تقدير دالة انحدار الحرف اللبي في الأنحدار المتعدد اللامعلمي بأستعمال المحاكاة

أ.م.د. لقاء علي محمد/ كلية الادارة والاقتصاد/ جامعة بغداد
الباحث/ صابرين حسين كاظم

تاريخ التقديم:2017/3/6
تاريخ القبول:2017/10/5

المستخلص

عادة ما يستعمل الباحثون بشكل عام و الأحصائيون بشكل خاص الأنحدار اللامعلمي عندما تعجز الطرائق المعلمية عن تحقيق غاياتهم في تحليل النماذج بدقة معينة ، و من ثم تكون هذه الطرائق غير مجدية لذلك يتم اللجوء الى الطرائق اللامعلمية لسهولة برمجتها حاسوبيا ، كما ويمكن أن تستعمل الطرائق اللامعلمية لأفتراض النموذج المعلمي للأنحدار لأستعماله لاحقآ ، و من ضمن استعمالات الطرائق اللامعلمية هي معالجة احدى مشاكل الأنحدار ، ألا وهي مشكلة التعدد الخطي **Multi-Colinearity Problem** بين المتغيرات التوضيحية عند اقترانها بمشكلة لاخطية البيانات **Nonlinear Data** ، و ذلك بإستعمال دالة انحدار الحرف اللبي **Kernel Ridge Regression (KRR)** ، والتي تعتمد على تقدير عرض الحزمة (او ما تسمى بمعلمة التمهيد **Bandwidth (smoothing parameter** و لذلك تم اللجوء الى طريقتين مختلفتين لتقدير المعلمة الأخيرة و هما طريقة الأمكان الأعظم للعبور الشرعي **Maximum Likelihood (MLCV** **Cross-Validation** و طريقة معيار **AKaike (AIC)** و المقارنة بين هاتين الطريقتين بأستعمال اسلوب المحاكاة و قد تم التوصل الى إن طريقة معيار **AKaike (AIC)** هي الأفضل بالنسبة لدالة **Gaussian** .

المصطلحات الرئيسة للبحث/انحدار الحرف اللبي (**KRR**) ، **MLCV** ، **AIC** ، معلمة التنظيم λ

مجلة العلوم
الاقتصادية والإدارية
العدد 103 المجلد 24
الصفحات 411-419

* بحث مستل من رسالة ماجستير

1- المقدمة : 1- Introduction :

تزداد الحاجة الى استعمال اساليب الأحصاء كونها ادوات ذات اهمية في التقدير و التنبوء ، و يعد تحليل الأنحدار احد اهم الأدوات الرئيسة و الفاعلة في اساليب التحليل الأحصائي لكثير من الباحثين و لكن يصاحب هذا التحليل عدد من المشاكل و منها مشكلة التعدد الخطي ، ويقصد بمشكلة التعدد الخطي هو ارتباط المتغيرات التوضيحية مع بعضها بعضا (كل المتغيرات او البعض منها) بعلاقة خطية يصعب فصلها .

و عادة ما يستعمل انحدار الحرف **Ridge Regression** لمعالجة هذه المشكلة وهو من الطرائق المستعملة لمعالجة مشكلة التعدد الخطي (شبه التام) [11] ، و يشترط لمعالجة تلك المشكلة بطريقة انحدار الحرف الاعتيادية ، خطية البيانات و عند عدم توفر ذلك الشرط يلجأ الباحثون الى معالجتها بأتباع احد اساليب الأنحدار اللامعلمي ، و هو اسلوب التقدير اللبي **Kernel** و الذي يعد اسلوبا لامعلميا تمهيديا لتقدير اية دالة احصائية عندما تكون البيانات لاخطية [2] .

2- هدف البحث :

أن الهدف من البحث هو تحليل و معالجة البيانات اللاخطية **Nonlinear Data** ولاسيما تلك التي تعاني مشكلة التعدد الخطي **Multi-CoLinearity Problem** بأستعمال دالة انحدار الحرف اللبي **Kernal Ridge Regression** ، و التعرف على الطريقة الأفضل لتقدير المعلمة التمهيدية و التوصل الى الحل الأمثل من خلال المقارنة بين طريقتي الأمكان الأعظم للعبور الشرعي **Maximum Likelihood Cross-Validation** و معيار **Akaike Inormation criterion (AIC)** وذلك بأستعمال اسلوب المحاكاة .

3- الجانب النظري :

إن الهدف الرئيس هو تقدير دالة الأنحدار المجهولة **m(.)** و التي تعبر عن التوقع الشرطي لمتغير الأستجابة Y_i بالنسبة الى X_i التي تشير الى مشاهدات العينة المدروسة [2].

$$m(x) = E(Y/X=x)$$

و من ضمن طرائق التمهيد المعروفة لتقدير دالة الأنحدار هي :-

3-1 الأنحدار الخطي المحدد (الموضعي) Local Linear Regression

يدعى احيانا بالانحدار متعدد الحدود الخطي الموضعي و قد اقترح من قبل الباحث **Fan** في عام **1993** و **Fan and Gijbels** [3] عامي **1992** و **1996** و يعد من افضل المقدرات في الانحدار اللامعلمي و ذلك لانه يصحح بعض العيوب في مقدرات **kernel** [3] .

و يعتمد هذا الممهد على افتراض ان المشتقة الثانية لدالة الأنحدار اللامعلمي المجهولة **m(x)** موجودة و على اساس ذلك و لتقدير المعالم **a , b** يتم تصغير المقدار الآتي[2] :-

$$\sum (Y_i - a - b\,(x - X_i))^2\, \mathrm{k}\left(\frac{\mathrm{x}-\mathrm{X_i}}{\mathrm{h}}\right)$$

و على فرض ان حل مسألة المربعات الصغرى الموزونة **(WLS)** يتمثل بالمقدرات $\hat{a}\,,\hat{b}$ و بأجراء بعض الحسابات البسيطة يكتب ممهد **LLS** بالشكل الآتي [2]:

$$\hat{a} = \widehat{m}(\mathrm{x}) = \frac{\sum_{i=1}^{n} w_i\, Y_i}{\sum_{i=1}^{n} w_i}$$

إذ تشير w_i الى دالة الوزن و تحسب بالشكل الآتي [2] :

$$w_i = \mathrm{k}\left(\frac{\mathrm{x}-\mathrm{X_i}}{\mathrm{h}}\right)\,(S_{n,2} - (\mathrm{x}-\mathrm{X_i})S_{n,1}\,)$$

مع الإشارة الى إن [2] :

$$S_{n,l}=\sum_{i=1}^{n} k\left(\frac{x-X_i}{h}\right)\ (x-X_i)^{l} \qquad l=1,2\ .$$

: k(u) تمثل دالة kernel.

h : تمثل معلمة التمهيد (عرض الحزمة) و التي يجب تقديرها بأحدى الطرائق الأحصائية المناسبة .

4- تقدير عرض الحزمة :-

تدعى احيانا بالمعلمة التمهيدية او سعة القيد و يرمز لها بالرمز (h) ، ولها علاقة طردية مع مقدار التحيز و علاقة عكسية مع التباين (اي بزيادة عرض الحزمة يزداد التحيز ويقل التباين و العكس صحيح) . لذلك يتوجب على الباحث اختيار عرض الحزمة بطرائق معينة للوصول الى مرحلة التوازن بين التحيز و التباين .

و من ضمن تلك الطرائق لأختيار عرض الحزمة هي :

1-4 طريقة الإمكان الأعظم للعبور الشرعي Maximum Likelihood Of Cross -Validation

اقترحت هذه الطريقة لأختيار عرض الحزمة h من قبل Habbema , Hermans و van den Broek في عام 1974 و Duin في عام 1976 و تكتب صيغتها العامة كما يأتي [7]،[4] :-

$$\mathbf{MLCV(h)}=\left(n^{-1}\sum_{i=1}^{n}\log\left[\sum_{j\neq i}K\left(\frac{x-Xi}{h}\right)\right]-\log[(n-1)h]\right)$$

ومن الطبيعي ان يتم اختيار عرض الحزمة الأعظم لل MLCV(h) اي أن :-

$$\mathbf{h=argmax}_{h>0}\ (\mathbf{MLCV(h)})$$

و يلاحظ انه يعطي نتائج زائفة عند افتراض ان h=0 كقيمه اولية [4] [7] .

2-4 معيار معلومات Akaike criterion Akaike-Information

و يرمز له بالرمز (AIC) اختصاراً و يعد تقدير غير متحيز تقريبي لمعلومات *Kullback-Leibler* المتوقعة [8] ، كما و تعتبر الأسلوب التقريبي لمعيار *LCV* و تكتب الصيغة العامة لها بالشكل الآتي :-

$$AIC(h) = -\sum_{i=1}^{n}\log\widehat{m}_h(x_i)-\sum_{i=1}^{n} infl\ k\left(\frac{x-X_i}{h}\right)$$

إذ ان :-

$$\widehat{m}_h(x_i)=\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{x-X_i}{h}\right)$$

$$infl\ k\left(\frac{x-X_i}{h}\right)=k(0)/(nh\widehat{m}_h(x_i))$$

إذ إن $infl\ k\left(\frac{x-X_i}{h}\right)$ يمثل مقياس الحساسية .

و بتقليل معيار AIC نجد إن [9] :

$$h = (AIC\ \min_{h>0}(h))$$

و يلاحظ استحالة إيجاد $\widehat{m}_h(x_i)$ إذا تم إفتراض إن *h=0* كقيمة أولية إذ لا يجوز أن يكون المقام مساوياً للصفر و في هذه الحالة فإن معيار *Akaike* يعطي قيماً زائفة .

5- معلمة التنظيم Regularization parameter

تدعى بمعلمة الضبط Tuning Parameter وهي تتحكم بكمية التنظيم و بحجم المعاملات و يرمز لها بالرمز λ .

و من الجدير بالذكر انه عندما تقترب قيمة هذه المعلمة من الصفر $\lambda \to 0$ يتم الحصول على حلول المربعات الصغرى و الخالية من التنظيم [14] و تكتب الصيغة العامة لها بالشكل الآتي [13] :-

$$\lambda_n = 4\sigma R\sqrt{\frac{\log p}{n}}$$

عندما [13] :

$$R = max_j \frac{\|x_j\|}{\sqrt{n}}$$

إذ إن p تمثل عدد المتغيرات التوضيحية X_j .

6 - انحدار الحرف اللبي Kernel Ridge Regression

هو اسلوب فاعل لبناء نماذج الأنحدار اللاخطي و الذي يعمل من خلال انحدار الحرف بطرائق Kernel و الذي يهدف الى تقدير الدالة غير المعلومة .

و يرمز له اختصارا بالرمز KRR و يمكن كتابة الصيغة العامة بالشكل الآتي [10] :

$$\hat{m}_{KRR} := \text{argmin}_{f \in H} \left(\frac{1}{n}\sum_{i=1}^{n} (Y_i - m^*(x_i))^2 + \lambda \|m^*\|_H^2\right)$$

حيث ان λ هي معلمة التنظيم .

و على فرض أن $x_1, x_2, \ldots, x_n$ هي متجهات عشوائية مستقلة تمتلك التوزيع (iid) نفسه حيث ان x_i هو متجه صفي من المتغيرات التوضيحية X_j [6] إي إن :-

$$x_i = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$$

إذ إن :-

i=1,2,….,n
j=1,2,…..p

عندما *p* تمثل عدد المتغيرات التوضيحية .

n *حجم العينة* .

و يشير الرمز $\|m\|$ الى طول المتجه m و الذي يحسب بالشكل الآتي على اساس الضرب الداخلي Inner product المعرف على فضاء هيلبرت *Hilbert Space* :-

$$\| m^* \| = \sqrt{< m^*, m^* >}$$

إذ ان [12] :

$$< m^*, m^* > = m^{*\prime} m^*$$

و من ثم فإن [10] :

$$\|m^*\|_H^2 = < m^*, m^* >$$

حيث ان الدالة m^* تستخرج بالشكل الآتي : -

$$m^*(.) = \sum_{i=1}^{n} \sum \alpha_i \; K(., X_i)$$

$$\alpha = (\alpha_1, \alpha_2, \ldots\ldots, \alpha_n)$$

تمثل متجهات معاملات انحدار الحرف و يمكن كتابة الصيغة العامة لها بالشكل الآتي [6]:

$$= (\mathrm{K} + \lambda I_n)^{-1}\, \mathrm{Y}\alpha^*$$

و يشير الرمز K الى مصفوفة **Kernel** و التي يمكن ان توصف بالشكل الآتي [6] :-

$$K_{ij} = k(x_i, x_j)$$

و يتم تكوينها بأستعمال دوال كيرنال و قد تم استعمال دالة **Gaussian** وكما في الجدول الآتي :

Range	K(u)	Kernel
I (l u l ≤ ∞)	$/2)(2\pi)^{-1/2} \exp(-u^2$	Gaussian

حيث ان :

$$\mathrm{u} = \left(\frac{x - X_i}{h}\right)$$

7- الجانب التجريبي :

تم أستعمال اسلوب المحاكاة لعرض الجانب النظري و تطبيق الطرائق المعروضة فيه و المقارنة بينها بأستعمال معيار الأختبار **Mean Square Error (MSE)** أستناداً الى حجوم عينات مختلفة ***(n=70,150)*** و قيم مختلفة للأنحراف المعياري ***(1،*** $0.5sd =$***)*** فضلا عن ابعاد مختلفة للمتغيرات التوضيحية ***(p=9,12)*** و عرض حزمة افتراضي كقيمة أولية و هو ***(h=0.1)*** و كذلك عندما يكون عدد مرات تكرار التجربة هو ***(r=1500)*** كما و تم احتساب المتغير المعتمد من خلال النموذج الآتي [5] :-

$$Y_i = \exp((x_i x_i')/2\mathrm{m}) + e_i$$

عندما :

$$\mathrm{m{=}mean}\ (x_i x_i')$$

إذ إن

$$e_i \sim N(0, \sigma^2)$$

كما و تم رسم بعض الحالات المفترضة للمتغير المعتمد Y_i مع دالة انحدار الحرف اللبي التقديرية بأستعمال برنامج **Excel** إذ لا مجال لعرض جميع الحالات وكما في الشكل (1)،(2) بينما كانت النتائج كما يأتي :-

جدول رقم (1) يوضح قيمة متوسط مربعات الخطأ **MASE** عندما تكون عدد المتغيرات التوضيحية (p=9) و لكافة الحالات المفترضة

Sd=1	Sd=0.5	Standard devation	sample
Gauss	Gauss	Function Method	
0.0245	0.0064	MLCV	n=70 P=9
0.0238	0.0063	AIC	
0.0161	0.0042	MLCV	n=150 P=9
0.0165	0.0043	AIC	

الجدول رقم (2) يوضح قيمة متوسط مربعات الخطأ MSE عندما تكون عدد المتغيرات التوضيحية (p=12) و لكافة الحالات المفترضة

sample	Standard deviation	Sd=0.5	sd=1
	Function Method	Gauss	Gauss
n=70	MLCV	0.0099	0.0374
p=12	AIC	0.0096	0.0363
n=150	MLCV	0.0059	0.0230
p=12	AIC	0.0060	0.0233

يتضح من الجدول رقم (1) و الجدول رقم (2) ما يأتي :-

- لوحظ أن طريقة (AIC) هي الأفضل عند حجم عينة n=70 وذلك عندما يكون عدد المتغيرات التوضيحية p=9,12 اثناء اختيار دالة Gaussian .
- بينما تتميز طريقة MLCV بالأفضلية عند حجم عينة n=150 وذلك عندما يكون عدد المتغيرات التوضيحية p=9,12 اثناء اختيار دالة Gaussian .
- يتضح أن قيمة MASE يقل بزيادة حجم العينة بينما يزداد بزيادة الإنحراف المعياري و بحسب النظرية الإحصائية.

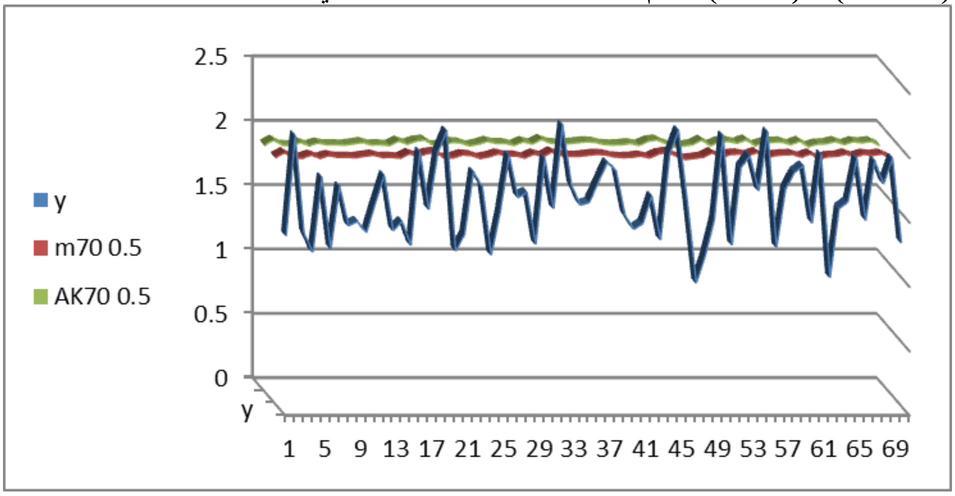
8- الأستنتاجات:-

- نستنتج من خلال نتائج المحاكاة بأن طريقة AIC هي الأفضل من طريقة MLCV عند حجم عينة n=70 بينما تكن طريقة MLCV الأفضل عند n=150 و للحالات المفترضة كافة.
- نستنتج بأن هنالك علاقة عكسية بين معيار الإختبار(Mean Average Square Error (MASE وحجم العينة (n) بينما تكن هناك علاقة طردية بينه و بين الإنحراف المعياري (sd).

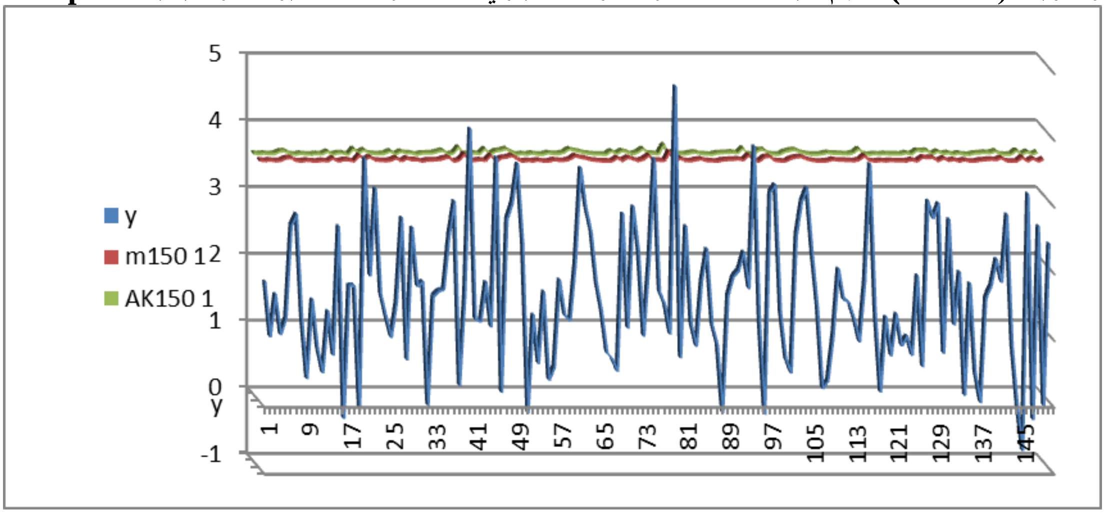
9- التوصيات :-

- نوصي بأستعمال طريقة AIC عند استعمال دالة Gaussian بدلا من طريقة MLCV لحجم عينة اقل بينما نوصي بعكس ذلك لحجم عينة اعلى.
- نوصي بأختيار قيمة اقل للأنحراف المعياري كما نوصي بأختيار حجم عينة اكبر لتقليل قيمة معيار الأختبار MSE .
- نوصي بأستخدام دوال اخرى مثل دالة Double Exponential .

الشكل رقم (1) يوضح المتغير المعتمد y مع دالة انحدار الحرف اللبي التقديرية KRR عند استعمال طريقتي (MLCV) و (AIC) لحجم عينة n= 70 و انحراف معياري sd=0.5 و عدد متغيرات توضيحية p=9



الشكل رقم (2) يوضح المتغير المعتمد Y مع دالة انحدار الحرف اللبي التقديرية عند استعمال طريقة (AIC) و طريقة (MLCV) لحجم عينة n=150 و انحراف معياري sd=1 و عدد متغيرات توضيحية p=12



المصادر:-

1- المصادر العربية:-

1-كاظم ، اموري هادي و الدليمي ، محمد مناجد "مقدمة في تحليل الأنحدار الخطي"1988 الطبعة 2001 ، طبع في مديرية دار الكتب للطباعة و النشر /جامعة الموصل .

2-حمود ، مناف يوسف "مقارنة مقدرات Kernel اللامعلمية لتقدير دوال الأنحدار " المجلة العراقية للعلوم الأحصائية كلية علوم الحاسبات و الرياضيات جامعة الموصل العدد 2 لعام 2001 ص ص 26-44 .

3-حمود ، مناف يوسف و عاشور ، مروان عبد الحميد " مقارنة بضعة مقدرات لاخطية لتقدير دالة الأنحدار " مجلة العلوم الأقتصادية و الأدارية المجلد 18 العدد 68 ص ص 359 – 372 .

4- حمود ، مناف يوسف (2005) " مقارنة المقدرات اللامعلمية لتقدير دوال الكثافة الأحتمالية " اطروحة دكتوراه مقدمة الى كلية الإدارة و الأقتصاد / جامعة بغداد .

2- **المصادر الأجنبية** :-

5- Rosipal, Roman and Trejo , Leonard J. (2001) " Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space" Journal of Machine Learning Research 2 , pp [97-123] .

6- Kasiviswanatham , Shiva & Rudelson , Mark (2015) " spectral norm of random kernel matrices with application to privacy " Algorithms and Techniques - 18th International Workshop, APPROX 2015 , and 19th International Workshop , RANDOM 2015 , vol (40) , No (17) , pp[898-914] .
rudelson@umich.edu.

7- Guidoum, A.(2015). " Kernel Estimator and Bandwidth Selection for Density and its Derivatives ". University of Science and Technology, The kedd Package, Version 1.0.3 .

8- Clifford M. Hurvich; Jeffrey S. Simonoff; Chih-Ling Tsai (1998) " Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion " vol (60) No (2) pp [271-293] .

9- Clive R. Loader (1999) " Bandwidth Selection: Classical Or Plug-In " The Annals of statistics ,No 27, pp [415-438].

10- Zhang, Yuchen & Duchi, John and Wainwright, Martin (2013) "Divide and Conquer Kernel Ridge Regression" University of California , JMLR: Workshop and Conference Proceedings vol(30) pp[1-26] .

11- Arthur E.Horl and Robert W.Kennard (1970) " Ridge Regression : Biased Estimation For Non orthogonal Problems " University of Delaware and E.I.dupont de Nemours &Co. vol(12) , No(1) pp[55-67].

12- Bierens , H.J. (2007) " Lecture : Introduction to Hilbert Spaces [PDF] " Retrieved from Pennsylvania State University , Time series econometrics , Theory and applications Blackboard. http://personal.psu.edu/hxb11/HILBERT.PDF.

13- Tomioka, R. (n.d) " Introduction To The analysis of Learning algorithms : ridge regression and lasso " University of Tokyo .
http://tomioka.dk/teaching/dtuphd13/dtuphd13.pdf

14- Quarter, Autumn (2006) " Lecture: Regularization: Ridge Regression and the Lasso [Pdf] " Lecture Note.
Blackboard:
http://statweb.stanford.edu/~owen/courses/3051314/Rudyregularization.pdf.

Estimate Kernel Ridge Regression Function in Multiple Regression

ABSTRACT :

In general, researchers and statisticians in particular have been usually used non-parametric regression models when the parametric methods failed to fulfillment their aim to analyze the models precisely. In this case the parametic methods are useless so they turn to non-parametric methods for its easiness in programming. Non-parametric methods can also used to assume the parametric regression model for subsequent use. Moreover, as an advantage of using non-parametric methods is to solve the problem of Multi-Colinearity between explanatory variables combined with nonlinear data. This problem can be solved by using kernel ridge regression which depend on what so-called bandwidth estimation (smoothing parameters). Therefore, for this purpose two different methods were used to estimate the smoothing parameter (Maximum Likelihood Cross-Validation (MLCV) and Akaike Information Criterion (AIC)). Furthermore, a comparision between the previouse methods had been provided using simulation technique , and the method of Akaike Information Criterion (AIC) has been found to be the best for the Gaussian function .

Keyword :kernel ridge regression KRR , MLCV, AIC , Regularization parameter λ .