AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

استعمال تحليل المركبات الرئيسة اللبية لمعالجة مشكلة التعدد الخطي لبيانات اعداد المصابين بفيروس كورونا المستجد (Coronavirus)•

أ.م.د اسماء غالب جابر	هبه مصطفى فوزي
drasmaa.ghalib@coadec.uobaghdad.edu.iq	almaroofhiba@gmail.com
قسم الاحصاء - كلية الادارة والاقتصاد - جامعة بغداد ، بغداد ، العراق	

المستخلص

ان الهدف الاساسي من استعمال طرائق التحليل متعددة المتغيرات هو تلخيص كمية البيانات الكبيرة والتي غالباً ما تكون متغيراتها التفسيرية مرتبطة مع بعضها البعض بعلاقات قوية ومعقدة، اي ان اغلبها طرق تبسيطية تساعد الباحث على تكوين فكرة واستنتاج حول هذه مجاميع المتداخلة . تستعمل طريقة تحليل المركبات الرئيسة الكلاسيكية لتحويل مجموعة المتغيرات المرتبطة الى مركبات متعامدة تدعى المركبات الرئيسة لكن في حال كانت مصفوفة البيانات لا خطية يكون من الصعب التعامل مع هذه البيانات بطريقة المركبات الرئيسة. تم استعمال (18) متغيراً تمثل محافظات العراق تضم بيانات اعداد المصابين بفيروس كورونا المستجد (Coronavirus) بالاعتماد على الموقف الوبائي اليومي لدائرة الصحة العامة لوزارة الصحة العراقية.

يهدف البحث الى استعمال طريقة تحليل المركبات الرئيسة اللبية (KPCA) التي تتعامل مع مجموعة البيانات غير الخطية وهي مشابهة لتحليل المركبات الرئيسة لكنها تؤدي الى اسقاط البيانات في فضاء عالي الابعاد يدعى بفضاء الميزة او فضاء الصفة (feature space)، و قد تبين من النتائج انه يمكن معالجة مشكلة التعدد الخطي باستعمال طريقة تحليل المركبات الرئيسة اللبية اذ تم تمثيل المتغيرات المرتبطة بعدد اقل من المركبات المتعامدة والتي بلغت (14) مركبة رئيسة فسرت نسبة (84%) من التباين الكلي.

معلومات البحث

تواريخ البحث:

تاريخ تقديم البحث: 2021/6/2
تاريخ قبول البحث: 2021/6/27
تاريخ رفع البحث على الموقع: 2022/6/25

الكلمات المفتاحية:

مشكلة التعدد الخطي، غير الخطي، الجذور الكامنة، المتجهات الكامنة، تحليل المركبات الرئيسة اللبية، PCA، KPCA

للمراسلة:

هبه مصطفى فوزي
almaroofhiba@gmail.com

DOI: https://doi.org/10.55562/jrucs.v51i1.519

1. المقدمة

ظهر التحليل الاحصائي متعدد المتغيرات (Multivariate Statistical Analysis) بعد الحاجة الى استعمال اساليب احصائية تشمل تحليل الظواهر بأنواعها المختلفة بما يناسب نوعية وحجم العينة المتوفرة. اذ يقدم التحليل المتعدد طرائق حديثة ومتطورة لوصف وتحليل واختبار البيانات المتعددة، كما يساعد في دراسة العلاقة بين المتغيرات للتنبؤ بسلوكها والتحكم بها. تحليل المركبات الرئيسة احد طرائق تحليل الاحصاء المتعدد، اذ اقترح (Karl person) في عام (1901)م اداة تستعمل لتلخيص وضغط البيانات والمعلومات ، كما تفسر اكبر قدر ممكن من التباين الاصلي لها . طورها (Hoteling) و اطلق عليها تحليل المركبات الرئيسة لكنها تتعامل مع البيانات الخطية ففي حال كانت مصفوفة البيانات غير خطية تصبح هذه الطريقة غير مجدية، لذلك طور (Schölkopf) عام (1998)م طريقة حديثة تتعامل مع البيانات غير الخطية تدعى تحليل المركبات الرئيسة اللبية اذ ان الفكرة الرئيسة لها مشابهة لفكرة تحليل المركبات الرئيسة الكلاسيكية لكنها تؤدي الى اسقاط البيانات في فضاء مختلف عالي الابعاد يدعى بفضاء الميزة او فضاء الصفة (feature space) ، يهدف البحث الى معالجة مشكلة التعدد الخطي التي تعاني منها البيانات

• بحث مستل من رسالة ماجستير

الخاصة بأعداد المصابين بفيروس كورونا من خلال تقليل ابعاد مصفوفة البيانات دون فقدان اي من المعلومات المهمة باستعمال طريقة تحليل المركبات الرئيسة اللبية على بيانات اعداد المصابين بفيروس كورونا.

عام (1999)م اقترح كل من (Bernhard Scholkopf & Alexander Smola) طريقة جديدة مشابهة لتحليل المركبات الرئيسة تتعامل مع البيانات غير الخطية تدعى طريقة المركبات الرئيسة تعتمد على دوال اللب (kernel) تحسب المركبات الرئيسة في فضاء عالي الابعاد يدعى بفضاء الصفة او الميزة واستنتجوا بعض المزايا لتطبيق هذه الطريقة هو ان اسلوب المركبات الرئيسة اللبية يمكن تطويره ، اعطت المركبات الرئيسة غير الخطية اداء افضل من المركبات الرئيسة الخطية ، امكانية استخدامها للتخلص من ضوضاء البيانات. في عام (2016)م تناول الباحثان لقاء علي و امير علي الى استخدام دالة (RBF KERNEL) اللبية في تحليل المركبات الرئيسية حيث قدرت معلمة التمهيد لها بعدة طرائق تقدير و تم المقارنة بينها من خلال عدد المركبات الرئيسية الفعالة التي تزيد قيم الجذور الكامنة لها عن الواحد الصحيح وما يقابلها من النسبة في تفسير التباين الكلي. في عام (2019)م استخدم (PALLE E.T. JORGENSEN & SOORAN KANG) تحليل المركبات الرئيسة لتقليل ابعاد مصفوفة البيانات غير الخطية باختيار المركبات الفعالة التي تقلل من الخطأ وتحسن تمثيل الرقمي للصور مما يؤدي الى تحسين النقل والتخزين، واثبتوا العديد من نظريات تقليل الابعاد الجديدة، وفي نفس العام قارن الباحثان اسماء غالب و اسيل مسلم طريقة تحليل المركبات الرئيسة في حال كون المركبات خطية مع طريقة تحليل المركبات الرئيسة اللبية في حال كانت المركبات لا خطية لاختيار الطريقة المناسبة لتقليل الابعاد الصورية حيث اظهرت النتائج ان طريقة المركبات الرئيسة اللبية (KPCA) هي الافضل .

2. الجانب النظري

2.1. مشكلة التعدد الخطي

عند ارتباط متغيرين او عدة متغيرات مع بعض بعلاقة خطية يقال عن مجموعة المتغيرات انها تعاني من مشكلة التعدد الخطي، و الارتباط الخطي المتعدد يعني انتهاك لإحدى فرضيات نموذج الانحدار والتي تنص على الا يكون المتغير المستقل دالة خطية في متغير او عدة متغيرات اخرى، و تميل المتغيرات التي تربطهما علاقة خطية قوية الى التحرك معاً مما يسبب صعوبة في تحديد اثر احد المتغيرات على الاخر[1].

تعرف مشكلة التعدد الخطي رياضياً هي عدم القدرة من حساب معكوس مصفوفة البيانات $X'X$ وذلك لان احد شروط نموذج الانحدار اختل وهو شرط الرتبة التامة $r(X) = K$ اذ ان K عدد المتغيرات [7].

يقسم التعدد الخطي الى نوعين :

- **التعدد الخطي التام (Perfect Multicollinearity):**

يظهر الارتباط الخطي التام عندما يكون احد المتغيرات دالة خطية تامة في متغير اخر او عدة متغيرات اخرى ، ويعني وجود علاقة خطية تامة بين متغيرين او اكثر عند تحقق الشرط الاتي :

$$\delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k = 0$$

علماً ان δ_k، ... ، δ_1 ثوابت على الاقل احدها لا يساوي الصفر تكون محددة مصفوفة البيانات مساوية صفر $|X'X| = 0$ اي انها مصفوفة مفردة (Singular Matrix) مما يؤدي الى اختلال شرط الرتبة ليصبح $r(X) < K$ فلا يمكن ايجاد معكوس المصفوفة البيانات في هذه الحالة .

- **التعدد الخطي غير التام (Non-perfect Multicollinearity):**

ويعني وجود علاقة خطية شبه تامة بين متغيرين او اكثر نتيجة لتحرك المتغيرات سوياً بالزيادة او النقصان وعند تحقق العلاقة الاتية :

$$\delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + \nu_i = 0 \qquad ، i = 1،..،k$$

علماً ان δ_k، ... ، δ_1 ثوابت على الاقل احدها لا يساوي الصفر و ν_i متغير عشوائي لا يساوي الصفر ايضاً . تكون قيمة محددة مصفوفة البيانات قريبة من الصفر $|X'X| \cong 0$ اي قيمتها صغيرة ينتج عنها مقدرات غير دقيقة لكبر قيم تباين المعلمات [6].

2.1.1. اكتشاف مشكلة التعدد الخطي (Detection of Multicollinearity)

هنالك عدة معايير للكشف عن وجود مشكلة التعدد الخطي ابسطها الاعتماد على معامل الارتباط البسيط حيث يعتبر هناك تعدد خطي بين المتغيرات المستقلة اذا كانت قيمة معامل الارتباط الخطي بين متغيرين كبيرة مساوية (0.8) او اكثر [7] .

الدليل الشرطي((Condition Index (CI): يعتمد معيار الدليل الشرطي على حساب الجذور الكامنة لمصفوفة البيانات X'X لمعرفة قوة العلاقة بين المتغيرات المستقلة ، يحسب الدليل الشرطي كجذر تربيعي لنسبة أكبر جذر كامن الى كافة جذور المصفوفة الكامنة (λ_K، ... ، λ_2، λ_1) بالتتالي وفق الصيغة الاتية:

$$K،...،CI_i = \sqrt{\frac{\lambda_{max}}{\lambda_S}} \qquad S=1 \qquad (1)$$

اذا تراوحت قيمة الدليل الشرطي CI بين (30-15) فأنه يوجد علاقة جدية بين المتغيرات المستقلة ، بينما اذا تجاوزت قيمته ال 30 فتعتبر العلاقة بين المتغيرات خطيرة [8].

2.2. مصفوفة اللب (KERNEL MATRIX)

يرمز لها (K) هي مصفوفة مربعة من الدرجة (N×N) متماثلة (Symmetric) شبه موجبة (Simi Positive) تساعد في حساب المسافات بين ازواج البيانات من خلال المعلومات التي تحتويها وصيغتها العامة:

$$\mathrm{K} = K_{\mathrm{ij}} = \Omega(x_i)\Omega(x_j) \quad i = 1، \dots، n، j = 1، \dots، n \tag{2}$$

علماً ان K: مصفوفة اللب . تحسب مصفوفة اللب من دوال اللب التي تعتمد في حسابها على معلمة التمهيد او عرض الحزمة (Bandwidth) ، كما لا تحدد كفاءة دالة اللب دون كفاءة مصفوفة اللب [2] [9].

2.3. معلمة التمهيد (Bandwidth)

تدعى ايضاً بعرض الحزمة تكون قيمتها اكبر من الصفر ومن الضروري اختيارها بشكل ملائم لأنها جزء مهم في تقريب دالة الانحدار اللا معلمي من الدالة الاصلية لذلك يعد اختيارها اهم من اختيار دالة اللب، وتكون معلمة عرض الحزمة ملائمة عندما يتوازن كل من التباين والتحيز يتحقق هذا التوازن من خلال استعمال دالة التمهيد عرض الحزمة [4].

2.4. دالة اللب (Kernel function)

تعرف الدالة اللبية بأنها دالة حقيقية متماثلة محدودة مستمرة تكاملها يساوي واحد تعتمد الدوال اللبية في حسابها على معلمة التمهيد h (Bandwidth) او عرض الحزمة، و لكل دالة لبية هنالك معلمة تمهيدية خاصة بها تختلف عن باقي الدوال اللبية [3]. الدالة اللبية المستعملة في البحث هي دالة (Gaussian kernel) وذلك لسهولة التعامل معها وتناسبها مع طبيعة البيانات وذلك بعد التجريب، وبما انه بحث مستل من الرسالة تم استعمال نفس الدالة المعتمدة في الرسالة (ان شاء الله سيعتمد دوال اخرى في البحوث والدراسات التالية) صيغتها هي:

$$\mathrm{k}(\mathrm{x_i x_j}) = \exp\left(-\frac{\left\|\mathrm{x_i} - \mathrm{x_j}\right\|^2}{2\mathrm{h}^2}\right) \tag{3}$$

2.5. تحليل المركبات الرئيسة (Kernel Principal Component Analysis)

من الشائع استخدام تحليل المركبات الرئيسة (Principal Component Analysis)(PCA) لمعالجة مشكلة التعدد الخطي (تقليل الابعاد) من خلال تحويل المتغيرات المرتبطة الى توليفات خطية متعامدة تدعى المركبات الرئيسة وبهذا نتخلص من مشكلة الارتباط الخطي التي قد تعاني منها البيانات (Multicollinearity) ، لكن في حال كون البيانات غير خطية يصبح من الصعب استعمال طريقة تحليل المركبات الرئيسة و التعامل معها . تعرف طريقة تحليل المركبات الرئيسة اللبية (KPCA) من الطرق الحديثة الشائعة الاستعمال مع مثل هذا النوع من البيانات تتشابه فكرة المركبات اللبية مع الفكرة العامة لطريقة المركبات الرئيسة فهي تحول مجموعة كبيرة من المتغيرات الى مركبات جديدة غير مرتبطة تفسر التباين الكلي للمصفوفة لكن يتم ذلك في فضاء جديد فضاء الميزة او الصفة (feature space) [10] .

2.5.1. الاسلوب الرياضي لتحليل المركبات الرئيسة اللبية:

لدينا مجموعة بيانات غير الخطية $\mathrm{X_n \epsilon R^d}$، ... ،$\mathrm{X_2}$،$X_1$ ونود تحويلها من فضائها الاصلي $\mathrm{R^d}$ الى فضاء ذو ابعاد عالية يدعى بفضاء الميزة ويرمز له $\mathcal{H}$ ($\mathcal{H}$: $highy\ dimensional\ feature\ space$) بواسطة تحويل لا معلمي (Ω)

$$\Omega: \mathrm{R^d} \rightarrow \mathcal{H}$$

اذ ان $\mathrm{R^d}$: فضاء البيانات الاصلي .

$\mathcal{H}$: فضاء الميزة ذو الابعاد العالية الذي يتم فيه اسقاط البيانات، Ω : دالة لا معلمية .

بعد نقل و تسليط البيانات الى الفضاء الجديد نطبق طريقة المركبات الرئيسة في الفضاء $\mathcal{H}$ اذ نحسب الجذور الكامنة والمتجهات الكامنة المقابله لها للحصول على المركبات الرئيسة ، و نفرض ان متوسط البيانات في فضاء الميزة الجديد مساوياً الى صفر اي بيانات متمركزة $E\big(\Omega(X)\big) = 0$ نحسب مصفوفة التباين والتباين المشترك $\Sigma_{\mathcal{H}}$ للبيانات في الفضاء الجديد [12] :

$$\Sigma_{\mathcal{H}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{j=1}}^{\mathrm{n}} \Omega\big(\mathrm{x}_j\big)\,\Omega(\mathrm{x_j})^{\mathrm{T}} \tag{4}$$

ونجد المتجهات الكامنة التي يقابل كل منها جذر كامن بحل المعادلة الاتية [11]:

$$\Sigma_{\mathcal{H}}\,\alpha = \alpha\lambda \tag{5}$$

علماً ان :

λ: الجذور الكامنة لمصفوفة التباين في الفضاء الجديد $\lambda_i \geq 0$.

α : متجهات كامنة متعامدة يقابل كل منها جذر كامن .

لدينا المتجهان α_i، α_j :

$$\alpha_i\alpha_j = \begin{cases} 0 & if\ i \neq j \\ 1 & if\ i = j \end{cases}$$

اذ حققت شرط التعامد للمركبات الرئيسة العادية فيمكن التعبير عن المتجهات الكامنة كالأتي :

$$\alpha = \sum_{i=1}^{n} \alpha_i \Omega(x_i) \tag{6}$$

علماً ان α_i : المتجه الكامن i ويمثل معامل المتغير $\Omega(x_i)$. نعوض (4) في المعادلة رقم (5) ينتج الاتي :

$$\frac{1}{n}\sum_{j=1}^{n} \Omega(x_j)\,\Omega(x_j)^{T}\alpha = \alpha\lambda$$

$$\frac{1}{n\lambda}\sum_{j=1}^{n} \Omega(x_j)\,\Omega(x_j)^{T}\alpha = \alpha$$

$$\sum_{j=1}^{n} \frac{\Omega(x_j)^{T}}{n\lambda}\Omega(x_j)\,\alpha = \sum_{i=1}^{n} \alpha_i \Omega(x_i)$$

نضرب طرفي معادلة رقم (5) ب($\Omega(x_k)$) فنحصل :

$$\Omega(x_k)\Sigma_{\mathcal{H}}\,\alpha = \Omega(x_k)\,\alpha\lambda \tag{7}$$

نعوض (4) و (6) في المعادلة *رقم* (7).

$$\Omega(x_k)\ \ \frac{1}{n}\sum_{j=1}^{n} \Omega(x_j)\,\Omega(x_j)^{T}\sum_{i=1}^{n} \alpha_i\Omega(x_i) = \Omega(x_k)\lambda\ \sum_{i=1}^{n} \alpha_i\Omega(x_i) \tag{8}$$

بإعادة ترتيب المعادلة (8) :

$$\lambda\ \sum_{i=1}^{n} \Omega(x_k)\alpha_i\Omega(x_i) = \frac{1}{n}\sum_{i=1}^{n} \alpha_i\,\langle\Omega(x_k)\sum_{j=1}^{n} \Omega(x_j)\rangle\,\Omega(x_j)\Omega(x_i) \tag{9}$$

نعتمد على مصفوفة اللب K (Kernel Matrix) لحساب α_i التي تمثل معاملات $\Omega(x_i)$ باستخدام الصيغة التالية :

$$K_{ij} = \Omega(x_i)\Omega(x_j)$$

تصبح المعادلة (9) كالآتي :

$$\lambda \sum_{i=1}^{n} \alpha_i K_{ki} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \alpha_i \sum_{\mathrm{j}=1}^{\mathrm{n}} K_{ki}\, K_{ij} \qquad (10)$$

وبما ان k=1،…،n تصبح المعادلة (7) كالآتي:

$$\lambda K\alpha = \frac{1}{n} K^2 \alpha$$

$$\lambda n K\alpha = K^2 \alpha \qquad (11)$$

نضرب المعادلة (11) ب(K^{-1}) :

$$\lambda n\alpha = K\alpha \qquad (12)$$

$$K\alpha - \lambda n\alpha = 0$$

$$(K - \lambda n)\alpha = 0$$

نفرض ($\lambda n = \tilde{\lambda}$) :

$$(K - \tilde{\lambda})\alpha = 0$$

يمكن جعل المصفوفة Kernel مركزية كالأتي :

$$K_{ctr} = K - UK - KU + UKU$$

حيث ان U مصفوفة من الدرجة $(n \times n)$ جميع عناصرها مساوية الى $(\frac{1}{\mathrm{n}})$.
وبجعل المتجهات الكامنة α لمصفوفة K normalize في الفضاء $\mathcal{H}$ كالأتي :

$$\sum_{،j=1}^{n} \alpha_i^k\, \alpha_j^k \Omega(x_i)\Omega(x_j) = 1$$

$$\sum_{،j=1}^{M} \alpha_i^k\, \alpha_j^k K_{ij} = 1$$

$$\alpha^k K \alpha^k = 1$$

$$\lambda_k \alpha^k \alpha^k = 1$$

لإيجاد المعاملات يجب تحقيق الشرط الاتي:

$$|K - \tilde{\lambda}| = 0 \qquad (13)$$

و هكذا نحصل على المركبات الرئيسة من الصيغة الاتية [13] :

$$p_{\omega} = \sum_{i=1}^{n} \alpha_{ij} K\left(x_i, x_j\right) \quad (14)$$

2.5.2. خوارزمية المركبات الرئيسة اللبية (Kernel Principal Component Algorithm)

- الخطوة الاولى : احسب معلمة التمهيد h .
- الخطوة الثانية : احسب معادلة اللب $k(x_i x_j)$ باستخدام المعادلة (3)

$$k(x_i x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2h^2}\right)$$

- الخطوة الثالثة : احسب مصفوفة اللب K من صيغتها العامة :

$$K = K_{ij} = \Omega(x_i)\Omega(x_j) \qquad i = 1, \dots, n, j = 1, \dots, n$$

- الخطوة الرابعة : اوجد الجذور الكامنة والمتجهات الكامنة بحل مشكلة الجذور الكامنة ل K .
- الخطوة الخامسة : اوجد المركبات الرئيسة من المعادلة (14)

$$p_{\omega} = \sum_{i=1}^{n} \alpha_{ij} K\left(x_i, x_j\right)$$

3. الجانب التطبيقي

في هذا الجانب اعتمد على مصفوفة بيانات تضم اعداد الاصابة بفيروس كورونا المستجد (Coronavirus) ، و لمدة (91) من 2021/1/15و لغاية 2021/4/14 لكافة محافظات العراق. تم اخذ البيانات من موقع دائرة الصحة العامة التابعة لوزارة الصحة العراقية بالاعتماد على الموقف الوبائي اليومي للفيروس.

جدول (3-1): وصف متغيرات اعداد الإصابات اليومية بفيروس كورونا لمحافظات العراق

Variable	City	Variable	City	Variable	City
X_1	Baghdad	X_7	Kakuk	X_{13}	Dewaniyah
X_2	Najaf	X_8	Dayala	X_{14}	Thekqar
X_3	Sulaymaniyah	X_9	Wast	X_{15}	Anbar
X_4	Arbil	X_{10}	Basra	X_{16}	Mothana
X_5	Dohouk	X_{11}	Mesan	X_{17}	Naynawa
X_6	Karbala	X_{12}	Babil	X_{18}	Salahaldeen

3.1. نبذة عن فيروس كورونا (Coronavirus):

ينتمي فيروس كورونا (covid-19) لعائلة الفيروسات التاجية تنشأ في الحيوان ومن ثم تنتقل منها الى الإنسان اكتشف اول نوع منها سنة (1960) وتم تسجيل اول حالة وفاة بسببها سنة (2012) في دولة السعودية [17] . ظهر فيروس كورونا المستجد (covid-19) في مدينة ووهان الصينية سنة (2019) ومن ثم تفشى بشكل واسع لتغطي الاصابات به كافة انحاء العالم . ينتقل الفيروس عبر الاتصال الشخصي بالشخص المصاب او عند ملامسة الأسطح الملوثة لقدرة الفيروس العيش عليها لبعض ساعات [16] .

3.2. اعراض فيروس كورونا:

تقسم اعراض فيروس كوونا الى ثلاثة مستويات مختلفة باختلاف درجة وفترة الإصابة ، فقد تظهر الاعراض بعد يومين او بعد 14 يوماً من الاصابة [15] .

جدول (3-2): مستويات اعراض فيروس كورونا

الاعراض الاكثر شيوعاً	الأعراض الأقل شيوعًا	الأعراض الخطيرة
حمّى	آلام وأوجاع التهاب الحلق إسهال التهاب الملتحمة صداع	صعوبة أو ضيق في التنفس
سعال جاف	فقدان حاسة التذوق أو الشم	ألم أو ضغط في الصدر
إرهاق	طفح جلدي، أو تغير في لون أصابع اليدين أو أصابع القدمين	فقدان القدرة على الكلام أو الحرك

3.3. تشخيص مشكلة التعدد الخطي

3.3.1. معاملات مصفوفة الارتباط

لتشخيص مشكلة التعدد الخطي تم إيجاد مصفوفة الارتباطات للمتغيرات المدروسة والتي يتضح منها ان اغلب متغيرات الدراسة ترتبط فيما بينها بعلاقة طردية قوية اذ تجاوزت قيمة الارتباط بين بعض المتغيرات (0.8) وهذا دليل على وجود تعدد خطي بين المتغيرات .

جدول (3-3): يمثل مصفوفة الارتباط بين المتغيرات

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18
x1	1	0.41	0.73	0.76	0.83	0.21	0.68	0.77	0.88	0.84	0.92	0.62	0.65	0.67	0.76	0.59	0.38	0.8
x2	0.41	1	0.04	0.32	-0.01	0.6	-0.12	0.07	0.31	0.64	0.2	0.6	0.73	0.72	0	0.52	-0.25	0.36
x3	0.73	0.04	1	0.88	0.77	-0.1	0.72	0.72	0.78	0.58	0.76	0.5	0.39	0.44	0.79	0.39	0.52	0.63
x4	0.76	0.32	0.88	1	0.64	0.03	0.6	0.66	0.79	0.74	0.76	0.54	0.62	0.54	0.71	0.53	0.35	0.76
x5	0.83	-0.01	0.77	0.64	1	-0.07	0.78	0.8	0.81	0.57	0.89	0.42	0.3	0.39	0.83	0.32	0.48	0.66
x6	0.21	0.6	-0.1	0.03	-0.07	1	-0.21	-0.02	0.15	0.29	0.05	0.45	0.48	0.6	-0.04	0.28	-0.19	0.1
x7	0.68	-0.12	0.72	0.6	0.78	-0.21	1	0.73	0.7	0.43	0.75	0.28	0.24	0.27	0.74	0.21	0.6	0.56
x8	0.77	0.07	0.72	0.66	0.8	-0.02	0.73	1	0.77	0.59	0.82	0.41	0.38	0.37	0.72	0.41	0.45	0.7
x9	0.88	0.31	0.78	0.79	0.81	0.15	0.7	0.77	1	0.79	0.9	0.61	0.59	0.59	0.81	0.54	0.4	0.77
x10	0.84	0.64	0.58	0.74	0.57	0.29	0.43	0.59	0.79	1	0.76	0.64	0.76	0.65	0.57	0.67	0.17	0.77
x11	0.92	0.2	0.76	0.76	0.89	0.05	0.75	0.82	0.9	0.76	1	0.45	0.52	0.49	0.84	0.53	0.39	0.77
x12	0.62	0.6	0.5	0.54	0.42	0.45	0.28	0.41	0.61	0.64	0.45	1	0.58	0.68	0.45	0.39	0.12	0.53
x13	0.65	0.73	0.39	0.62	0.3	0.48	0.24	0.38	0.59	0.76	0.52	0.58	1	0.7	0.44	0.63	0.01	0.52
x14	0.67	0.72	0.44	0.54	0.39	0.6	0.27	0.37	0.59	0.65	0.49	0.68	0.7	1	0.37	0.55	0.08	0.48
x15	0.76	0	0.79	0.71	0.83	-0.04	0.74	0.72	0.81	0.57	0.84	0.45	0.44	0.37	1	0.45	0.45	0.61
x16	0.59	0.52	0.39	0.53	0.32	0.28	0.21	0.41	0.54	0.67	0.53	0.39	0.63	0.55	0.45	1	0.02	0.45
x17	0.38	-0.25	0.52	0.35	0.48	-0.19	0.6	0.45	0.4	0.17	0.39	0.12	0.01	0.08	0.45	0.02	1	0.25
x18	0.8	0.36	0.63	0.76	0.66	0.1	0.56	0.7	0.77	0.77	0.77	0.53	0.52	0.48	0.61	0.45	0.25	1

3.3.2. معيار الدليل الشرطي ((Condition index (CI)):

جدول (3-4): قيم الدليل الشرطي

Dimension	Eigenvalue	Condition Index
1	14.738	1
2	1.343	3.313
3	0.536	5.246
4	0.261	7.507
5	0.223	8.126
6	0.186	8.906
7	0.114	11.385
8	0.108	11.655
9	0.098	12.292
10	0.084	13.208
11	0.08	13.568
12	0.056	16.278
13	0.05	17.244
14	0.035	20.47
15	0.027	23.278
16	0.026	23.747
17	0.02	27.36
18	0.016	30.736

من الجدول رقم (3-4) نجد ان قيم الجذور الكامنة (من الجذر الكامن رقم (1) الى الجذر الكامن رقم (11)) لم تظهر قيمة الدليل الشرطي المقابلة لها اكبر من (15) اي انها لا تعاني من تعدد خطي، في حين كانت قيم الدليل الشرطي لبقية المتغيرات بين $15 \leq CI \leq 30$ اذ بلغت قيمة المعيار للمتغير رقم (12) (16.278) وتصاعدت قيمته حتى بلغ عند المتغير(18) (30.736) مما يدل على وجود مشكلة تعدد خطي اي ان هنالك علاقة جدية بين المتغيرات . لمعالجة مشكلة التعدد الخطي . تم تطبيق طريقة تحليل المركبات الرئيسة اللبية على مصفوفة البيانات للتخلص من مشكلة التعدد الخطي .

3.4. طريقة تحليل المركبات الرئيسة اللبية

طبقنا طريقة المركبات الرئيسة اللبية على مصفوفة البيانات للحصول على مركبات خطية متعامدة و لمعرفة نسبة المساهمة في تفسير التباين الكلي لمصفوفة البيانات. لاختيار المركبات تم الاعتماد على الجذور الكامنة التي تتجاوز قيمتها الواحد الصحيح .

جدول (3-5): قيم الجذور الكامنة و نسب تفسير التباين حسب معيار Kaiser

	eigenvalue	percentage of variance	cumulative percentage of variance		eigenvalue	percentage of variance	cumulative percentage of variance
comp1	1.715	9.527	9.527	comp10	1.012	5.62	61.995
comp2	1.255	6.971	16.497	comp11	1.011	5.618	67.614
comp3	1.066	5.922	22.419	comp12	1.011	5.618	73.232
comp4	1.026	5.702	28.122	comp13	1.011	5.618	78.85
comp5	1.022	5.677	33.799	comp14	1.01	5.612	84.462
comp6	1.02	5.666	39.465	comp15	0.994	5.524	89.985
comp7	1.017	5.651	45.116	comp16	0.907	5.038	95.024
comp8	1.014	5.635	50.751	comp17	0.65	3.612	98.636
comp9	1.012	5.625	56.375	comp18	0.245	1.364	100

يضم جدول رقم (3-5) قيم الجذور الكامنة و نسب تفسير التباين حسب معيار Kaiser فأن المركبات الداخلة في التحليل هي المركبات التي تتجاوز قيمتها الواحد الصحيح، كما اشار Morrison ان نسبة تفسير 75% او اكثر من التباين تكون كافية . من الجدول اعلاه نجد ان (14) جذراً كامناً تجاوزت قيمته الواحد الصحيح لذلك سيكون هنالك (14) مركبة رئيسة ، اذ فسر اول جذر كامن نسبة (9.527) من التباين الكلي ومن ثم يتناقص لتبلغ نسبة مساهمة الجذر الرابع عشر في التفسير مساوية الى (5.612) . كما نلاحظ ان المقدار التراكمي لنسبة التباين المفسر بواسطة المركبات الداخلة مساوية (84%) بينما المركبات المتبقية فسرت اقل من (20%) من التباين الكلي فيمكن الاعتماد على المركبات الرئيسة الـ(14) الاولى .

جدول (3-6): يمثل المركبات الرئيسة اللبية

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9
Baghdad	0.745	0.714	0.67	0.341	0.499	0.685	0.576	0.635	**0.728**
Najaf	0.741	0.712	0.666	0.295	0.444	0.057	0.508	0.129	**0.047**
Sulaymaniyah	0.011	0.498	0.013	0.154	0.123	0.054	0.107	0.104	**0.015**
Arbil	-0.02	0.498	0.009	0.118	0.118	0.034	0.091	0.012	**0.014**
Dohouk	-0.02	-0.055	0.006	0.062	0.048	0.019	0.087	0.009	**0.014**
Karbala	-0.02	-0.055	0.005	0.06	0.046	0.018	0.087	0.008	**0.014**
Kakuk	-0.02	-0.055	-0.049	0.03	0.044	0.017	0.087	0.008	**0.001**
Dayala	-0.02	-0.056	-0.049	0.029	0.044	0.017	0.01	0.008	**0**
Wast	-0.027	-0.056	-0.049	0.028	0.044	0.017	-0.001	0.004	**0**
Basra	-0.027	-0.056	-0.051	0.028	-0.002	0.013	-0.002	0	**0**
Mesan	-0.031	-0.074	-0.051	0.028	-0.002	0.005	-0.003	0	**-0.004**
Babil	-0.031	-0.075	-0.084	-0.002	-0.007	-0.001	-0.019	0	**-0.004**
Dewaniyah	-0.032	-0.086	-0.086	-0.002	-0.026	-0.001	-0.042	-0.003	**-0.007**
Thekqar	-0.032	-0.087	-0.121	-0.004	-0.04	-0.013	-0.043	-0.003	**-0.012**
Anbar	-0.033	-0.088	-0.124	-0.01	-0.04	-0.013	-0.134	-0.01	**-0.019**
Mothana	-0.038	-0.088	-0.149	-0.045	-0.188	-0.051	-0.286	-0.026	**-0.02**
Naynawa	-0.535	-0.091	-0.163	-0.045	-0.472	-0.078	-0.29	-0.086	**-0.024**
Salahaldeen	-0.541	-0.107	-0.239	-0.884	-0.532	-0.735	-0.451	-0.764	**-0.696**

تكملة الجدول (3-6)

	Dim.10	Dim.11	Dim.12	Dim.13	Dim.14
Baghdad	0.42	0.715	0.384	0.821	**0.712**
Najaf	0.382	0.011	0.384	0	**0**
Sulaymaniyah	0.366	0.011	0.384	0	**0**
Arbil	0.366	0.011	0.009	0	**0**
Dohouk	0.366	0.001	0	0	**0**
Karbala	0.003	0	0	0	**0**
Kakuk	-0.003	0	0	0	**0**
Dayala	-0.003	0	0	0	**0**
Wast	-0.003	0	-0.003	0	**0**
Basra	-0.043	0	-0.003	0	**0**
Mesan	-0.043	0	-0.006	0	**0**
Babil	-0.098	0	-0.006	0	**0**
Dewaniyah	-0.106	0	-0.007	0	**0**
Thekqar	-0.13	0	-0.008	0	**0**
Anbar	-0.142	0	-0.009	0	**0**
Mothana	-0.169	0	-0.014	0	**0**
Naynawa	-0.287	-0.031	-0.528	-0.409	**-0.002**
Salahaldeen	-0.344	-0.706	-0.539	-0.412	**-0.71**

صفوف الجدول رقم (3-6) تمثل المتغيرات الاصلية لأعداد الاصابة بجائحة كورونا بينما تمثل الاعمدة المركبات الرئيسة لأعداد الاصابة، اذ تم تمثيل (18) متغيراً مرتبطاً في عدد اقل من المركبات المتعامدة والتي بلغت (14) مركبة رئيسة .

جدول (3-7): يبين قيم معيار الدليل الشرطي حسب طريقة المركبات الرئيسة

Condition index (CI)									
comp1	1	comp5	1.295	comp9	1.302	comp13	1.302	comp17	1.624
comp2	1.169	comp6	1.297	comp10	1.302	comp14	1.303	comp18	2.646
comp3	1.268	comp7	1.299	comp11	1.302	comp15	1.314		
comp4	1.293	comp8	1.301	comp12	1.302	comp16	1.375		

من الجدول اعلاه نلاحظ اختفاء مشكلة التعدد الخطي اذ كانت قيم معيار الدليل الشرطي اقل من ال (15) .

4. الاستنتاجات والتوصيات

4.1. الاستنتاجات

1. هنالك العديد من الطرائق للكشف عن مشكلة التعدد الخطي ومعالجتها، تعتبر طريقة المركبات الرئيسة اللبية اختياراً مناسباً في حالة البيانات لا تتبع التوزيع الخطي او تعاني من مشاكل .
2. تحليل المركبات الرئيسة اللبية اكثر كفاءة للتعامل مع المتغيرات غير الخطية في تقليل عدد المتغيرات المؤثرة .
3. تحليل المركبات الرئيسة اللبية حول المتغيرات المرتبطة الى مركبات متعامدة .
4. نسبة التباين المفسر كانت عالية عند استعمال دالة (Gaussian) وهي احدى دوال اللب شائعة الاستعمال .
5. اقتراح استعمال طريقة المركبات الرئيسة اللبية في التعامل مع البيانات غير الخطية اعطى اداء افضل مما لو تم استعمال طريقة المركبات الرئيسة الكلاسيكية.

4.2. التوصيات

1. استعمال الدوال اللبية في تحليل المركبات الرئيسة اللبية في حال كانت البيانات متعددة المتغيرات غير خطية وتعاني من مشاكل لحل مشكلة التعدد الخطي .
2. توظيف طريقة المركبات الرئيسة اللبية لمعرفة المتغيرات المؤثرة على الظاهرة المدروسة من خلال اعتماد الجذور الكامنة التي تزيد قيمتها عن الواحد الصحيح .
3. استعمال الطريقة المقترحة لتحويل المتغيرات المرتبطة الى متغيرات جديدة متعامدة تدعى المركبات .
4. استعمال المركبات الرئيسة اللبية بالاعتماد على دالة (Gaussian) .

5. اعطاء موضوع فيروس كورونا اهتماماً اكبر ، اذ يجب اخذ الحيطة والحذر في حال ظهور احد الاعراض المذكورة سابقاً لان من الممكن ان يسبب الفيروس مضاعفات تؤدي الى الموت .

المصادر

[1] السواعي، خالد محمد ، مبادئ الاقتصاد القياسي، طبعة (3) دار الكتاب الثقافي، الاردن، (2018).

[2] محمد ، لقاء علي و عبود، امير علي، "مقارنة مقدرات عرض الحزمة (معلمة التمهيد) باستخدام الدوال اللبية في تحليل المركبات الرئيسية"، مجلة كلية التراث الجامعة، عدد (20)، (2016)، ص 436-412 .

[3] حمود، مناف يوسف، "تقدير دالة الانحدار اللامعلمية باستخدام دوال لب قانونية "، مجلة العلوم الاقتصادية و الادارية، المجلد (17) ، عدد(61)، (2011)، ص 225-212 .

[4] الصفاوي، صفاء يونس و متي، نور صباح، "تقدير دوال الانحدار اللامعلمي باستخدام بعض أساليب التمهيد"، المجلة العراقية للعلوم الاحصائية، وقائع المؤتمر العلمي الرابع لكلية علوم الحاسبات والرياضيات، (2011)، ص 392-373

[5] الراوي، اسماء غالب و عيسى، اسيل مسلم، "مقارنة بين طريقتي تحميل المركبات الرئيسية والمركبات الرئيسية اللبية لتقليل الابعاد الصورية"، المجلة العراقية للعلوم الاحصائية، عدد (29)، (2019)، وقائع المؤتمر الطلابي الاول ص 12-24.

[6] حياوي، هيام عبد المجيد، "تقدير نماذج فضاء الحالة باستخدام أسلوب انحدار الحرف مع التطبيق"، المجلة العراقية للعلوم الاحصائية المجلد (10)، العدد (18)، (2010)، ص 176-155 .

[7] Widiyanto, Ibnu, "Multicollinearity: Does It Really Matter?", Fokus Ekonomi – Vol. 5, No.2, (2006), pp. 110-129.

[8] Senaviratna, Namr, "Diagnosing Multicollinearity of Logistic Regression Model", Asian Journal of Probability and Statistics, Vol. (5), (2019), pp. 1-9.

[9] Ho , Tu Bao, Nguyen, Canh Hao, "Kernel Matrix Evaluation", Japan Advanced Institute of Science and Technology, (2007), pp. 923-1292 .

[10] Maestri, Mauricio L., Cassanello, Miryan C., "Kernel PCA Performance in Processes with Multiple Operation Modes", The Berkeley Electronic Press., Vol. (4), (2009), pp. 1-16.

[11] Scholkopf, Bernhard, Smola, Alexander, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", Tubingen: Max-Planck-Institut fur biologische Kybernetik, (Tech. Rep. No. 44), (1996), pp. 1-18.

[12] Scholkopf, Bernhard, Smola, Alexander, "Kernel Principal Component Analysis", spring verlag, Vol. (1327), (1997), pp. 583-588.

[13] Samuel, Raphael Tari, Cao, Yi, "Nonlinear process fault detection and identification using kernel PCA and kernel density estimation", Systems Science & Control Engineering, 4:1, (2016), pp. 165-174.

[14] Mezaache, Hatem, Bouzgou, Hassen, "Kernel Principal Components Analysis with Extreme Learning Machines for Wind Speed Prediction", Seventh International Renewable Energy Congress, (2016).

[15] Mayo Clinic, coronavirus disease 2019, available at: https://www.mayoclinic.org/ar/diseases-conditions/coronavirus/symptoms-causes/syc-20479963.

[16] Medicine Sans Frontier, Coronavirus COVID-19 pandemic, available at: https://www.unicef.org/iraq/ar .

[17] World Health Organization, Middle East respiratory syndrome coronavirus (MERS-CoV), available at: http://www.aun.edu.eg/arabic/society/may_2014.html .

Journal of AL-Rafidain University College for Sciences (2022); Issue 51; 12-22

AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

Using Kernel Principal Component Analysis of Treating Linear Multiplicity Problem for Covid-19 injured Data

Dr. Asmaa G. Jaber	Hiba M. Fawzi
drasmaa.ghalib@coadec.uobaghdad.edu.iq	almaroofhiba@gmail.com

Department of Statistics - College of Administration and Economics - University of Baghdad, Baghdad - Iraq.

Article Information

Article History:
Received: June, 2, 2021
Accepted: June, 27, 2021
Available Online: June, 25, 2022

Keywords:
Non-linear, Eigenvalues, Eigenvectors, Kernel Principal Component Analysis, PCA, KPCA.

Correspondence:
Hiba M. Fawzi
almaroofhiba@gmail.com

Abstract

The main goal of using multivariate analysis method is to summarize the large number of data whose explanatory variables are correlated with each other by strong and complex relationships, i.e. most of them are simplistic methods that help the researcher form an idea and conclusion about these overlapping groups. The classical principal component analysis method is used to turn a set of related variables into orthogonal components called principal components, but it is difficult to deal with this data in the principal component method if the data matrix is nonlinear. We used (18) variables representing the governorates of Iraq and data about the number of people infected with the new Coronavirus, based on the daily epidemiological situation of the Public Health Department of the Iraqi Ministry of Health

This research aims to use Kernel Principal Component Analysis (KPCA) to deal with the Non-linear data set. It is similar to Principal Component Analysis but it's mapping the data in high dimensional space called feature space. The results show that the problem of linear multiplicity can be addressed by using principal component analysis method, where the correlated variables represented with a smaller number of orthogonal components, which amounted to (14) principal component that explained a percentage (84%) of the total variance.

DOI: https://doi.org/10.55562/jrucs.v51i1.519