin, Mosteller, and Tukey, 1983). However, this does not hold in our simulation results and we are not quite sure why it's so. Our guess is that the behavior of the statistics $(Y_i-\hat{m}(X_i))^2/(1-S_{ii})^2$ of (2.4) is quite different from the simple random samples from a single population. That is, the statistics of (2.4) would not have very extreme values.

<Table 2> Monte Carlo MISE of each CV score function based on 5,000 replications

| | *CV* | *ACV* | *BCV* | *HCV* | *LCV* | *MCV* |
|---|---|---|---|---|---|---|
| $N(0,1)$ | 0.1427 | 0.1431 | 0.1624 | 0.1558 | 0.1613 | 0.1616 |
| $t(4)$ | 0.2010 | 0.1867 | 0.1905 | 0.1865 | 0.1901 | 0.1913 |
| $t(3)$ | 0.2278 | 0.2058 | 0.2056 | 0.2033 | 0.2045 | 0.2060 |
| $CN(.1,5)$ | 0.2170 | 0.1845 | 0.1816 | 0.1771 | 0.1827 | 0.1823 |
| $CN(.2,5)$ | 0.2662 | 0.2309 | 0.2136 | 0.2083 | 0.2137 | 0.2135 |
| $CN(.3,5)$ | 0.3238 | 0.2924 | 0.2818 | 0.2754 | 0.2802 | 0.2809 |
| $CN(.1,10)$ | 0.2447 | 0.1999 | 0.1845 | 0.1784 | 0.1869 | 0.1826 |
| $CN(.2,10)$ | 0.3171 | 0.2817 | 0.2408 | 0.2150 | 0.2593 | 0.2260 |
| $CN(.3,10)$ | 0.3624 | 0.3311 | 0.3644 | 0.3101 | 0.3295 | 0.3360 |
| $Cauchy(0,.1)$ | 0.3234 | 0.2605 | 0.2566 | 0.2526 | 0.2582 | 0.2538 |

Wang and Scott (1994) proposed ACV and argued that ACV appeared to be a good option for robust smoothing parameter selection, but its performance in this simulation was not so good. ACV showed competitive performance at $N(0,1)$ and $t(4)$, but its performance was getting worse as the tail of error distribution became heavier. BCV, LCV, and MCV showed almost identical performance. They were competitive at some cases, but worse than HCV anyway.

## 4. Conclusions

We have considered the problem of estimating the underlying regression function from a set of noisy data which is contaminated by a long tailed error distribution. Robust smoothing techniques can reduce the influence of the outliers, but they must be based on the robust smoothing parameter selection rule. However, relatively less attention has been made for a robust cross validation score function. In this paper, we have adopted the idea of the robust location parameter estimation method and proposed the robust cross validation score functions. It has been turned out that the robust cross validation score function based on the Huber estimator is a very good option for the robust smoothing parameter selection rule. Only empirical evidences were provided and the theoretical backbones were not derived here.

Open research directions include the derivation of the theoretical properties of proposed methods and the generalization to other smoothing techniques and more complex models.

## References

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

[2] Cantoni, E. and Ronchetti, E. (2001), Resistant selection of the smoothing parameter for smoothing splines, *Statistics and Computing*, 11, 141-146

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

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

[5] Fan, J. (1993), Local linear regression smoothers and their minimax efficiency, *The Annals of Statistics*, 21, 196-216

[6] Fan, J. and Gijbels, I. (1992), Variable bandwidth and local linear regression smoothers, *The Annals of Statistics*, 20, 2008-2036

[7] Fan, J. and Gijbels, I. (1995), Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of Royal Statistical Society*, Series B, 57, 371-394

[8] Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman

& Hall, London

[9] Gasser, T. and Müller, H. (1979), Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics, 757, 23-68, Springer-Verlag, New York

[10] Hoaglin, D., Mosteller, F. and Tukey, J. (1983), *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York

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

[12] Loader, C. (1999a), Bandwidth selection: classical or plug-in?, *The Annals of Statistics*, 27, 415-438

[13] Loader, C. (1999b), *Local Regression and Likelihood*, Spring-Verlag, New York

[14] Park, D. (2004), Robustness weight by weighted median distance, *Computational Statistics*, 19, 367-383

[15] Rice, J. (1984), Bandwidth choice for nonparametric regression, *The Annals of Statistics*, 12, 1215-1230

[16] Rousseeuw, P. and Leroy, A. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York

[17] Ruppert, D. and Wand, M. (1994), Multivariate locally weighted least squares regression, *The Annals of Statistics*, 22, 1346-1370

[18] Wand, M. and Jones, M. (1995), *Kernel Smoothing*, Chapman & Hall, London

[19] Wang, F. and Scott, D. (1994), The $L_1$ method for robust nonparametric regression, *Journal of the American Statistical Association*, 89, 249-260

[ Received February 2005, Accepted March 2005 ]