
A COMPARISON BETWEEN GLOBAL AND ADAPTIVE BANDWIDTHS USING (DPI) AND (RSC) SELECTION METHODS

A.K.Abdlkder* A.M.Mami**
*Dep. Of St. Fac. Of Science – Sebha U.
** Dep. Of St. Fac. Of Science – Benghazi U.

Abstract

Non-parametric regression is type of regression analysis in which functional form of the relationship response variable and the associated predictor variable dose not to be specified in order to fit a model to a set of data. There are many different methods for non-parametric regression. We have used the local polynomial kernel estimator with the optimal choice of the smoothing parameter(s).Choosing the optimal smoothing parameter(s), which is usually called the bandwidth(s), is considered to be one the most important issue when using the kernel-based estimator. Also, the optimal bandwidth(s) played crucial role to uncover the structure of the underlying data. In this paper, we use simulation approach to make comparisons between two different strategies for selecting the optimal bandwidth(s), namely the Direct plug-In (**DPI**) selection method and the Residual Square Criterion (**RSC**) method. Within the context of these two strategies of selecting the optimal bandwidth(s), there are two different settings of choosing the smoothing parameter: global (single bandwidth) or adaptive (variable bandwidths). Moreover, four different example-regression models have been used in order to smooth the mean regression functions. Several statistical properties have been investigated in the simulation study. Such study must be restrictive because of the many possibilities to be consider, we decide to consider the following three elements:
The size of sample (n). Distribution of the error's (normal, exponential). The kind of Design (fixed and random).
Key words: Non-parametric regression, local polynomial kernel estimator, Direct plug-In (DPI) selection, Residual Square Criterion.

Introduction

As one basic form of statistical inference, regression analysis has been usually used in discovering the relationship between one quantity (called dependent variable) and several other quantities (called explanatory variables). Classic regression analysis concerns mainly the inference on a fixed number of parameters, e.g., slope and intercept of a regression line, such regression relationships are called parametric as they are clearly determined by those fixed number of parameters on which statistical inferences are made. Scientific data from many disciplines, however, exhibit strong nonconformity to parametric models.

In a well-known motorcycle crash test data, for instance, the acceleration of the dummy head after impact follows a complicated rather than a simple polynomial time trend. Effective tools for extracting information from such complicated regression data have to be nonparametric in nature; that is, no particular formulae are imposed on the regression structures, allowing “the data to speak for itself ”.

The limitations of parametric models have led to the enormous development of nonparametric regression models in the recent two decades (Yang (2006)).

Experimental

A **simulation** is an imitation of some real things, or process. The act of simulating something generally entails representing certain key characteristics or behaviors of a selected statistical model. In this paper the main purpose is to conduct a simulation study in order to evaluate and compare the estimated regression curves. These estimated regression curves are resulted from utilizing the two bandwidth selection methods: namely the Direct Plug-In (**DPI**) selection method, and Residual Square Criteria (**RSC**) selection method in two different design-settings namely: global bandwidth selecting and the adaptive bandwidths selection. Such a study is necessarily to be restrictive, because there are many possibilities for the choices of example regression functions g(x), sample sizes (n), bandwidth selection setting of either global or adaptive, and the distribution of the errors term.

Since the regression analysis can be applied to both the fixed design and the random design, in the next sections, we will deal with each design case separately.

Description of the Experiment in the Fixed and Random Design Case:-

We have conducted 48 simulation studies, using four different example regression functions, and two selection methods of the bandwidth selection settings (either global bandwidth or adaptive bandwidths), and three different sample sizes n, and two different types of the errors' distribution. Each simulation study involves (L=500) repetitions.

We have generated the random design through the uniform distribution on the interval [a,b].We consider four different example regression function as present in (Table 1).The shape of these functions exhibited several features such as concave, convex, and flat parts. We have also chosen three different sample sizes of n=50, n=100 and n=300.

Table 1 The Example Regression Functions used in the simulation study

Example	$g(x)$	Interval	σ
1)	$g_1(x)= 0.3\, e^{(-16(x-0.25)^2} + 0.7\, e^{(64(x-0.75)^2)}$	$x \in [0,1]$	0.15
2)	$g_2(x) = \frac{3}{4} Sin(\frac{3\pi x}{2} + 1.25)$	$x \in [0,1]$	0.6
3)	$g_3(x)=Sin(2x)+2\, e^{-16x^2}$	$x \in [-1,1]$	0.3
4)	$g_4(x)=1+Sin(x)+2\, Cos(x)+3\, Sin(5x)$	$x \in [-1,1]$	1.2

Numerical Summary

Having conducted the simulation runs using the (L=500) simulated data sets, we obtained the results that are tabulated in Table 2, and Table 3. These tables displays the numerical summary of results obtained using two different bandwidth selection procedures (DPI and RSC) when the bandwidth selection setting is either global or adaptive. Moreover, two different errors' distributed, are used once as normaly distributed and other as exponentially distributed with three different sample sizes of n=50, n=100, and n=300. The numerical summary includes three statistical properties, namely the Average Square Bias (ASB), the Average Variance (AVAR) and the Square Root of Average Mean Square Errors (RAMSE). The formulae that we have been used to calculate ASB, AVAR, and RAMSE are as follows:

$$ASB = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{L}\sum_{j=1}^{L} g(x_{ij}) - g(x_i)\right]^2$$

$$AVAR(g) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{L}\sum_{j=1}^{L} g^2(x_{ij}) - \left(\frac{1}{L}\sum_{j=1}^{L} g(x_{ij})\right)^2\right]$$

$$RAMSE(g) = \sqrt{ABS(g) + AVAR(g)}$$

The main program and the other four auxiliary sub-programs are used to calculate the bandwidths as well as obtaining the corresponding fitted values of the regression curves through utilizing two different criteria, namely: (**DPI**) selection method and (**RSC**) method when the bandwidth is either in global setting or adaptive setting. All computations have been carried out using the R statistical package plus two different libraries LOKERN and KernalSmooth. (The programs are available upon request through the first author)

RESULTS and DISCUSSION

Table 2 displays the numerical summary that results from using two different bandwidth selections the (**DPI**) selection method and the (**RSC**) method in two settings global and adaptive from the simulation study in fixed design when the errors' distributed as normal. The numerical summary includes **RAMSE** for 500 simulation runs using 4 example regression functions $g_1(x)$, $g_2(x)$, $g_3(x)$, and $g_4(x)$.

From Table 2 , we have made several comparisons as following: The less **RAMSE** in each experiment:

At the sample size n=50, we have noticed that the values of **RAMSE** are smaller when using the (**DPI**) bandwidth selection (in global setting) and this can be seen in the 2nd, and 4th example regression functions. Also the values of **RAMSE** are smaller when using the (**DPI**) bandwidth selection (in adaptive setting) as seen in the 1st and 3rd example regression functions. At the sample size n=100, we have noticed that the smallest values of **RAMSE** are alternating between the two selection criteria and the two settings. At the sample size n=300, we have noticed that the values of **RAMSE** obtained using the (**RSC**) bandwidth selection (in global setting) are the smallest among all.

Comparison between (**DPI**) and (**RSC**) bandwidth selection:

An interesting feature to note is that the values of **RAMSE** obtained using the (DPI) bandwidth selection are smallest when sample size n=50. However, the values of **RAMSE** obtained using the in (**RSC**) bandwidth selection are the smallest when n=300. There is no clear cut which criteria is better when n=100.

Comparison between global setting and adaptive setting in (**DPI**) bandwidth selection:

We have noted that the (**DPI**) bandwidth selection favor the global setting (since it gives smaller values of the RAMSE) when n=50 or n=300.

Comparison between global setting and adaptive setting in (**RSC**) bandwidth selection:

It is worth noting that the (**RSC**) bandwidth selection favors the global setting for all choices of n in all example regression functions.

Table 3 displays the numerical summary that results from using two different bandwidth selections the (**DPI**) selection method and the (**RSC**) method in two settings global and adaptive from the simulation study in fixed design when the errors' distributed as exponential. The numerical summary includes **RAMSE** for 500 simulation runs using 4 example regression functions $g_1(x)$, $g_2(x)$, $g_3(x)$, and $g_4(x)$.

(Table 2) The numerical summary of results obtained using two different bandwidth selections (DPI and RSC) in two settings global and adaptive from the simulation study in fixed design when the errors' distributed as normal. That includes RAMSE for 500 simulation runs using 4 example regression functions g_1 (x), g_2 (x), g_3 (x), and g_4 (x).

	Global		Adaptive		Global		Adaptive		Global		Adaptive	
	Sample size n=50		*Sample size n=50*		*Sample size n=100*		*Sample size n=100*		*Sample size n=300*		*Sample size n=300*	
	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC
The 1st Example Regression Function $g_1(x)$												
RAMSE	**0.06589**	**0.06711**	**0.06572**	**0.06924**	**0.05049**	**0.05007**	**0.05155**	**0.05138**	**0.03218**	**0.03145**	**0.03329**	**0.03154**
The 2nd Example Regression Function $g_2(x)$												
RAMSE	**0.20387**	**0.22205**	**0.23709**	**0.23817**	**0.14957**	**0.15269**	**0.16698**	**0.16139**	**0.08949**	**0.08877**	**0.11025**	**0.09189**
The 3rd Example Regression Function $g_3(x)$												
RAMSE	**0.14526**	**0.14198**	**0.13754**	**0.14692**	**0.10939**	**0.10603**	**0.10351**	**0.10921**	**0.06845**	**0.06655**	**0.06977**	**0.06687**
The 4th Example Regression Function $g_4(x)$												
RAMSE	**0.51211**	**0.52115**	**0.53265**	**0.54551**	**0.38892**	**0.38705**	**0.39103**	**0.40506**	**0.24082**	**0.23738**	**0.25550**	**0.24405**

From Table 3, we have also made several comparisons as following:

The less **RAMSE** in each experiment:

We have noticed that the values of **RAMSE** , at the sample size n=50, are smaller when using the (**DPI**) bandwidth selection (in global setting) for all choices example regression functions. At the sample size n=100, we have noticed that the smallest values of **RAMSE** are alternating between the two selection criteria and the two settings. At the sample size n=300, we have noticed that the values of RAMSE obtained using the (RSC) bandwidth selection (in global setting) are the smallest among all.

Comparison between (global and adaptive setting) in (**DPI**) bandwidth selection:

The (**DPI**) bandwidth selection favors the global setting (since it gives less values of the RAMSE) for all choices of the sample sizes n. . However, the value of RAMSE obtained using the in (RSC) bandwidth selection are the smallest when n=300.

There is no clear cut which criteria is better when n=100.

Comparison between (global and adaptive setting in (**RSC**) bandwidth selection:

Once more, the (**RSC**) bandwidth selection favors the global setting for all choices of n and all example regression functions.

Lastly, having examined carefully Table 2 and Table 3, we have noticed the following final remarks: First, the values of **RAMSE** decrease as the sample size increases in both cases when the errors' distributed as either normal or exponential. Second, we noticed that the values of **RAMSE** in (**DPI**) are smaller than their corresponding counterparts of (**RSC**) when the sample size n=50. While the values of **RAMSE** in (**RSC**) are better than those obtained from (**DPI**) when the sample size n=300. Third, we noticed too that the values of RAMSE are smaller in the global setting comparing with those obtained from the adaptive setting. The complete numerical results.

Table 4 displays the numerical summary that results from using two different bandwidth selections the (DPI) selection method and the (RSC) method in two settings global and adaptive from the simulation study in random design when the errors' distributed as normal. The numerical summary includes RAMSE for 500 simulation runs using 4 example regression functions $g_1(x)$, $g_2(x)$,$g_3(x)$, and$g_4(x)$.

(Table 3) The numerical summary of results obtained using two different bandwidth selection (DPI and RSC) in two settings global and adaptive from the simulation study in fixed design when the errors' distributed as exponential. That includes RAMSE for 500 simulation runs using 4 example regression functions g_1 (x), g_2 (x), g_3 (x), and g_4 (x).

	Global		Adaptive		Global		Adaptive		Global		Adaptive	
	Sample size n=50		*Sample size n=50*		*Sample size n=100*		*Sample size n=100*		*Sample size n=300*		*Sample size n=300*	
	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC
The 1st Example Regression Function g_1(x)												
RAMSE	**0.10888**	**0.12195**	**0.14742**	**0.14375**	**0.06725**	**0.06769**	**0.08126**	**0.07724**	**0.02369**	**0.0222982**	**0.03650**	**0.02365**
The 2nd Example Regression Function g_2(x)												
RAMSE	**0.10204**	**0.11501**	**0.14907**	**0.13501**	**0.05991**	**0.06213**	**0.07688**	**0.07128**	**0.02125**	**0.0208903**	**0.03718**	**0.02251**
The 3rd Example Regression Function g_3(x)												
RAMSE	**0.15467**	**0.15798**	**0.16443**	**0.18008**	**0.09270**	**0.08612**	**0.08510**	**0.09627**	**0.03512**	**0.0325779**	**0.04112**	**0.03397**
The 4th Example Regression Function g_4(x)												
RAMSE	**0.19502**	**0.19955**	**0.21599**	**0.22288**	**0.11022**	**0.10786**	**0.11526**	**0.11851**	**0.0443441**	**0.04325**	**0.05195**	**0.04571**

From the Table 4, we have made several comparisons as following:

The less RAMSE in each experiment:

At the sample sizes n=50 and n=300, we have noticed that the values of RAMSE are smaller when using the (DPI) bandwidth selection (in global setting) and this can be seen in all example regression functions. However, when the sample size n=100, we have noticed that there is no clear conclusion can be drawn of which criterion (RSC or DPI) or setting (global or Adaptive) give the smallest values of RAMSE.

Comparison between (DPI) and (RSC) bandwidth selection:

It is obvious from the results obtained that the values of RAMSE obtained using the (DPI) bandwidth selection are smallest than their corresponding counterparts obtained using the in (RSC) bandwidth selection for all sample sizes used (n=50, 100, and 300).

Comparison between global setting and adaptive setting in (DPI) bandwidth selection:

We have noted that the (DPI) bandwidth selection favor the global setting (smaller values of the RAMSE). It is very clear for choices of the sample sizes n=50, and n=300.

Comparison between global setting and adaptive setting in (RSC) bandwidth selection:

Again, we have noticed that the (RSC) bandwidth selection favors the global setting for all choices of n and all example regression functions used in the study

(Table 4) The numerical summary of results obtained using two different bandwidth selection (DPI and RSC) in two settings global and adaptive from the simulation study in random design when the errors' distributed normal. That includes RAMSE for 500 simulation runs using 4 example regression functions g_1 (x), g_2 (x), g_3 (x), and g_4 (x).

	Global		Adaptive		Global		Adaptive		Global		Adaptive	
	Sample size n=50		*Sample size n=50*		*Sample size n=100*		*Sample size n=100*		*Sample size n=300*		*Sample size n=300*	
	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC
The 1st Example Regression Function g_1(x)												
RAMSE	**0.01147**	**0.01325**	**0.01202**	**0.01387**	**0.00510**	**0.00567**	**0.00460**	**0.00609**	**0.004252**	**0.004456**	**0.00438**	**0.00451**
The 2nd Example Regression Function g_2(x)												
RAMSE	**0.05693**	**0.08174**	**0.07582**	**0.08889**	**0.02674**	**0.03670**	**0.03212**	**0.04001**	**0.009538**	**0.011104**	**0.014019**	**0.011627**
The 3rd Example Regression Function g_3(x)												
RAMSE	**0.17742**	**0.18803**	**0.17463**	**0.18976**	**0.06276**	**0.06084**	**0.06157**	**0.06236**	**0.018674**	**0.019466**	**0.019091**	**0.019778**
The 4th Example Regression Function g_4(x)												
RAMSE	**0.69313**	**0.76657**	**1.38086**	**0.79522**	**0.54678**	**0.57609**	**0.55142**	**0.59526**	**0.110432**	**0.11657**	**0.121526**	**0.120768**

From the Table 5, we have also made several comparisons as following:

The less **RAMSE** in each experiment:

We have noted that the values of **RAMSE** are the smallest when using the (**DPI**) bandwidth selection (in global setting). It is very obvious for choices of the sample sizes n=50, and n=300, for all choices example regression functions.

Comparison between (**DPI**) and (**RSC**) bandwidth selection:

For most example regression functions used in the study, we have noticed that the values of **RAMSE** obtained using the (**DPI**) bandwidth selection are smaller than their corresponding counterparts obtained using the (**RSC**) bandwidth selection among the all choices of sample size.

Comparison between global setting and adaptive setting in (**DPI**) bandwidth selection:

We have noted that the (**DPI**) bandwidth selection in the global setting gives the smallest values of the **RAMSE** for all choices of the sample sizes n, and the majority of example regression function used in the study.

Comparison between global setting and adaptive setting in (**RSC**) bandwidth selection:

It is worth noting that the (**RSC**) bandwidth selection prefers the global setting for all choices of sample sizes and all example regression functions. This were exhibited by the numerical results obtained.

Generally speaking, we noticed through examining Table 4 and Table 5 three important points: First, the values of **RAMSE** decrease as the sample sizes increase regardless the errors' are distributed either as normal or exponential. Second, we noticed too that the values of **RAMSE** in obtained using the (**DPI**) bandwidth selection criterion are smaller the corresponding counterparts obtained using the (**RSC**) bandwidth selection criterion. Third, we noticed also that the values of **RAMSE** are better (smaller) when the bandwidth selection setting is global.

(Table 5) The numerical summary of results obtained using two different bandwidth selection (DPI and RSC) in two settings global and adaptive from the simulation study in random design when the errors' distributed exponential. That includes RAMSE for 500 simulation runs using 4 example regression functions $g^{1}(x)$, $g^{2}(x)$, $g^{3}(x)$, and $g^{4}(x)$.

	Global		Adaptive		Global		Adaptive		Global		Adaptive	
	Sample size n=50		*Sample size n=50*		*Sample size n=100*		*Sample size n=100*		*Sample size n=300*		*Sample size n=300*	
	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC	DPI	RSC
The 1st Example Regression Function $g_1(x)$												
RAMSE	**0.10638**	**0.15550**	**0.18565**	**0.18040**	**0.07196**	**0.09011**	**0.08251**	**0.10002**	**0.023480**	**0.02630**	**0.03681**	**0.02788**
The 2nd Example Regression Function $g_2(x)$												
RAMSE	**0.13614**	**0.20063**	**0.19770**	**0.23153**	**0.08912**	**0.11187**	**0.10166**	**0.12539**	**0.024426**	**0.02670**	**0.04637**	**0.02885**
The 3rd Example Regression Function $g_3(x)$												
RAMSE	**0.23965**	**0.29563**	**0.27825**	**0.32781**	**0.11001**	**0.12828**	**0.10818**	**0.13832**	**0.04409**	**0.05041**	**0.04962**	**0.05227**
The 4th Example Regression Function $g_4(x)$												
RAMSE	**0.43654**	**0.48990**	**0.46456**	**0.53265**	**0.51506**	**0.53036**	**0.52178**	**0.53981**	**0.18669**	**0.19485**	**0.19285**	**0.19755**

Summary and Conclusions

The main results from our study can be summarized in the following 5 remarks:

- The values of **RAMSE** decreases as the sample size (n) increases in both strategies (**DPI** or **RSC**) of selecting the optimal bandwidth regardless of the the errors' distribution (normal or exponential), type of design is (fixed or random), and the setting (global or adaptive).
- The (**DPI**) bandwidth selection criterion appears to give better (smaller) estimates of **RAMSE** when the sample sizes (n) are small. However (**RSC**) bandwidth selection

criterion gives better (smaller) estimates of **RAMSE** when the sample sizes (n) are large, especially when the design of x-values is fixed.

- The (**DPI**) selection criterion in general seems to work in much better manner than their corresponding counterparts of (**RSC**) when the design of x-values is random regardless of the errors' distribution.

- The values of **RAMSE** obtained using the (**DPI**) bandwidth selection criterion (in global setting) are smaller (more consistency) than their corresponding counterparts of (in Adaptive setting) regardless the type of the design (fixed or random), and errors' distribution (normal or exponential).

- The values of **RAMSE** obtained using the (**RSC**) bandwidth selection are smaller (more consistency) than their corresponding counterparts of (**DPI**) bandwidth selection (in global setting), when the design is either fixed or random, and the errors' distribution is either normal or exponential.

مقارنة بين معلمة التنعيم الثابتة والمتغيرة باستخدام طريقتي التعويض المباشر ومعيار مربع البواقي

عبدالسلام كامل بشير*   أحمد محمد مامي**
*قسم الإحصاء -كلية العلوم- جامعة سبها
** قسم الإحصاء -كلية العلوم- جامعة بنغازي

الملخص

الانحدار اللامعلمي هو نوع من تحليل الانحدار والتي تكون فيه الدالة الرياضية للعلاقة بين المتغير التابع ونظيره المتغير المستقل ليست محددة الملامح وبالتالي لا يتلائم مع البيانات بصدد الدراسة . يوجد هنالك العديد من الطرق لتنعيم البيانات بطريقة الانحدار اللامعلمي . في هذا البحث سوف نستخدم local polynomial kernel estimator مع الاختيار الأمثل لمعلمة التنعيم . اختيار القيمة المثلى لمعلمة التنعيم kernel-based estimator تعتبر من إحدى المواضع الهامة عند استخدامنا للمقدرات المبنية على دالة الوزن kernel . كذلك القيمة المثالية لمعلمة التنعيم تلعب دورا رئيسيا في الكشف عن هيكلية البيانات بصدد الدراسة . في هذا البحث سوف نستخدم أسلوب المحاكاة للمقارنة ما بين إستراتيجيتان مختلفتان لاختيار القيمة المثالية لمعلمة (معالم) التنعيم bandwidth(s) وهما طريقتي Direct plug -In (DPI) and Residual Square Criterion (RSC). عندما معلمة التنعيم تكون ثابتة (قيمة وحيدة) أو متغيرة (أكثر من قيمة)، في هذا البحث اخترنا أربعة نماذج انحدار مختلفة تحت الاختبار وذلك لتنعيم دالة الانحدار ثم تم التحقيق في تأثير ثلاثة عناصر مهمة وهي

1- حجم العينة
2- توزيع الأخطاء (طبيعي ، أسي)
3- نوع التصميم (ثابت ، عشوائي)

References

1. Abdelkder A. K. (2009). A Comparison Between Global and Adaptive Bandwidths Using(DPI) and (RSC) Selection Methods. Unpublished M.Sc thesis Dept. of Statistics, University of Garyounis Libya.
2. Fan, J.,and Gijebels, I. (1996) Local polynomial Modeling and Its Applications. Chapman&Hall. London.
3. Mami, A.(2002). Local Polynomial Regression with Applications To both Independent and Longitudinal

Data. Unpublished Ph.D thesis Dept. of Statistics, University of Manchester United Kingdom.

4. R Development Core Team, R: A Language and environment for statistical computing, R Foundation for statistical computing Vienna, ISBN 3-900051-07-0, 2005

5. Wand, M. P. and Jones, M. C. (1995). Kernel Smooting .Chapman&Hall London.

6. Yang, L. (2006). Have you considered nonparametric regression? Michigan State University 1,1-2