استخدام المحاكاة للمقارنة بين الطريقة التقليدية وطريقة Kernel لتحليل الارتباط القويم

طاهر ريسان دخيل
جامعة القادسية
قسم الإحصاء
tahir_reisan@yahoo.com

صفاء كريم كاظم
جامعة القادسية
قسم الفيزياء
xxe1988@yahoo.com

الخلاصة

في تحليل الارتباط القويم والذي يدرس الارتباطات بين مجموعتين من المتغيرات العشوائية يفترض إن تكون متغيرات المجموعتين هذه ذات تركيبة خطية ولكن عند عدم تحقق هذا الافتراض أي عندما تكون التركيبة لا خطية لذلك يكون استخدام الطريقة التقليدية غير ملائم لهذا التحليل وبالتالي يلجأ إلى استخدام طريقة Kernel, والتي تكون قادرة على التعامل مع هكذا حالة . في هذا البحث تم استخدام عدة دوال من الدوال التابعة لهذه الطريقة وهذه الدوال هي دالة Quartic (أو تسمى أحيانا بدالة Biweight) ودالة Epanechnikov لغرض إجراء المقارنة بينها وبين الطريقة التقليدية في محاولة لمعرفة أفضلية الطرائق وذلك من خلال استخدام المحاكاة .

Abstract

In canonical correlation analysis which studies the correlations between two sets of random variables, assume that these sets have a linear structure, but if this assumption dose not achieve, so the using of classical canonical correlation method isn't suitable thus we can turn into kernel method which is able to deal with this case. In this paper many of kernel functions is used (Quartic (Biweight) and Epanechnikov functions) in purpose of comparing between classical and kernel methods by using simulation.

المقدمة

ان العلاقة التي تربط مجموعتين يمكن دراستها باستخدام تحليل الارتباط القويم هاتين المجموعتين هما مجموعة المتغير المكون من p من الأبعاد هو X والمتغير المتكون من q من الأبعاد هو Y , والهدف هو الحصول عل تركيبات خطيه $a^T X$, $b^T Y$ من المتغيرات ألاصلية تمتلك اعظم تباين , ويمكن التعبير عن ذلك بصيغة رياضية بالشكل التالي:

$$(\alpha,\beta) = MAX\left|corr(a^T X, b^T Y)\right| \quad \ldots\ldots\ldots\ldots\ldots\ldots..(1)$$

حيث يتم الحصول على المتغيرات التالية $Z=a^TX$, $W=b^TY$ والتي تسمى بالمتغيرات القويمة حيث نلاحظ ان α و β متجهات محددة بثابت كما في المعادلة رقم (١) .

ان الارتباط القويم الأول ρ يعرف كقيمه مطلقه لأعلى ارتباط بين مجموعتين من المتغيرات القويمة وكما في العلاقة رقم (1) , حيث ان المتغير القويم من الرتبة k , $1 < k < Min(p,q)$ يجب ان يكون مستقلا عن المتغيرات القويمة الاخرى ذات الرتب الدنيا .

ويمكن وضع Σ والتي تمثل مصفوفة التباين المشترك للمجتمع للمتغير العشوائي U حيث ان $U=(X^T,Y^T)^T$ بالشكل التالي :

$$\Sigma=\begin{bmatrix}\Sigma_{XX} & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_{YY}\end{bmatrix}$$

ان المتجهات α و β تمثل متجهات مميزه مترافقة مع اكبر قيمه مميزه للمصفوفات :

$$..\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\,,\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\,\ldots\ldots\ldots\quad(2)$$

إن اكبر القيم المميزة للمصفوفتين في (٢) يمثل الارتباط القويم الأول , ولتقدير α و β و ρ فانه يجب أولا تقدير مصفوفة التباين المشترك للعينة $\hat{\Sigma}$ وذلك من خلال العينة $U_1,U_2,..,U_n$ حيث إن $U_i=(X_i^T,Y_i^T)^T\in IR^P*IR^q$.
إن حساب المتجهات المميزة والقيم المميزة المحسوبة في المصفوفات في العلاقة (2) يعطي تقديرات غير كفؤة وذلك بسبب ان العلاقة تكون غير خطية وبالتالي فان التركيبات تكون غير خطية مما يجعل الطريقة التقليدية قاصرة عن ايجاد تقديرات ذات الدقة المطلوبة وهذا ما يدعو الى استخدام دوال kernel لحساب الارتباطات القويمة بدلا من استخدام الطريقة التقليدية.

طريقة kernel لتحليل الارتباط القويم:

ان أسلوب تقدير kernel والذي هو احد طرق تمهيد (smoothing) البيانات يمكن ان يعطى بالمعادلة التالية:

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

حيث ان h يسمى بمعلمة الـ bandwith
اما k(u) فتسمى بدالة kernel او منفذ kernel
ان دالة kernel هي دالة وزن والتي تضع اوزان مختلفة في نقاط مختلفة حيث ان اعلى الاوزان تعطى عند النقاط القريبة من x وتقل هذه الاوزان كلما ابتعدت X_i عن x حيث ان البعد والقرب من x يحدد من قبل المعلمة h فاذا كانت كبيرة فاننا نختار قيما كبيرة حول x والعكس بالعكس .
ان دوال kernel يجب ان تحقق الشروط الاتية:

١. يجب ان تكون k(u) مستمرة
٢. ان k(u) تكون متماثلة حول الصفر أي إن k(u)=k(-u)
٣. $\int_{-\infty}^{\infty}k(u)du=1$

٤. $0 < \int u^2 k(u) du < \infty$

إن هناك بعض دوال كيرنل والتي تحقق الخواص أعلاه والتي تم اقتراحها من العديد من الباحثين وسوف نعرض صيغ بعض الدوال والتي استخدمت في هذا البحث وهي:

- دالة Quartic (Biweight) وصيغتها بالشكل التالي:

$$k(u) = \begin{cases} \frac{15}{16}(1-u^2)^2 & \text{if } (|u| \le 1) \\ 0 & \text{otherwise} \end{cases}$$

- دالة Epanechnikov وصيغتها بالشكل التالي:

$$k(u) = \begin{cases} \frac{3}{4}(1-u^2) & \text{if } (|u| \le 1) \\ 0 & \text{otherwise} \end{cases}$$

إن عملية اختيار قيمة h وكذلك اختيار الدالة المناسبة يتم من خلال الحصول على اقل خطأ ممكن , وفي دراستنا هذه يتم اختيار الدالة الأفضل من خلال اختيار الارتباط القويم الأعلى كونه أفضل وأقوى ارتباط والذي يعطي أفضل نتيجة كون إن دراستنا تعتمد على تحليل الارتباط القويم الأول والذي هو بدوره أعلى من جميع الارتباطات القويمة الأخرى .

المحاكاة

لقد تمت المقارنة باستخدام المحاكاة بين مقدرات الارتباط القويم والتي تم ايجادها باستخدام الطريقة التقليدية وطريقة kernel بدالتيها المستخدمة وهي دالة Quartic (Biweight) و دالة Epanechnikov وذلك لغرض إيجاد الطريقة الأفضل ذات المقدرات التي تحمل اعلى ارتباط قويم حيث تم كتابة برنامج باستخدام لغة (V.B) خاص بالتجربة والتي تم تكرارها (1000) مرة حيث يمكن أدراج خطوات , أجراء عملية المحاكاة كالتالي :

١ــنولد قيم متغيرين (x_1,x_2) لتكوين تركيبة خطية هي Za_i حيث ان

$$Za_i = a_1x_1 + a_2x_2$$

واخرى لاخطية هي Zk_i حيث ان

$$Zk_i = a_1k(x_1) + a_2k(x_2)$$

حيث ان K(.) تمثل دالة كيرنل وان

$$x = (e^{\theta})\sin(6\theta) + \varepsilon_1$$

حيث ان θ متغير يتبع التوزيع المنتظم المستمر على الفترة [-π, π] وان ε_1 يمثل حد الخطأ العشوائي والذي يتبع التوزيع الطبيعي القياسي مرة والتوزيع الوغارتمي الطبيعي القياسي مرة وتوزيع كامبل مرة اخرى حيث كانت احجام العينات المستخدمة هي (n=10,20,40,60,80,100).

٢ــيتم توليد متغيرين(y_1,y_2) لتكوين تركيبة خطية هي Wb_i حيث ان

$$Wb_i = b_1y_1 + b_2y_2$$

واخرى لاخطية هي Wk_i حيث ان

$$Wk_i = b_1k(y_1) + b_2k(y_2)$$

حيث ان K(.) تمثل دالة كيرنل وان

$$y = (e^{\theta/4})\cos(4\theta)\sin(\theta) + \varepsilon_2$$

حيث ان θ متغير يتبع التوزيع المنتظم المستمر على الفترة [-π, π] وان ε_2 يمثل حد الخطأ العشوائي والذي يتبع التوزيع الطبيعي القياسي مرة والتوزيع الوغارتمي الطبيعي القياسي مرة وتوزيع كامبل مرة اخرى حيث كانت احجام العينات المستخدمة هي (n=10, 20, 40, 60, 80,100)

٣- يتم حساب الارتباط القويم باستخدام الطريقة التقليدية وطريقة كيرنل بالدالتين الانفتي الذكر وتتم المقارنة بالافضلية على اساس اكبر ارتباط قويم بين Za_i و Wb_i وكذلك بين Zk_i و Wk_i .

تحليل النتائج :

لقد تم اجراء عملية المحاكاة والتي حصلنا من خلالها على النتائج الموضحة فـي الجداول رقم (١) و (٢) و(٣)

١- من خلال الجدول رقم (١) والذي يمثل قيم الارتباطـات القويمـة باسـتخدام الطريقة التقليدية وطريقة كيرنل عندما يتوزع ε_1 و ε_2 توزيعا طبيعيـا قياسـيا نلاحظ ان هناك افضلية لطريقة كيرنل باستخدام دالة Epanechnikov وتـاتي

بعدها طريقة كيرنل باستخدام دالة Quartic ومن ثم الطريقة التقليدية ولكن عند زيادة حجم العينة تصبح الافضلية للطريقة التقليدية.

جدول رقم (١)

و يمثل قيم الارتباطات القويمة عندما يتوزع ε_1 و ε_2 توزيعا طبيعيا قياسيا

n	CCQ	CCE	CCA
١٠	.211	.220	.200
٣٠	.215	.223	.207
٥٠	.184	.185	.١٨٧
٨٠	.201	.201	.264
١٠٠	.٢٠٢	.٢٠٤	.٢٦٥

٢-من خلال الجدول رقم (٢) والذي يمثل قيم الارتباطات القويمة باستخدام الطريقة التقليدية وطريقة كيرنل عندما يتوزع ε_1 و ε_2 توزيعا لوغارتميا طبيعيا قياسيا نلاحظ ايضا ان هناك افضلية لطريقة كيرنل باستخدام دالة Epanechnikov وتاتي بعدها طريقة كيرنل باستخدام دالة Quartic ومن ثم الطريقة التقليدية وعند زيادة حجم العينة تصبح الافضلية للطريقة التقليدية ايضا.

جدول رقم (٢)

و يمثل قيم الارتباطات القويمة عندما يتوزع ε_1 و ε_2 توزيعا لوغارتميا طبيعيا قياسيا

n	CCQ	CCE	CCA
١٠	.215	.226	.201
٣٠	.216	.223	.211
٥٠	.183	.183	.١٨٤
٨٠	.209	.210	.260
١٠٠	.٢٠٧	.٢٠٨	.٢٦٠

٣-من خلال الجدول رقم (٣) والذي يمثل قيم الارتباطات القويمة باستخدام الطريقة التقليدية وطريقة كيرنل عندما يتوزع ε_1 و ε_2 توزيع كامبل نلاحظ ان هناك افضلية لطريقة كيرنل باستخدام دالة Epanechnikov ولجميع احجام العينات المستخدمة وتاتي بعدها دالة Quartic ومن ثم الطريقة التقليدية والتي كانت لها اقل ارتباطات قويمة.

جدول رقم (٣)

و يمثل قيم الارتباطات القويمة عندما يتوزع ε_1 و ε_2 توزيع كامبل

n	CCQ	CCE	CCA
١٠	.228	.226	.213
٣٠	.191	.194	.171
٥٠	.164	.164	.١٤٥
٨٠	.214	.261	.245
١٠٠	.214	.215	.240

حيث تعني CCQ طريقة الارتباط القويم باستخدام دالة Quartic

وتعني CCE طريقة الارتباط القويم باستخدام دالة Epanechnikov
اما CCA فتعني طريقة الارتباط القويم التقليدية

الاستنتاجات

١- نلاحظ من خلال النتائج ان طريقة كيرنل هي افضل من الطريقة التقليدية ولكن الأخيرة تصبح افضل في حالة زيادة حجم العينة وذلك عندما يكون توزيع حد الخطأ العشوائي لمجموعتي المتغيرات المستخدمة توزيعا طبيعيا قياسيا وتوزيعا لوغارتميا طبيعيا قياسيا وكانت ايضا طريقة كيرنل باستخدام دالة Epanechnikov هي الأفضل .

٢- نلاحظ من خلال الجدول الأخير وهو جدول رقم (٣) بان هناك افضلية مطلقة لدالة كيرنل باستخدام دالة Epanechnikov ومن ثم تاتي دالة كيرنل باستخدام دالة Quartic وكانت الطريقة التقليدية غير فعالة عندما يكون توزيع حد الخطأ العشوائي لمجموعتي المتغيرات المستخدمة توزيع كامبل.

المصادر

١- كاظم,علي جواد (٢٠٠٦)"تحليل الارتباط القويم اللاخطي باستخدام بعض دوال كيرنل" مجلة القادسية للعلوم الادارية والاقتصادية , المجلد (٨) , العدد (٢).
2- Akaho,S. (2001) "A kernel method for canonical correlation analysis",international meeting of psychometric socity ,Osaka.
3- Donald F. (1988)" Multivariate statistical methods" second edition , McGraw Hill series in probability and statistics.
4- Florian ,M (2003) " canonical correlation analysis with kernels" ,computational diagnostics group seminar .Berlin.
5- Lai, P. and C. Fyfe (2000) " kernel and nonlinear canonical correlation analysis " , International Journal of neural systems 10(5),365-377.
6- Rmanazzi,M (1992) " Influence in canonical correlation analysis ", psychometrika , 57,237-259.