On the Degree of Polynomial for Robust Locally Weighted Polynomial Regression

HAFED M. MOTAIR

Department of Mathematics, Distinguished Secondary School

Diwaniya, Ministry of Education

IRAQ

Abstract: Local Polynomial Regression is one of the important methods in estimation nonparametric regression curve, furtherwere containing outliers in the data sample will affect the estimated parameters and will also affect the shape of the estimated curve. Therfore, many robust methods have been appeared which will belittle the influence of outliers by down the weighting given for outliers in data samples. One of these methods is LOWESS method. In this paper we studied the influence of degree of polynomial in finding estimators using LOWESS method. Depending on the generated data through simulation study and by using one of error criterion which is integrated squared error (ISE), we found out quadratic polynomial was better than constant and linear polynomial when estimating regression curve using LOWESS method.

Keywords: Nonparametric Regression, Local Regression, Robust, Outliers, LOWESS.

1 Introduction

Given scatterplot data $(x_i, y_i), 1 \leq i \leq n$, the traditional way to summarize the relationship between x_i and y_i is ti fit a linear model. In smoothing framework, we extend this linear restriction and assume that there is a smoothing functional relationship between x_i and y_i as follows:

$$y_i = f(x_i) + \varepsilon_i \ , \ 1 \leq i \leq n$$

where f is regression function and ε_i random errors. The objective of nonparametric regression is to estimate regression function f directly rather than estimate parameters. The nonparametric regression statistical methods are not assumption free as in sometime asserted, the function f assumed to belong to some collection of function possibly infinite dimention that share certain properties, it may be required that f is differentiable for example, also the errors are independent normally distributed with constant variance (i.e var $(\varepsilon_i) = \sigma^2$). There are several methods appeared in literature to estimate the regression function nonparametrically, some of these methods depend on spline function such as smoothing spline, regression spline [1], penelized spline [2],[3], the other are depend on kernel function such as Nadaraya Watson estimator [4], Local polynomial regression estimator [5]. An outlier is any observation which different so much from the other observation. [6]. Containing outliers in data sample will affect the estimated parameters also the shape of the estimated curve [7], Therefore, many robust methods were proposed to avoid the negative effects of outliers in the data. The effect of outliers are reduced using these methods without ignoring them [8] . In [9] Cleveland compute the robust estimators by fitting the weighted polynomial regression model, Cleveland [10] described the distributional properties of the local polynomial regression, Boente and Fraiman [11] proposed a nonparametric robust estimation of the conditional expectation by defining a robust conditional location functional. Wang and Scott [12] propose a robust nonparametric estimator by replacing the error criterion (L2) by criterion error (L1) of the parametric least squares regression. Cazals et all [13] propose The order-m technique, this estimator using the idea of expected minimum input function with varying degrees of robustness. By and Heng [14] propose a nonparametric robust kernel regression estimator by propose proposed a robust nonparametric cross validation to estimate the bandwidth for the procedure. Cai and Zhou [15] studied asymptotic equivalence for nonparametric robust regression with unbounded loss functions. Kai et all [16] proposed local polynomial (CQR), and showed that the estimator improve the efficiency of the estimator for for common non normal errors. Zheng, et all [17] propose a robust nonparametric kernel regression estimator by using a convex combination of different loss functions. Sun, et all proposed a weighted local polynomial (WCQR) which extends the local polynomial to asymmetric error distributions. Kennedy, et all, proposed a novel kernel smoothing procedure, also they derive asymptotic properties and propose data driven bandwidth selection.

The rest of the paper is organized as follows: In section 2, we give the basic idea of local polynomial regression. In section 3 we give the basic idea of robust local polynomial regression. In section 4

the simulation study and the results in section 5.

2 Local Polynomial Regression

The basic idea of this method is to estimate regression function locally rather than using global polynomial of p degree using all data to estimate f by estimating $p+1$ parameters as in parametric regression. To estimate the function f locally at the point x, a neighborhood of the form $(x-h, x+h)$ determined, where h is a bandwidth or smoothing parameter which determines a neighborhood around x. Only points of data withen the neighborhood are used to estimate the function f, also to ensure the smoothing, observations y_i that are points x_i of data close to the point x is has given more weight than those observations that are points of data further, also the observations that are data points outside the neighborhood given a zero weight. If we assume that the function f has derivatives of order $p+1$ at the point x, then the estimation of the function f using local polynomial of degree p is to find a solution β where $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$[4], using the weighted least squars by minimize the criterion:

$$Q = \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j (X_j - x)^j \right)^2 k\left(\frac{X_i - x}{h}\right)$$

Where k is kernel function, then

$$Q = (Y - X\beta)^T W (Y - X\beta)$$

X is designe matrix defined as follows:

$$X = \begin{pmatrix} 1 & (X_1 - x) & \dots & (X_1 - x)^p \\ 1 & (X_2 - x) & \dots & (X_2 - x)^p \\ & & \ddots & \\ 1 & (X_n - x) & \dots & (X_n - x)^p \end{pmatrix}$$

$$W = diag\left(k(\frac{X_j - x}{h})\right)$$

$$= \begin{pmatrix} k(\frac{X_1 - x}{h}) & & \\ & \ddots & \\ & & k(\frac{X_n - x}{h}) \end{pmatrix}_{n \times n}$$

by defferentiation with respect to β then:

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y$$

hence:

$$\hat{f}(x) = \hat{\beta}_0 = e_1^T (X^T W X)^{-1} X^T W Y$$

where $e_1^T = (1, 0, \dots, 0)_{1 \times n}$.

If $p = 0$ we get local constant estimator (Nadaraya Watson estimator), if $p = 1$ we get local linear regression estimator, and if $p = 2, p = 3$ we get local quadratic and local cubic regression estimators. The degree of of polynomial play a role in influencing of trade-off between biase and variance, for example, when choosing constant polynomial the estimator suffers from the problem of higher biase especially at the boundary (boundary biase), as well as suffering from increase in biase in the middle of the range especially in the case where the distribution of points irregularly [5]. If we choose linear polynomial, then the estimator adapts directly with the biase at the boundary, but in the case of data that contain rapid change in the slope, the estimator suffering from the increase in biase and can avoid this bby increase the degree of polynomial, but there is a price where the variance increase due to the increasing of the number of the estimated parameters [18].

3 Robust Local Polynomial Regression

Cleveland [9] [19] suggesting LOWESS algorithm (Locally Weighted Scatter plot Smoothing) to obtain an estimated curve that is robust to outliers and fit local polynomial of degree p for each observation using weights wj, according the size of the estimated residuals he found new weights to get robust estimates, then compute the robust estimators by fitting the weighted polynomial regression estimator. These weights in this estimator ensure that the adjacent points remain strongly weighted, while other points with high residuals have less influence over the final fit. This keeps ensuring a smooth estimates.

LOWESS Algorithm [19]:

1. Assign to all observations y_i weights $v_i = 1$.
2. Smooth estimate of data using local polynomials, with robust weight $v_i k_i(x)$ which is the production of robustness weight and localization weight.
3. Calculate the values $\hat{\varepsilon} = y_i - \hat{f}(x_i)$, $1 \le i \le n$ which are the residuals and estimate the s as the median of the absolute value residuals
4. Assign to the observations the robustness weight $v_i = k(\frac{\hat{\varepsilon}_i}{s})$.
5. Repeat steps 2,3,and 4.

The main advantage of LOWESS method that it does not need the function specification to fit a model to all data in the sample and the only needing are the value of smoothing parameter and the local polynomial degree.

4 Simulation Studies

For the purpose of achieving the objectives of this research and for more inclusive analysis, a large number of samples of different sizes and different values of standard errors were used. We perform 500 reblications generating independent sample of size n=30 for small sized problem, n=100 for moderate sample sizes and n=300 for large sample sizes. Data points are generated uniformally (i. e , $x_i \sim u(0,1)$).

We use the following modefied test functions:

$$f_1(x) = \sin(2\pi x^3) + 0.5(x - 0.75)^2,$$

$$f_2(x) = \frac{\sin(12(x+0.2))}{(x+0.2)} + \exp(-12x).$$

ε_i are iid and independent of x_i , where

$$\varepsilon_i \sim (1-\alpha)N(0,3) + \alpha N(0,3).$$

We considered four contamination properties $\alpha = 0\,, 0.1\,, 0.2$ and 0.3, the first one corresponding to the central normal model.

The aim of this study is to compare the behavior of three kinds of estimators which are robust local constant estimator denoted RLCE, robust local linear estimator denoted RLLE and robust local quadratic estimator denoted RLQE.

We use the integrated squared errors ISE as comparison criterion, where

$$\text{ISE}=\int [f(x) - \hat{f}(x)]^2 \, dx$$

which can be approximated by the formula

$$\text{ISE}=\left(1/t\right) \sum_{i=1}^{t} \left[f(z_i) - \hat{f}(z_i)\right]^2 ,$$

$z_i = 1/t$, $1 \le i \le t$, t=300.

5 Results

1. According to the values of ISE (Table 1) for the first test function f_1 we found
   the follwing results:
   a) For 0% contamination data:
      - If n=30 ,n=100, the arrangments of the estimators as follows:
        RLQE , RLLE, RLCE
      - If n=300 the arrangment of the estimatorls as follows:
        RLLE, RLQE, RLCE
   b) For 10% contamination data:
      - If n=30, n=300 the arrangments of the estimators as follows:
        RLQE, RLCE , RLLE
      - If n=100 the arrangments of estimators as follows:
        RLQE, RLLE, RLCE
   c) For 20% contamination data the arrangment of the estimators for every n as follows:
      RLQE, RLCE , RLLE
   d) For 30% contamination data:
      - If n=30, n=300 the arrangment of the estimators as follows:
        RLCE, RLQE, RLLE
      - If n=100 the arrangments of estimators as follows:
        RLQE, RLCE, RLLE
2. According to the values of ISE (Table 2) for the first test function f_2 we found
   the follwing results:

a) For 0% contamination data, the arrangments of estimators for every n as follows: RLQE, RLLE, RLCE
b) For 10% contamination data, the arrangments of estimators for every n as follows: RLQE, RLLE, RLCE
c) For 20% contamination data:
   - If n=30, n=300 the arrangments of the estimators as follows:
     RLQE, RLLE, RLCE
   - If n=100 the arrangments of the estimators as follows:
     RLQE, RLCE , RLLE
d) For 30% contaminations data:
   - If n=30, n=100 the arrangments of the estimators as follows:
     RLCE, RLQE, RLLE
   - If n=300 the arrangment of the estimators as follows:
     RLQE, RLCE, RLLE

6 Concolusions

1. The robust local quadratic estimator has best performance compared with the rest of the estimators in this study, especially when contamination rates 0%, 10%, and 20%, followed by the robust local linear estimator then robust local constant estimator.
2. Robust local constant estimator give resonable performance with different rates of contamination.

7 Recommendations

1. We recommends using robust RLCE especially when there is little contamination in data.
2. In the case of existance of many outliers in the data, we recommends the use of RLQE.

Estimator	n	ISE	ISE	ISE	ISE
		0%	**10%**	**20%**	**30%**
RLCE	**30**	**1.370**	**2.112**	**2.530**	**3.865**
	100	**1.384**	**2.423**	**3.098**	**4.105**
	300	**1.568**	**2.607**	**3.475**	**4.937**
RLLE	**30**	**1.233**	**2.115**	**2.636**	**4.246**
	100	**1.241**	**2.346**	**3.489**	**4.659**
	300	**1.366**	**2.608**	**4.162**	**6.500**
RLQE	**30**	**1.167**	**2.039**	**2.447**	**3.908**
	100	**1.235**	**2.179**	**3.009**	**4.093**
	300	**1.397**	**2.358**	**3.326**	**5.058**

Table (1) The results of the test function f_1

Estimator	n	ISE	ISE	ISE	ISE
		0%	**10%**	**20%**	**30%**
RLCE	**30**	**2.893**	**3.609**	**4.187**	**5.040**
	100	**3.241**	**4.243**	**4.279**	**5.288**
	300	**5.089**	**5.571**	**5.095**	**4.815**
RLLE	**30**	**2.738**	**3.532**	**4.096**	**5.403**
	100	**2.813**	**3.848**	**4.388**	**5.400**
	300	**3.251**	**4.345**	**4.945**	**5.023**
RLQE	**30**	**2.730**	**3.520**	**4.097**	**5.125**
	100	**2.642**	**3.685**	**4.085**	**5.310**
	300	**2.765**	**3.843**	**4.349**	**4.359**

Table (2) The results of the test function f_2.

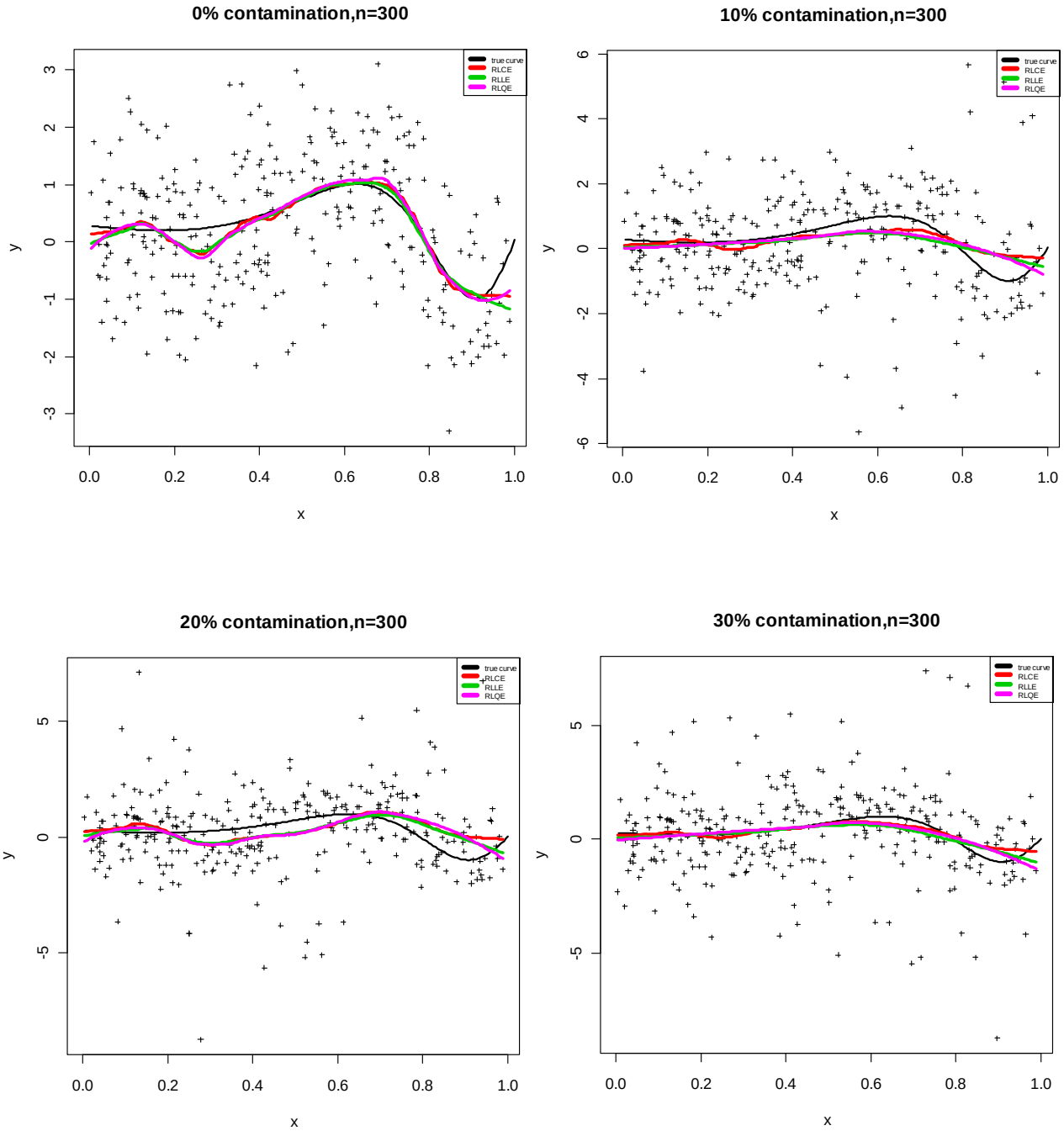


Figure (1) Graphics of estimators and the test function f_1 for n=300 and different contamination rates

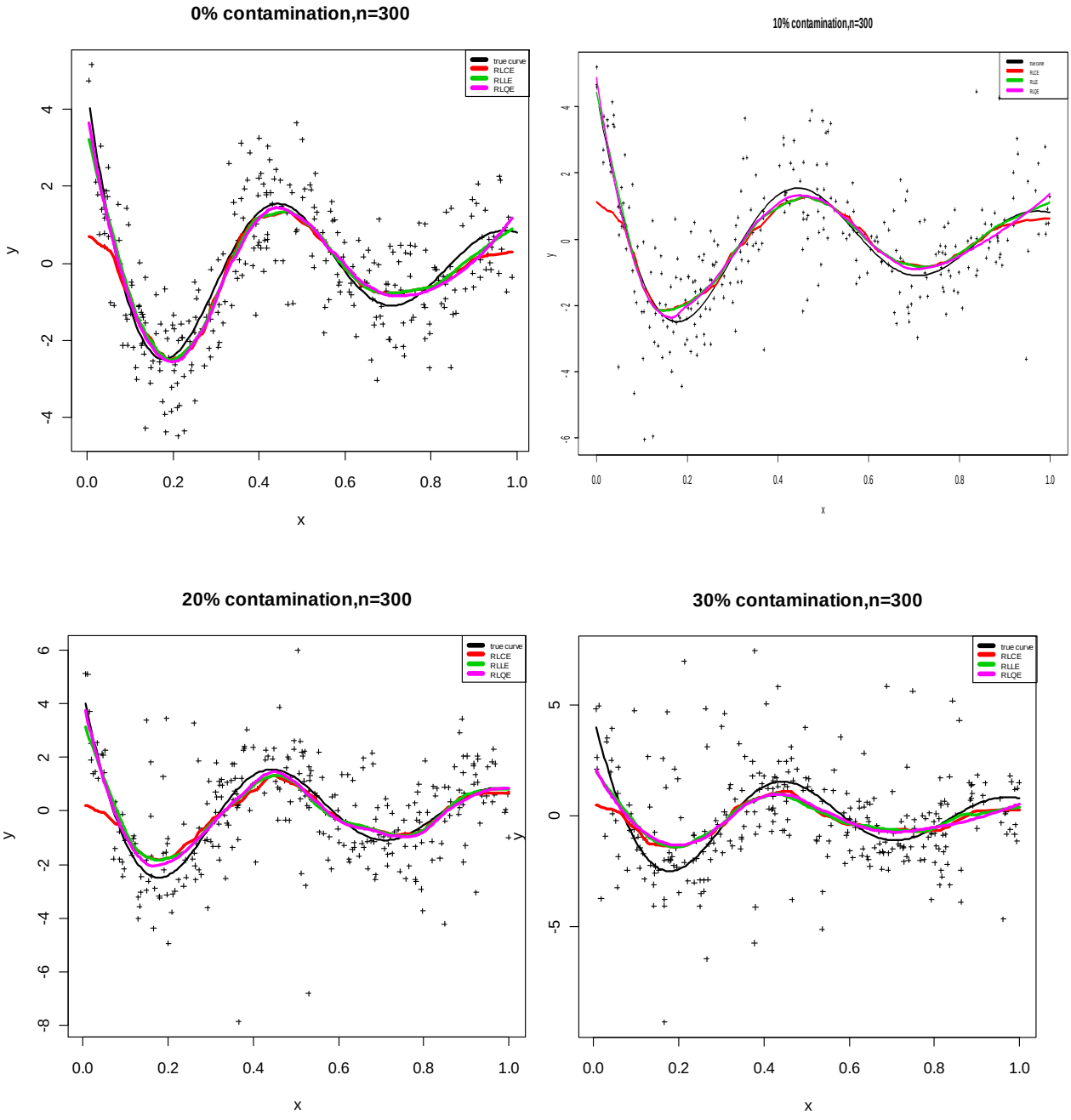
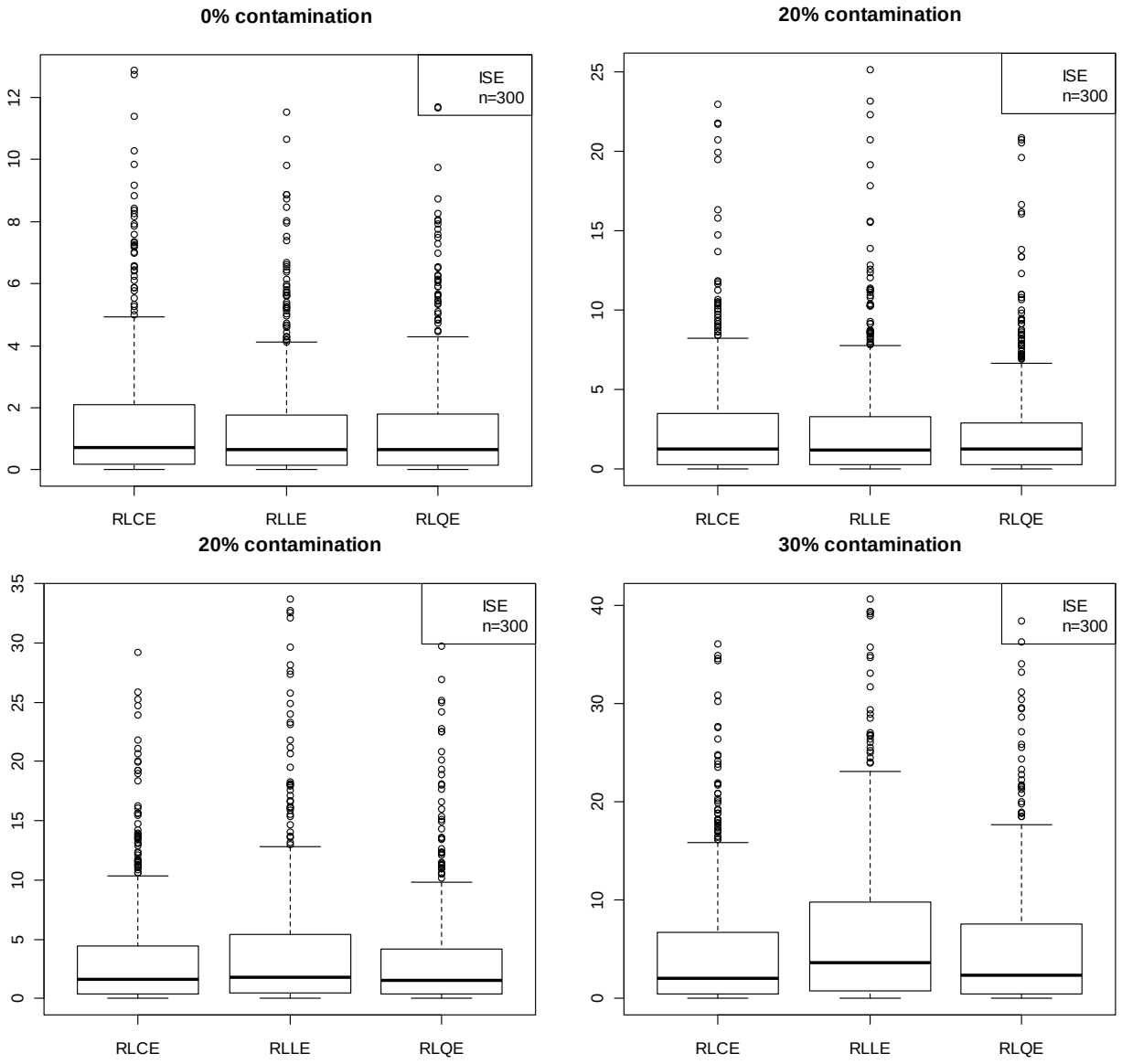
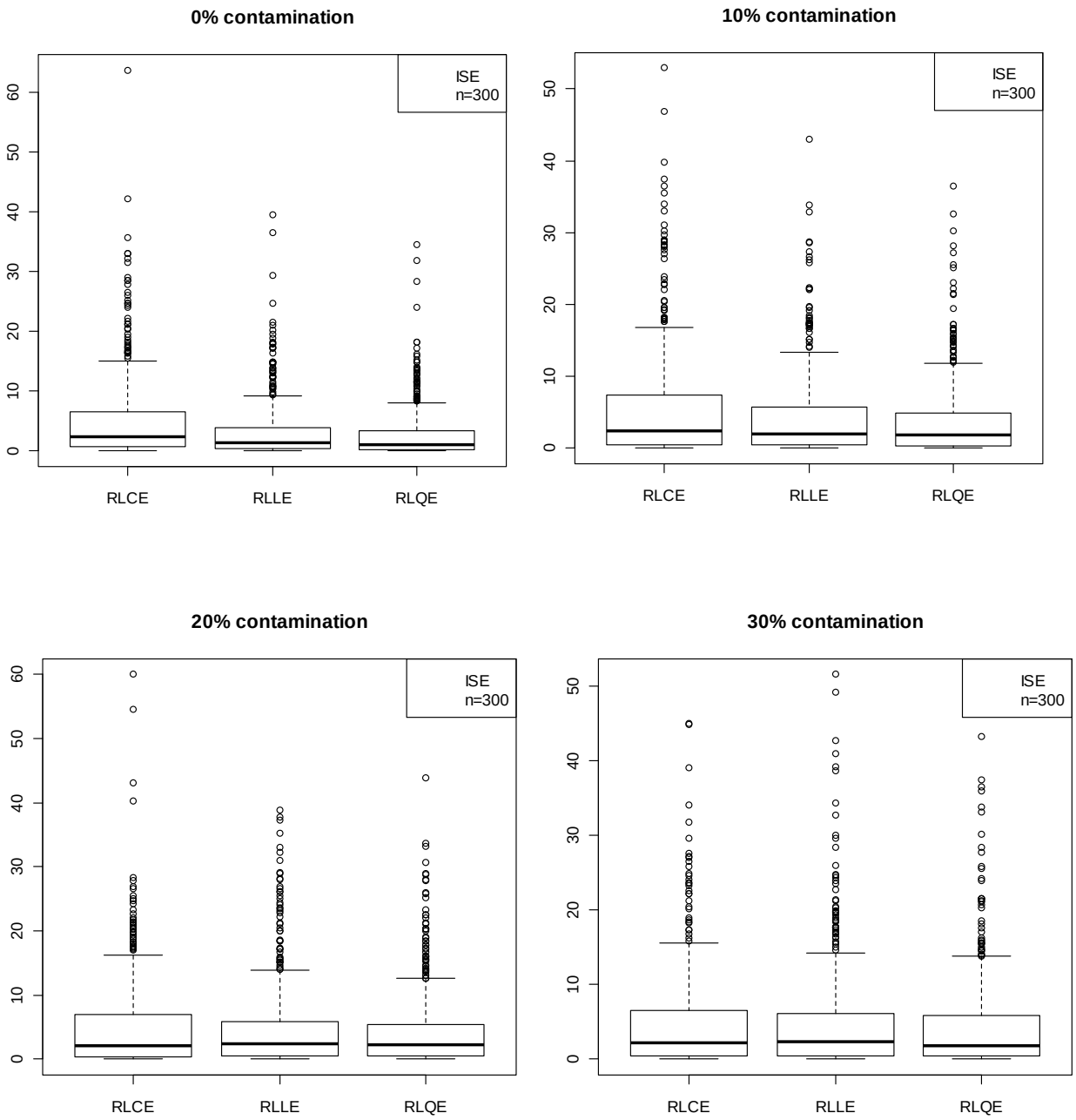


Figure (2) Graphics of estimators and the test function f_2 for n=300 and different contamination rates



Figure(3) Boxplots of the values of ISE for n=300, the function f_1 and different contamination rates.



Figure(4) Boxplots of the values of ISE for n=300, the function f_2 and different contamination rates.

8 References

[1] Eubank R.L., "Nonparametric Regression and Spline Smoothing". Marcel Dekker, NewYork, NY,1999.

[2] Eilers P.H.C., Marx, B.D., " Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder)". Statistical Science, 1996; 11(2): 89-121

[3] Loader C. R., " Local Regression and Likelihood". Lucent Technologies,ISBN 0-387-98775-4 Springer, 1996.

[4] Fan J., Gasser T., Gijbel I., Brockman M., Engel J., "Local Polynomial Fitting : A Standard for Nonparametric Regression". Mineo Series 2302, Department of Statistics,chapel, 1993.

[5] Fan J., " Design adaptive nonparametric regression". J. Am. Statist. Assoc.1992; 87, 998-1004

[6] Charu C. Aggarwal " Outlier Analysis". Second Edition, Charu C. Aggarwal, ISBN 978-3-319-47577-6, Springer International, 2017.

[7] Huber P.J., Ronchetti M.E., "Robust Statistics". second ed. Wiley, New York, 2009.

[8] Pimpan A., Prachoom S., "A Modefied Influence Function for Estimation of Regression Coefficient with Outliers". Chiang Mai J.Sc; 2011; 38(3):346-359.

[9] Cleveland W.S., Devlin S.J.,"Robust Locally weighted regression: an approach to regression analysis by local fitting". Journal of the American Statistical Association, 1979; 74, 829-836.

[10] Cleveland W., "Regression by local fitting methods, properties, and computational algorithms". Journal of Econometrics, 1988; 37(1), 87–114.

[11] Boente G., Fraiman R., " Robust Nonparametric Regression Estimation". Journal of Multivariat Analysis, 1989; 29, 180-198.

[12] Ferdinand T. Wang., David W. Scott., "The L1 Method for Robust Nonparametric Regression". Journal of the American Statistical Association, 1994; 89:425, 65-76.

[13] Cazals C., Florens J., Simar L., "Nonparametric frontier estimation: a robustapproach". Journal of Econometrics, 2002; 106 (1), 1-15.

[14] By D., Heng Y., "Cross validation in nonparametric regression with outliers", The Annals of Statistics, 2005; 33(5), 2291–2310.

[15] Cai T. T. Zhou H.H., " Asymptotic equivalence and adaptive estimation for robust nonparametric regression". The Annals of Statistics, 2009; 37, 3204-3235.

[16] Kai B., Li R., Zou H., "Local Composite Quantile Regression Smoothing": An Efficient and Safe Alternative to Local Polynomial Regression’. Journal of the Royal Statistical Society Series B, 2010; 71, 49–69.

[17] Qi Z, Colin G., Kulasekera K.B., "Adaptively weighted kernel regression". Journal of Nonparametric Statistics, 2013; 25:4, 855-872.

[18] Hurdle W., " Applied Nonparametric Regression" ,Hill,North Carolina Oxford University Press, Oxford, 1990.

[19] Cleveland., William S., "LOWESS: A program for smoothing scatterplots by

robust locally weighted regression". The American Statistician, 1981; 35 (1): 54.