في تحليل المركبات الرئيسة اللبية

الباحث/ حيدر يحيى محمد / كلية الادارة والاقتصاد / قسم الاحصاء
أ.د. لقاء علي محمد / كلية الادارة والاقتصاد /جامعة بغداد / قسم الاحصاء

P:ISSN 1813 - 6729
E:ISSN 2707 - 1359
http://doi.org/10.31272/JAE.43.2020.123.21

تاريخ استلام البحث 2019/10/22 **مقبول للنشر بتاريخ 2019/11/19**

المستخلص

عند التعامل مع البيانات متعدد المتغيرات ذات الابعاد العالية غالبا نستخدم طريقة المركبات الرئيسية (principal component analysis(pca)) لتقليص الابعاد ، ولكن في حالة البيانات غير الخطية يصبح من غير الممكن التعامل بالمقدرات التقليدية بسبب الحصول على نتائج مضللة لذلك يتم اللجوء الى الاساليب اللبية ، لمعالجة مشكلة البيانات اللاخطية بإستخدام الدوال اللبية لمعرفة المتغيرات الاكثر تأثيرا على الظاهرة المدروسة ، نسب التباين المفسر عند إستعمال الدوال اللبية (Gaussian) و (Laplacian) ترتفع بزيادة حجم العينة وزيادة عدد المتغيرات ولجميع تجارب المحاكاة المستعملة.

الكلمات المفتاحية : تحليل المركبات الرئيسة ، بيانات عالية الابعاد ، تحليل المركبات الرئيسة اللبية.

مجلة الادارة والاقتصاد
العدد 123 / **اذار** / 2020
الصفحات 394 -376

* **بحث مستل من اطروحة دكتوراه**

1- المقدمة Introduction

طريقة المركبات الرئيسة تستعمل لاكتشاف وتفسير الاعتمادية الظاهرة بين المتغيرات وفحص العلاقة التي قد تكون موجودة بين المشاهدات. [9;pp.445]

الهدف الرئيسي لطريقة المركبات الرئيسة هو تصغير ابعاد البيانات المدخلة التي تتألف من مجموعة كبيرة من المتغيرات المرتبطة ، ويتم ذلك من خلال التحويل الى مجموعة جديدة من المتغيرات تسمى المركبات وتكون غير مرتبطة مع بعضها.[5;pp1,4;pp215]

إن وجود عدد كبير من المتغيرات الداخلة أو المؤثرة في تكوين أية ظاهرة يجعل من الصعب تفسير هذه المعاملات بسهولة وكفاءة لسببين أولهما كثرة هذه المعاملات وثانيهما إن هذه المعاملات تقيس درجة العلاقة ونوعها بين متغيرين فقط وهذا بالنتيجة يؤدي إلى خلق العلاقات المتداخلة مع المتغيرات الأخرى ، ومن أبرز طرق التحليل للتعامل مع هكذا نوع من البيانات هو التحليل العاملي Factor analysis.[1;pp.18]

2- هدف البحث

دراسة وتحليل البيانات اللاخطية بأستخدام اساليب لامعلمية منها تحليل المركبات الرئيسة اللبية (Kernel Principal Component Analysis)(KPCA) والتوصل الى افضل النتائج من خلال المقارنة بين دوال اللب بأجراء تجارب محاكات.

3- الجانب النظري

1-3 تحليل المكونات الرئيسة Principle Component Analysis

إن اول من اقترح فكرة المكونات الرئيسة هو(Karl person عام1901)، وذلك حين استخدمها وسيله للوصول إلى ما سماه حينها بالمربعات الصغرى المتعامدة (orthogonal least squares). وبعد ذلك بسنوات وابتداءاً من عام 1933 قام (Haroid Hoteling) بتطوير هذه الطريقة تطويراً لافتاً للنظر ليكون اساس عمل هذه الطريقة كما هو بين أيدينا.

إن طريقة المركبات الرئيسة هي طريقة استكشافية يمكن الإفادة منها للتوصل الى تفسير او فهم العلاقات المتداخلة بين المتغيرات ، وهي تعالج مجموعة المتغيرات المرتبطة بتحويله الى متغيرات غير مترابطة فيما بينها (متعامدة orthogonal) والمتغيرات الاخيرة تدعى بالمركبات او المكونات الرئيسة ، ويكون عددها بعدد المتغيرات المدروسة ، وإن كل مركبة رئيسة هي عبارة عن تداخل خطي للمتغيرات المدروسة يكون تباينها بمثابة مؤشر لتفسير جزء من التباين الكلي، لذلك فان الباحث حين يرغب في تقليل أبعاد المشكلة، كتقليل عدد المتغيرات المدروسة من دون فقدان لذلك فان الباحث يرغب في تقليل ابعاد المشكلة كتقليل عدد المتغيرات المدروسة من دون فقدان كمية كبيرة من المعلومات يقوم باختيار المكونات الرئيسة الأولية ، يستطيع الباحث تحليل عدد قليل من المكونات الرئيسة المستقلة بدلاً من تحليل عدد كبير من المتغيرات الأصلية المرتبطة فيما بينها بعلاقات معقدة.[1;PP.26-27]

2-3 نموذج المكونات الرئيسة Principal component model

يهدف الى حساب العلاقة الرياضية الخطية بين جميع المتغيرات المدروسة $x_{j\ (j=1,2,\ldots,p)}$ تمثل مجموع المتغيرات المدروسة بعد ضربها بالمعاملات a_{ij} ويمكن تمثيلها كالاتي :

$$Y_i = a_{1i}x_1 + \cdots + a_{pi}x_p$$

$$Y_i = \sum_{j=1}^{p} a_{ij}x_j \qquad i=1,\ldots,p$$

إذ ان :

Y_i : المكون الرئيس i.

a_{ij} : معامل المتغير j في المكون i وتمثل قيمة المتجهات المميزة (a_i) (Characteristic Vectors) المرافق للجذور المميزة (Characteristic Roots) للمصفوفة المستخدمة.

وحيث إن العلاقة السابقة يمكن ايجادها عن طريق مصفوفة التباين والتباين المشترك في حالة كون المتغيرات المدروسة لها وحدات القياس نفسها ، اما اذا كانت وحدات القياس مختلفة فتحول المتغيرات الاصلية الى متغيرات جديدة معيارية (اي تتوزع توزيعاً طبيعياً قياسياً بوسط حسابي مقداره صفر وتباين مقداره واحد) عن طريق مصفوفة الارتباط .[1;pp.27]

المركبة الرئيسة الاولى First principal component هي تركيب خطي للمشاهدات x ومعادلتها هي :

$$Y_1 = a_{11}x_1 + \cdots + a_{p1}x_p \qquad (1)$$
$$Y_1 = a_1'x$$

Y_1 : تمثل المركبة الرئيسة الاولى.

a_1 : تمثل المتجه المميز الاول Eigen vector المرافقه للجذور المميز الاول Eigen value.

ويكون تباين المركبة الاولى هو :

$$S_{Y1}^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{i1}a_{j1}S_{ij} \qquad (2)$$
$$= a_1'Sa_1 = l_1$$

حيث l_1 هو الجذر المميز الاول First Eigen value.

S : مصفوفة التباين والتباين المشترك للمتغيرات x_j.

ويمثل (S_{Y1}^2) التباين الاكبر مقارنة بالتباينات الاخرى ولتحقيق ذلك نستعين بمضروب لاكرانج بالقيد :

$$a_1'a_1 = 1$$

ونشتق بالنسبة لـ a_1'

$$\frac{\partial}{\partial a_1}\left(S_{Y1}^2 + l_1(1 - a_1'a_1)\right) =$$
$$2(S - l_1I)a_1 \qquad (3)$$

وبمساوات المعادلة (3) بالصفر :

$$(S - l_1I)a_1 = 0 \qquad (4)$$

نحصل على المتجه المميز المرافق لأكبر جذر مميزة للمعادلة:

$$(S - l_1I) = 0 \qquad (5)$$

وبضرب المعادلة (4) بـ a_1' نحصل على :

$$l_1 = a_1'Sa_1 = S_{Y1}^2 \qquad (6)$$

إذ أن l_1 تمثل الجذر المميز الاكبر ويمتلك اكبر جزء من التباين المفسر الكلي.

المركبة الرئيسة الثانية معادلتها هي :

$$Y_2 = a_{12}x_1 + \cdots + a_{p2}x_p \qquad (7)$$

وباستعمال مضروب لاكرانج بالقيدين :

$$a_2'a_2 = 1 \qquad a_1'a_2 = 0$$

$$\frac{\partial}{\partial a_2'}(a_2' S a_2 + l_2(1 - a_2' a_2) + M a_1' a_2) = 2(S - l_2 I)a_2 + M a_1' \qquad (8)$$

وبنفس اسلوب حل المعادلة (3) نحصل على :

$$(S - l_2 I)a_2 = 0 \qquad (9)$$

والذي يمثل المتجه المميز الثاني المرافق لثاني اكبر جذر مميز.

والان يمكن اعطاء الصيغة العامة لإيجاد باقي الجذور المميزة والمتجهات المرافقة لها.

$$(S - l_j I)a_j = 0 \qquad \text{ولكل } a_j \qquad (10)$$

وأن $l_1 > l_2 \geq \cdots \geq l_j > 0$ وكذلك إن :

$$\sum_{j=1}^{p} l_j = \sum_{j=1}^{p} S_{Yj}^2 = tr(s) \qquad (11)$$

3-3 طريقة اختيار المكونات الرئيسة

Method of selection principal component

يتم اتخاذ قرار بعدد المكونات الرئيسة المؤثرة لتلخيص البيانات عملياً وسوف نذكر بعض هذه الطرق :

1. ان عدد المركبات الرئيسة المختارة يكون بعدد الجذور المميزة الاكبر من واحد $(l>1)$.
2. نحتفظ بالمكونات التي تفسر (80%)من التباين الكلي.
3. أشار Morrison (1976) إلى أن تفسير 75% من التباين الكلي يكون كافياً ، وعموماً كلما كانت نسبة التباين المفسر عالية وعدد المكونات المختارة قليلة كان ذلك أفضل من ناحية سهولة مناقشة النتائج وتفسيرها.[1;pp.31]

4-3 تحليل المركبات الرئيسة اللبية

Kernel Principal component Analysis

في البداية سنشرح مفهوم تمهيد كيرنل ومصفوفة اللب :

1. تمهيد كير نل kernel smooth

تقدير كثافة اللب هو احد الاساليب المهمة في تمهيد كيرنل خاصة عندما تكون البيانات ذات تشتت عالي او تكون البيانات لاخطية يكون التقدير غير مستقر وبالتالي نلجأ اليها.

مقدر الكثافة اللبي في الحالة البسيط للعينة العشوائية $x_1, x_2, \ldots, x_n$ هو :

$$\hat{f}_h(x) = \frac{-1}{n} \sum_{i=1}^{n} K_h(x - X_i)$$

حيث أن

h : تسمى معلمة التمهيد Bandwidth (عرض الحزمة او النافذة)

K : تمثل دالة اللب kernel function

وتمتلك عادة دالة كيرنل الافتراضات التالية :

1. دالة كيرنل تكون متماثلة $K(u) = K(-u)$ ، $K(u) \geq 0$
2. هي دالة كثافة احتمالية اي أن تكامل الدالة يساوي واحد $\int_{-\infty}^{\infty} K_h(u)\, du = 1$
3. $\int_{-\infty}^{\infty} u^j K_h(u)\, du = 0 \ , for\ j = 1, \ldots, k-1$

4. $\int_{-\infty}^{\infty}|u^j\ K_h(u)|\ du \neq 0\ ,for\ j = k$

ونستدل من الافتراضات ان دالة كيرنل تكون من الرتبة الثانية من خلال قابلية التكامل او الاشتقاق. [10;pp.2,3;pp.2]

وتم استعمال دالتي اللب:

Gaussian kernel $K(X,Y) = \exp(-(X-Y)^2/\sigma)$

Laplacian Kernel $K(X,Y) = \exp(-(X-Y)/\sigma)$

اما معلمة التمهيد Bandwidth (عرض الحزمة) فهي معلمة حرة ونرمز لها h حيث ان اهمية اختيارها اهم من عملية اختيار دالة كيرنل الملائمة K ، والقيم الصغيرة لـ h تؤدي الى زيادة في التباين وتقليل التحيز اما في حالة قيمتها كبيرة تؤدي الى قلة التباين وزيادة في التحيز، فيجب اختيار قيمة تحقق التوازن بين التحيز والتباين ، وتبين مدى اقتراب المنحنى الحقيقي من المنحنى الجديد المستخرج بالاعتماد على قيمة معلمة التمهيد.[10;pp.3]

2. قاعدة القياس الطبيعي (Normal scale rules) :

طريقة قاعدة القياس الطبيعي من اكثر الطرق استعمالاً في تقدير قيمة معلمة التمهيد (h) كما انها من الطرائق البسيطة و السريعة لإختيار معلمة التمهيد اذ ان هذه الطريقة تعتمد في حسابها على مجموع مربعات الخطاء التكاملي التقريبي Asymptotic Mean Integrated Squared Error (AMISE) عند استعمال دالة (Normal).

$$\mathrm{AMISE(h)} = (\mathrm{nh})^{-1}\mathrm{R(K)} + \mathrm{h}^4\left(\frac{\mathrm{M_2(K)}}{2}\right)^2 \mathrm{R(f^2)} = \frac{\mathrm{R(K)}}{\mathrm{nh}} + \frac{\mathrm{h}^4\mathrm{d_v^2}}{4}\mathrm{R(f^{"})}$$

علما إن $\mathrm{R(K)}$، $\mathrm{M_2(K)}$ هي ثوابت kernel اذ ان :

$$\mathrm{M_2(K)} = \int_{-\infty}^{\infty} x^2\, g(x)dx$$

$$R(K) = \int_{-\infty}^{\infty} K^2(z)\ dz$$

وإن v تمثل درجة المشتقة لـ kernel أو تسمى درجة kernel.

وان المعلمة التمهيدية المثلى التي تعمل على تقليل $AMISE\,(\hat{f}(x))$ تكون كما في المعادلة الاتية :

$$h_{AMISE} = \left(\frac{R(K)}{\mathrm{M_2(K)}\ R(f^{"})}\right)^{\frac{1}{5}} n^{-\frac{1}{5}}$$

وبما إن الدالة المفروضة هي دالة (Normal) فأن:

$$\boldsymbol{R(f^{"}) = \int_{-\infty}^{\infty} (f^{"}(x))^2\, dx = \frac{3\sigma^{-5}}{8\pi^{1/2}}}$$

وبتعويض المعادلة الاخيرة في المعادلة السابقة لها نحصل على :

$$h_{AMISE} = \left(\frac{8\pi^{1/2}R(K)}{3\mathrm{M_2(K)}\ n}\right)^{\frac{1}{5}} \hat{\sigma}$$

وبتعويض كل من $R(K)$، $M_v^2(K)$ نحصل على صيغة معلمة التمهيد وفق قاعدة القياس الطبيعي. [11;pp.60-61]

3. مصفوفة اللب (Kernel Matrix)

لتعريف مصفوفة اللب K (Kernel Matrix) ذات قياس $m \times m$ $[K]_{ij}$

$$[K_{ij}] = \langle \phi(x_i) ، \phi(x_j) \rangle$$

$$K = [K_{ij}] = \begin{bmatrix} K(x_1 ، x_1) & K(x_1 ، x_2) & \cdots & K(x_1 ، x_m) \\ K(x_2 ، x_1) & K(x_2 ، x_2) & \cdots & K(x_2 ، x_m) \\ \vdots & & \ddots & \vdots \\ K(x_m ، x_1) & K(x_m ، x_2) & \cdots & K(x_m ، x_m) \end{bmatrix}$$

$K(x_i ، x_j)$: تمثل دالة اللب المستعملة

وتحتوي المصفوفة اعلاه المعلومات لحساب ازواج المسافات للبيانات.[2;pp.226]

وسنشرح الان بشكل مفصل تحليل المركبات الرئيسة اللبية KPCA من خلال المعطيات الوارده اعلاه حيث يتم استعمال هذا الاسلوب عندما تكون المشاهدات لا معلمية حيث يتم استعمال الدوال اللبية التي تحول البيانات المدخلة x من فضائها الاصلي R^d الى فضاء عالي الابعاد يرمز له F بواسطة تحويل لامعلمي $\emptyset(x) \leftarrow x$ حيث ان $\emptyset$ دالة لامعلمية.

إن صيغة مصفوفة التباين المشترك لمصفوفة المدخلات x هي :

$$C = \frac{1}{m} \sum_{j=1}^{m} x_j \, x_j^T \qquad (12)$$

نحصل منها على مصفوفة ذات قياس $m \times m$ وبالتالي يتم ايجاد الجذور المميزة من المعادلة $V=CVl$ نتائج الجذور المميزة $l \geq 0$ ، وكما شرحنا سابقا التحويل اللامعلمي $\emptyset$ يحول البيانات المدخلة الى فضاء مميز عالي الابعاد F ، ثم طريقة المكونات الرئيسة يتم تطبقها على الفضاء المميز.

وصيغة مصفوفة التباين المشترك في الفضاء المعدل $\emptyset$ هي:

$$C^F = \frac{1}{m} \sum_{j=1}^{m} \emptyset(xj) \, \emptyset(xj)^T \qquad (13)$$

لجعل مصفوفة التباين المشترك قطرية يجب حل مشكلة الجذور المميزة في الفضاء المستقبلي

$$lV = C^F V \qquad (14)$$

حيث ان $l \geq 0$ والمتجه المميز V المرافق لأكبر جذر مميز l يمثل المركبة الاولى ، والمتجه المميز المرافق لأصغر جذر مميز يمثل المركبة الاخيرة ، $C^F V$ يمكن التعبير عنها كالاتي:

$$C^F V = \left(\frac{1}{m} \sum_{j=1}^{m} \emptyset(xj) \, \emptyset(xj)^T\right) V \qquad (15)$$

$$\lambda < \emptyset(Xk, V) >= (\emptyset(Xk) C^F V) \qquad (16)$$

ثم نوجد معامل الجذور المميزة α_i كالاتي :

$$V = \sum_{i=1}^{m} \alpha_i \emptyset(xj) \qquad (17)$$

بتعويض المعادلة (16) في (17) نحصل على:

$$\sum_{j=1}^{m} \alpha_i < \emptyset(Xk), \emptyset(xi) > \; =$$

$$\frac{1}{m}\sum_{i=1}^{m} \alpha_i < \emptyset(Xk), \sum_{j=1}^{m} \emptyset(xj) >< \emptyset(xj), \emptyset(xj) > \quad (18)$$

ثم يمكن كتابة الطرف الايسر من المعادلة (18) كالاتي :

$$l\sum_{j=1}^{m} \alpha_i < \emptyset(Xk), \emptyset(xi) > \; = l\sum_{j=1}^{m} \alpha_i K_{ki} \qquad (19)$$

بما ان k=1,…,m المعادلة رقم (19) تصبح كالاتي:

$$\lambda\sum_{j=1}^{m} \alpha_i K_{ki} = \propto lk$$

ويمكن كتابة الجانب الايمن من المعادلة (18) كالاتي :

$$\frac{1}{m}\sum_{i=1}^{m} \alpha_i < \emptyset(Xk), \sum_{j=1}^{m} \emptyset(xj) >< \emptyset(xj), \emptyset(xj) >= \frac{1}{m}\sum_{i=1}^{m} \alpha_i \frac{1}{m}\sum_{j=1}^{m} K_{kj}K_{ji} \qquad (20)$$

وبما ان k=1,…,m المعادلة رقم (21) تصبح

$$\frac{1}{m}\sum_{i=1}^{m} \alpha_i \frac{1}{m}\sum_{j=1}^{m} K_{kj}K_{ji} = \frac{1}{m}K^2\alpha$$

بجمع المعادلتين 19 و 20 نحصل على :

$$l\mathrm{m}K\alpha = K^2\alpha \qquad (21)$$

حيث ان $\alpha = [\alpha_1, \dots, \alpha_m]$

لإيجاد حل للمعادلة (21) نحل مشكلة الجذور المميزة

$$lm\alpha = K\alpha \qquad (22)$$

بالإضافة لذلك يمكن جعل مصفوفة kernel matrix (K) مركزية centralized لكي نحصل على مصفوفة ثابته invariant كالاتي :

$$K_{ctr} = K - 1_m K - K1_m + 1_m K 1_m$$

حيث إن 1_mتمثل مصفوفة كل عنصر من عناصرها يساوي $1/m$.

وللحصول على معاملات الجذور المميزة$(\alpha_1, \dots, \alpha_m)$ normalize نجعل المتجهات المميزة v normalize في الفضاء المستقبلي F وكالاتي:

$$(V^k . V^k) = 1$$

$$\sum_{i,j=1}^{m} \propto_i^k \propto_j^k < \emptyset(xi) . \emptyset(xj) > = 1$$

$$\sum_{i,j=1}^{m} \propto_i^k \propto_j^k K_{ij} = 1$$
$$(\alpha^k . K\alpha^k) = 1$$
$$l(\alpha^k . \alpha^k) = 1$$

وبالتالي نحصل على المركبات الرئيسية بالفضاء المستقبلي كالاتي :

$$(kPC)_x(n) = \left(V^n \emptyset(X)\right) = \sum_{i=1}^{m} \alpha_i^n \, k(x_i, x) \quad (23)$$

[8;pp.2-5,7;pp.66-67,2;pp.24-25]

4-الجانب التجريبي

1-4 المقدمة

يتضمن الجانب التجريبي مراحل مختلفة وهي مرحلة توليد البيانات، مرحلة إختيار معلمات دوال الرابطة، ومرحلة تقدير القيم المميزة (eigenvalues) ومتجهات التحميلات (Loadings) ومرحلة المقارنة، وذلك بإستعمال برنامج مكتوب بلغة (R3.5.1) ، وفيما ياتي وصفا للمراحل والخطوات التي تم إتباعها في الجانب التجريبي:

2-4 مراحل المحاكاة

المرحلة الاولى: توليد المتغيرات

المرحلة الأولى هي مرحلة توليد البيانات وإختيار القيم الإفتراضية لموجه المتوسطات $\underline{\mu}$ ومصفوفات التباين والتباين المشترك $\sum$ للمتغيرات التي تتبع التوزيع الطبيعي متعدد المتغيرات ، وذلك بالإعتماد على طريقة (Box-Muller) والتي تتلخص بالخطوات التالية :

1- توليد المتغيرين العشوائيين U_1, U_2 اللذان يتبعان التوزيع المنتظم $U(0,1)$.

2- إيجاد المتغير Z الذي يتبع التوزيع الطبيعي القياسي $N(0,1)$ من خلال التحويل التالي :

$$Z = \sqrt{-2\, Ln(U_1)} \, . Cos(2\pi\, U_2)$$

3- إيجاد المتغير (X) الذي يتبع التوزيع الطبيعي بمتوسط (μ) وتباين (σ^2) من خلال الصيغة التالية :

$$X_i = \mu + Z_i\, \sigma \qquad , \quad i = 1,2,..,5$$

4- يتم إيجاد المشاهدات للمتغيرات(X_i) سيكون كما يأتي:

$$X_i = (0.75)\ (N(0\ ,1))\text{^}(2) + (0.25)\exp(\ N(-2\ ,4))$$

اي *0.75%* من المتغير تتوزع $(N(0\ ,1))\text{^}(2)$ و *0.25%* تتوزع $\exp(\ N(-2\ ,4))$ ولجميع احجام العينات بنفس النسب.[6;pp112]

5- تم توليد بيانات الدراسة لتشمل $(p = 10, 15, 20, 25)$ من المتغيرات وبأحجام عينات $(n = 15, 22, 30, 50)$ لكل متغير، عليه فإن التجارب التي تم تصميمها وكما ياتي:

- التجربة الاولى $(n = 15, p = 10)$.
- التجربة الثانية $(n = 22\ , p = 15)$.
- التجربة الثالثة $(n = 30\ , p = 20)$.
- التجربة الرابعة $(n = 50, p = 25)$.

المرحلة الثانية: التقدير

في هذه المرحلة تم تقدير الجذور المميزة، الأهمية النسبية، الأهمية النسبية التراكمية، والتحميلات وذلك من خلال البيانات التي تم توليدها في المرحلة الاولى ولكافة التجارب وحسب المعطيات والخطوات والصيغ الخاصة بكافة طرائق التقدير التي تم التطرق لها في الجانب النظري، وكما ياتي :

1- طريقة المركبات الرئيسة الإعتيادية (PCA) وذلك بالإعتماد على مصفوفات الإرتباط للبيانات التي توليدها في المرحلة الإولى.

2- طريقة المركبات الرئيسة اللبية (KPCA) وذلك بالإعتماد على مصفوفات الإرتباط للبيانات تم توليدها في المرحلة الإولى ومن خلال مصفوفات اللب شبه الموجبة المتماثلة وإن الدالة اللبية المستعملة للحصول على مصفوفات اللب هي دالتي كاوسين ولابلاس بمعلمات تمهيد محسوبة بقاعدة التوزيع الطبيعي (قاعدة الإبهام).

يتم تكرار كل تجربة للمراحل والخطوات السابقة (r=2000) ، وعليه فإن المقارنة تتم بين معدلات تلك التقديرات أي معدلات القيم المميزة، الأهمية النسبية للتباين، الأهمية النسبية التراكمية، التحميلات.

نتائج المركبات الرئيسة

نتائج التجارب التي تم تصميمها بإستعمال المركبات الرئيسة الاعتيادية PCA واللبية KPCA مع معايير المفاضلة لكل حالة وكمايلي :

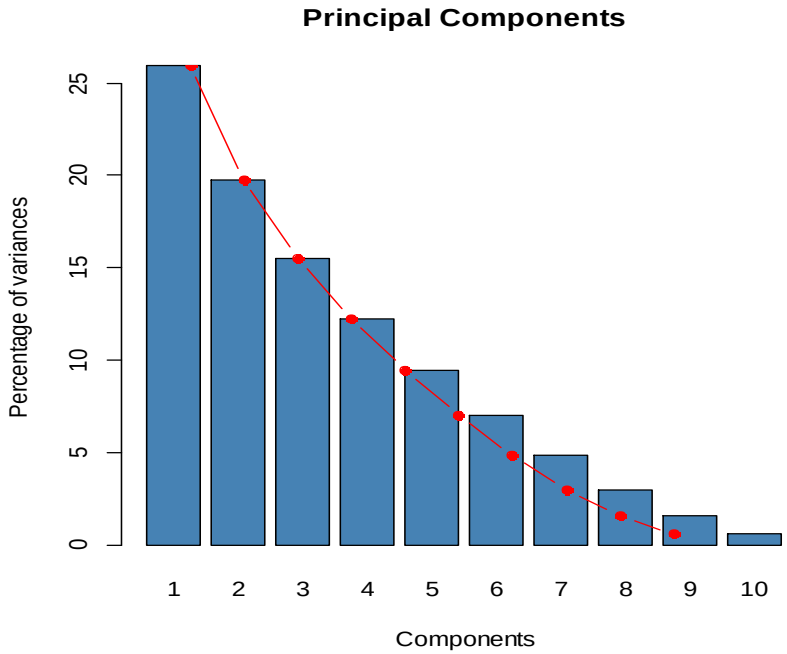
1. النتائج عند عندما يكون عدد المشاهدات n=15 و عدد المتغيرات p=10

اظهرت النتائج ان طريقة المركبات الرئيسة اللبية بأستعمال دالة لب Gaussian بنسبة تفسير (87.8) ثم طريقة المركبات الرئيسة اللبية بأستعمال دالة لب Laplacian بنسبة تفسير (83.4) واخير فسرت طريقة المركبات الرئيسة (75.0) ، حيث ان جدول رقم (1) يبين القيم المميزة و نسبة تفسير كل عامل من التباين المفسر بالإضافة الى التباين المفسر التجميعي للعوامل المؤثرة.

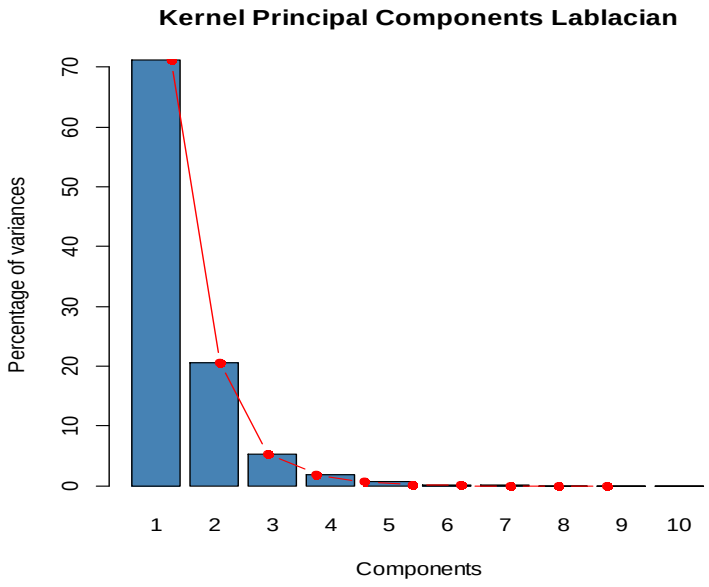
جدول رقم (1)

يمثل القيم المميزة ونسبة تفسير التباين لكل عامل من التباين الكلي ولجميع أساليب المركبات الرئيسة المستخدمة (الاعتيادية ، اللبية) عندما يكون عدد المشاهدات n=15 و عدد المتغيرات p=10

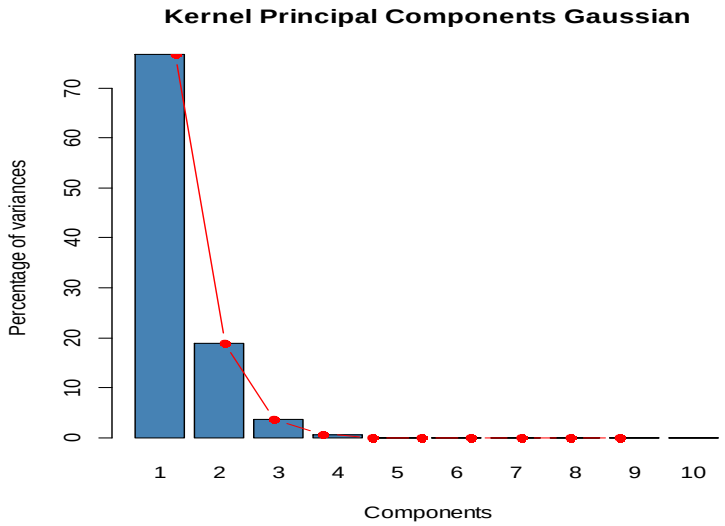
PC	Eigen Value			Proportion Variance			Cumulative Variance		
	PCA	Kpc Laplace	KPC Gaussian	PCA	KPC Laplace	KPC Gaussian	PCA	KPC Laplace	KPC Gaussian
Pc1	2.59	7.11	7.67	25.9	71.2	76.7	25.9	71.2	76.7
Pc2	1.98	2.05	1.89	19.8	20.5	18.9	45.7	91.7	95.6
Pc3	1.55			15.5			61.2		
Pc4	1.22			12.2			73.5		



شكل رقم (1)
scree plot لطريقة المركبات الرئيسة الاعتيادية عندما n=15 و p=10



شكل رقم (2)
scree plot لطريقة المركبات الرئيسة اللبية بأستعمال دالة لب Laplacian عندما n=15 و p=10



شكل رقم (3)

scree plot لطريقة المركبات الرئيسة اللبية بأستعمال دالة لب Gaussian عندما n=15 و p=10

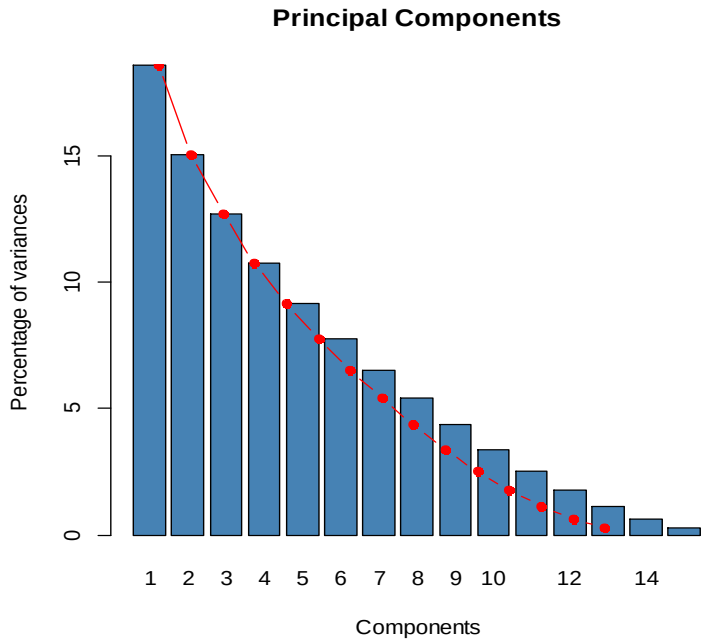
2. النتائج عند عندما يكون عدد المشاهدات n=22 و عدد المتغيرات p=15

اظهرت النتائج ان طريقة المركبات الرئيسة اللبية بأستعمال دالة لب Gaussian تمتلك نسبة التباين المفسر الاكبر (94.1) ثم طريقة المركبات الرئيسة اللبية بأستعمال دالة لب Laplacian بنسبة تفسير (91.8) واخير فسرت طريقة المركبات الرئيسة (73.6) ، حيث ان جدول رقم (2) يبين القيم المميزة و نسبة تفسير كل عامل من التباين المفسر بالإضافة الى التباين المفسر التجميعي للعوامل المؤثرة.

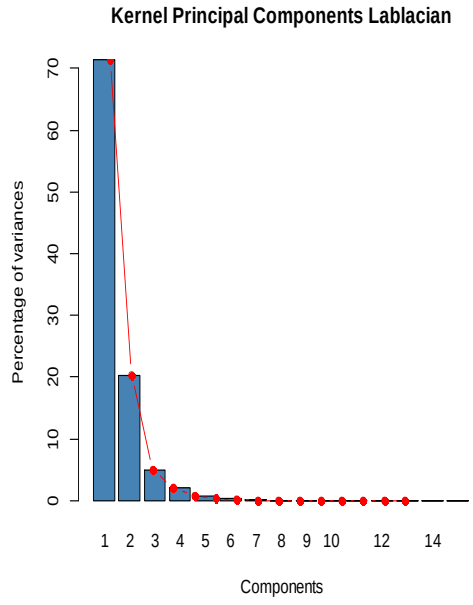
جدول رقم (2)

يمثل القيم المميزة ونسبة تفسير التباين لكل عامل من التباين الكلي ولجميع أساليب المركبات الرئيسة المستخدمة (الاعتيادية ، اللبية) عندما يكون عدد المشاهدات n=22 و عدد المتغيرات p=15

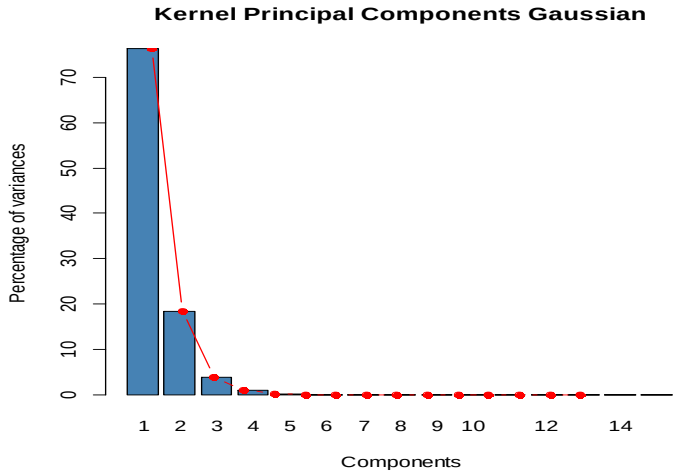
PC	Eigen Value			Proportion Variance			Cumulative Variance		
	PCA	KPC Laplace	KPC Gaussian	PCA	KPC Laplace	KPC Gaussian	PCA	KPC Laplace	KPC Gaussian
Pc1	2.78	10.7	11.4	18.6	71.4	76.3	18.6	71.4	76.3
Pc2	2.26	3.05	2.77	15.1	20.3	18.5	33.6	91.7	94.8
Pc3	1.9			12.7			46.3		
Pc4	1.61			10.8			57.1		
Pc5	1.37			9.15			66.2		
Pc6	1.16			7.77			74		



شكل رقم (4)
scree plot لطريقة المركبات الرئيسة الاعتيادية عندما n=22 و p=15



شكل رقم (5)
scree plot لطريقة المركبات الرئيسة اللبية بأستعمال دالة لب Laplacian عندما n=22 و p=15



شكل رقم (6)

scree plot لطريقة المركبات الرئيسة اللبية بأستعمال دالة لب Gaussian عندما n=22 و p=15

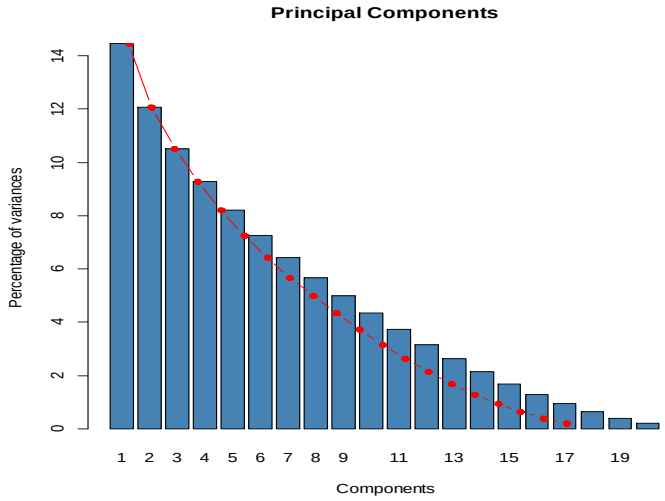
3. النتائج عند عندما يكون عدد المشاهدات n=30 و عدد المتغيرات p=20

اظهرت النتائج ان طريقة المركبات الرئيسة اللبية بأستعمال دالة لب Gaussian بنسبة تفسير (92.6) ثم طريقة المركبات الرئيسة اللبية بأستعمال دالة لب Laplacian بنسبة تفسير (91.0) واخير فسرت طريقة المركبات الرئيسة (77.5) ، حيث ان جدول رقم (3) يبين القيم المميزة و نسبة تفسير كل عامل من التباين المفسر بالإضافة الى التباين المفسر التجميعي للعوامل المؤثرة.

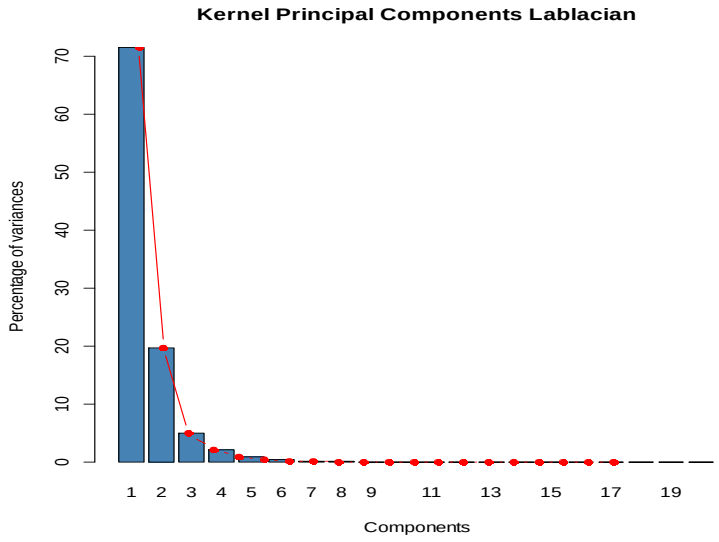
جدول رقم (3)

يمثل القيم المميزة ونسبة تفسير التباين لكل عامل من التباين الكلي ولجميع أساليب المركبات الرئيسة المستخدمة (الاعتيادية ، اللبية) عندما يكون عدد المشاهدات n=30 و عدد المتغيرات p=20

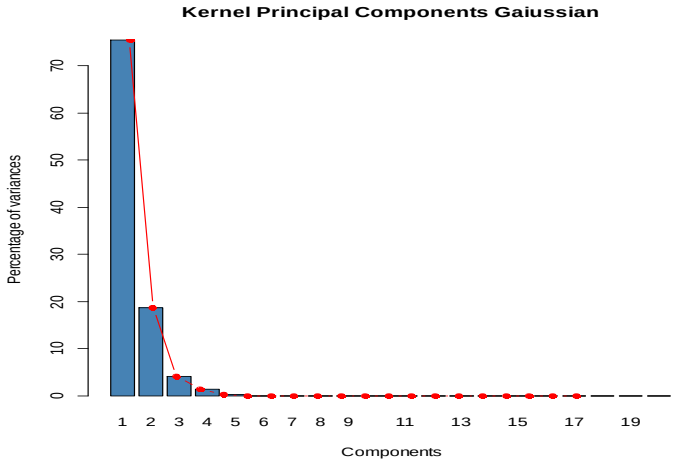
PC	Eigen Value			Proportion Variance			Cumulative Variance		
	PCA	Kpc Laplace	KPC Gaussian	PCA	KPC Laplace	KPC Gaussian	PCA	KPC Laplace	KPC Gaussian
Pc1	2.89	14.3	15.1	14.5	71.5	75.5	14.5	71.5	75.5
Pc2	2.41	3.49	3.73	12.1	19.7	18.6	26.5	91.2	94.1
Pc3	2.1			10.5			37		
Pc4	1.86			9.28			46.3		
Pc5	1.64			8.2			54.5		
Pc6	1.45			7.26			61.8		
Pc7	1.28			6.42			68.2		
Pc8	1.13			5.66			73.9		



شكل رقم (7)
scree plot لطريقة المركبات الرئيسة الاعتيادية عندما n=30 و p=20



شكل رقم (8)
scree plot لطريقة المركبات الرئيسة اللبية بأستعمال دالة لب Laplacian عندما n=30 و p=20



شكل رقم (9)

scree plot لطريقة المركبات الرئيسة اللبية بأستعمال دالة لب Gaussian عندما n=30 و p=20

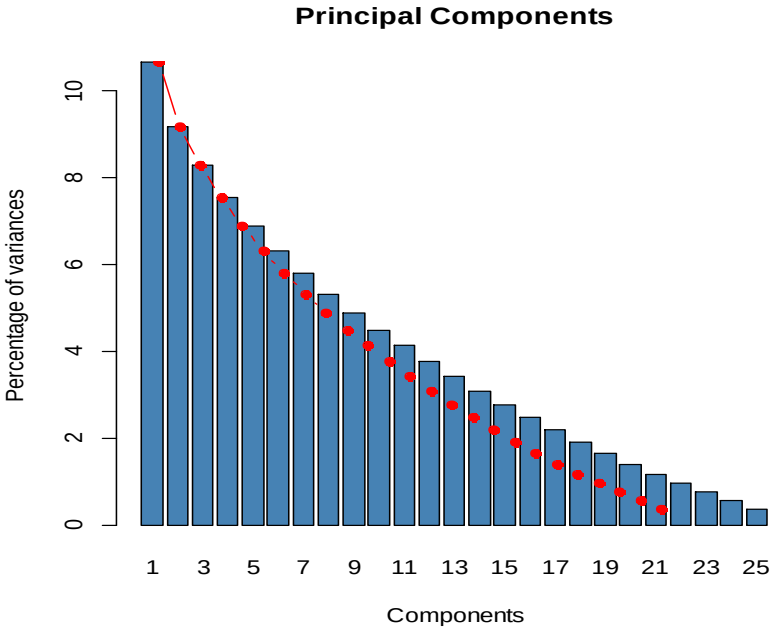
4. النتائج عند عندما يكون عدد المشاهدات n=50 و عدد المتغيرات p=25

اظهرت النتائج ان طريقة المركبات الرئيسة اللبية بأستعمال دالة لب Gaussian تمتلك نسبة التباين المفسر الاكبر (97.4) ثم تليها طريقة المركبات الرئيسة اللبية بأستعمال دالة لب Laplacian بنسبة تفسير (91.7) واخير فسرت طريقة المركبات الرئيسة (72.0) ، حيث ان جدول رقم (4) يبين القيم المميزة و نسبة تفسير كل عامل من التباين المفسر بالإضافة الى التباين المفسر التجميعي للعوامل المؤثرة.

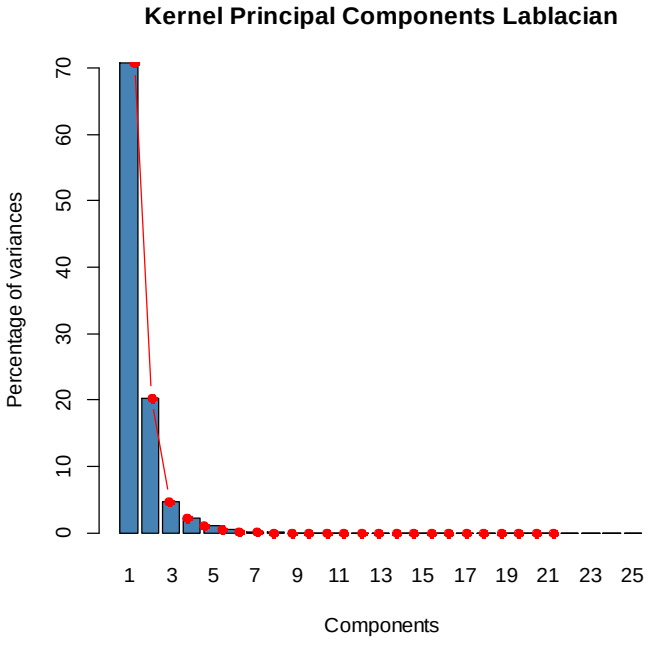
جدول رقم (4)

PC	Eigen Value			Proportion Variance			Cumulative Variance		
	PCA	Kpc Laplace	KPC Gaussian	PCA	KPC Laplace	KPC Gaussian	PCA	KPC Laplace	KPC Gaussian
Pc1	2.66	17.7	18.6	10.65	70.7	74.5	10.7	70.7	74.5
Pc2	2.3	5.07	4.75	9.19	20.3	19	19.8	91	93.5
Pc3	2.07	1.2	1.05	8.28	4.78	4.19	28.1	95.8	97.7
Pc4	1.88			7.54			35.7		
Pc5	1.72			6.88			42.5		
Pc6	1.58			6.31			48.8		
Pc7	1.45			5.8			54.7		
Pc8	1.33			5.33			60		
Pc9	1.22			4.89			64.9		
Pc10	1.12			4.5			69.4		
Pc11	1.03			4.13			73.5		

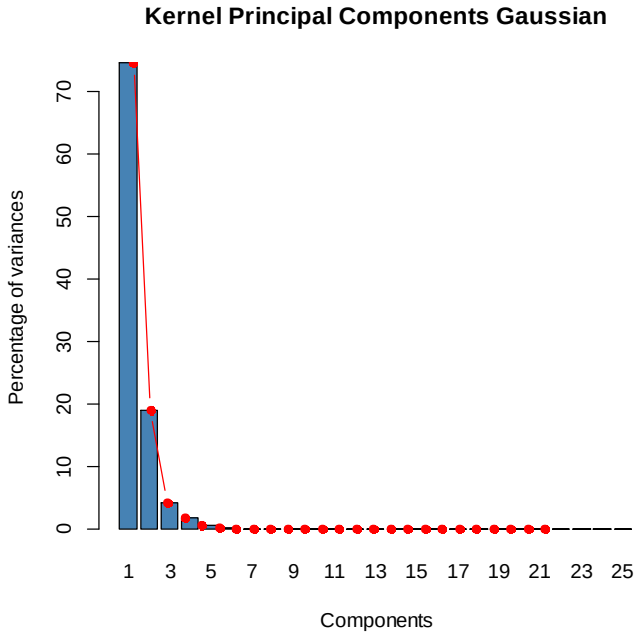
يمثل القيم المميزة ونسبة تفسير التباين لكل عامل من التباين الكلي ولجميع أساليب المركبات الرئيسة المستخدمة (الاعتيادية ، اللبية) عندما يكون عدد المشاهدات n=50 و عدد المتغيرات p=25



شكل رقم (10)
scree plot لطريقة المركبات الرئيسة الاعتيادية عندما n=50 و p=25



شكل رقم (11)
scree plot لطريقة المركبات الرئيسة اللبية بأستعمال دالة لب Laplacian عندما n=50 و p=25



شكل رقم (12)
scree plot لطريقة المركبات الرئيسة اللبية بأستعمال دالة لب Gaussian عندما n=50 و p=25

5- الاستنتاجات والتوصيات

5-1 الاستنتاجات

1. ان التحليل العاملي اللبي اكثر كفاءة من التحليل العاملي في حالة البيانات اللامعلمية.
2. استخدام تحليل المركبات الرئيسة اللبية باستخدام دالة Gaussian اكفاء من دالة Laplacian ولجميع احجام العينات من خلال النتائج المبينة في الجداول (3-1 , 3-2 , 3-3, 3-4).

5-2 التوصيات

1. استخدام التحليل العاملي اللبي في حالة البيانات اللامعلمية.
2. استخدام تحليل المركبات الرئيسة اللبية باستخدام دالة Gaussian.
3. استخدام الدوال اللبية في الظواهر الانسانية والاقتصادية والسلوكية لتمتعه بمرونة كبيرة بالتعامل مع الظواهر المدروسة.

المصادر

1. محمد ، حيدر يحيى ، (2013) ، دراسة العوامل المرتبطة بمرض السكري من خلال مواءمة اسلوبي التحليل العاملي وتحليل المسار ، رسالة ماجستير ، كلية الأدارة والأقتصاد ، الجامعة المستنصرية.
2. Alamsyah, M., Nafisah, Z., Prayitno, E., Afida, A. M., & Imah, E. M. (2018, January). The Classification of Diabetes Mellitus Using Kernel k-means. In Journal of Physics: Conference Series (Vol. 947, No. 1, p. 012003). IOP Publishing.
3. Duong, T. (2004). Bandwidth selectors for multivariate kernel density estimation. University of Western Australia.

4. Härdle, W., & Simar, L. (2007). Applied multivariate statistical analysis (Vol. 22007, pp. 1051-8215). Berlin: Springer.
5. Jolliffe, I. (2011). Principal component analysis (pp. 1094-1096). Springer Berlin Heidelberg.
6. Rosipal, R., & Trejo, L. J. (2001). Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of machine learning research*, *2*(Dec), 97-123.
7. Samuel, R. T., & Cao, Y. (2016). Nonlinear process fault detection and identification using kernel PCA and kernel density estimation. Systems Science & Control Engineering, 4(1), 165-174.
8. Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. Neural computation, 10(5), 1299-1319.
9. Timm ,H.N. (2002) , " Applied multivariate analysis " , Springer , New York.
10. Turlach, B. A. (1993, January). Bandwidth selection in kernel density estimation: A review. In CORE and Institut de Statistique.
11. Wand, M. P., & Jones, M. C. (1994). Kernel smoothing. Chapman and Hall/CRC.
12. Wu, W., Massart, D. L., & De Jong, S. (1997). Kernel-PCA algorithms for wide data Part II: Fast cross-validation and application in classification of NIR data. Chemometrics and Intelligent Laboratory Systems, 37(2), 271-280.
13. Wu, W., Massart, D. L., & De Jong, S. (1997). The kernel PCA algorithms for wide data. Part I: theory and algorithms. Chemometrics and Intelligent Laboratory Systems, 36(2), 165-172.

On kernel principal component analysis

Lekaa Ali Muhamed[a], Hayder Yahya mohammed[b]

[a]. Prof. Department of Statistics, College of Management and Economics Baghdad University, Baghdad, Iraq
[b] PhD student. Department of Statistics, College of Management and Economics Baghdad University, Baghdad, Iraq
Corresponding author: haiderstatistic@yahoo.com

Abstract

When dealing with multivariate data with higher dimensions, we often use principal component analysis (pca) to reduce the dimensions, but in the case of nonlinear data it is not possible to deal with classic estimated because of obtaining misleading results and therefore using kernel methods , The aim of this research is used KPCA to solve nonlinear data using kernel function to know the most effect variables on the phenomenon.

Keywords: principal component analysis ,High-dimensional data, kernel principal component analysis (KPCA).