الخوارزمية الاسرع لتحليل المركبات الرئيسة الحصينة مع تطبيق عملي على المتغيرات المؤثرة على ارتفاع نسبة الالومينيوم في الدم

ظافر حسين رشيد و خلود يوسف خمو
كلية الادارة والاقتصاد- جامعة بغداد

الخلاصة

ان احد التقنيات الشائعة في تحليل البيانات متعددة المتغيرات هي تحليل المركبات الرئيسة PCA الذي يحول عدد كبير من المتغيرات المرتبطة الى عدد اقل من المركبات غير المرتبطة، وفي حالة وجود القيم الشاذة والتي يكشف عنها في طرق عديدة فان اعتماد مصفوفة التباين والتباين المشترك ومنه مصفوفة الارتباط الاعتياديتين سيؤدي الى نتائج مظللة لتحليل المركبات الرئيسة.يهدف هذا البحث الى تناول خوارزمية جديدة وسريعة في تحليل المركبات الرئيسة الحصينة عند احتواء البيانات على قيم شاذة، والتي تفشل فيها الطرائق التقليدية في الكشف عن الشواذ وبالتالي الحصول على نتائج مظللة، وقد تم تطبيق الطريقة لبيان مدى كفاءتها على بيانات واقعية لمتغيرات تؤثر على ارتفاع نسبة الألومينيوم في الدم.

المقدمة وهدف البحث

ان وجود القيم الشاذة في البيانات في الحالة متعددة الأبعاد يؤثر على مصفوفة التباين والتباين المشترك ومنه مصفوفة الارتباط وبالتالي فان اعتماد الأسلوب التقليدي لتحليل المركبات الرئيسة يؤدي الى عدم الدقة في تفسير العلاقة بين المتغيرات قيد الدراسة.من هنا فان هدف البحث هو تناول طريقة جديدة وسريعة في تحليل المركبات الرئيسة الحصينة وهي قليلة التداول في البحوث لأسباب التعقيدات البرمجية والحاجة إلى مستوى برامجيات عالية الكفاءة. تم تطبيق الطريقة لبيان مدى كفاءتها على بيانات واقعية لمتغيرات تؤثر على ارتفاع نسبة الألومينيوم في الدم.ان التطبيق يعتبر ذات أهمية كون الأسباب المؤدية إلى ارتفاع نسبة الألومينيوم في الدم غير معروفة من قبل الأطباء إذ لم تتضح معظم أسباب ارتفاع نسبة الألومينيوم وما سينيج من تبعات لامراض خطيرة. وما توصل إليه الباحثين في دراساتهم العديدة هو ما زال في طـــور البحـــث بخصوص تبعات ومسببات المرض.إذ بينوا من أسباب ارتفاع نسبة الألومينيوم في الدم قد تكون

أمراض منها ما تعيق التركيز الذهني للشخص نتيجـــة الإصابـــة بالخرف المبكر(فقدان الذاكرة) إضافة إلى صعوبة مزاولة الأعمال نتيجة لمرض خطير آخر هو الشلل الرعاشي.

الجانب النظري

1- تحليل المركبات الرئيسة Principle Component Analysis PCA

الانموذج العاملي(1976,Harmen) لـــ k من المتغيرات المشاهدة ولعينة ذات حجم n يفسر على اساس دالة خطية لـــ m من العوامل المشتركة، حيث k>m و k من العوامل الوحيدة لكل متغير اي ان

$$\underline{X} = A\underline{F} + \underline{U} \quad \dots \quad (1)$$

حيث :

$\underline{X}$ الموجه للمتغيرات من درجة (k×1) .

A مصفوفة تحميلات العوامل من درجة (k×m) .

$\underline{F}$ موجه العوامل المشتركة من درجة(m×1).

$\underline{U}$ موجه العوامل الوحيدة من درجة(k×1) .

إن طريقة المركبات الرئيسة PC تعتبر طريقة رئيسة في التحليل العاملي إذ تقوم بتفسير ظاهرة تعتمد على عدد كبير من المتغيرات غير المستقلة لغرض الوصول الى اعلى درجة من المعلومات وبعوامل مستقلة تكون اقل من المتغيرات المستخدمة والتي تعبر عن العلاقات الموجودة بين المتغيرات.إن فكرة (1976,Marrison,1986,Norusis) الجذور والمتجهات الذاتية (eigen values & vector) وقبل التطرق الى تحليل المركبات الرئيسة هي وبافتراض المصفوفة X درجتها p للحصول على متجه عمودي غير صفري $\underline{a}$ عدد عناصره p فان

$$\mathbf{X\underline{a}_i = \lambda_i \underline{a}_i \quad \dots (2)}$$

قيمة λ_i التي تحقق هذه المعادلة تسمى الجذور الذاتي للمصفوفه X اما المتجهات $\underline{a}_i$ والتي تناظر هذه الجذور تسمى المتجهات الذاتية (eigen vectors) للمصفوفة X.

$$\mathbf{(X-\lambda_i I)\underline{a}_i = 0 \quad \dots (3)}$$

فاذا كانت المصفوفة $(X-\lambda_i I)$غير احادية فيمكن ايجاد $\underline{a}_i$ بالضرب المسبق للمعادلة (3) في معكوس هذه المصفوفة وفي هذه الحالة تكون $\underline{a}_i$ متجها صفرياً وهذا ما يتعارض مع كون $\underline{a}_i$ متجه غير صفري، لذلك فان الشرط اللازم لايجاد المتجه a_i ان تكون المصفوفة $\mathbf{|X-\lambda_i I|}$

احادية اي ان قيمة محددتها تساوي صفر

$$| X - \lambda_i | = 0 \quad \dots (4)$$

وبحل المعادلة (4) يمكن ايجاد قيم λ_i وباستخدام المعادلة (3) يمكن ايجاد المتجهات الذاتية المناظرة لتلك الجذور بحيث تكون هذه المتجهات متعامدة فيما بينها.فلو كان لدينا p من المتغيرات العشوائية $X_1,...,X_n$ بمتوسط مجتمع $\underline{\mu}=0$ ومصفوفة تباين مشتركة $\sum$ ، وبافتراض ان المصفوفة S تمثل تقدير لمصفوفة التباين المشترك للمجتمع بدرجة حرية n=N-1 وهي متماثلة وموجبة التحديد (p.d.) او شبه موجبة (p.s.d.)، وان من اهم خواص الجذور والمتجهات الذاتية للمصفوفة S كون الجذور الذاتية للمصفوفة S موجبة او غير سالبة $\lambda_1>\lambda_2>...>\lambda_p>0$ وبافتراض ان المتجهات الذاتية المناظرة للجذور الذاتية هي $\underline{a}^*_1,\underline{a}^*_2,...,\underline{a}^*_p$ على الترتيب تكون المتجهات الذاتية المتعامدة المعدلة للمصفوفة S هي $\underline{a}_1,\underline{a}_2,...,\underline{a}_p$ ، وان $\underline{a}_1$ المتجه الذاتي المناظر للجذر الذاتي λ_i وبافتراض ان $\mathbf{V_1 \approx N_p(0,\underline{a}'_1 S\underline{a}_1)}$ وبما ان $\mathbf{\underline{a}'_1 S\underline{a}_1 = \lambda_1}$ فان $\mathbf{V_1 = N_p(0,\lambda_1)}$ اي ان الجذر الذاتي الاكبر للمصفوفة S يستخدم لتقدير التباين الاعظم للمركبة الرئيسة الاولى والمتجه الذاتي a_1 يستخدم لتقدير المعاملات للمركبة الرئيسة الاولى وبالطريقة نفسها تكون المكونة الرئيسة الثانية حيث ان ثاني اكبر جذر للمصفوفة S يستخدم لتقدير التباين الاعظم للمركبة الرئيسة الثانية والمتجه الذاتي الثاني يستخدم لتقدير معاملات المركبة الرئيسة الثانية والتباين المشترك بين V_1 و V_2 هو $\text{cov}(V_1,V_2)$ اي ان الارتباط بين المركبة الرئيسة الاولى والمركبة الثانية يساوي صفر.

2- خوارزمية C-R Croux & Ruiz -Gazen (2002, Hubert , 1985,Li)

بفرض $X_{n,p}$ مصفوفة البيانات الاصلية تسمى خوارزمية Croux & Ruiz - PP Gazen بخوارزمية C-R وهي كالآتي:

الخطوة الابتدائية البيانات تتمركز حول الوسيط $\mathbf{L^1}$ $\mathbf{\hat{\mu}^R}$ وهذا مقدّر حصين بدرجة عالية وتعامدياً مكافىء الى مقدّر الموقع والذي يعرف بالوسيط الحيزي (Spatial Median) ويعرف عند نقطة θ التي تقلل مجموع المسافات لكل المشاهدات حيث:

$$\mathbf{\hat{\mu}^R = arg_\theta min \sum_{i=1}^{n} || x_i - \theta ||} \quad \dots (5)$$

وغالباً المقدّر يتطلب تساوي التغاير (Affine Equivariant) والذي يؤكد انه المقدّر يملك خاصية التحويل عندما تدور المحاور السينية ويعاد قياسها.

لتبسيط الخطوات سنرمز للمشاهدات المركزية بـــ

$$\mathbf{X^{(1)}_i = X_i - \hat{\mu}^R \; : \mu^R = I_n \, (\hat{\mu}^R)' \quad ...(6)}$$

حيث $I_n=(1,...,1)'$ متجه عمودي وكل قيمة من قيمه تساوي واحد ، اما صفوف المصفوفة M^R مساوية الى $\left(\hat{\mu}^R\right)'$، ومصفوفة البيانات المركزية $\mathbf{(X^{(1)}_1,...,X^{(1)}_n)'}$ تساوي $\mathbf{X_{n,p}-M^R}$.

وبفرض r رتبة مصفوفة البيانات المركزية وان $r \le \min(n-1,p)$، توجد الابعاد الجزئية r متمثلة بالمتجهات الاحادية العمودية $\mathbf{v_\tau \;\; \tau = 1,...,r}$ بحيث ان التباين الحصين (Robust Scale) S_i للمشاهدات البارزة (Projected Observations) في v_ι هي الاعظم وهذا مشابه لطريقة PCA ، وان الاسقاط في البيانات في v_1 له اعظم تباين حصين s_1 .

اولاً يتـــم ايجاد المتجه الذاتي الاول v_1 والمشاهدات تسقط في المكملة العمودية الى v_1 .

المتجه الثاني v_2 يعود الى المكملة العمودية والتباين الحصين s_2 للمشاهـــدات المسقطة فـــي v_2 ويجب ان تكون الاعظم وهذا يكرر حتى إيجاد r من المتجهات.

قد برهن Li & Chen بان الطريقة تكون حصينة بدرجة عالية لانها ورثت نقطة انهيار مقدّر التباين الحصين. ان نقطة الانهيار تبين نسبة نقاط البيانات التي تلوث بينما تنتج مقدّر محدد، كما توصل كل من Croux & Ruiz - Gazen الى ان مقدّر Q_n يعطي افضل النتائج.

لاي مجموعة بيانات احادية $Z_1,\ldots,Z_n$ مقدّر Q_n يعرف كالآتي :

$$\mathbf{Q_n(Z_1,...,Z_n) = 2.2219 * C_n * \{|Z_i - Z_j| \; ; i < j\}_{(k)} \quad ...(7)}$$

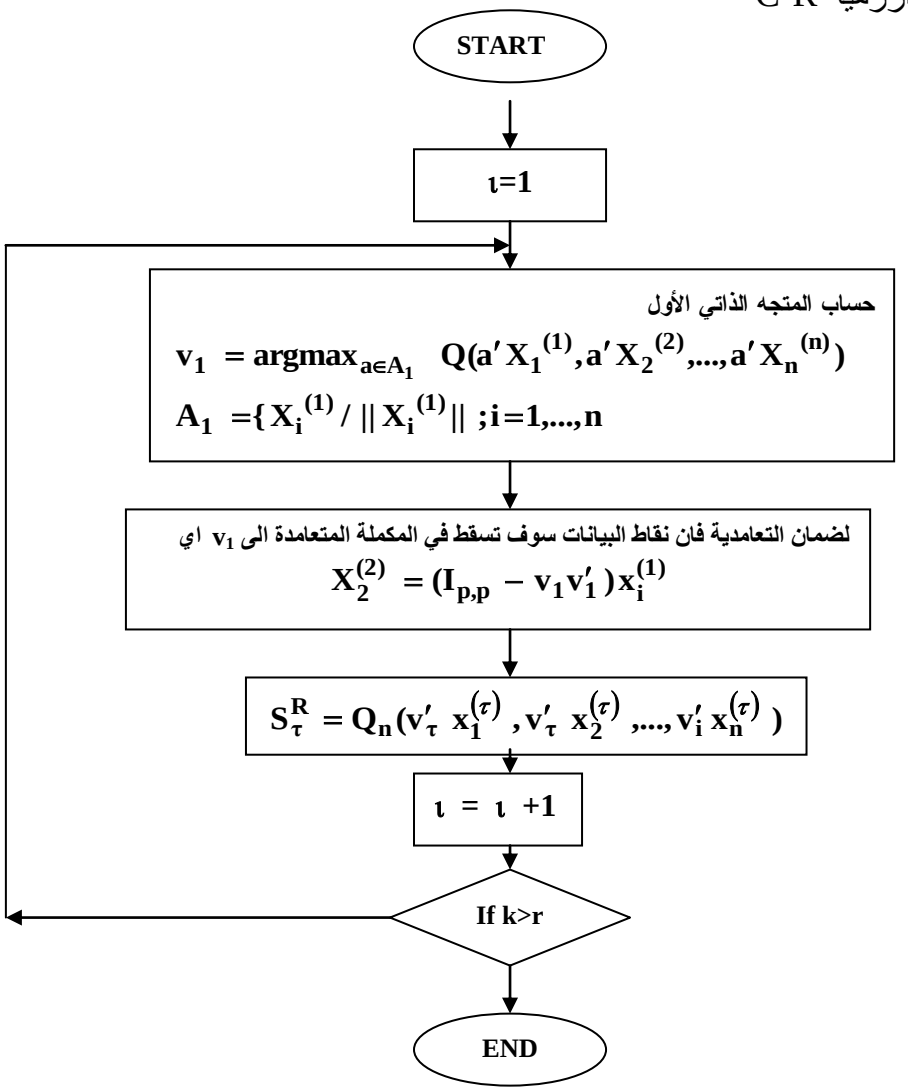
ومع

$$\mathbf{k = \binom{h}{2} \approx \binom{n}{2}/4; \quad k = (n/2)+1}$$

الثابت C_n هو عامل تصحيح القيمة الصغيرة Small-Sample Correction Factor والذي يجعل Q_n مقدّر غير متحيز (ان C_n تعتمد فقط على حجم العينة n وان $C_n \to 1$ عندما تزداد n)قيمة الانهيار الى Q_n هي 50% حيث الكفاءة الاحصائية للمتجهات الذاتية الناتجة ولتوزيع كاوس الطبيعي Gaussian تساوي 67% . لجعل حسابات الخوارزمية ممكنة فان Croux & Ruiz - Gazen قيدوا مجاميع الاتجاهات لبحث كل الاتجاهات ونقاط البيانات التي تمر خلال $\hat{\mu}^R$.

لتوضيح خطوات برمجة خوارزمية C-R والتي تحسب القيم والمتجهات الذاتية وهي من تصميم الباحث وكالتي:

مخطط (1) خوارزمية C-R



3-الخوارزمية المحسنة(Improved Algorithim Rousseeuw,1999,Hubert,2002)

السبب في عدم الاستقرارية العددية لخوارزمية C-R هو عند انجاز الحسابات في مجال الابعاد العالية، فان تتابع الاسقاطات يقود الى تراكم الاخطاء وبالتالي الحصول على مقدّرات غير معول عليها، لذا تم تحسين الخوارزمية من خلال تخفيض الابعاد.بافترض ايجاد المتجه الذاتي الأول v_1 سوف يطبق تحويل البيانات المتعامدة U_1 بحيث ان v_1هو متجه الاساس الأول وان $e_1=(1,0,\ldots,0)'$،ان الطريقة الاسرع لعمل ذلك ليست ببناء مصفوفة تعامدية كبيرة لكنبواسطة الانعكاس ولبناء الانعكاس U_1 نحتاج فقط حساب المتجه الطبيعي الاحادي(Unit Normal Vector) أي $n_1=(e_1-v_1) / \| e_1 - v_1 \|$ وان كل نقطة بيانات مركزية $x_i^{(1)}$ تحول الى

$$\mathbf{x_i^{(2)} = U_1(x_i^{(1)}) = x_i^{(1)} - 2\langle x_i^{(1)}, n_1\rangle n_1 \ \ldots\ (8)}$$

وان $U_1(v_1)=e_1$ كما ان $\langle a,b\rangle$ تمثل الضرب الداخلي للمتجهين والذي يحسب بالشكل $a'b$. واذا اردنا تسقيط البيانات بالمكملة العمودية الى $U_1(v_1)$ يجب ان نحرك الاحداثي الأول الى $x_i^{(2)}$ لكل i ، وهذا ينقص الفضاء الاصلي ببعد واحد وبذلك فان كل العمليات تكون دقيقة، كما وان الانعكاس (8) يتطلب مجاميع أولية وهذا الاجراء يتكرر حتى إيجاد r من المتجهات حيث $\mathbf{r = rank(X_{n,p} - M^R)}$ وفي كل خطوة نبحث عن الاتجاه الجديد v_1 في تخفيض الفضاء $\mathbf{(p-\tau+1)}$_ ، كما ان المتجه v_1 يعكس المتجه الاساسي الأول في المجال والاسقاطات للبيانات المعكوسة نحصل عليها مباشرة بواسطة حذف الاحدائي الأول .ولاجل تفسير النتائج يجب تحويل كل اتجاه v_1 خلفياً الى فضاء البعد الأصلي p وهذه العملية ستكون مستقرة عددياً طالما ان نظير الانعكاس هو الانعكاس نفسه لذا لا نحتاج الى مصفوفة الانقلاب ويشار الى هذه العملية بطريقة الخطوة R R-Step . بتعريف $P_{p,r}$ مصفوفة المتجهات الذاتية المحولة للخلف كأعمدة سوف نحصل على التجزئة الحصينة لمصفوفة $X_{n,p}$ الى مصفوفة الهدف $T_{n,r}$ ومصفوفة التحميـــــلات $P_{p,r}$ ونحصـــل على

$$\mathbf{X_{n,p} - M^R = T_{n,r}\ P'_{r,p} \ \ldots(9)}$$

ان $T_{n,r}$ تتضمن احداثيات نقاط البيانات في المجال المقاس بواسطة المركبات الرئيسة .

أخيراً يمكن تخفيض العدد الأصلي للمتغيرات p بواسطة اعتبار اول k من الاعمدة الى p ، والعدد k يمثل عدد المكونات الرئيسة التي نود بقائها.

وكما موضح في مخطط (2) طريقة R-Step والتي وضعت من قبل الباحث

وبعد إيجاد $\mathbf{v_1,\ldots,v_{\tau-1}}$ من المتجهات الذاتية ، الانعكاس $U_{\iota-1}$ ، التحويل واسقاطات نقاط البيانات يمكن ان تعرف كالآتي:

$$\mathbf{U_{\tau-1}(\tilde{v}_{\tau-1}) = \tilde{e}_{\tau-1} = (1,0,\ldots,0)' \in \Re^{p-\tau+2} \ \ldots\ (10)}$$

$$\mathbf{x_i^{(\tau)} = U_{\tau-1}(\tilde{x}_i^{(\tau-1)} \in \Re^{p-\tau+2} \ \ldots\ (11)}$$

$$\mathbf{\tilde{x}_i^{(\tau)} = (x_{i,2}^{(\tau)},\ldots,x_{i,p-\tau+2}^{(\tau)}) \in \Re^{p-\tau+1} \ \ldots\ (12)}$$

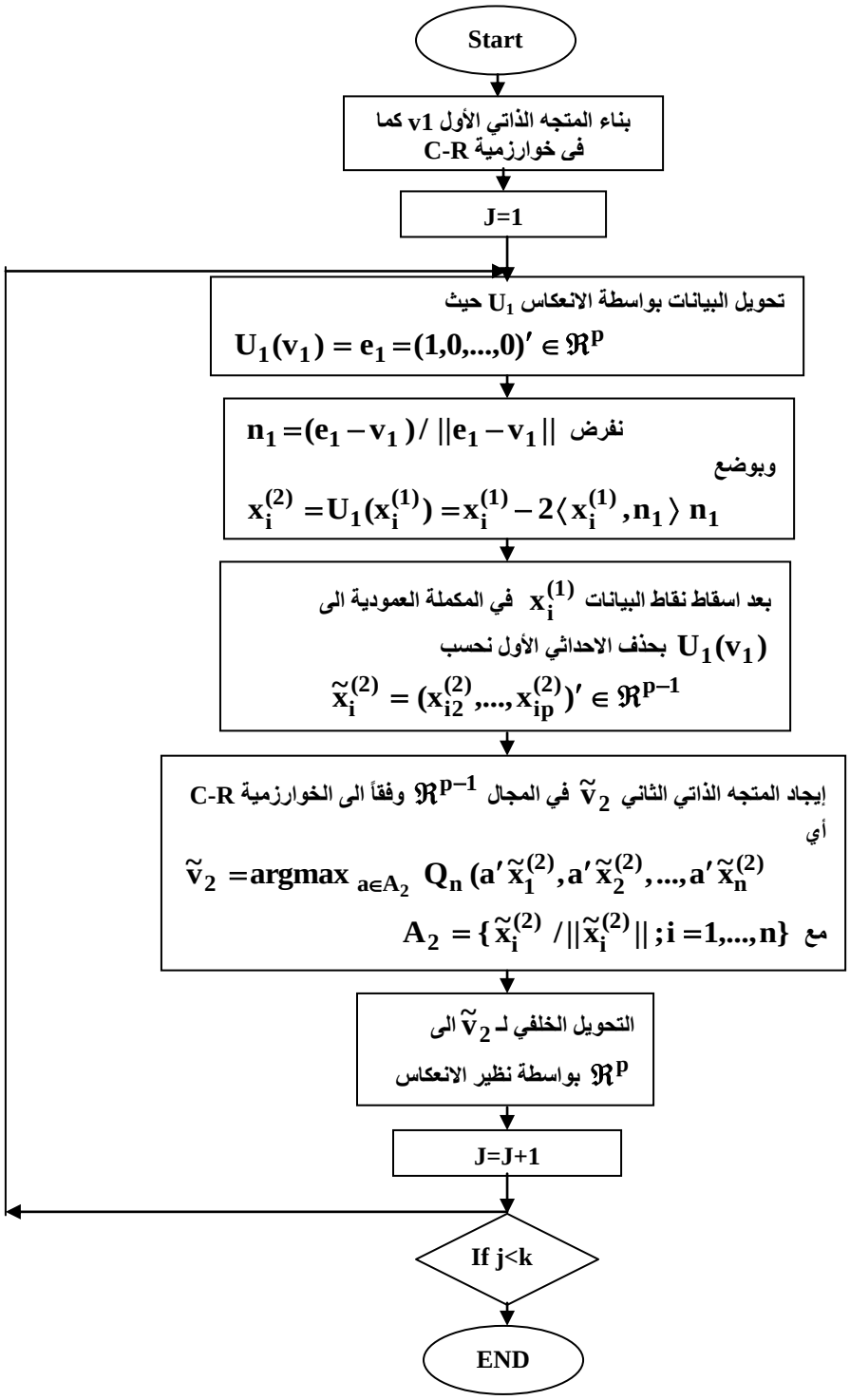
$$\mathbf{\tilde{v}_\tau = argmax_{a\in A_1}\ Q_n(a'\tilde{x}_1^{(\tau)}, a'\tilde{x}_2^{(\tau)}, \ldots, a'\tilde{x}_n^{(\tau)}) \ldots (13)}$$

مع

$$\mathbf{A_1 = \{\tilde{x}_i^{(\tau)} / \,||\, \tilde{x}_i^{(\tau)} \,||\,;\ i = 1,\ldots,n\}}$$

المتجه الذاتي v_ι نحصل عليه من $\tilde{v}_\tau$ بواسطة التحويل الخلفي.

مخطط (2) طريقة R-Step



4-خوارزمية أسرع خطوتين(1999,2002,Rousseeuw,2002,Hubert)AFaster Two Step

أن طريقة R-Step عددياً أكثر استقراراً من C-R ، ولكن تأخذ وقت طويل للحسابات اي أبطىء ومع زيادة عدد المركبات الرئيسة التي تحسب بالفعل تزداد بطئاً لذا لزم تسريعها. طالما ان وقت الحسابات الكبير هو بسبب زيادة قيمة p ، والخزارزمية الجديدة ذات خطوتين:

أولاً تخفيض مجال البيانات الى مجال جزئي بواسطة n من المشاهدات ، وهذا يعمل اسرع وادق بواسطة الأساليب التقليدية PCA بدون خسارة في المعلومات فيما يخص البيانات لذا نجزء $X_{n,p}$ كالآتي:

$$\mathbf{X_{n,p} - M^C = \tilde{T}_{n,r}\ \tilde{P}'_{r,p}} \quad \text{... (14)}$$

حيث $\tilde{\mu}^C$ متجه الوسط التقليدي وان $\mathbf{M^C = I_n(\hat{\mu}^C)'}$ كما ان $r = \text{rank}(X_{n,p} - M^C) \le n-1$

هذه الخطوة مفيدة طالما $p > r$ وعندما $p >> n$ نحصل على تخفيض ضخم.

ثانياً R-Step تنجز لمصفوفة الهدف $\tilde{T}_{n,r}$ ووفقاً الى (9) نحصل على

$$\mathbf{(\tilde{T}_{n,r} - M_T^R\ \tilde{\tilde{P}}'_{k,r}) = T_{n,k}} \quad \text{... (15)}$$

ومع $k \le r$ وهي عدد المتغيرات التي نرغب بحسابها ، كما ان مصفوفة المراكز الحصينة هي $\mathbf{M_T^R}$ الى $\mathbf{\tilde{T}_{n,r}}$، وباستخدام المعادلات (9) و (15) وبواسطة الحسابات المتعامدة الى الوسيط L^1 نحصل على

$$\mathbf{T_{n,k} = (X_{n,p} - M^R)\ P_{p,k}} \quad \text{... (16)}$$

ومع $\mathbf{P_{p,k} = \tilde{P}_{p,r}\ \tilde{\tilde{P}}_{r,k}}$ و $\mathbf{M^R = M^C + M_T^R \tilde{P}'_{r,p}}$ وان الاعمدة $P_{p,k}$ تبقى متعامدة طالما ان مضروب المصفوفة $\mathbf{P_{p,k} = \tilde{P}_{p,r}\ \tilde{\tilde{P}}_{r,k}}$ يحافظ على التعامدية.ان خوارزمية الخطوتين تسمى Improved Reflection-Based Algorithm for Principal Component Analysis (IRAPCA) وهي نسخة أسرع من طريقة R-Step.

الجانب التطبيقي

١- عينات البحث والمتغيرات المستخدمة

العينة الأولى جمعت من أشخاص يعملون في معامل الألومينيوم إذ اخذت عينة من دمهم وبعد تحليل عينات الدم من قبل أطباء اختصاص حددوا نسبة الألومينيوم لكل حالة، اما بقية المعلومات تم الحصول عليها من خلال استمارة تضمنت المتغيرات ذات العلاقة بارتفاع

نسبـــة الألومينيوم في الدم وباستشارة اختصاصيين إذ اعطيت لكل شخص تم أخذ عينة من دمه وكان حجم العينة 50 اما المتغيرات المستخدمة فهي:

1 . تركيز الألومينيوم في الدم ويمثل متغير الاستجابة y.

2 . المتغيرات التوضيحية وهـــي: X_1 العمر، X_2 سنوات العمل في معامل الألومينيوم، X_3 الأواني المستخدمة في اعداد الشاي وهي(الستيل،الألومينيوم)، X_4 الاطعمة التي يعاد طبخها (Aubergine ,Okra)،X_5 كمية شرب الشاي، X_6 طريقة أعداد الشاي وهي(.Normal,Rep)، X_7 نسبة تركيز الشاي، X_8 الاواني المستخدمة في الطبخ وهي(الستيل،الألومينيوم)، X_9 الامراض المصاب بها وهي(Epilepsy, Alzheimer, Parkinson, no disease)، X_{10} الجنس.

العينة الثانية أخذت من اشخاص عاديين لا يعملون في معامل الألومينيوم وكانت العينة بحجم 70 وإخذت بنفس أسلوب العينة الأولى من تحليل الدم بالاضافة الى الاستمارة، اما المتغيرات المستخدمة فهي:

1. تركيز الألومينيوم في الدم ويمثل متغير الاستجابة y.

2. المتغيرات التوضيحية وهي: X_1 العمر،X_2 الجنس،X_3 المهنة، X_4 الأواني المستخدمة في اعداد الشاي وهي(الستيل،الألومينيوم)،X_5 كمية شرب الشاي،X_6 نسبة تركيز الشاي، X_7 طريقة أعداد الشاي وهي (.Normal, Rep)، X_8 الاواني المستخدمة في الطبخ وهي (الستيل ،الألومينيوم)، X_9 الاطعمة التي يعاد طبخها(Aubergine ,Okra)، X_{10} الامراض المصاب بها وهي (Epilepsy, Alzheimer, Parkinson, no disease).ان المتغيرات أخذت بناءاً على ما توصل اليه الباحثين لعوامل تؤثر على ارتفاع نسبة الألومينيوم في الدم كالأواني المستخدمة في الطبخ وكذلك طريقة أعداد الشاي ونوع الأواني المستخدمة في أعداده، إضافة الى الطريقة الخاطئة في اعداد بعض الاطعمة مثل (Aubergine,Okra) .

٢- الاختبارات

أولاً: تم اختبار البيانات لمعرفة هل ان المتغيرات تتوزع طبيعياً أم لا من خلال جودة توفيق البيانات في برنامج Matlab **(1997,Hahn)** والتي تعتبر افضل اختبارات لجودة التطابق مقارنة مع اختبارات لتطبيقات اخرى، والاختبارات هي (Iillietest) فعندما H = 0 تقبل الفرضية حيث العينة تتوزع توزيع طبيعي اما H = 1 فتعني رفض الفرضية كما يعطي الاختبار قيمة p التقريبية approximate p-value والذي هو مستوى المعنوية عندما ترفض

H_o ، الاختبار الاخر (jbtest) ويعطي قيمة p p-value ، حسب الاختباران للعينتين الأولى والعينة الثانية إذ أظهـــرت قيم الاختبار ابتعاد بعض المتغيرات عن التوزيع الطبيعي نتيجة لوجود القيم الشاذة.

ثانياً : تم الكشف عن وجود القيم الشاذة بشكل أولي بطريقة Box-and Whisher Plot في البرنامج statgraph حيث أظهرت النتائج بأن المتغيرات أعلاه تحتوي على قيم شاذة و بأعداد متفاوتة و من طرف واحد و هناك متغيرات لم تظهر فيها قيم شاذة .

٣- تفسير النتائج لعينات البحث

تم تنفيذ التحليل على العينات قيد البحث باستخدام برنامج Matlab الذي يعتبر من البرامجيات القابلة للبرمجة و ذو إمكانية عالية في التحليلات الاحصائية المتقدمة بدرجة عالية .

1. لتشخيص القيم الشاذة في مجموعة البيانات تم استخدام نقاط المسافات و لكل نقطـــة بيانـــات قيست المسافة العادية و الحصينة و التي تعطى بالشكل

$$CD_i = \sqrt{\sum_{j=1}^{k}\left(\frac{t^{c}_{ij}}{S^{c}_{j}}\right)^2} \quad ; \quad RD_i = \sqrt{\sum_{j=1}^{k}\left(\frac{t^{R}_{ij}}{S^{R}_{j}}\right)^2} \quad ...(17)$$

k تمثل عدد المكونات المختارة، و أن المشاهدات التي فيها المسافات تتجـــاوز نقطـــة القطـــع $\sqrt{\chi^2_{k,0.975}}$ يشار إليها كشواذ.

2. بالنسبة للعينة الأولى نجد أن مقدّر IRAPCA كشف عن 8 من المشاهدات الشـــاذة ضـــمن المسافة الحصينة، أما العينة الثانية نجد مقدّر IRAPCA كشف عن 13 مشاهدة شـــاذة ضـــمن المسافة الحصينة، أما مقدّر PCA فلم يكشف عن أية مشاهدة شاذة ضمن المسافة العادية وعـــن مشاهدتين شاذة و للعينتين الأولى و الثانية على التوالي.

3. بالنسبة إلى العينة الأولى لوحظ أن هناك خمسة عوامل تفسر العلاقة بين المتغيرات باستخدام مقّدر IRAPCA و بعد استخراج مرتبة العامل لكل مشاهدة تم حساب الارتباط بين كل واحـــد من العوامل الخمسة الحصينة مع المتغير المعتمد نسبة الألومينيوم في الدم و كما يلي :

$$r_{1y} = 0.9278,\ r_{2y} = -0.8763\ ,\ r_{3y} = 0.9031,\ r_{4y} = 0.7899\ ,\ r_{5y} = 0.8632$$

أي أن العوامل الخمسة هي مهمة في تفسير العوامل المسببة لارتفاع نسبة الألومينيوم في الدم.

أما العينة الثانية لوحظ أن هناك أربعة عوامل تفسر العلاقة بين المتغيرات باستخدام مقدّر IRAPCA و بعد استخراج مرتبة العامل لكل مشاهدة تم حساب الارتباط بين كل من العوامل

الاربعة الحصينة مع المغير و كما يلي:

$$r_{1y} = 0.9082 \; , \; r_{2y} = 0.8261 \; , \; r_{3y} = 0.9108 \; , \; r_{4y} = 0.8621$$

أما بالنسبة لمقدّر PCA لوحظ أن هناك ثلاثة عوامل رئيسية تفسر العلاقة بين المتغيرات للعينة الأولى و بعد استخراج مرتبة العامل لكل مشاهدة تم حساب الارتباط بين كل واحد من العوامل الثلاثة مع المتغير المعتمد نسبة الألومينيوم في الدم و كما يلي :

$$r_{1y} = 0.410 \; , \; r_{3y} = 0.509 \; , \; r_{2y} = -0.364$$

أي أن تفسير العوامل الثلاثة للمتغيرات المسببة لارتفاع نسبة الألومينيوم هو ضعيف و هذا يرجع إلى أن مقدّر PCA غير مقاوم للقيم الشاذة و النتائج التي يعطيها غير كفوءة مقارنة مع مقدّر IRAPCA .

و للعينة الثانية لوحظ أن هناك أربعة عوامل رئيسة تفسر العلاقة بين المتغيرات باستخدام مقدّر PCA وبعد استراج مرتبة العامل لكل مشاهدة تم حساب الارتباط بين كل من العوامل الأربعة مع المتغير المعتمد و كما يلي:

$$r_{1y} = -0.349,0.277 \; , \; r_{1y} = 0.476 \; , \; r_{1y} = 0.533$$

ايضاً تفسير العوامل الأربعة للمتغيرات المسببة لارتفاع نسبة الألومينيوم هو ضعيف و لنفس السبب أعلاه.

الاستنتاجات

من خلال عينات البحث تم التوصل للاستنتاجات التالية:

1 . أعطى مقدّر IRAPCA اعلى نسبة تباين من التباين الكلي مقارنة مع مقدّر PCA ويلاحظ ان نسبة الانهيار فيما يخص المقدر الأول بلغت 50% وللعينتين.

2 . ان مقدّر IRAPCA استطاع ان يكشف عن عدد اكبر من الشواذ وهي ثمانية وثلاثة عشر ضمن المسافة الحصينة للعينتين الأولى والثانية على التوالي اما مقدّر PCA فلم يكشف عن اية مشاهدة شاذة في العينة الأولى بينما كشف عن مشاهدتين ضمن العينة الثانية وباستخدام المسافة الاعتيادية.

3 . بالنسبة للتباينات المحسوبة مع $IRAPCA_{full}$ هي اكبر من PCA_{red} وهذا بسبب الشواذ هي أكثر تشتتاً من البيانات الاصلية حيث تزداد Q_n عند مجموعة البيانات الكاملة.

اما $IRAPCA_{red}$ وعند تطبيقها على مجموعة البيانات المختزلة (غير الملوثة) يلاحظ ان التباينات في هذه الحالة تكون قريبة من التي حصلنا منها من PCA_{red} .

4 . باعتماد مقدّر IRAPCA وجد ان كل من العوامل المدورة الحصينة مرتبطة بقوة مع المتغير المعتمد وللعينتين، اما الارتباطات بين كل من العوامل المدورة مع المتغير المعتمد وباعتمـــاد المقدّر الاعتيادي PCA فكانت ضعيفة بسبب عدم مقاومة الاخير للقيم الشاذة.

التوصيات

من خلال الاستنتاجات يمكن ادراج التوصيات التالية:

1 . اعتماد مقدّر IRAPCA بدلاً من مقدّر PCA في حالة احتواء البيانات علـــــى قيم شاذة كون الاعتماد على الاخير في تفسير النتائج سيعطي نتائج مظللة.

2 . الاعتماد على الخوارزمية المحسنة كونها تعمل جيداً في حالة الابعاد العالية وذلك من خلال تخفيض الابعاد الى مجال جزئي كذلك فانها تحقق الاستقرار العددي.

جدول (1) : نتائج التباين الحصين لمجموعة بيانات العينة الأولى لكل من PCA_{full} ، $IRAPCA_{full}$ ،**PCA_{red}** ،**$IRAPCA_{red}$**

	PCA_{full}	**$IRAPCA_{full}$**	**PCA_{red}**	**$IRAPCA_{red}$**
s_1^2	**336.421**	**3.048**	**2.227**	**2.693**
s_2^2	**6.546**	**2.346**	**1.102**	**1.724**
s_3^2	**0.851**	**0.776**	**0.225**	**0.308**

جدول (2) : نتائج التباين الحصين لمجموعة بيانات العينة الثانية لكل من PCA_{full} ، $IRAPCA_{full}$ ،PCA_{red} ،$IRAPCA_{red}$

	PCA_{full}	**$IRAPCA_{full}$**	**PCA_{red}**	**$IRAPCA_{red}$**
s_1^2	**245.81**	**4.327**	**3.321**	**3.687**
s_2^2	**5.317**	**2.282**	**2.001**	**2.169**
s_3^2	**1.485**	**1.103**	**0.359**	**2.225**

جدول (3): نتائج العينة الأولى

Criteria	**PCA**	**IRAPCA**
α	-	**0.50**
Mean(λ_i)	**1.312**	**1.692**
$\sum(\lambda_i)$	**3.937**	**8.462**
total $\sum_{i=1}^{k}\lambda_i$	**50.6**	**78.6**

جدول (4): نتائج العينة الثانية

Criteria	**PCA**	**IRAPCA**
α	-	**0.50**
Mean(λ_i)	**1.332**	**1.516**
$\sum(\lambda_i)$	**3.995**	**6.065**
total $\sum_{i=1}^{k}\lambda_i$	**53.1**	**79.0**

جدول (5): نتائج المسافات للعينة الأولى

المشاهدة	IRAPCA (RD)	PCA (CD)
1	2.369	0.862
2	6.193	3.687
3	3.753	2.432
4	3.792	3.224
5	2.846	2.003
6	4.445	2.521
7	3.805	3.306
8	4.227	0.441
9	9.759	1.102
10	3.003	1.004
11	2.156	2.167
12	2.334	1.156
13	3.257	0.981
14	5.406	1.555
15	3.499	1.361
16	33.534	2.151
17	3.361	2.838
18	35.419	2.104
19	4.968	2.242
20	44.724	3.275
21	34.217	2.066
22	3.475	0.879
23	2.521	1.016
24	7.126	2.541
25	29.368	3.063
26	5.700	3.149
27	3.826	3.545
28	6.634	1.414
29	18.824	2.530
30	7.484	1.437
31	4.324	3.294
32	4.724	1.380
33	28.557	0.029
34	2.683	2.910
35	4.205	2.609
36	9.575	2.636
37	46.955	5.877
38	3.403	2.461
39	5.986	2.589
40	4.768	3.089
41	4.500	3.759
42	3.719	2.686
43	6.697	4.402
44	8.216	2.790
45	9.058	3.095
46	4.749	2.203
47	3.353	0.606
48	3.468	1.206
49	4.111	2.311
50	7.961	3.022

جدول(6): نتائج المسافات للعينة الثانية

المشاهدة	IRAPCA RD	PCA CD	المشاهدة	IRAPCA RD	PCA CD
1	4.724	3.142	54	6.853	1.514
2	3.775	2.292	55	4.625	3.242
3	6.112	4.349	56	6.791	6.331
4	5.683	0.655	57	5.957	0.874
5	5.608	2.311	58	5.835	3.219
6	7.934	4.587	59	7.835	4.581
7	24.093	5.574	60	6.950	2.895
8	8.169	2.507	61	39.492	2.639
9	3.825	4.210	62	3.748	2.093
10	4.896	3.686	63	5.579	1.847
11	30.248	1.052	64	3.303	2.924
12	2.568	2.448	65	4.134	6.752
13	4.684	3.031	66	5.938	2.986
14	3.617	1.826	67	5.177	3.264
15	17.293	4.412	68	5.598	3.286
16	4.807	3.315	69	6.367	2.958
17	18.876	0.743	70	2.347	3.310
18	3.725	2.713			
19	3.778	4.131			
20	51.397	2.312			
21	2.552	2.760			
22	3.466	5.756			
23	4.625	5.065			
24	4.019	3.383			
25	24.969	14.020			
26	3.925	4.395			
27	59.802	3.052			
28	9.725	1.525			
29	3.229	0.552			
30	3.475	2.146			
31	4.521	3.317			
32	39.126	1.479			
33	7.668	3.777			
34	3.826	2.381			
35	3.771	2.701			
36	23.182	1.388			
37	2.284	0.594			
38	7.324	3.060			
39	7.971	0.347			
40	4.819	3.585			
41	3.279	1.166			
42	2.392	3.762			
43	4.449	0.821			
44	6.747	3.838			
45	4.791	4.759			
46	43.847	9.563			
47	5.723	5.416			
48	15.775	3.634			
49	39.904	3.961			
50	4.689	1.656			
51	5.248	2.583			
52	4.568	3.239			
53	3.617	2.985			

المصادر

- Hahn, B. D.,(1997): *Essential matlab for scientists and Engineers*, Wily , New York.
- Harmen, H.,(1976): Modern Factor Analysis, The university of Chicago,Press, London.
- Hubert, M., Rousseeuw, P.J. & Verboven, S., (2002): A fast method for Robust Principal components with applications to chemometrics, *Chemometrics and Intelligent Laboratory Systems*,Vol.60,pp.101-111.
- Li, G. & Chen, Z., (1985): Projection-Puruit Approach to Robust Dispersion Matrices and Principal Components: Primary theory & Monte Carlo, *JASA*, Vol. 80, pp. 759-766.
- Marrison, D.F., (1976): *Multivariate Statistical Method* , McGraw Hill New York.
- Norusis, M.,(1986): User Guide SPSS/PC^{+} for IBM, Chicago (Manual)
- Rousseeuw, P.J. & Katrin van Driessen (1999): A fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrices* Vol.41, pp. 212-233.
- Rousseeuw, P.J. & Mia Hubert (2002): ROBPCA: A New Approach to Robust Principal Component Analysis , *Submitted*.

The fastest algorithm for analyzing robust principal components with application on variables affecting the increase of aluminum level in blood Abstract

Khlod Y.Khmo and Dhafer H.Rashid
***College* of Administration and Economics**

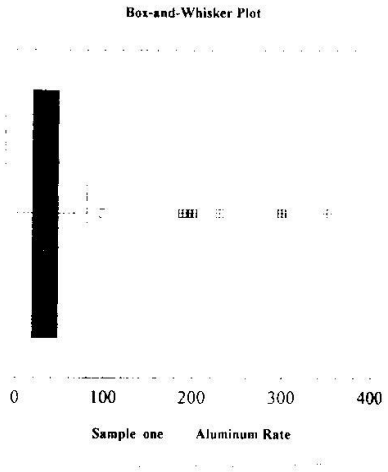
Abstract

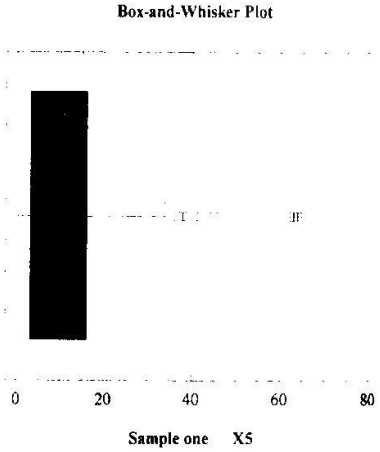
One of the common techniques in analyzing multivariate data is analyzing of the principal components. This transforms large number of related variables into lesser number of no related components (PCA).

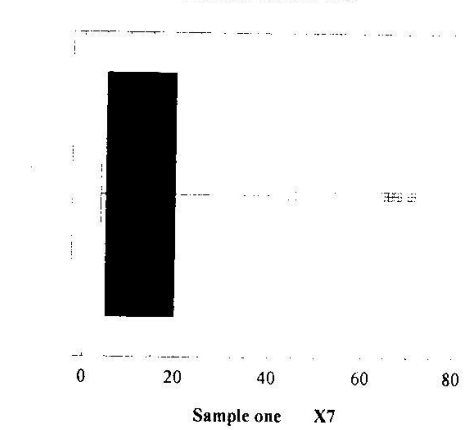
In case of existence of outliers, which can be detected in several ways, then the dependence of variance and ordinary covariance matrices, also correlation matrix, would lead to misleading results in analyzing the principal components.

The aim of this research is to introduce a new and fast algorithm in analyzing the robust principal components; when data contain outliers, while conventional methods fall in detecting outliers in data; then the results are misleading. The method is implemented to show its real effectiveness on variables affecting the increase of aluminum level in blood.

شكل (1) : الكشف عن وجود القيم الشاذة بطريقة Box &Whisher plot لبعض متغيرات العينة الأولى







شكل (2) : الكشف عن وجود القيم الشاذة بطريقة Box &Whisher plot لبعض متغيرات العينة الثانية