Converges in L_1 of nearest neighbor regression function estimate for strong mixing processes

Dr. Mohamed M. Deribati*
Dr. Ahmad Younso **
Dema Al-shakh ***

(Received 1 / 6 / 2021. Accepted 23 / 12 /2021)

□ ABSTRACT □

In this paper, we will study the issue of estimating regression function using k-nearest neighbors method (knn) for α mixing processes. We will extend the convergence in L_1 for knn regression from independent case to the dependent case; In addition, we will conduct a simulation study using R software program to display the importance and influence of choosing number of neighbors (k) and the sample size (n) on behavior of the estimator. For this purpose, the mean squares error criterion (MSE) was used. The high variability of the MSE results shows that knn estimators are very sensitive to the choice of the number of neighbors. More results show that the higher value of n, the more accurate and effective the estimator.

Key words: Nonparametric regression, mixing concepts, strong mixing, k-nearest neighbor estimator, Cross validation method, convergence.

* **Associate Professor, Depart. Of Mathematical Statistics, Faculty of Science, Tishreen University, Lattakia, Syria.** DRDribatem@gmail.com
** **Associate Professor, Depart. Of Mathematical Statistics, Faculty of Science, Damascus University, Damascus, Syria.** ahyounso@yahoo.fr
*** **Postgraduate student, Depart. Of Mathematical Statistics, Tishreen University, Lattakia, Syria**. dema.srmad@gmail.com

التقارب في L_1 لمقدّر الجوارات الأكثر قرباً لدالّة الانحدار لعمليات مزوجة بقوة

د. محمد مزيد دريباتي*

د. احمد يونسو**

ديمه الشاخ***

(تاريخ الإيداع 1 / 6 / 2021. قُبِل للنشر في 23 / 12 /2021)

□ ملخّص □

قمنا في هذه الورقة بدراسة مسألة تقدّير دالّة الانحدار باستخدام مقدّر الجوارات الـ k الأكثر قرباً اللاوسيطي في حالة الارتباط. تم توسيع نتائج تقارب مقدّر الجوارات لدالّة الانحدار في L_1 من الحالة المستقلّة إلى الحالة المرتبطة بمعامل المزج α. بالإضافة لدراسة محاكاة باستخدام الحزمة الإحصائية **R** لمعرفة مدى أهمية وتأثير كلّ من عدد الجوارات (k) وحجم العيّنة (n) على سلوك هذا المقدّر حيث تمَّ استخدام معيار (MSE) متوسط مربّعات الخطأ (Mean Squares Errors) لهذا الغرض. تشير نتائج دراسة المحاكاة إلى مدى حساسيّة مقدّر الجوارات الـ k الأكثر قرباً لاختيار عدد الجوارات k. كما أظهرت النّتائج أنّه كلّما زاد حجم العينة، كلّما زادت دقة التقدّير وفعاليّته.

الكلمات المفتاحيّة: الانحدار اللّاوسيطي، معاملات المزج، المزج القوّي، مقدّر الجوارات الـ k الأكثر قربًا، طريقة الّتحقق المتبادل، الّتقارب.

* أستاذ مساعد – قسم الإحصاء الرياضي–كلية العلوم– جامعة تشرين–اللاذقية–سورية. DRDribatem@gmail.com

** أستاذ مساعد – قسم الإحصاء الرياضي– كلية العلوم–جامعة دمشق–دمشق– سورية.ahyounso@yahoo.fr

*** طالبة دكتوراه–قسم الإحصاء الرياضي– كلية العلوم–جامعة تشرين–اللاذقية–سورية. dema.srmad@gmail.com

مقدّمة

يعتبر التنبؤ من الموضوعات الإحصائيّة المهمّة، حيث يلعب دورا بارزاً في تحليل الانحدار والسّلاسل الزّمنية ونحتاج الى إجراء التّنبؤات لقيم ظاهرة عشوائية ما استناداً إلى القيم الحالية والماضية على فرض وجود نوع من انواع الارتباط بين قيم الظاهرة في كثير من الظواهر الاقتصادية أو الاجتماعية أو الطبية أو الطبيعية.

يوجد العديد من طرق الانحدار الوسيطيّة (parametric regression) وغير الوسيطيّة (nonparametric regression) التي تستخدم في تحليل الانحدار.

تركيزنا في هذه الورقة على الطريقة اللّاوسيطيّة التي تفترض وجود نوع من انواع الارتباط بذاكرة قصيرة والذي يُسمّى الارتباط بمعامل المزج α.

قبل التطرق الى هذا المفهوم لا بد من التعرف على بعض المفاهيم الأساسية في نظرية الاحتمالات التي تمثل عادة أيّة ظاهرة عشوائية تتأثر قيمها بالزمن بطورية عشوائية. ركزنا على الطّوريّات بمتغيرين (X_t,Y_t) .

نعتبر أنّ (X_i,Y_i) كثنائية قيم مماثلة التوزيع لشعاع عشوائي (X,Y) حيث (X_i,Y_i) نسخة الشعاع في اللحظة i .

نسمي Y المتغير التابع(dependent variable) و X المتغير المستقل (independent variable) حيث أنّ (X,Y) معرّف على الفضاء الاحتماليّ (Ω,F,p) ويأخذه قيمه في $R^d \times R$.

يُعرف نموذج الانحدار لـ Y على X بالشكل:

$$Y = m(X) + \varepsilon \qquad (1)$$

حيث ε يمثّل الخطأ العشوائي (random error) وهو مستقل عن X فرضاً و $m(X)$ دالّة الانحدار (regression function) وتعُطى بالعلاقة:

$$m(x) = E(Y|X = x) \qquad (2)$$

أي القيمة المتوّقعة لـ Y عندما يأخذ X القيمة x.

نرغب في تقدّير دالّة الانحدار $m(x)$ وإجراء التّنبؤات باستخدام النّموذج المقدّر. إذ عادةً ما يتم تقدّير دالّة الانحدار $m(x)$ باستخدام أساليب وسيطيّة وأخرى غير وسيطيّة. في الانحدار الوسيطي تكون $m(x)$ في فضاء منتهٍ درجته محدّدة بعدد الوسطاء الداخلة في النموذج والتي يلزم تقدّيرها. حيث يتم تشخيص نموذج معين لدالّة الانحدار يصف العلاقة بين المتغير التابع والمتغيرات المستقلة إذا كان هذا التشخيص صحيحاً وما تبّقى من افتراضات غاوص ماركوف محققة فإننا سنحصل على مقدّرات غير متحيّزة وذات كفاءة، ولكن خطأ التّشخيص احياناً نتيجة عدم وضوح غيمة الانتشار أو تعقيدها الذي يؤدي إلى نموذج بعيد عن الواقع غير صالح للحصول على تنبؤات موضوعيّة لقيم الظّاهرة. وبالتالي يتم اللجوء إلى طرائق الانحدار اللّاوسيطي التّي تؤمّن بدائل عن الطرائق الوسيطيّة التقليدية، حيث يتم فيها تقدّير دالّة الانحدار ككل أي أنّه لا يعتمد على افتراضات قوّية فيما يتعلق بشكل العلاقة بين المتغير التابع والمتغيرات المستقلة وبدلاً من ذلك فإنّه يسمح للبيانات بالتحّدث عن نفسها في تحديد شكل دالّة الانحدار المقدّرة. من أكثر الطرق اللاوسيطيّة طريقة الجوارات الأكثر قرباً (k-nearest neighbors) وطريقة النّواة (kernel method).

اقتصرنا في دراستنا على مقدر الجوارات الأكثر قرباً في حالة الارتباط بمفهوم المزج لما له من مزايا بين الطرائق اللاوسيطيّة فهو أسهل للفهم والتنفيذ. تعد طريقة الجوارات الأكثر قرباً من أقدم الطرق اللاوسيطيّة في تحليل الانحدار والتصنيف. تمّ اقتراح الفكرة الاساسية لهذه الطريقة من قبل (Fix;Hodges,1951)، ثمّ عممت من قبل

(Royall,1966) تحت فرضية الاستقلال (i.i.d). وهناك العديد من الدراسات حول خصائص طريقة الجوارات الـ k– الأكثر قرباً حيث قام Royall بدراسة لكل من MSE (Mean Squares Error) و MISE (Mean Integrated Squares Error) والتقارب الطبيعي للمقدّر مع دالّة وزن منتظمة. فيما بعد قام (Mack,1981) بتوسيع هذه النتائج لتشمل الحالة التي تكون فيها الأوزان غير منتظمة، حيث درس التحيز والتقارب الطبيعي، أيضاً تحت شرط الاستقلال.

اثبت (Stone,1977) التقارب في L_p من أجل أنواع مختلفة لمقدرات الجوارات الـk– الأكثر قرباً. تمت دراسة التقارب شبه الاكيد للمقدّر في الحالة التي تكون فيها Y محدودة من قبل(Devroye,1981)، درس أيضاً (Devroye,1982) التقارب القوي والتقارب المنتظم لنفس المقدّر. أنواع أخرى للتقارب تم دراستها من قبل (Collomb,1980) كالتقارب بالاحتمال والتقارب شبه الاكيد والتقارب شبه التام.

ايضاً تمّ دراسة معدل التقارب المنتظم القوّي لمقدّر الجوّارات الأكثر قرباً لدالّة الانحدار من قبل (Silverman,1982) و(View ;Kudraszow,2013). وقام (Biau,2010) بدراسة بعض أشكال التقارب تحت (L_p $risk$)، (Deveroy; Gyorfi; Krzyzak;Lugosi,1994) قاموا بدراسة التقارب القوي لمقدّر الجوارات الـk– الأكثر قرباً (under L1 risk) ايضاً تحت فرضية الاستقلال. إنّ رغبتنا في هذه الورقة بتوسيع بعض النتائج التي توصل إليها Deveroy من الحالة المستقلة (Deveroy et al.,1994) إلى الحالة المرتبطة لمقدّر الجوارات الأكثر قرباً لدالّة الانحدار ودراسة التقارب في L_1 لهذا المقدّر لعمليات مرتبطة بمفهوم المزج القوي($\alpha - mixing\ processes$).

أهمية البحث وأهدافه

تأتي أهمية هذا البحث لإمكانية تطبيقه عندما لا يتحقق شرط الاستقلال للبيانات. يهدف هذا البحث إلى تعميم تقارب مقدّر الجوّارات الـ k الأكثر قرباً لدالّة الانحدار من الدالّة الحقيقية في L_1 من الحالة المستقلة إلى الحالة المرتبطة، وتطبيق المقدّر على عينات مزوجة مولدة باستخدام الحزمة الاحصائية **R** لدراسة سلوك هذا المقدّر.

طرائق البحث ومواده

طرحنا في هذه الورقة مقدّر الجوارات لدالّة الانحدار لبيانات مزوجة بقوة. وقمنا بصياغة نموذج محاكاة للتحقق من أداء هذه الطريقة باستخدام مقياس متوسط مربعات الخطأ (**MSE**) ، فضلاً عن مقارنة تقارب هذه الطرائق من المنحني الحقيقي لدالّة الانحدار من خلال الرسوم (**Plots**) التي تعتمد على نتائج مستخلّصة من تجارب المحاكاة.

طريقة الجوّارات الـ k–الأكثر قرباً:

لتكن $\mathcal{D}_n = \{(X_1,Y_1),..,(X_n,Y_n)\}$ عينة حجمها n من مشاهدات للشعاع العشوائي (X,Y). يعطى المقدّر التقليدي لدالّة الانحدار $m(x)$ عندما $X = x$ المعطاة بالعلاقة (2) باستخدام العينة $\mathcal{D}_n$ بالشكل التالي:

$$m_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i \qquad \forall x \in \mathbb{R}^d \qquad (3)$$

حيث $W_{ni}(x)$ تابع احتمال يسمى تابع وزن مترافق مع الجوارات، يعطى في الحالة العامة بالشكل:

$$w(x-X_i)=\frac{K(\frac{x-X_i}{R_n})}{\sum_{j=1}^{n}K(\frac{x-X_j}{R_n})} \qquad (4)$$

حيث R_n يشير إلى المسافة بين x والمجاورة الـ k الأقرب. حيث $K:\mathbb{R}^d\to\mathbb{R}$ تابع كثافة احتمالية محدود.

وقد يعطى تابع الوزن بالشكل:

$$W_{ni}(x)=\begin{cases}\frac{1}{k}, & \|x-X_i\|\le R_n\\ 0, & otherwise\end{cases} \qquad (5)$$

ويسمى الوزن المنتظم لأنه يعطي نفس الوزن لجميع المجاورات الـk– الأقرب. حيث $\sum_{i=1}^{n}W_{ni}(x)=1$

ركزنا في هذه الدراسة على الوزن المعطى بالعلاقة (5) حيث سنقوم بدراسة التقارب في L_1 للمقدار

$$J_n=\int|m_n(x)-m(x)|\mu(dx) \qquad (6)$$

تحت شرط المزج القوي(strong mixing). حيث $m_n(x)$ معرفة كالتالي:

$$m_n(x)=\frac{1}{k}\sum_{i=1}^{k}Y_{(i)}$$

حيث $(X_{(1)},Y_{(1)}),..,(X_{(n)},Y_{(n)})$ هي القيم المرتّبة بحسب تزايد القيم $\|x-X_{(i)}\|$ أي أنّ:

$$\|x-X_{(1)}\|\le\cdots\le\|x-X_{(n)}\|$$

إذا كان X_i و X_j متساويا البعد عن x أي أنّ:

$$\|x-X_i\|=\|x-X_j\|$$

يكون لدينا ربطة، حيث يتم إزالة الرّبطات بمقارنة الأدّلة، أي أنّ X_i أقرب إلى x من X_j إذا كان $i<j$.

فيما يلي سنتعرف على بعض انواع الارتباط بمفهوم المزج.

3.شروط المزج: (Mixing conditions)

قبل التطرق لمفهوم المزج لا بد من التعريف بمفهوم **استقلال المتغيرات العشوائية:**

نقول عن المتتالية من المتغيرات العشوائية $(Z_1,Z_2,..,Z_n)$ والمعرفة على الفضاء الاحتمالي (Ω,F,p) انها مستقلة إذا كانت الجبور التامة $\sigma(Z_i)$ مستقلة، حيث $\sigma(Z_i)$ هو الجبر التام المولد بـ Z_i $(\forall\, i=1,2,..,n)$.

المزج: نقول عن المتتالية $(Z_i,i\ge1)$ أنها ألفا مزوجة $(\alpha-mixing)$ أو (مزوجة بقوة) إذا تحقق :

$$\alpha(n)=\sup_{l\ge1}\ \sup_{A\in\mathcal{F}_1^l,B\in\mathcal{F}_{n+l}^{\infty}}|P(B\cap A)-P(A)P(B)|\to0\ as\ n\to\infty$$

حيث $\mathcal{F}_1^{\infty},\mathcal{F}_{l+n}^{\infty}$ الجبور الجزئية المولدة بـ $(Z_i,i=1,..,l)$ و $(Z_i,i=l+n,\dots)$ على التتالي. يعتبر معامل المزج القوي من أكثر معاملات المزج شيوعاً (Bradely,2005) ، (Rio,2017) مع تطبيقات واسعة في الإحصاء (Bosq,2012).

$$\alpha(n)=O(n^{-\rho})\qquad for\ \rho>0\,. \qquad (7)$$

أحد معاملات المزج الأخرى هو معامل المزج بيتا $\beta-mixing$ حيث نقول عن المتتالية $(Z_i,i\ge1)$ أنها بيتا مزوجة إذا تحقق:

$$\beta(n)=\sup_{l\ge1}E(\sup_{A\in\mathcal{F}_1^l}|P(B\cap A)-P(A|B\in\mathcal{F}_{n+l}^{\infty})|)\to0\ as\ n\to\infty$$

يمكن التحقق من أنّ $2(n) \leq \beta(n)$ والتي تعني أن كلّ متتالية بيتا مزوجة هي متتالية ألفا مزوجة (Rio,2017).
سنفرض

$$\beta(n) = O(n^{-\rho}) \quad for\ \rho > 0\,. \qquad (8)$$

النتائج والمناقشة

قمنا في هذه الفقرة بدراسة التقارب في L_1 للمقدار J_n المعرف بالعلاقة (6)، ثمّ قمنا بدراسة محاكاة باستخدام لغة البرمجة الإحصائية R (Crawley,2013).

في المبرهنة التالية افترضنا أنّ Y محدودة كما في (Deveroy et al.,1994)

مبرهنة:

لتكن $\mathcal{D}_n$ متتالية من المشاهدات العشوائية الـ α –مزوجة للزوج العشوائي (X,Y) وتحقق (7) مع $\rho > 1$.

بفرض أنّ Y محدودة وأنّ الشروط التالية محققة:

$$H1: \quad k/n \to 0\ as\ n \to \infty$$

$$H2: \quad k \to \infty\ as\ n \to \infty$$

$$H3: \quad \frac{k}{\sqrt{n}} \longrightarrow \infty\ as\ n \longrightarrow \infty$$

$$H4: \quad \sum_{t=1}^{\infty} \alpha(t) = \sum_{t=1}^{\infty} t^{-\rho} < \infty\ ;\ \rho > 1$$

فإنّ:

$$E(J_n) \to 0 \quad as\ \ n \to \infty$$

إنّ الشرط $H3$ هو أضعف من الشرط الموجود في (النظرية 11.3)(Bosq; Lecoutre ,1987)

البرهان:

لإثبات المبرهنة سنعرف $\hat{m}\ (x) = \frac{1}{k}\sum_{i=1}^{n} Y_i I_{\{x_i \in S_{(x,r_n)}\}}$ حيث $r_n = r_n(x)$ تحقق

$$\mu\left(S_{(x,r_n)}\right) = {}^{k}/_{n} \qquad (9)$$

حيث $I_{\{.\}}$ تابع الاشارة ويحقق:

$$I_A(w) = \begin{cases} 1, & w \in A \\ 0, & w \notin A \end{cases}$$

و $S_{(x,r_n)}$ هي الكرة المغلقة التي مركزها x ونصف قطرها r_n.

لنكتب الحد $|m_n(x) - m(x)|$ على الشكل التالي:

$$\int_{R^d} |m_n(x) - m(x)|\mu(dx)$$

$$\leq \int_{R^d} |m_n(x) - E\hat{m}_n(x)|\mu(dx)\mu(dx)$$

$$+ \int_{R^d} |E\hat{m}_n(x) - m(x)|\,\mu(dx). \qquad (10)$$

وبالتالي لإثبات أن الحد في الطرف الأيسر يسعى للصفر يكفي أن نثبت أنّ كل حد من الحدود في الطرف الايمن يسعى إلى الصفر.

من (9) لدينا:

$$k/n \to 0 \implies r_n \to 0$$

وبالتالي بحسب مبرهنة كثافة لوبيغ ((Wheeden; Zygmund,1977) يكون لدينا:

$$E\hat{m}_n(x) = \frac{1}{\mu(S_{(x,r_n)})}\int_{S_{(x,rn)}} E(Y|X=\acute{x})\mu(d\acute{x}) \to E(Y|X=x) = m(x)$$

وبما أنّ Y محدودة، يكون لدينا بحسب مبرهنة التقارب المحدود :

$$\int_{R^d} |m(x) - E\hat{m}_n(x)|\mu(dx) \to 0 \quad (11)$$.

وبالتالي بحسب (10) بقي أن نثبت أنّ $\int_{R^d} |m_n(x) - E\hat{m}_n(x)|(dx) \to 0$.

لدينا:

$$E\int_{R^d} |m_n(x) - E\hat{m}_n(x)|\mu(dx)$$

$$\leq E\int_{R^d} |m_n(x) - \hat{m}_n(x)|\mu(dx)$$

$$+ E\int_{R^d} |\hat{m}_n(x) - E\hat{m}_n(x)|\mu(dx) := I + II \quad (12)$$

سنثبت أنّ الحدّين I وII في الطرف الأيمن من العلاقة (12) يسعيان إلى الصفر عندما $n \to 0$.

لدينا أولاً بحسب متراجحة كوشي شوارتز:

$$II = E\int_{R^d} |\hat{m}_n(x) - E\hat{m}_n(x)|\mu(dx)$$

$$\leq \left(\int_{R^d} 1^2\mu(dx)\right)^{\frac{1}{2}}\left(\int_{R^d} (\hat{m}_n(x) - E\hat{m}_n(x))^2\mu(dx)\right)^{\frac{1}{2}}$$

$$= \int_{R^d} \sqrt{E\left(\hat{m}_n(x) - E\hat{m}_n(x)\right)^2}\mu(dx)$$

$$= \int_{R^d} \sqrt{var\left(\hat{m}_n(x)\right)}\mu(dx)$$

$$\leq \int_{R^d} \sqrt{\frac{n}{k^2} var\left(YI_{\{X\in S_{(x,r_n)}\}}\right) + C_n(x)}\mu(dx)$$

حيث $C_n(x) = \frac{1}{k^2}\sum_{i\neq j}\left|cov\left(Y_iI_{\{X_i\in S_{(x,r_n)}\}}, Y_jI_{\{X_j\in S_{(x,r_n)}\}}\right)\right|$

لدينا من جهة: $\frac{n}{k^2} var\left(YI_{\{X\in S_{(x,r_n)}\}}\right) \leq \frac{n}{k^2}M^2EI_{\{X\in S_{(x,r_n)}\}}$

$$\leq \frac{n}{k^2}M^2\mu\left(S_{(x,r_n)}\right) \quad \leq \frac{n}{k^2}M^2\frac{k}{n} = M^2\frac{1}{k}$$

ومن جهة أخرى: $C_n(x) \leq \frac{4}{k^2}\sum_{i\neq j}\left\|Y_iI_{\{X_i\in S_{(x,r_n)}\}}\right\|_\infty \left\|Y_jI_{\{X_j\in S_{(x,r_n)}\}}\right\|_\infty \alpha(|i-j|)$

$$\leq \frac{4M^2}{k^2}\sum_{i\neq j}\alpha(|i-j|)$$

$$\leq \frac{4nM^2}{k^2}\sum_{t=1}^{\infty}\alpha(t)$$

من الفرض لدينا:

$$\sum_{t=1}^{\infty} \alpha(t) = \sum_{t=1}^{\infty} t^{-\rho} \leq \infty \; ; \; \rho > 1$$

وبالتالي: $C_n(x) \leq \frac{cn}{k^2} \sum_{t=1}^{\infty} t^{-\rho} \leq \frac{cn}{k^2}$

$$E \int_{R^d} |\hat{m}_n(x) - E\hat{m}_n(x)| \mu(dx) \longrightarrow 0 \qquad (13)$$

بقي أن نثبت أنّ:

$$I = E \int_{R^d} |m_n(x) - \hat{m}_n(x)| \mu(dx) \longrightarrow 0 \qquad (14)$$

لتكن $X_{(k)}(x)$ الجوار الـ k–الأكثر قرباً لـ x ولنعرف $\rho_n = \|X_{(k)}(x) - x\|$، وبالتالي:

$$|m_n(x) - \hat{m}_n(x)| = \frac{1}{k} \left| \sum_{j=1}^{n} Y_j I_{\{X_j \in S_{(x,r_n)}\}} - \sum_{j=1}^{n} Y_j I_{\{X_j \in S_{(x,r_n)}\}} \right|$$

$$\leq \frac{M}{k} \sum_{j=1}^{n} \left| I_{\{X_j \in S_{(x,r_n)}\}} - I_{\{X_j \in S_{(x,\rho_n)}\}} \right|$$

$$= M \left| \frac{1}{k} \sum_{j=1}^{n} I_{\{X_j \in S_{(x,r_n)}\}} - 1 \right|$$

$$= M |\widetilde{m_n}(x) - E\widetilde{m_n}(x)|$$

حيث أنّ $\widetilde{m_n}(x)$ تعرف على أنها $\hat{m}_n(x)$ باستبدال كل Y بمتغير عشوائي ثابت $Y = 1$. وبالتالي لإثبات أنّ I تسعى للصفر يكفي اثبات أنّ الحد

$$\int_{R^d} |\widetilde{m_n}(x) - E\widetilde{m_n}(x)| \mu(dx) \to 0 \qquad (15)$$

يتم بنفس الطريقة التي تم فيها اثبات (14) وبهذا يكون قد تمّ البرهان.

دراسة محاكاة:simulation study

من أهم المشاكل في طريقة الجوارات هو تعيين k (عدد الجوارات). من أجل تعيين قيمة k المناسبة يجب أن نقوم بسلسلة من الاختبارات مع مقادير مختلفة لـ k لتعيين القيمة المناسبة.

أولا قدمنا دراسة محاكاة توضح مدى تأثير اختيار عدد المجاورات في سلوك مقدّر الجوارات الـ k – الأكثر قرباً. تم تقييم سلوك المقدّر باستخدام معيار متوسط مربعات الخطأ (Mean Square Error) والذي يعطى بالعلاقة التالية:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{m}(X_i))^2 \qquad (16)$$

من البديهي أنه كلما كان هذا المقدار أقرب الى الصفر كلما كانت دقة التقدّير أكبر.

تمّ استخدام نموذج الانحدار المعطى بالصّيغة التالية:

$$Y_i = X_i^3 - 2\sin(X_i^2) + \varepsilon_i \qquad (17)$$

حيث ε_i لها توزيع طبيعي بمتوسط صفر وتباين 0.25 ودالّة تغاير $cov(X_i, Y_j) = (|i-j|)^{-2}$

من الجدير بالذكر أنّه في التوزيع الطبيعي يوجد تكافؤ بين مفهوم الارتباط ومفهوم المزج. كما يمكن تطبيق مقدّر انحدار الجوارات الأكثر قرباً علي أي نموذج انحدار تحقق بياناته شرط المزج α.

لتطبيق مقدّر الجوارات تم توليد عينة $\rho-$ مزوجة والتي بدورها تكون $\alpha-$مزوجة لنموذج الانحدار المعطى بالعلاقة (17) بحجم (100) ومن ثم تطبيق قيم مختلفة لـ k (5, 10, 20, 50) ومقارنة النتائج من خلال قيم MSE لتوضيح أهمية اختيار النافذة ومدى تأثيرها على دّقة المقدّر.

الجدول(1): قيم متوسط مربعات الخطأ المقابل لكل قيمة من قيم k

قيمة k	MSE
5	1.156199
10	3.982364
20	8.821905
30	10.83534
50	15.73393

من الجدول نجد أنّ التباين الكبير في قيم متوسط مربعات الخطأ **MSE** يوضح حساسية طريقة الجوارات لاختيار k (عدد الجوارات) وتدل على أن الطرق التجريبية غير مجدية وأنّ هذه القيم غير مناسبة بما فيه الكفاية، لذلك يجب إيجاد قيمة لـ k تجعل قيمة **MSE** صغيرة.

عادةً في التطبيقات العملية عندما يكون حجم العينة n منتهٍ، يستخدم عدد الجوارات $k = \lfloor \sqrt{n} \rfloor$ ، غالبا ما تكون هذه الطريقة غير مجدية ولا تعطي نتائج مرضية. لذلك ينصح باختيار k باستخدام **طريقة التحقق المتبادل (Cross Validation)**. حيث يتم الحصول على القيمة الأفضل لـ k من خلال تقليل معيار التحقق المتبادل (CV Criterion).

تعتمد طريقة التحقق المتبادل على حذف المشاهدة (j) من مجموعة المشاهدات ثم حساب مقدّر دالّة الانحدار عندها بالاعتماد على الـ($n-1$) مشاهدة الباقية من مشاهدات العينة، ويعطى بالعلاقة التالية:

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - \hat{m}_{-j}(X_j)\right)^2 \qquad (18)$$

حيث:

$$\hat{m}_{-j}(X_j) = \sum_{i \neq j} w_{ni}(X_j)Y_i \qquad (19)$$

حيث $\hat{m}_{-j}(X_j)$ مقدّر الجوارات الاكثر قرباً لدالّة الانحدار عند (X_j, Y_j) بالاعتماد على الـ($n-1$) مشاهدة الباقية من مشاهدات العينة

في المثال التالي تم حساب قيمة k الامثلية من خلال معيار التحقق المتبادل وتمت دراسة تأثير حجم العينة على فعالية المقدّر.

من أجل ذلك قمنا بتوليد عينات عشوائية $\rho-$مزوجة بحجوم مختلفة [n=100,300,500,800,1000,2000] لنموذج المحاكاة المعطى بالعلاقة(17). لزيادة قوة النتائج (robustness of results) تمّ تكرار كل تجربة 100 مرة. تمّ تلخيص النتائج في الجدول التالي:

الجدول(2): قيم متوسط مربعات الخطأ المقابل لكل قيمة من قيمn

حجم العينة n	MSE
100	0.582578
300	0.3443516
500	0.2248572
800	0.1320578
1000	0.07060497
2000	0.0638995

من الجدول السابق نجد أن قيم MSE تنخفض بازدياد حجم العينة وبالتّالي ستزداد دقة المقدّر وفعاليته بازدياد حجم العينة وهذا يتوافق مع النتائج النظرية التي توصلنا لها والتي تنص على أنه كلّما زاد حجم العينة كلّما اقترب الخطأ من الصفر.

تطبيق مقدّر الجوّارات على نماذج الانحدار الذاتي من المرتبة p:

تبنى نماذج الانحدار الذاتي من المرتبة p على أساس أنّ القيم الحالية للسلسلة الزمنية تفسر كدالّة بـ p قيمة ماضية للسلسلة $z_{t-1}, \dots, z_{t-p}$ حيث p يمثل عدد الخطوات من الماضي التي تسمح لنا ببناء نموذج يمكننا من توقع القيم المستقبلية للسلسلة ونرمز لنموذج الانحدار الذاتي بـ z_t ويكتب بالشكل التالي:

$$z_t = \phi_1 z_{t-1} + \dots + \phi_p z_{t-p} + \varepsilon_t \quad , \quad t = 1, \dots, n$$

حيث $\phi_1, \dots, \phi_p$ ثوابت غير معدومة و ε_t سلسلة الضجيج الأبيض (white noise series) وهي سلسلة من المتغيرات العشوائية غير المرتبطة ذات متوسط صفر وتباين منتهٍ.

تنبؤات السلسلة الزمنية باستخدام مقدّر الجوّارات:

لتكن لدينا السلسلة الزمنية المستقرة بالتوزيع الاحتمالي $\{Z_t\}_{t=1}^{n}$ ولنعرف المتجه $X_t = (Z_{t-1}, .., Z_{t-p})$ عندها يمثل نموذج الانحدار للسلسلة بشكل عام بالعلاقة:

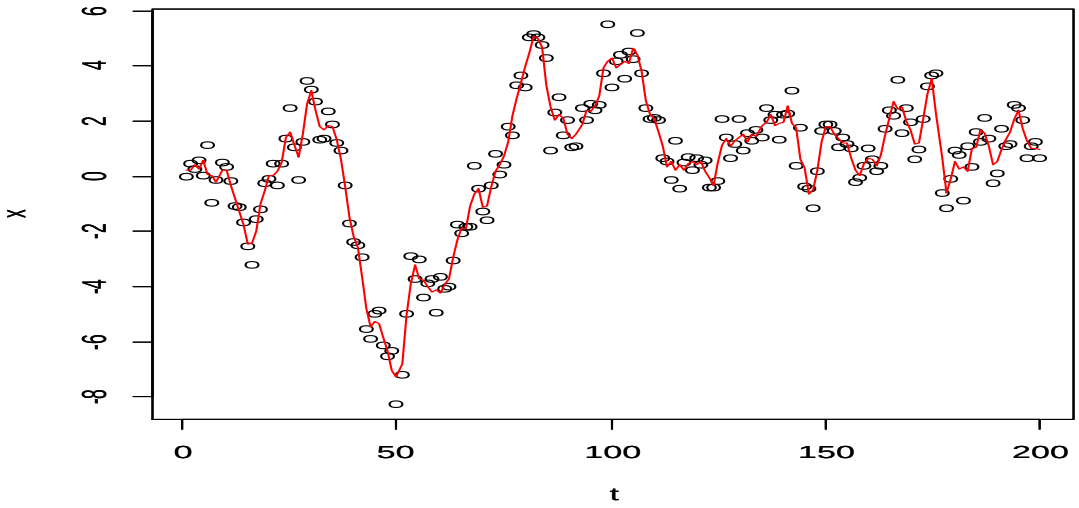
$$Z_t = m(X_t) + \varepsilon_t \quad , \quad t = p+1, p+2, \dots, n$$

من ابسط نماذج الانحدار الذاتي نماذج $AR(1)$

$$Z_t = \phi_1 Z_{t-1} + \varepsilon_t$$

فيما يلي قمنا بتطبيق مقدّر الجوارات لدالّة الانحدار على سلسلة زمنية مزوجة بقوة موّلدة عشوائيا باستخدام لغة البرمجة الإحصائية R وذلك بعد إيجاد قيمة k الأمثلية بطريقة التحقق المتبادل.

تم توليد سلسلة زمنية مزوجة بقوة بحجم n=200 لمشاهدات نموذج انحدار ذاتي $AR(1)$،ثم تم إيجاد قيمة k الأمثلية باستخدام طريقة التحق المتبادل فكانت قيمتها $k = 2$ ومن ثمّ تم تطبيق مقدّر الجوارات لحساب القيم المقدّرة. تم رسم مخطط الانتشار للسلسلة الزمنية و رسم الخط البياني للنتائج الحاصلة.



شكل1: مقدّر الجوارات للسلسة الزمنية مع $k = 2$

يبين الشكل(1) مدى موافقة مقدّر الجوارات لحزمة انتشار السلسلة الزمنية المدروسة وبالتالي يمكن التنبؤ من خلال مقدّر الجوارات للسلسلة الزمنية بقيم مستقبلية صحيحة ودقيقة الى حد كبير.

ملاحظة: يبين نموذج الانحدار الذاتي أن شكل العلاقة خطي، ولكن لا يتضح ذلك في الرسم البياني لأنّ العلاقة خطية بين المشاهدة الحالية t والمشاهدة التي تسبقها $(t-1)$ فقط.

الاستنتاجات والتوصيات

تم تقديم مقدّر الجوارات الـ k الآكثر قرباً لدالّة الانحدار في حالة الارتباط كما تم تعميم تقارب هذا المقدّر في L_1 من الحالة المستقلة الى الحالة المرتبطة بمعامل المزج α. نتائج دراسة المحاكاة التي أجريت لتقييم أداء المقدّر بينت مدى أهمية اختيار k (عدد الجوارات) لمقدّر الجوارات الـ k الآكثر قرباً، كما بينت الدراسة أنه كلما زادت قيمة n كلما زادت دقة وفعالية التقدّير.

في ضوء النتائج سابقة الذكر نوصي بما يلي:

- تطوير النتائج إلى الحالة التي تكون فيها y غير محدودة.
- توسيع النتائج إلى حالة الحقول العشوائية والطوريات الفراغية.
- توسيع النتائج لتشمل انواع اخرى من الارتباط تشمل الارتباط بذاكرة طويلة وماركوف وغيرها.

References

[1] Bosq, D. *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*, Lecture Notes in Statistics, Springer-Verlag, 2nd eds, Berlin(1998)232.
[2] Bosq, D. ; Lecoutre, J. P. *Théorie de l' Estimation Fonctionnelle*. Economica, Paris(1987)
[3] Bosq, D., *Nonpaametric statistics for stochastic processes estimation and prediction*, Springer Science & Business Media,2012.
[3] Bradley, R. C. *Basic properties of strong mixing conditions. A survey and some open questions*, *Probab. Surv.* 2,2005,107–144.
[4] Cohen, J.Y.*Statistics and Data with R: An Applied Approach Through Examples*. A John Wiley & Sons, Ltd.,2008,618.
[5] Crawley J. M. *The R book.* 2nd. ed., John Wiley & Sons, Ltd.,2013, 1060.
[6] Devroye, L.P. *Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. Z.* Wahrscheinlichkeneitstheorie Verw Gebiete 61, ,1982, 467-81.
[7] Devroye, L.P. *The uniform convergence of nearest neighbor regression function estimators and their application in optimization*, IEEE Trans. Inform. Theory, 24 ,1978, 142–151.
[8] Devroye L.P. ; Györfi, L. *Nonparametric Density Estimation*: the L1 View. Wiley: New York,1985,
[9] Devroye, L., *On the almost everywhere convergence of nonparametric regression function estimates*. Annals of Statistics, 9 ,1981a,1310–1319.
[10] Devroye, L., *On the asymptotic probability of error in nonparametric discrimination.* Annals of Statistics, 9 ,1981b,1320–1327.
[11] Devroye, L., Györfi, L., Krzyżak, A., and Lugosi, G. *On the strong universal consistency of nearest neighbor regression function estimates*. Annals of Statistics, 22 ,1994,1371–1385.
[12] Devroye, L.; Ferrario, P., Györfi, L., and Walk, H. *Strong universal consistent estimate of the minimum mean squared error*. In Empirical Inference - Festschrift in Honor of Vladimir N. Vapnik, Springer ,2013, 143–160.
[13] Györfi, L.; Kohler M.; Krzyzak, A.; Walk, H., A *Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, New York,2002,664.
[14] Fix, E. ; Hodges, J.L, *Discriminatory analysis, nonparametric discrimination*, USAF School of Aviation Medicine, Randolph Field, Texas, Project 21-49-004, Report 4, Contract AD41(128)31, 1951..
[15] Hardle, W. *Applied Nonparametric Regression.* Cambridge University Press, Cambridge ,1990,433.
[16] Hardle, W.; Muller, M.; Sperlich, S.;Werwatz,A., *Nonparametic and Semiparametric Models*, ser. Springer Series in Statistics. NewYork: Springer,2004,300.
[17] Kudraszow NL, Vieu P.*Uniform consistency of kNN regressors for functional variables*. Stat Probab Lett 83(8),2013,1863–1870.
[18] Mack, Y.P., *Local properties of k-NN regression estimates*, SIAM Journal on Algebraic and Discrete Methods 2,1981, 311-323.
[19] Mack, Y. P.; Rosenblatt M. *Multivariate k-nearest neighbor density estimates.* Journal of Multivariate Analysis 9 ,1979,1–19.

[20] Rio, E., *Asymptotic theory of weakly depended random process*, . Probability Theory and Stochastic Modelling, vol 80. Springer, Berlin, Heidelberg.,2017,211
[21] Rosenblat, M., *Remarks on some nonparametric estimates of a density function*, Ann. Math. Statist. 21,1956,832-837.
[22] Royall, R. M., *A class of nonparameteric estimates of a smooth regression function*,Ph.D. dissertation, Stanford University, Stanford,1966,68.
[23] Rozanova, Y.; Volkorski, V. *Some limit theorem for random functions* . Theory Probab. Appl. 4,1959,178-197.
[23] Stone, C. J. *Cross-validatory choice and assessment of statistical predictions (with discussion).*Journal of Royal Statistical Society 36 ,1974,111–47.
[24] Stone, C.*Consistent non parametric regression*, Annals of Statistics 5,1977, 595-645.
[25] Wheeden, R.; *Zygmund, A., Measure and Integral*, Deeker, New York,1977,285.