١ -المقدمــة

واحد من اكثر التقنيات شائعة الاستخدام في تحليل البيانات متعددة المتغيرات هو تحليل المركبات الرئيسية Principal Component analysis (PCA) والذي يستخدم لتحليل عدد كبير من المتغيرات المرتبطة بهدف اختصارها الى عدد اقل من العوامل غير المرتبطة يعزى لها تباين تلك المتغيرات، وهي تتضمن حساب القيم الذاتية Eigen Values لمصفوفة التباين المشترك Covariance matrix او مصفوفة الارتباط Correlation Matrix مع ان هذا التحليل يكون حساس بتطرف للمشاهدات الشاذة outliersوبالتالي فان النتائج المسحوبة لتحليل المركبات الرئيسية (PCA)سوف تكون مضللة وغير دقيقة.

لتلافي هذه المشكلة ولتعزيز الثقة بالنتائج والمقدرات المسحوبة يتم الاعتماد على مقدرات حصينة تحسب من خلال خوارزميات معينة تساعدنا في التغلب على مشكلة وجود الشواذ او الملوثات ضمن مجاميع العينات المدروسة وخاصة عندما يكون حجم العينة صغير وبالتالي فان نسبة صغيرة من الشواذ تؤدي الى تشويه النتائج وخاصة بالنسبة لمقدرات المعالم وخاصة عند تقدير معلمتي التموضع Location والقياس Scale حيث يتم التأثير على مصفوفة التباين المشترك ومنها الى مصفوفة الارتباط. وذلك يعني ان وجود الشواذ في مجموعة البيانات data set يؤدي الى تشويه العلاقة بين المتغيرات المدروسة.

تجدر الاشارة هنا الى ان هناك عدة باحثين اهتموا بدراسة المقدرات الحصينة امثال Groux , Ruiz – Gazen , Li & Chen وآخرون، ولكن بداية كان الاهتمام منصب على الحالة العامة للبيانات أي الحالة التي يكون فيها عدد المشاهدات (n) اكبر من عدد المتغيرات (P) أي (n > p) اما مايسمى بالبيانات تامة الرتبة Full rank data set .

حقيقة غالبا" مانواجهها عندما نتعامل مع بيانات بابعاد عالية وهي ان يكون عدد الابعاد او المتغيرات اكبر من عدد المشاهدات (p > n)ففي هذه الحالة سوف يتم اسقاط كل الطرق العادية والحصينة لانها بشروطها الموضوعة تنتج مالانهاية من الحلول لذا اصبحت عملية البحث عن طريقة تضع مقدرات حصينة الموقع والقياس والتشتت في حالة كون (p > n) اما مايسمى بالبيانات الناقصة او مجموعة البيانات غير تامة الرتبة Redused data or less than full rank data set نتيتجه الدراسة والاهتمام المتواصل تم التوصل الى عدة خوارزميات طورت جيل بعد آخر لتستطيع التعامل مع بيانات بهكذا ابعاد الى ان تم التوصل الى احدث خوارزمية تتعامل مع هذه البيانات وهي خوارزمية اسرع خطوتين – A fates two step algorithm والتي عالجت هذه المشكلة بجدارة.

مع دراستنا المستمرة واهتمامنا بالابعاد العالية وخاصة انها مفتاح البحوث في جميع المجالات الرياضية والطبية والاجتماعية والاقتصادية وغيرها ومن خلال الاتصال بالباحثين Prof. Croux, C. , prof. Haesbroek , G . & prof, Verboven, S. تم التوصل الى طريقة حديثة في التعامل مع البيانات الناقصة وخاصة بوجود الملوثات. تكمن هذه الطريقة في توظيف خوارزمية اسرع خطوتين كمرحلة اولى ضمن التعامل مع البيانات الناقصة النظيفة والملوثة لنحصل منها على مابسمى بمصفوفة العلامة Score matrix والتي من خلال واجراء التحليلات العادية والحصينة عليها يتم الحصول على نتائج ذات وثوقية اكبر وهي فكرة بحثنا.

٢- التحليل العاملي الحصين

يعمد التحليل العاملي الى تحليل البيانات المتعلقة بمجموعة من المتغيرات كأن تكون اتجاهات الافراد المستخلصة من صحيفة الاستبيان، حيث ترتب مشاهدات العينة بمصفوفة تسمى مصفوفة البيانات data matrix تمثل صفوفها الافراد وتمثل اعمدتها المتغيرات المدروسة. في هذا البحث سنعتمد على احدى اهم طرق التحليل العاملي وهي طريقة تحليل المركبات الرئيسية (PCA) التي تسمح بايجاد ابعاد واطئة مهمة تمثل المتغيرات المدروسة هذه الابعاد توضح هيكل المجتمعات للاشياء المعتمدة.

مثل بقية التقنيات العديدة فان كل من المدخلات والنتائج لتحليل المركبات الرئيسية ممكن ان تتأثر بالمشاهدات الشاذة القليلة كون القيم الذاتية التي يمكننا من خلالها ان نقيس كمية المعلومات الموضحة بواسطة المركبات الرئيسية تكون حساسة تجاه الشواذ او القيم المتطرفة Extreme Values.

نرى الان ان اختيار مصفوفة تباين مشترك حصينة Robust Covariance matrix اصبح ضروريا" لتحليل المركبات الرئيسية، علما" ان لأي اختيار لهذه المصفوفة يقود الى تحليل آخر للمركبات الرئيسية.

لاكتشاف الشواذ يتم الاعتماد اما على دالة التأثير Influence Function او على المسافات الحصينة Robust distance ولكن من الجيد معرفة ان دالة التأثير تقيس الحصانة فيما يتعلق بالشواذ الفردية أي انها تفيد في تحديد أي النقاط هي اكثر تأثيرا" عن البقية وهذا غير ملائم لأكتشاف الشواذ، لذا يبدو من الافضل استخدام المسافات الحصينة والتي هي نسخ حصينة لمسافات مهلونوبس التربيعية Squared Mahalonobis distances حيث

$$RDi = \sqrt{(x_i - t_n)' C_n^{-1} (x_i - t_n)} \qquad \text{.....(1)}$$

مع التقديرات الحصينة Robust Estimators للموقع والتشتت (t_n, C_n)مع ملاحظة ان مسافات مهلونوبس التقليدية هي ليست ملائمة لاكتشاف الشواذ لانها تعتمد على التقديرات العادية (غير الحصينة). معظم الطرق الحصينة لتحليل متعدد المتغيرات اقترحت من قبل (Jackson, pp 17-365) عام ١٩٩١. خلال هذه التقنية تبدل مصفوفة التباين المشترك العادية او مصفوفة الارتباط بالتقدير الحصين، وهذا الاجراء ربما يكون الاكثر شيوعا" في التطبيق.

٢-٢ مقدرات M الحصين Robust M estimator

واحدة من اكثر الطرق استخداما" في ايجاد المقدر الحصين في مجال تحليل متعدد المتغيرات هو مقدر (M) الحصين (Maronna 1976) والذي سنوجز طريقة ايجاده بما يلي:

لتكن مجموعة المشاهدات او العينة المدروسة بالابعاد (P) والمشاهدات (n) $x_1, x_2, \ldots, x_n$ معرفة بواسطة:

$$d(x_i, t, c) = \left\{ (x_i - t)' \, C^{-1} (x_i - t) \right\}^{1/2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

المسافة الاحصائية بين x_i , t وهي مقاسة بواسطة المصفوفة C موجبه التحديد. تقدير (M) يعرف بالشكل التالي:

$$t_n = \frac{\sum_{i=1}^{n} w_1\{d(x_i, t_n, c_n)\} x_i}{\sum_{i=1}^{n} w_1\{d(x_i, t_n, c_n)\}} \qquad \ldots\ldots\ldots\ldots\ldots(3)$$

$$C_n = \frac{1}{n} \sum_{i=1}^{n} w_2 \left\{ d^2(x_i, t_n, c_n) \right\} (x_i - t_n)(x_i - t_n)' \qquad \ldots\ldots\ldots\ldots\ldots(4)$$

حيث w_1 وw_2 هي دوال الاوزان الخاصة بهذا المقدر.

(Maronna 1976) بين ان طريقة ايجاد المقدر (M) تكون اقل واقل حصانة بزيادة الابعاد (وهذا ماسنثبته ضمن الجانب العملي لهذا البحث) لان نقطة الانهيار لهذا المقدر هي على الاغلب مساوية لـ $\frac{1}{p+1}$ (عرف Groux عام ١٩٩٩ نقطة الانهيار لتقدير التشتت في متعدد المتغيرات على انها اصغر كسر للمشاهدات والذي نحتاجه لابدال مواقع اعتباطية قبل تقدير التشتت أي اكبر قيمة ذاتية ترسل الى المالانهاية او اصغر قيمة ذاتية ترسل الى الصفر).

بالنسبة لهذا المقدر فان دوال الاوزان معطاة من قبل Huber بالشكل التالي:

$$w_1(y) = \varphi_H\left(y, q_t^{1/2}\right)/y \qquad .....(5)$$
$$w_2(y) = \varphi_H\left(y, q_t\right)/\beta y \qquad .....(6)$$

$$\varphi_H(y,k) = Max\{-k, Min(y,k)\} \qquad .....(7)$$

حيث:

β :هـي ثابـت بجعـل مقـدرات التشـتت متسـقة للنمـاذج الطبيعيـة. علمـا" ان التطبيـق والاجـراء التكراري في حساب هذا المقدر معطاة من قبل (Devlin 1981).

الملاحظ عند تطبيق خوارزمية ايجاد المقدر (M) الحصين في حالة كون عدد المتغيرات اكبر من $n(1-\alpha)$ مع نقطة انهيار مساوية الى (α) فسوف تكون هذه الخوارزمية غير قابلة للتطبيق مثالها في ذلك مثال مقدر (S) الحصين ومقدر Minimum Covariance determent estimator(MCD) الحصين ومقدر Minimum volume ellipsoid estimator (MVE) الحصين وغيرها ولذلك ففي حالة التعامل مع البيانات ذات الابعاد العالية وخاصة في حالة كون عدد المشاهدات عدد صغير لابد من ايجاد اجراء معين نستطيع من خلاله تحويل تلك البيانات الى صيغة معينة تخضع لشروط ايجاد المقدرات الحصينة كافة (او اغلبها) نستطيع من خلالها الوصول الى نتائج يمكن الوثوق بها والاعتماد عليها من صياغة مجموعة من التوصيات تكون ذات اهمية للعاملين في هذا المجال، ولهذا فقد انصب اهتمامنا في هذا البحث على هذه الحالة وهي الوصول الى خوارزمية تستطيع التعامل مع البيانات الناقصة لتعكس صورة معينة تكون مفتاح الدخول الى الخوارزميات الحصينة.

٢-٣ مقدر (S) الحصين Robust S estimator

يعد مقدر (S) من اهم المقدرات الحصينة للموقع والتشتت حيث انه يعود الى تصنيفات عالية الانهيار لمواقع متعددة المتغيرات والتشتت.

اوضح Croux في ابحاثه للاعوام ٢٠٠٠-٢٠٠٧ ان اهم طرق التقديرات الحصينة هي تقديرات (S) حيث انها تدمج بين الكفاءة العالية والخصائص الحصينة الجيدة (الانهيار العالي مع خاصية تساوي التغاير والاتساق). نقطة الانهيار لمقدر (S) هي ٥٠% (مع ملاحظة ان الكفاءة تتناقص عندما تزداد نقطة الانهيار) ان المقدرات التي تتمتع بخاصية تساوي التغاير فهي المقدرات التي لاتتغير بالتحويلات التي تجري على الموقع والقياس معا" بالشكل $\underline{x} \to A\underline{x} + \underline{b}$ وان ذلك يقود الى جسم قطع ناقص حجمه مساوي الى الحجم الاصلي مضروبا" في محددة المصفوفة A.

مقدرات (S) معرفة كحلول للموقع والشتت (t_n, c_n) لمسألة تقليل محددة مصفوفة التباين المشترك c وفقا" الى

$$\frac{1}{n}\sum_{i=1}^{n}\rho\{d(x_i,t,c)\}=b_0 \quad \text{.....(8)}$$

خلال كل $(t,c)\in R^p XSPD(p)$ حيث $SPD(p)$ مجموعة كل المصفوفات المتماثلة الموجبة قطعا" symmetric positive definite ذات البعد (p x p)، وان $d(x_i,t,c)$ هي المسافات العادية للمشاهدات في العينة.

شروط الدالة ρ هي ان تكون متماثلة وتزداد في الفترة R^+ أي انها غير متناقصة في الفترة $]0,\infty]$ واخيرا" فان $\rho(0)=0$

الدالة ρ اخذت كدالة وزن توكي Tuckey Biweight function وهي بصيغة Croux(2001):

$$\rho(x)=\min\left(\frac{x^2}{2}-\frac{x^4}{2C^2}+\frac{x^6}{6C^4}\,,\,\frac{C^2}{6}\right) \quad \text{.....(9)}$$

ولتحقيق نقطة الانهيار الى ٥٠% فان C يكون

$$\rho(C)=b/r \quad \text{.....(10)}$$

حيث $r = 0.5$.

علما" ان طريقة ايجاد هذا المقدر هي طريقة تكرارية معقدة نوعا" ما ولكن تبقى الخصائص الحصينة الجيدة التي تمتلكها هذه المقدرات سببا" في جاذبية هذه الطريقة.

٣- طريقة الانعكاس في تحليل المركبات الرئيسية

Reflecation Method in Principal Component Analysis

في الفقرات السابقة لهذا البحث كنا قد ذكرنا ان هناك عدة طرق تعمل على تحصين مصفوفة التباين المشترك التابعة لمجاميع البيانات الملوثة، ولكن للاسف فان هذه الطرق جميعا" تستخدم في حالة كون عدد المشاهدات (n) اكبر من عدد المتغيرات (p) او البيانات تامة الرتبة، اما في حالة التعامل مع مجموعة بيانات لاتمتلك هذه الصفة فسوف يتم اسقاط كل الطرق آنفة الذكر والبحث عن طريقة توجد مقدرات حصينة للموقع والشتت في هذه الحالة.

بداية التعامل مع البيانات الناقصة كان العمل بخوارزمية Croux & Ruiz-Gazen وبعدها تم العمل بالخوارزمية المحسنة An Improved Algorithm لنصل الى آخر ماتم

العمل به في حالة التعامل مع البيانات غير تامة الرتبة وهي خوارزمية اسرع خطوتين A fast two-step Algorithm.

طالما ان وقت الحسابات العالي يحدث بسبب عدد الابعاد (قيمة p) فان خوارزمية اسرع خطوتين تبدأ بعملية تخفيض الابعاد للبيانات الى فضاء جزئي بواسطة المشاهدات (n) وهذا نستطيع عمله اسرع وبأتقان اكثر بواسطة تحليل المركبات الرئيسية التقليدية (PCA) أي استخدام جوهراولب النتيجة لتحليل القيمة الذاتية وبعدم فقدان أي معلومات في البيانات.

الخطوة الأولى ضمن هذه الخوارزمية ان البيانات سوف تتمركز Centered حول $L'-Median\ \hat{\mu}^R$ والذي هو مقدر موقع متساوي التغاير متعامد حصين بدرجة عالية highly robust orthogonally equivariant location estimator ويعرف بالوسيط الخاص spatial median والذي ستحدد قيمته وفق θ التي تقلل مجموع المسافة الكلية لكافة المشاهدات

$$\hat{\mu}^R = \arg\min_{Q} \sum_{i=1}^{n} \|x_i - \theta\| \quad .....(11)$$

غالبا" هذا المقدر يمتلك صفة تساوي التغاير والذي سوف ترث خصائصه المحاور المشتقة عندما تختار تلك المحاور وتدور.

لتبسيط الخطوات سنعرف من الان المشاهدات المركزية Centered doservations بـ $X_i^{(1)}$ حيث ان:

$$X_i^{(1)} = X_i - \hat{\mu}^R \quad .....(12)$$

وسوف يكون:

$$M^R = 1_n \left(\hat{\mu}^R\right)' \quad .....(13)$$

حيث $1_n = (1,1,...,1)'$ متجه عمودي مركباته الـ (n) مساوية للواحد الصحيح.

لتكن (r) رتبة مصفوفة البيانات المركزية $r \le \min(n-1, p)$ واذا عرفنا $\rho_{P\times r}$ مصفوفة المتجهات الذاتية لتحويلات الوراء (الخلف) back transformed eigen vector سوف نحصل على التمثيل decomposition الحصين لـ $X_{n,P}$ (مصفوفة البيانات الاصلية) في مصفوفة العلامة Score matrix $T_{n\times r}$ ذات (n×r) و $P_{P\times r}$ مصفوفة المحملات matrix of Loadings ذات (P×r) يكون

$$T_{n\times r} = \left(X_{n,P} - M^R\right)P_{P\times r} \quad .....(14)$$

حيث

$$X_{n,P} - M^R = T_{n,r}\ P'_{r,P} \quad .....(15)$$

بهذا يمكننا تخفيض عدد المتغيرات الاصلية (P) بأعتبار فقط اول الاعمدة (k) من (P) حيث ان (k) يمثل عدد المركبات الرئيسية التي نود الاحتفاظ بها.

هذا النوع من الخوارزمية يكون مفيدا" اكثر في حالة تزايد الابعاد (P كبيرة) طالما نحن نحتاج فقط الى ايجاد V_K بدلا" من حساب كل (r) من المتجهات الذاتية وهذا يحفظ وقت كبير للحسابات.

وبسبب هذا الوقت العالي للحسابات فقد طورت هذه الخوارزمية حيث تتم العملية هنا بالشكل التالي:

$$X_{n,P} - M^C = \tilde{T}_{n,r}\ \tilde{P}'_{r,P} \quad \text{.....(16)}$$

حيث μ^C هو متجه الاوساط التقليدي:

$$M^C = 1_n \left(\hat{\mu}^C\right)' \quad \text{.....(17)}$$

وان:

$r = rank\left(X_{n,P} - M^C\right) \le n-1$. هذه الخطوة مفيدة جدا" حيث (P > r). وعندما (P >> n) فأننا نحصل على تخفيض كبير للابعاد. ونسبة الى (١٥) فأننا نحصل على:

$$\tilde{T}_{n,r} - M_R^T = T_{n,k}\ \tilde{\tilde{P}}'_{k,r} \quad \text{.....(18)}$$

مع $k \le r$ عدد المتغيرات اللاتينية Latent variabels التي نرغب في حسابها و T_R^T هي مصفوفة المراكز الحصينة لـ $\tilde{T}_{n,r}$ وبتعويض (١٨) في (١٦) ينتج اخيرا"

$$X_{n,P} = \left(T_{n,k}\ P'_{k,r}\ \tilde{\tilde{+}} M_T^R\right) P' \approx'_{r,P} + M^C \quad \text{.....(19)}$$

$$X_{n,P} = T_{n,k}\ \tilde{\tilde{P}}'_{k,r}\ \tilde{P}'_{r,P} + M^R \quad \text{.....(20)}$$

$$X_{n,P} = T_{n,k}\ P'_{k,P} + M^R \quad \text{.....(21)}$$

$$P_{P,k} = \tilde{P}_{P,r}\ P''_{r,k}$$

وفي الحقيقة من (١٦) ومن خاصية تساوي التغاير المتعامد للوسيط L' هذا ينتج:

$$M^R = M^C + M_T^R\ \tilde{P}'_{r,P} \quad \text{.....(22)}$$

كذلك نعرف ان اعمدة $P_{P,k}$ تبقى متعامدة طالما ان مضروب المصفوفة $P_{P,k} = \tilde{P}_{P,r}\ \tilde{\tilde{P}}_{r,k}$ يحافظ على التعامدية. هذه الخوارزمية عرفت لاحقا" بـ (RAPCA) والتي تمثل Reflecation-Based Algorithm for Principal components analysis وللتحقق من وجود الشواذ في مجموعة البيانات سوف تحسب المسافات العادية والحصينة المعرفة بواسطة $\left(T^R, S_j^R\right), \left(T^C, S_j^C\right)$ مصفوفة العلامات والجذر التربيعي للقيمة الذاتية j+h الحاصل عليها بواسطة (PCA) العادية وطريقة (RAPCA) الحصينة لذلك لكل نقطة في البيانات i فان المسافة التقليدية CDi والمسافة الحصينة RDi معطى بواسطة:

$$CDi = \sqrt{\sum_{i=1}^{k}\left(\frac{t_{ij}^{C}}{s_{j}^{C}}\right)^{2}} \quad \text{.....(23)}$$

$$RDi = \sqrt{\sum_{i=1}^{k}\left(\frac{t_{ij}^{R}}{s_{j}^{R}}\right)^{2}} \quad \text{.....(24)}$$

علما'' ان المشاهدات التي تزيد قيمة المسافة لها عن قيمة القطع cut-off value والتي تكون مساوية لـ $\sqrt{\chi^2_{k,0.975}}$ تعتبر مشاهدات شاذة. اما بالنسبة لـ t_{ij} في الـ (PCA) العادية فنحصل عليها من:

$$t_{ij} = \left(X_{n,P} - M^{C}\right)P_{p,r} \quad \text{.....(25)}$$

ان الفكرة التي طرحت في هذا البحث تقوم على اساس استخدام خوارزمية الانعكاس في تخفيض ابعاد البيانات غير تامة الرتبة للحصول منها على المصفوفة T والتي تكون بابعاد عادية أي ان عدد الصفوف فيها اكبر من عدد الاعمدة وبعد ذلك يتم تسخير الطرق العادية والحصينة المعروفة في حساب المقدرات لتلك المصفوفة. اننا بهذه الطريقة فقد قمنا بتجاوز عدة مشاكل قد يمر بها الباحث اولها التعامل مع بيانات يكون فيها عدد الصفوف اقل من عدد الاعمدة وهي مشكلة تقف كل الاجراءات والحسابات العادية عاجزة عن التعامل معها، ثانيها هو في حالة وجود الشواذ او الملوثات في تلك البيانات حيث وكما هو معروف كلما صغر حجم العينة كان تأثير الشواذ واضح اكبر وذا تأثير تشويهي على النتائج المستحصل عليها، وكيف اذا الحال اذا كان عدد المشاهدات صغير وتتعامل مع متغيرات كبيرة العدد أي بيانات ناقصة وملوثة !!!

٤- الجانب التطبيقي

لغرض تسليط الضوء على الافكار التي وردت ضمن الجانب النظري لهذا البحث فقد اعتمدنا على اسلوب المحاكاة Simulation او توليد البيانات التي تصب في مشكلة هذا البحث.

لقد اصبح الاعتماد على اسلوب المحاكاة امرا" مألوفا" في المجال الاحصائي عامة وفي مجال الحصانة والتشخيص خاصة ومن العوامل التي تزيد من اهمية استخدام المحاكاة في الدراسات المتعلقة بالحصانة والتشخيص ان المحاكاة تساعد في دراسة سلوك الطرائق الاحصائية في حالة التعامل مع العينات ذات الاحجام الصغيرة (والتي هي اساس مشكلة هذا البحث) وكذلك فان المحاكاة تساعد على التحكم بنسبة الشواذ $(\varepsilon;\ 0 \langle \varepsilon \langle 1)$ التي قد يرغب الباحث في درجها ضمن مفردات تلك العينات.

في الواقع العملي فان هناك مشاكل او دراسات تتعلق ببيانات يكون فيها عدد المتغيرات اكبر من عدد المشاهدات، ولاجل الحصول على بيانات لهكذا اوصاف فقد كان في النية دراسة سلوك العاملين ضمن منظمة الامم المتحدة United Nation (UN) العاملين في العراق لما تمتلكه بيانات هذه المجموعة في شروط تكون ملائمة لخواص البيانات اللازمة لدراستنا كون عدد العاملين ضمن تلك المنظمة عدد محدد من الاشخاص مع فرصة امكانية دراسة عدد كبير من المتغيرات التي من الممكن ان يتعرضوا لها مع مايمتلكونه من خصائص شخصية.

وبعد اعداد استمارة الاستبيان التي تشمل (٢٠) شخص مع (٤٠) متغير هي عبارة عن الاسئلة الشخصية والاسئلة الاخرى الموجهة اليهم فقد تعذر علينا الحصول على تلك البيانات بسبب رفض المبحوثين الاجابة على الاسئلة وملىء الاستمارات، لذا فقد تم اللجوء الى اسلوب المحاكاة في توليد البيانات.

ولتحقيق هدف البحث في التحقق من اداء الطرائق الحصينة وطريقة الانعكاس في التعامل مع البيانات الناقصة فقد تم توليد عينات بأبعاد $n_{25,30}$ أي بـ (٣٠) عمود مع (٢٥) صف. كذلك فقد اعتمدنا انموذجين مختلفين لهذه البيانات، الاول هو بيانات ذات توزيع طبيعي قياس standard normal distribution وهو الحالة المثلى للبيانات والانموذج الثاني هو الانموذج الملوث وبنسبة تلوث $(\varepsilon = 0.25)$ والجدول (١) ادناه يوضح الحالات المدروسة

جدول رقم (١)

تشكيلات الحالات التي افترضت في عملية المحاكاة

ε	**N**	**P**
0.0	25	30
0.25	25	30

ومنه نستطيع ان نرى اننا قمنا بدراسة اكثر من حالة ضمن عملية المحاكاة تم تنفيذ جميع الطرائق الاحصائية التقليدية والحصينة سابقة الذكر في كل حالة ليتسنى لنا عملية المفاضلة بين تلك المقدرات ضمن الخوارزميات المدروسة.

٥- سير عملية المحاكاة

لقد صيغت عملية المحاكاة المستخدمة في هذا البحث لتسير وفق الخطوات التالية:

١- تجري عملية تغذية برنامج المحاكاة بالمدخلات التي تحدد وفق التشكيلة المختارة من جدول (١).

٢- يجري توليد البيانات وفقا" للحالة المفترضة التي حددت سابقا" وفقا" لبرامج فرعية ثم ترسل البيانات المولدة الى البرنامج الرئيسي.

٣- مقدرات RAPCA للموقع والقياس تحسب من خلال تحديد نقطة الانهيار وعدد المركبات الرئيسية (k)المختارة من قبلنا. عملية اختيار القيمة (k) كان ليست بالطريقة العشوائية ولكن بصورة مدروسة حيث اننا ولاجل عدم التضحية بعدد كبير من المتغيرات نحدد k=n-1 أي اقل من عدد المشاهدات بعمود واحد فقط مع تحديد نقطة الانهيار بـ (٠.٥) لاستخراج النتائج. ان عملية اختيار k=n-1 تزيد من وقت الحسابات بأربعة اضعاف الوقت الذي تحسب قيمة مقدرات RAPCA الاصلية ولكن هدفنا من ذلك هو عدم اهمال أي متغير ممكن ان يكون ذا تأثير على طبيعة العلاقة بين المتغيرات المدروسة.

٤- تحسب التقديرات الحصينة للموقع والقياس كما يلي:

I. بالنسبة لمقدرات (S) الحصينة يتم اختيار (١٠٠٠) عينة عشوائية جزئية من مجموعة البيانات الاصلية تكون بحجم (P+1) يتم من خلالها الوصول للتقديرات الحصينة للموقع والقياس ضمن الخوارزمية العاملة لهذا المقدر حيث قمنا بتحديد نقطة الانهيار بـ (٠.٥).

II. بالنسبة لمقدر (M) الحصين فان اختيار العينات العشوائية الجزئية يتم بشكل تلقائي ضمن البرنامج الذي يقوم بحساب هذه المقدرات. نقطة الانهيار لمقدر (M) وكما قلنا سابقا" تصغر كلما كبر عدد الابعاد المتعامل بها والتي تؤول الى الصفر عندما يؤول حجم العينة الى المالانهاية.

٥- تكرار الخطوات ٢و٣ عدد N Run من التكرارات المحاكاة حيث استخدم التكرار (N Run =100).

٦- تحسب التقديرات الحصينة للموقع والقياس والتي تشتق من خلالها المركبات الرئيسية فضلا" عن المسافات التربيعية ليجري بعدها تصنيف المشاهدات.

وبعد الوصول الى هدف البحث في الوصول الى نتائج حصينة للبيانات غير تامة الرتبة (النظيفة والملوثة) فقد تم اختيار عدد من المعايير للمفاضلة بين الطرائق الحصينة المستخدمة وهي:

١. محددة المصفوفة Dterminat Matrix او D(S).
٢. مجموعة القيم الذاتية الفعالة التي تزيد قيمتها على الواحد الصحيح $\sum_{i=1}^{k}\lambda_i$.
٣. عدد المركبات الرئيسية الفعالة ومايقابلها من النسبة في تفسير التباين الكلي Tolerance.
٤. من معايير المفاضلة بين الطرائق الحصينة المختلفة نقطة الانهيار والمسجلة لكل مقدر لذا ستثبت كمعيار للمفاضلة.

٥-١ مناقشة نتائج المحاكاة للبيانات النظيفة

بالنسبة للانموذج الأول وهو انموذج البيانات الطبيعية القياسية (النظيفة) , N=25 P=30 فقد تم العمل وفقا" للمراحل التالية:

١) الخطوة الاولى لعملنا كانت اجراء تحليل المركبات الرئيسية العادي على مجموعة البيانات والذي اوضح ان محددة مصفوفة التباين لمجموعة البيانات تلك مساوية الى الصفر والذي ادى الى ايقاف العمل بخوارزميات ايجاد المقدرات (S) و (M) الحصينة.
٢) تحليل المركبات الرئيسية الذي اجري ضمن الخطوة الأولى اوضح لنا وجود (١٢) مركبة فعالة فسرت ٧٨.٨٢٧٠% من اجمالي التباين.
٣) بعد المرحلتين اعلاه تم العمل بخوارزمية RAPCA والتي وجدت للتعامل مع بيانات يكون فيها عدد المتغيرات اكبر من عدد المشاهدات.

العمل بخوارزمية RAPCA يتم اولا" بتحديد نقطة الانهيار وكما ذكرنا فقد حددناها بـ (٠.٥) والشئ الثاني هو تحديد (k) أي عدد اعمدة مصفوفة العلامة T. بهدف الابقاء على اكبر عدد من الاعمدة فقد حددنا (k=n-1=24) لتصبح المصفوفة T بعد اجراء تحليل RAPCA بدرجة 24 × 25 ونجد ايضا" ان محددة مصفوفة التباين لتلك المصفوفة مساوية الى (١.٨٣٢٧).
النتائج فسرت بـ (١٠) مركبات فعالة (القيمة الذاتية لها اكبر من الواحد الصحيح) وفسرت ٧٠.٣٩٢% من اجمالي التباين.

٤) هذه الخطوة هي مغزى هذا البحث وهي توظيف الطرائق الحصينة المعرفة على مجموعة البيانات الناقصة.

بعد اجراء تحليل RAPCA والحصول على المصفوفة T تم في هذه الخطوة اجراء تحليل (S) على تلك المصفوفة لنحصل على مقدرات (S) الحصينة للموقع والقياس يتم بعدها توظيف تلك المقدرات في عمل (PCA) والحصول على نتائج يعتمد عليها في صياغة القرارات المتعلقة بالبيانات المدروسة.

وجد ان det (S_s)=2.0720 اما النتائج فقد فسرت من خلال (١١) مركبة فعالة فسرت ٦٩.٨٧% من اجمالي التباين.

٥) نفس الخطوة السابقة تم اجراؤها هنا ولكن بدل العمل مع برنامج (S) تم العمل بخوارزمية (M) والوصول الى مقدرات (M) الحصينة للموقع والقياس. بداية وجد ان det (S_M)=2.1306 وان هناك (١٠) مركبات فعالة فسرت ٦٧.١٤٢٣% من اجمالي التباين.

٦) نقطة الانهيار لمقدر (S) كانت مساوية الى (٠.٥) وهي اعلى نقطة انهيار ممكن ان يصل لها المقدر لذا تسمى هذه المقدرات بمقدرات الانهيار العالي High Break down estimators. اما نقطة الانهيار لمقدر (M) فهي كما اوضحنا ضمن الجزء النظري تكون مساوية $\frac{1}{P+1}$ وهي تكشف عن ضعف هذا المقدر في هذه الخاصية وخاصة في حالة التعامل مع البيانات عالية الابعاد.

٧) لاكتشاف المشاهدات الشاذة فقد تم الاعتماد على مسافات مهلونوبس العادية والحصينة. بداية ولحساب مسافات مهلونوبس العادية فانه لم يتضح لنا وجود أي مشاهدة شاذة ضمن مجموعة البيانات لذلك كان تحليل المركبات الرئيسية مشابه في نتائجه لم تم التوصل اليه في الخطوة (٢).

٨) الخطوة المفترضة هنا في سياق هذا البحث هو اجراء جميع التحليلات سابقة الذكر بداية (PCA) العادي و RAPCA على مجموعة البيانات بعد استبعاد اثر المشاهدات الشاذة. ولكن وبما ان التحليل لم يكشف لنا عن أي مشاهدة شاذة فقد استمرت النتائج بنفس القيم (كما موضح في جدول ٢) ولم تختلف في شيء عن حالة التعامل مع البيانات ضمن المرحلة الاولى.

النتيجة التي تم الوصول لها هنا تكشف عن كفاءة مقدر RAPCAووصوله متفقا" مع المقدر العادي في عدم الكشف عن أي مشاهدة شاذة في حالة التعامل بالبيانات النظيفة وهذا ان دل على شيء فهو يدل على كفاءة تلك الخوارزمية في التعامل مع البيانات الناقصة والوصول بها الى نتائج تكون مشابهة لنتائج طرق التحليل التقليدية في حالة التعامل مع البيانات النظيفة

ولنتائج مغايرة لتلك الحاصل عليها في التحليل العادي في حالة التعامل مع البيانات غير تامة الرتبة والملوثة في آن واحد.

جدول رقم (٢)

نتائج المحاكاة للبيانات النظيفة

الطريقة	المعيار	n=25	P=30		ε=0.0
		PCA(full)	RAPCA(full)	PCA(red)	RAPCA(red)
Class	det(s)	0.0	1.8327	0.0	1.8327
	α	—	0.5	—	0.5
	$\sum_{i=1}^{k}\lambda_i$	38.2685	11.0301	38.2685	11.301
	$to\left(\sum_{i=1}^{k}\lambda_i\right)$	78.8270	70.3921	78.8270	70.3921
S	det(s)	—	2.0720	—	2.0720
	α	0.5	0.5	0.5	0.5
	$\sum_{i=1}^{k}\lambda_i$	—	10.3682	—	10.3682
	$to\left(\sum_{i=1}^{k}\lambda_i\right)$	—	69.8703	—	69.8703
M	det(s)	—	2.1306	—	2.1306
	α	$\frac{1}{13}$	$\frac{1}{25}$	$\frac{1}{13}$	$\frac{1}{25}$
	$\sum_{i=1}^{k}\lambda_i$	—	9.9113	—	9.9113
	$to\left(\sum_{i=1}^{k}\lambda_i\right)$	—	67.1423	—	67.1423

٢- مناقشة نتائج المحاكاة للبيانات الملوثة

كما ذكرنا سابقا" فأنه في حالة التعامل مع عينات صغيرة تكون هناك حساسية معينة في حساب المقدرات والمؤشرات اللازمة لاستخلاص النتائج. الحالة الاعقد اذا كان عدد المتغيرات لهذه العينات المدروسة اكبر من عدد المشاهدات والتي تقف عندها كل الطرق التقليدية العادية والحصينة عاجزة في حساب مؤشراتها. والاكثر تعقيدا" اذا كانت مجموعة البيانات الناقصة هذه تحوي على الشواذ ضمن مشاهداتها.

وايضا" وكما ذكرنا فان هدف هذا البحث هو التعامل مع هكذا انواع من البيانات والتغلب على الصعوبة المعروفة في حساب مؤشراتها العادية والحصينة والوصول الى نتائج في المفاضلة تكون قريبة الى ارض الواقع.

ولغرض تحقيق هذا الهدف فقد تم التعامل في المرحلة الثانية ضمن الجزء الكلي لهذا البحث مع بيانات غير نظيفة لوثت بنسبة ٢٥% في المشاهدات الاصلية، أي ان ٧٥% من البيانات ولدت بتوزيع طبيعي قياسي N(0,1) وان ٢٥% ولدت ضمنها توزيع طبيعي N(0.10).

ان الاستقلالية بين المفردات المولدة تضمن لنا ان يتم التلويث للبيانات بنفس التوزيع المولد وهو التوزيع الطبيعي بقيمة تباين تكون مغايرة للجزء المولد الأول، علما" ان تغيير الوسط الحسابي ضمن مرحلة التوليد تضمن ان يتم تحريك البيانات كليا" من مكان الى آخر ضمن الابعاد المدروسة اما تغيير التباين في مرحلة توليد البيانات تضمن ان يكون هناك تغيير في شكل انتشار البيانات وبالتالي تلويثها وهذا هو هدفنا في هذه المرحلة.

مراحل العمل مع البيانات الملوثة كانت بالشكل التالي:

١- بداية يتم عمل (PCA) العادي على البيانات الملوثة وظهر لنا ان محددة مصفوفة التباين المشترك مساوية الى الصفر وهو شيء طبيعي في حالة التعامل مع البيانات غير تامة الرتبة ولذلك فأننا لم نستطع اجراء تحليل S او M الحصين على البيانات.

٢- تحليل المركبات الرئيسية العادي اوضح لنا وجود (١٢) مركبة فسرت ٦٥.٣٠٢% من اجمالي التباين.

٣- في هذه المرحلة جاء دور خوارزمية RAPCA في التعامل مع البيانات الناقصة الملوثة وقد قمنا بتحديد k=n-1=24 واجراء التحليل واستخراج مصفوفة العلامة T ذات الدرجة (25 x 24) وبعد حساب محددة تلك المصفوفة كانت مساوية الى (0.9032) وقد فسرت النتائج من قبل (١٠) مركبات رئيسية كانت القيمة الذاتية لها اكبر من الواحد الصحيح وفسرت مانسبته ٧٥.٧٣٩١% من اجمالي التباين.

٤- الخطوة هنا هي هدف بحثنا وهي توظيف الطرائق الحصينة على البيانات الناقصة الملوثة. فبعد الحصول على المصفوفة T تم اجراء تحليل (S) الحصين على هذه المصفوفة لنحصل

على مصفوفة تباين حصينة تم اجراء تحليل المركبات الرئيسية عليها (PCA) فوجد ان det (Ss)=0.8101 وكما نلاحظ فأنها اصغر من محددة التباين لـ RAPCA الاصلي وكذلك فسرت النتائج من خلال (١١) مركبة فسرت ٨٠.٨٤٨١% من اجمالي التباين.

٥- نفس الخطوة السابقة ولكن بدل العمل ببرنامج (S) تم العمل ببرنامج (M) على المصفوفة T وتبين ان $(S_M)=0.9001$ det وفسرت النتائج من خلال (١١) مركبة فسرت ٧٨.٨٩٠١% من اجمالي التباين.

٦- نقطة الانهيار لمقدر (S) مساوية الى (٠.٥) اما نقطة الانهيار لمقدر (M) وكانت مساوية الى $\frac{1}{p+1}$.

٧- بعد حساب مسافات مهلونوبس العادية اتضح لنا وجود (٧) مشاهدات شاذة قيمة المسافة المحسوبة لها تزيد عن قيمة القطع والمساوية الى $\sqrt{\chi_p^2,0.975}$ وبأستبعاد هذه المشاهدات لتصبح (n=18) واجراء (PCA) العادي عليها كانت محددة مصفوفة التباين لها ايضا" مساوية للصفر وقد فسرت النتائج بـ (١٠) مركبات فسرت ٧٠.٨٧٣١% من اجمالي التباين.

٨- باللجوء الى خوارزمية (RAPCA) في التحليل والوصول الى المصفوفة T والتي كانت محددة مصفوفة التباين لها مساوية الى ٠.٨٧٢٩ وتسخير هذه المصفوفة في حساب مسافات مهلونوبس فقد اتضح وجود خمسة مشاهدات شاذة قيمتها تزيد على قيمة القطع وبأستبعاد هذه المشاهدات فقد كان لدينا (n=20) وتحديد k=n-1=19 لأكمال العمل ببقية الخوارزميات.

٩- بعد الحصول على المصفوفة T ذات الدرجة (19 x 20) تم اجراء تحليل (S) الحصين عليها ووجد ان $(S_s)=0.6305$ det وبعمل (PCA) عليها اتضح وجود (٨) مركبات فعالة فسرت ٨٠.٣٤٩١% من اجمالي التباين.

١٠- اجراء تحليل (M) الحصين على مصفوفة العلامة (T) اوضح لنا ان det $(S_M)=0.7293$ وبعمل (PCA) على تلك المصفوفة الحصينة اتضح لنا وجود (٨) مركبات فعالة فسرت ٧٦.٤٦٧١% من اجمالي التباين.

جدول رقم (٣)

نتائج المحاكاة للبيانات الملوثة

الطريقة	المعيار	ε = 0.25			
		PCA(full)	RAPCA(full)	PCA(red)	RAPCA(red)
Class	det(s)	0.0	0.9032	0.0	0.8729
	α	—	0.5	—	0.5
	$\sum_{i=1}^{k}\lambda_i$	28.0219	14.9812	25.0191	15.1088
	$to\left(\sum_{i=1}^{k}\lambda_i\right)$	65.302	75.7391	70.8731	75.2890
S	det(s)	—	0.8101	—	0.6305
	α	0.5	0.5	0.5	0.5
	$\sum_{i=1}^{k}\lambda_i$	—	16.6322	—	16.6738
	$to\left(\sum_{i=1}^{k}\lambda_i\right)$	—	80.8481	—	80.3491
M	det(s)	—	0.9001	—	0.7293
	α	$\frac{1}{13}$	$\frac{1}{25}$	$\frac{1}{13}$	$\frac{1}{20}$
	$\sum_{i=1}^{k}\lambda_i$	—	15.8267	—	14.3877
	$to\left(\sum_{i=1}^{k}\lambda_i\right)$	—	78.8901	—	76.4671

٦- الاستنتاجات والتوصيات

كان هدف بحثنا هذا هو الوصول الى مقدرات نستطيع الوثوق بها والاعتماد عليها في حال التعامل مع بيانات يكون منها عدد المشاهدات اقل من عدد المتغيرات فضلا" عن كون هذه

المجموعة من المشاهدات تحوي على مفردات شاذة او ملوثة. ولأجل المفاضلة بين هذه المقدرات فقد تم الاعتماد على جملة من المعايير تصب جميعها ضمن معنى خصائص الحصانة من حيث نقطة الانهيار ثم كفاءة المقدر وصولا" الى تمتعه بخاصية تساوي التغاير ومنه نستطيع القول:

ان اعلى نقطة انهيار سجلت كانت (٠.٥) سجلتها المقدرات (RAPCA , S) حيث تأتي جميع الطرق الحصينة في التقدير في المراتب الاخرى بعد هذين المقدرين ضمن هذه الخاصية.

ومن الجدير بالذكر هنا انه ليس فقط مقدر (M) المستخدم في هذا البحث يسجل نقطة انهيار اوطأ من (٠.٥) ولكن هناك جملة من المقدرات الحصينة المعروفة تقف بعد هذين المقدرين في تسجيل اعلى نقطة انهيار منها المقدرات (…,R, Areraged MVE, RMCD, MVE, MCD) بالاضافة الى كون مقدر (S) قد سجل اعلى نقطة انهيار فأنه من المقدرات التي تتمتع بخاصية تساوي التغاير وكذلك فقد لوحظ انه سجل اعلى نسبة للتباين المفسر من خلال مركباته المشتقة من مصفوفة التباين المشترك الحصينة التابعة له.

اثبت المقدر (M) عجزه عن التعامل مع البيانات ذات الابعاد العالية اضافة الى نقطة انهيار واطئة نستطيع من خلالها القول ان هذا المقدر مقدر ناجح في التعامل مع البيانات ذات الابعاد الواطئة.

بالنسبة لمقدر (RAPCA) فقد برهن عن جدارة ووثوق في التعامل مع البيانات غير تامة الرتبة والوصول بها الى مقدرات جيدة من حيث الخصائص في حالة التعامل مع البيانات الملوثة اما في حالة التعامل مع البيانات النظيفة فقد اجتاز هذا المقدر عقبة المحددة الصفرية ووصل الى نتائج قريبة الى نتائج المقدر التقليدي وهو الشيء الافضل.

من خلال ماموضح اعلاه يمكن القول انه يمكن الاعتماد على مقدر (RAPCA) في التعامل مع البيانات الناقصة واجراء كافة تحاليل متعدد المتغيرات وليس فقط تحليل المركبات الرئيسية مع امكانية المفاضلة بين المقدرات الحصينة المعروفة مثل المقدرات (…,R,L,RMCD,MVE,MCD).

المصادر

1- Croux, C.& Haesbroeck, G.(1999), " Influence function and Efficiency of the Minimum covariance Determinant scatter matrix Estimators"

Journal of Multivariate analysis, 71,161-190.

2- Croux, C.& Dehon, C.(2001), " Robust Lineor Discriminante Analysis using S-Estimators" the Canadian Journal of Statistics, vol.29, 473-492.

3- Hubert, M., Verboven, S.& Vanden Branden, K.(2005) " Robust Calibration Based on a Robust Covariance Estimator in High Dimensions" chemometrics and Intelligent laboratary system volume 75,issue 2, 28 February 2٠٠5 pages127-136.

4- Jockson, J.E. (1991), " Auser's Guide to principal components", New York, Wiley.

5- Rousseeuw, P.J, Hubert M.& verboven, S.(2002), " Fast Method for robust Principal components with Application to Chemometrics" August 16.2005, , 453K.

6- Rousseeuw, P.J Hubert, M. (2002) " ROBCA:A New Approach to Robust Principal component Analysis".