مقارنة بعض الطرائق الحصينة لتقدير معلمات انحدار المربعات الصغرى الجزئيه*

أ. م. د. سجى محمد حسين م. رباب عبد الرضا صالح
كلية الادارة والاقتصاد/ جامعة بغداد- قسم الاحصاء

المستخلص

تعد تقنية تخفيض الابعاد واختيار المتغيرات من المواضيع المهمة في التحليل الاحصائي لنماذج متعدد المتغيرات، فعندما يرتبط إثنان او اكثر من المتغيرات التوضيحية في الإنحدار بعلاقة اوعدة علاقات تامة او غير تامة ، تحدث مشكلة التعدد الخطي والتي فيها خرق لأحد الفروض الأساسية لطريقة المربعات الصغرى الاعتيادية مما يؤدي الى تقديرات غير دقيقة .

هناك طرائق عدة اقترحت لمعالجة هذه المشكلة نذكرمنها طريقة المربعات الصغرى الجزئية (PLS) والتي تستعمل لتخفيض الأبعاد في تحليل الإنحدار، بإستعمال تحويلات خطية تقوم بتحويل مجموعة من المتغيرات المرتبطة إرتباطاً عالياً، الى مجموعة من المتغيرات المستقلة الجديدة وغير المرتبطة تعرف بالمكونات ، وتكون هذه المكونات خطية متعامدة ومستقلة بعضها عن البعض الآخر.

ان طريقة PLS تفشل في التعامل مع البيانات التي تتضمن وجود القيم الشاذة ،وعليه فان نجاح هذه الطريقة يتوقف على عدم وجود هذه القيم الشاذة التي لها تأثير غير مرغوب على النتائج، وللحد من تواجد هذه القيم نلجأ الى استعمال الطرائق الحصينة .

في هذا البحث أستعمل خوارزمية PLSKURSD، والتي تطبق خوارزمية SIMPLS على مصفوفة التباين والتباين المشترك الحصين ،فضلاً عن الطريقة المقترحة MPLSKURSD وهي تعديل الى طريقة PLSKURSD .

جرى مقارنة بين معلمات انموذج الانحدار الخطي بطريقة المربعات الصغرى الجزئية مع الطرائق الحصينة للمربعات الصغرى الجزئية من خلال تجارب محاكاة، اعتمدت على وجود أنواع عدة من القيم الشاذة من البيانات وبنسب مختلفة من التلوث ولحجوم عينات وابعاد متغيرات مختلفة.

المصطلحات الرئيسية للبحث: المربعات الصغرى الجزئية- مصفوفة التباين المشترك الحصينة- التفرطح- الاسقاطات.

مجلة العلوم
الاقتصادية والإدارية
المجلد 20
العدد ٧٥
لسنة ٢٠١٤
الصفحات ٤١٣- ٤٣١

*مستل من اطروحة دكتوراه

١-١ المقدمة

أن أغلب التقنيات الاحصائية التقليدية تصمم بصورة خاصة لأبعاد قليلة من البيانات عندما يكون عدد المشاهدات أكبر من عدد المتغيرات لكن في بعض المسائل الخاصة بموعد البقاء أو تصنيف الورم لتنبؤ المريض كما في تحليل بيانات التعبير الجيني التي تتضمن اعداد كبيرة من المتغيرات (الجينات) التي ليس لها صلة بالتنبؤ اذ أن ادراج هذه المتغيرات في الانحدار والتصنيف يؤدي الى خسارة في الاداء، [9]. حتى لو كان اسهام هذه المتغيرات قليلة وغيرها من التطبيقات التي تعتمد على الابعاد العليا من البيانات فالتطبيقات الاحصائية في مثل هذه الحالات قليلة ومن هذه التطبيقات طريقة المربعات الصغرى الجزئية التي تحلل مثل هكذا البيانات .

تعد طريقة أنحدار المربعات الصغرى الجزئية partial least square (pls) أحدى تطبيقات الانحدار الخطي لحل مشكلة تخفيض الابعاد ومن ثم التخلص من مشكلة التعدد الخطي ،وأول من أستعمل هذة التطبيقات الباحث بعلم الاقتصاد Herman Wold عام (١٩٦٦) ثم طورت من قبل العالم نفس عام(1975) وقد تم تطبيق خوارزمية(NIPALS (Non-linear Iterative partial least squares (PLS1,PLS2) لمعالجة سلسلة من المصفوفات والتطبيقات في الاقتصاد القياسي ذات الابعاد العليا لواحد او اكثر من متغيرات الاستجابة فهي طريقة تتصل بالمصفوفة x والمتجه y او المصفوفة Y . وقد طورت هذة الطريقة واصبحت مشهورة ليس في مجال البحوث القياسية وانما في بقية الحقول وخصوصا في مجال علوم الكيمياء (chemometrics) ، وبعدها استخدمت في مجالات الطب والفيزياء والعلوم الاجتماعية ،علوم صناعة الاغذية وغيرها من العلوم الاخرى [13] .

وقد قدم De Jong عام ١٩٩٣ خوارزمية بديلة عن هذه الخوارزمية وهي statistically inspired modification of the partial least squares(SIMPLS) وهي خوارزمية سريعة وكفوءة [8,3] بالمقارنة مع خوارزمية NIPALS (pls1,pls2) ، اذ انها تعطي نتائج pls1 نفسها في حالة متغير الاستجابة واحد ونتائجها تختلف عن pls2 في حالة وجود عدة متغيرات استجابة.

وتوالت الدراسات ضمن طريقة المربعات الصغرى الجزئية فقد قدم الباحث Andersson بحثاً عن أستخدام خوارزميات عدة الى المربعات الصغرى الجزئية فضلا عن الخوارزميات المذكورة انفا، نذكر منها خوارزميةKERNAL المنسوبة الى Dayal ،وخوارزمية PLS-F المنسوبة الى Manne. [1] وغيرها من الخوارزميات. ان طريقة المربعات الصغرى الجزئيه PLS حساسة لوجود القيم الشاذة في مجموعة البيانات والتي تؤدي الى عدم دقة النتائج ولمعالجة القيم الشاذة ، تم استخدام الطرائق الحصينة للمربعات الصغرى الجزئية .

وقد كان اول اقتراح للطرق الحصينة للمربعات الصغرى الجزئية من قبل العالم Wakelling and Macfie عام ١٩٩٢ حيث تضمن تطبيقات الانحدارالحصين الى المربعات الصغرى الجزئية باستعمال خوارزمية biweight وهي طريقة تكرارية تعتمد على أعادة الاوزان لكل مشاهدة بالاعتماد على حجم بواقي الانحدار لطريقة المربعات الصغرى الجزئية [16] وتوالت الدراسات حول حصانة المربعات الصغرى الجزئية فقد قدم الباحث(Gonzales Javier. & et al عام ٢٠٠٩) أسلوبا جديدا لإزالة تأثير القيم الشاذة من نقاط البيانات لطريقة المربعات الصغرى الجزئية وتقدير حصين لمصفوفة التباين المشترك يتم حسابه بواسطة البحث عن الشواذ لاسقاطات المتغيرالاحادي للبيانات يجمع بين الاتجاه العشوائي الى (Stahel _ Donoho) والاتجاه المحددالحاصل من تعظيم وتصغير معامل التفرطح او(التفلطح) لـ Pena (and Prieto) وقد اوضحت هذه الطريقة كفائتها واعطائها افضل النتائج من الطرائق الأخرى وقد تم استعمال بيانات حقيقية اضافة الى اسلوب المحاكاة عند ظهورعدة انواع من الشواذ في البيانات وكذلك تمت المقارنة بين التقديرات تحت النموذج الذي يتوزع توزيع طبيعي وتوزيعات مختلفه من الأخطاء .

٢-١ هدف البحث

ان الهدف من البحث هو الحصول على افضل تقدير لمعلمات الانحدار في حالة وجود التعدد الخطي بين المتغيرات التوضيحية فضلا عن وجود القيم الشاذة في البيانات باستعمال طريقة المربعات الصغرى الجزئية PLS بخوارزمية SIMPLS ومقارنتها مع الطرائق الحصينة باستعمال خوارزميتي PLSKURSD , MPLSKURSD بتجارب محاكاة لعدة انواع من الشواذ وبنسب وحجوم عينات وابعاد مختلفة .

٢-١ الجانب النظري

ان طريقة المربعات الصغرى الجزئية PLS تعد اداة لاسلوب الانحدار الخطي وضعت لربط العديد من الانحدارات لواحد او اكثر من متغيرات الاستجابة بواسطة متغيرات كامنة Latent variable عندما يكون للمتغيرات التوضيخية ارتباطاً عالياً وكذلك عندما يكون عدد المتغيرات التوضيحية تفوق عدد المشاهدات.
وسيتم الاعتماد في البحث على خوارزمية SIMPLS حيث ان العوامل لـ PLS باستعمال خوارزمية SIMPLS تحدد باستعمال معيار تعظيم مصفوفة التباين المشترك.

٢-٢ انحدار المربعات الصغرى الجزئية Partial Least Squares Regression

في هذا البحث يتم الفرض بان المصفوفة $X_{n,p}$ لها **n** من المشاهدات و**p** من المتغيرات ويرمز لها $X_{n,p}=(X_1,X_2,\cdots,X_n)'$ حيث $X_i=(X_{i1},X_{i2},\cdots,X_{ip})$ الى i^{th} من المشاهدات.
والمصفوفة **Y** تتكون من **n** من المشاهدات و**q** من متغيرات الاستجابة ونرمز لها بالرمز $Y_{n,q}=(Y_1,Y_2,\cdots,Y_n)'$ اما مجموعة البيانات المدمجة ($X_{n,p}$ ، $Y_{n,q}$) نشير اليها بالرمز $Z_{n,m}$ [8] حيث **m= p+q**.
فنموذج الانحدار الخطي :

$$y_i=\beta_0+\beta'_{p,q}\,x_i+e_i \qquad \cdots(2-1)$$

e_i : حد الخطأ الذي يشترط فيه:
$\mathrm{cov}(e_i)=\Sigma_e,\ \mathrm{E}(e_i)=0$ لحجم **q**
β_0 = (β_{01} , ... , β_{0q}) ، حد ثابت غير معلوم ببعد **q**
β_{pq} تمثل المعالم غير المعلومة وهي مصفوفة الميل ببعد **pxq**.

٢-٣ خوارزمية The SIMPLS algorithm SIMPLS [14,8,3,2]

ان خوارزمية SIMPLS تفرض المتغيرات x_i وy_i حيث i= 1,2,...,n، لها علاقة من خلال الانموذج الثنائي الاتي:

$$x_i=\bar{x}+P_{p,k}\tilde{t}_i+g_i \qquad \ldots(2-2)$$

$$y_i=\bar{y}+A'_{q,k}\tilde{t}_i+f_i \qquad ..(2-3)$$

$\bar{x}, \bar{y}$:الوسط الحسابي للمتغيرات x_i و y_i

$\tilde{t}_i$: القياسات اوالمركبات او النقاط (scores or component) ببعد k حيث k << p

$p_{p,k}$: مصفوفة التحميل (x- loading)

$A'_{k,p}$: مصفوفة الميل في انحدار y_i على $\tilde{t}_i$

f_i , g_i : البواقي.

تفترض الخوارزمية اولا تشكيل او بناء المركبات حيث يتم الحصول على **k** من المركبات التي تكون كتوافيق خطية من **x** من المتغيرات والتي لها اكبر تباين مشترك مع التوافيق الخطية من المتغيرات **y** وبصورة ادق نفرض $\tilde{Y}_{n,q}, \tilde{X}_{n,p}$ تشير الى مصفوفة البيانات الممركزة ***(mean-centered data matrices)*** حيث ان:-

$$\tilde{x}_i = x_i - \bar{x}$$

$$\tilde{y}_i = y_i - \bar{y} \qquad \cdots(2-4)$$

r_a، q_a : متجهات اوزان pls الطبيعية (normalized) حيث ($||r_a||=||q_a||=1$) تعرف على انها المتجهات التي تعظم لكل a حيث a= 1…k.

$$\operatorname{cov}(\tilde{Y}_{n,q}\, q_a\, , \tilde{X}_{n,p}\, r_a\,) = q'_a\, \frac{\tilde{Y}'_{q,n}\, \tilde{X}_{n,p}}{n-1}\, r_a = q'_a\, S_{y,X}\, r_a \qquad \cdots \quad (2-5)$$

حيث ان $S'_{yx}=S_{xy}$ هي مصفوفة التباين المشترك التجريبية **empirical cross-covariance matrix** بين المتغيرات X,Y وعناصر الـ $\tilde{t}_i$ تعرف على انها التوافيق الخطية للبيانات الممركزة **(mean-centered data)** اوتكافئ

$$\tilde{t}_{ia} = \tilde{\chi}'_i\, r_a$$

$$\tilde{T}_{n,k} = \tilde{X}_{n,p}\, R_{p,k} \qquad \cdots(2-6)$$

حيث $R_{p,k} = (r_1 \ \ldots \ r_n)$

ان تعظيم المعادلة (2-5) لها حل واحد ان اول زوج لمتجه الاوزان لـ simpls $(r_a، q_a)$ يتم الحصول عليه من يسار ويمين المتجه المفرد المميز الى S_{yx}. وهذا يعني بانه r_1 المتجه المميز الى $S_{xy} * S_{yX}$ حيث ابعادها **(p×p)** و q_1 هو المتجة المميز الى S_{yx}*S_{xy} حيث ابعادها **(q×q)**.وللحصول على حل أكثر من واحد ،المركبات $\tilde{T}_{n,k}$ تتطلب أن تكون غير مرتبطة ***orthogonal*** :

$$T'_a T_j = 0 \quad \forall\, j \neq a$$

القيد المذكور انفا يفرض توليد سلسلة من الحلول المختلفة للمعادلة (٥-٢) هذا فضلا عن تجنب التعدد الخطي بين الانحدارات في الخطوة الثانية من الخوارزمية ، وبوجود هذا القيد اوزان المتجهات لخوارزمية ***Simpls*** الباقية وهي r_a ,q_a ($2 \le a \le k$)، يتم الحصول عليها بواسطة المتجهات المميزة لـ $S^a_{yx}S^a_{xy}$ و $S^a_{xy}S^a_{yx}$ وبشكل متتالي حيث ان مصفوفة التباين –التباين المشترك الجديدة **(deflated)** يتم الحصول عليها من المعادلة الاتية :

$$S^a_{x,y} = (I_p - V_{a-1}\, V'_{a-1})\, S^{a-1}_{x,y} \qquad \cdots \quad (2-7)$$

تبدأ هذه الخوارزمية مع $S^1_{xy} = S_{xy}$ وتكرر العملية حتى نحصل على **k** من المركبات والتي يتم بعدها تحديد عدد المركبات **k** بحسب طريقة **cross validation**

وتمثل (V_1 ,...,V_{a-1})بالمتعامد الطبيعي **orthonormal base** الى مصفوفة التحميل **X** ,(P_{a-1}...) (P_1) حيث ان مصفوفة التحميل P_j تصف العلاقة الخطية بين المتغيرات **X** و j^{th} من المركبات وكالاتي :

$$P_j = \tilde{X}'\, T_j \,/ T'_j\, T_j \qquad \cdots (2-8)$$

وبالتعويض عن T_j نحصل على الاتي :-

$$P_j = \tilde{X}'\, \tilde{X}\, r_j \,/\, (r'_j\, \tilde{X}'\tilde{X}\, r_j) = S_x\, r_j \,/\, (r'_j\, S_x r_j) \qquad \cdots (2-9)$$

والتي تمثل معامل انحدار المربعات الصغرى في انحدار $\tilde{X}$ على المركبة T_j

حيث: S_x هي مصفوفة التباين التجريبية للانحدارات

الخطوة الثانية في الخوارزمية هي انجاز الانحدار المتعدد الخطي **(MLR)** لانحدار المركبات المستخلصة T_1 ... T_k على متغيرات **y** الاصلية حيث صيغة انموذج الانحدار كالاتي:

$$y_i = \alpha_0 + A'_{q,k}\, \tilde{t}_i + f_i \qquad \cdots (2-10)$$

حيث $\mathrm{E}(f_i) = 0 \quad and \quad \mathrm{cov}(f_i) = \sum_f$

تقدير الانحدار الخطي المتعدد **MLR**

$$\hat{\alpha}_0 = \bar{y} - \hat{A}'_{q,k}\, \bar{\tilde{t}} \qquad ..(2-12)$$

$$\hat{A}_{k,q} = (S_t)^{-1} S_{ty} = (R'_{k,p} S_x R_{p,k})^{-1} R'_{k,p} S_{xy} \qquad ...(2-11)$$

$$S_f = S_y - \hat{A}'_{q,k} S_t \hat{A}_{k,q} \qquad ..(2-13)$$

حيث S_y & S_t :مصفوفة التباين الاولية للمتغيرات **y & t**

نلاحظ ان الانحدار الخطي المتعدد (multiple linear regression MLR) يشير الى انحدار المربعات الصغرى الكلاسيكية لمتغيرات x المتعددة ،وعندما عدد المتغيرات المعتمدة أكبر من الواحد لمتغيرات y المتعدد يعرف بالانحدار الخطي لمتعدد المتغيرات (multivariate multiple linear regression)

وبسبب $\tilde{\bar{t}} = 0$ الحد الثابت α_0 يقدر بواسطة $\bar{y}$

وعند تعويض $\tilde{t}_i = R'_{k,p}(x_i - \bar{x})$ في المعادلة (2-3) نحصل على تقديرات المعالم لنموذج الانحدار الاصلي في المعادلة (2-1) وكالأتي :

$$\hat{\beta} = R_{p,k}(R'_{K,p}\, S_x R_{p,K})^{-1} R'_{K,p}\, S_{xy} \qquad \cdots(2-14)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}'_{q,p}\, \bar{x} \qquad \cdots(2-15)$$

واخيرا التقدير الى Σ_e هو

$$S_e = S_y - \hat{\beta}\, S_x \hat{\beta}' \qquad \cdots(2-16)$$

في حالة الاستجابة واحدة (*q=1*) التقدير للمعالم يكون $\hat{\beta}_{p,1}$ وتكتب على شكل متجه ،بينما تباين الخطأ S_e يتحقق كالأتي $\hat{\sigma}^2{}_e = S_e^2$

٤-2 انحدار المربعات الصغرى الجزئية الحصينة Robust Partial Least Squares Regression

على الرغم من ان المربعات الصغرى الجزئية (PLS) تعالج مشكلة تقليل الابعاد في متعدد المتغيرات الا انها تفشل في التعامل مع بيانات متعدد المتغيرات والتي تتضمن وجود القيم الشاذة التي تكون بعيدة عن تمركزالبيانات او تكون غير منسجمة او مختلفة عن بقية البيانات وتحدث نتيجة اخطاء في القياس اوالتسجيل او تنتمي بعض المشاهدات الى مجتمع اخر ومن ثم ينتج تقدير غير دقيق ومن المهم جدا اكتشاف القيم الشاذة لان وجودها لها تاثير غير مرغوب به على التقديرات ولازالة هذه التاثيرات الغير مرغوب بها فقد افترض الباحثون استبدال الطرائق الاعتيادية بالطرائق الحصينة وهناك عدة طرائق لحصانة المربعات الصغرى الجزئية منها الطرائق التي تعتمد على حصانة مصفوفة التباين والتباين المشترك والطرائق التي تعتمد على الانحدار الحصين . [25]

5--2 خوارزمية *PLSKURSD* The PLSKURSD algorithm [12,11,10,5]

تعتمد خوارزمية PLSKURSD على تطبيق خوارزمية SIMPLS على مصفوفة التباين والتباين المشترك الحصين ،ان تقدير مصفوفة التباين يتم بواسطة اسقاط البيانات في بعض الاتجاهات وايجاد القيم الشاذة على هذه الاتجاهات وحذفها من العينة واستعمال البيانات النظيفة لحساب مصفوفة التباين والتباين المشترك.

اما الية عمل هذة الطريقة فانها تكون على ثلاث خطوات ،حيث تفترض وبدون فقدان العمومية ان البيانات الاصلية لها وسط حسابي صفر ومصفوفة التباين Sz حيث تحول البيانات باستعمال المعادلة الآتية لمصفوفة البيانات Z=[X,Y] :-

$$\tilde{Z}_i = S_z^{-1/2}\, Z_i \qquad , i = 1,2,\ldots,n \qquad \cdots(2-17)$$

وعندما تكون مصفوفة التباين مفردة (singular) نستخدم المعكوس العام(generalized inverse)

وبمكن تطبيق هذه الخوارزمية في حالة عدد المتغيرات أكبر من المشاهدات بأستعمال خوارزمية svd [7]

اما خطوات هذه الطريقة فهي كالاتي:

الخطوة الاولى STEP ONE

يتم حساب الاتجاه الذي يعظم معامل التفلطح (Kurtosis) للاسقاطات وكذلك يتم حساب المسافات ($r_i^{(j)}$) حيث j= 1,2 لكلا الاتجاهين (تعظيم وتصغير) للمتغير الاحادي الطبيعية للبيانات
اول اتجاه نحصل علية من حل الصيغة التالية

$$D1 = \arg\max 1/n \sum_{i=1}^{n} (d'\tilde{Z}i\ )^4 \qquad \cdots(2-18)$$
$$s.t \quad d'd = 1$$

ونفس العملية تكرر لحساب الاتجاه الذي يصغر معامل التفرطح

$$D2 = \arg\min 1/n \sum_{i=1}^{n} (d'\tilde{Z}i\ )^4 \qquad \ldots(2-19)$$
$$s.t \quad d'd = 1$$

نفرض ان

$$Pi^{(j)} = d'_j \tilde{z}_i \qquad i = 1,2,\ldots,n \quad j = 1,2 \qquad \cdots(2-20)$$

الذي يمثل اسقاطات القيم على هذين الاتجاهين وثم تحسب قياس الشواذ للمتغير الاحادي او المسافات $ri^{(j)}$ للمتغير الاحادي الطبيعية كالاتي

$$ri^{j} = 1/\beta_p \frac{\left| Pi^{(j)} - median_i(Pi^{(j)}) \right|}{MAD_i(Pi^{(j)})} \qquad j = 1,2 \qquad \ldots(2-21)$$

حيث ان

$$MAD_i(pi^{(j)}) = median_i \left| pi^{(j)} - median_i(pi^{(j)}) \right| \qquad \ldots(2-22)$$

حيث ان MAD : وسيط الانحراف المطلق Median Absolute Devation
و ان قيمة القطع Bp تعتمد على عدد المتغيرات (p) ويتم الحصول عليها من تجارب محاكاة لضمان في حالة غياب الشواذ بان نسبة المشاهدات الصحيحة التي صنفت بانها شواذ تقريبا هي 0.05 .

الخطوة الثانية STEP TWO

حساب الاتجاهات العشوائية أيجاد الاتجاه العشوائي وذلك من خلال اسلوب المعاينة الطبقية ومن ثم البحث عن القيم الشاذة في هذه الاتجاهات العشوائية كل اتجاه يولد بخطوتين وهي

1- يتم اختيار مشاهدتين عشوائيا من العينة ثم تحسب الاتجاه المعرف لهاتين المشاهدتين (اما ان يكون على اساس المسافة او الفرق بين المشاهدتين) ومن ثم اسقاط المشاهدات على هذا الاتجاه وتكرر الى h من المرات حيث p *10,... ,h= 1,2 حيث pهو(عدد المتغيرات) .

٢- في هذه الخطوه يتم بناء مجموعة من k من العينات الطبقية حيث يتم ترتيب الاسقاطات وتقسم الى k من الفترات حيث k تؤخذ اما 3 او 5 تحدد مسبقا وكل فترة سيكون حجمها n/k لكل k من الفترات، 1< k<K، سيتم اختيار عينة جزئية تتكون من p من المشاهدات تنتخب وبدون الارجاع ويحسب الاتجاة dj الذي يستخدم لحساب الشواذ كما في الخطوة الاولى حيث ان الاسقاطات $p_i^{\sim(j)}$ هي:-

$$\tilde{P}_i^{(j)} = \tilde{d}_j \tilde{Z}$$

ومن خلالها سوف تحسب المسافات $\tilde{r}_i$ للمتغير الاحادي الطبيعية :

$$\tilde{r}_i^{\,j} = 1/\beta_p \frac{\left|\tilde{P}i^{(j)} - median_i(\tilde{P}i^{(j)})\right|}{MAD_i(\tilde{P}i^{(j)})} \qquad j = 1,2,\ldots,h*K \qquad \ldots(2-23)$$

الخطوة الثالثة (التحقيق) STEP THREE (checked)

لكل مشاهدة (i) فان قياس الشواذ الطبيعي لها يتم الحصول عليه كما يلي

$$r_i = \max\{ r_i^1 , r_i^2 , \tilde{r}_1 \quad \cdots \tilde{r}_i^{\,j} \} \qquad \cdots(2-24)$$
$$j = 1,2, \quad \cdots h*K$$

والتي تم الحصول عليها بموجب المعادلات (2-21) ، (2-23)

هذه المشاهدات اذا كانت ri > 1 فانها تصنف على انها قيمة شاذة وتزال من العينة اذا كان عددهم اصغر من [n – [n+ p + 1] /2]عدا ذلك فان المشاهدات التي عددها [n- [n + p +1]/2] والتي لها اكبر قيم الى ri فهي تعتبر شواذ وتزال من البيانات.

واخيرا نجد المجموعة U والتي تمثل المشاهدات التي لاتتضمن القيم الشاذه بعدها يتم حساب مسافة مهالنوبس بالاعتماد على المشاهدات الجيدة وكما يلي:-

$$\tilde{m} = \frac{1}{|U|}\sum_{i\in U} Z_i \qquad \cdots(2-25)$$

$$\tilde{S}_z = \frac{1}{|U|-1}\sum_{i\in U}(Z_i - \tilde{m})(Z_i - \tilde{m})' \qquad \ldots(2-26)$$

$$V_i = (Z_i - \tilde{m})'\tilde{S}_Z^{-1}(Z_i - \tilde{m}) \qquad \ldots(2-27)$$

حيث ان $|U|$ تمثل عدد المشاهدات في المجموعة U

فاذا كانت $V_i < \chi^2_{p,1-.99}$ فانها لا تصنف على انها قيم شاذه وتوضع في مجموعة U وفي حالة $\widetilde{S}_z^{-1}$ غير موجودة سوف نستعمل المعكوس العام لها.

وهذة الخطوات تكرر حتى لاتبقى قيم شاذه (او تصبح U مجموعة كل البيانات) .

ويمكن ان تستعمل هذه الخوارزمية عندما p > n حيث يتم تطبيق تجزئة القيم المفردة (SVD) على البيانات لاجل تقليص مجال البيانات الى مجال جزئي مقسم بواسطةn [28,19]

ولايوجد معيار قيمة القطع للبيانات التي لاتتوزع توزيعا طبيعيا وانما يتم استعمال معيار بديل لقيم القطع (الحصينة) تكون بالاعتماد على الجذر التربيعي لمسافة مهالنوبس $\sqrt{\widetilde{V}_i}$ الذي يكون توزيع متماثل اكثر من $\widetilde{V}$ وان المشاهدة تعتبر شاذه اذا تحقق مايلي:-

$$\sqrt{\widetilde{V}_i} \geq median_i\,(\sqrt{\widetilde{V}_i}) + 4.5\,MAD_i\,(\sqrt{\widetilde{V}_i}) \qquad \cdots(2-28)$$

و ان MAD_i عرف في المعادلة (2-22)

٦-٢ الطريقة المقترحة THE PROPSED ALGORITHM MPLSKURSD

يتم في هذه الخوارزمية ايجاد مصفوفة التباين والتباين المشترك الحصينة والتي توظف في خوارزمية SIMPLS لإيجاد مقدرات المعلمات في حالة وجود القيم الشاذة فضلاً عن وجود مشكلة التعدد الخطي ،أذ تعتبر هذه الطريقة كتعديل لطريقة المربعات الصغرى الجزئية الحصينة لخوارزمية PLSKURSD حيث ان هذه الطرقة تعتمد على اربع خطوات وكالاتي :

١- في الخطوة الاولى :يتم ايجاد قيم $\widetilde{X}$ الحصينه وذلك أولا بحساب تقدير الوسيط متعدد الابعاد Column-wise median أو L1-median (وهو تقدير حصين للموقع في متعدد المتغيرات وله خصائص جيدة ويسمى بالوسيط المكاني (spatial median) وقد تم تعريفه من قبل العالم weber 1909(بأنه أي نقطه لها أقل مجموع لمسافة أقليديس من كل النقاط في مجموع البيانات [4])وتوجد عدة خوارزميات لإيجاد l1-median وقد تم الاعتماد على خوارزمية Hossjer and Croux (1995) [6] فالوسيط L1- او (spatial median) الى مجموعة من البيانات { $x_1,\ldots,x_n$ } X= لكل $x_i \in R^p$ هو متجه يرمز له بـ $\hat{\mu}_{sm}$ والذي يحقق أقل مجموع وكالاتي :

$$\hat{\mu}_{sm} = \arg\min_{\eta \in R^P} \sum_{i=1}^{n} \| x_i - \eta \|$$ حيث ان $\|\,.\,\|$ يشير الى الاقليديس الطبيعي.

ويتم إيجاد قيم $\widetilde{y}$ الحصينة بأستعمال الوسيط حيث يعتبر من المقايس الحصينة أيضاً.

ثانيا المتغيرات التوضيحية (X) يطرح منها الوسيط الخاص L 1_median والمتغير (y) يطرح منه الوسيط، المتغيرات الحصينة $Z = [\widetilde{X}, \widetilde{y}]$ يتم أيجاد

$$\widetilde{Z}_i = S_z^{-1/2}\, Z_i \qquad , i = 1,2,\ldots,n$$

٢-في هذه الخطوة :يتم ايجاد الاتجاه الذي يعظم ويصغر معامل التفرطح للبيانات الحصينة وحسب المعادلات (٢-١٨)،(٢-١٩)، ثم يتم أيجاد أسقاطات القيم الحصينة على هذين الاتجاهين وبعدها يتم ايجاد قياس الشواذ للمتغير الاحادي $ri^{(j)}$ وحسب المعادلة (٢-٢١)

٣- حساب قياس الشواذ للمتغير الاحادي $\tilde{r}_i$ *من خلال أسقاطات القيم الحصينة على الاتجاه العشوائي* من خلال اسلوب المعاينة الطبقية ومن ثم البحث عن القيم الشاذة في هذه الاتجاهات العشوائية وحسب المعادلة(٢٣-٢)

٤- *في هذه الخطوة يتم التدقيق* لكل مشاهدة (i) بعد ايجاد قياس الشواذ للمتغير الاحادي r_i

فاذا كانت ri > 1 فان المشاهده i تصنف على انها قيمة شاذة وتزال من العينة اذا كان عددهم اصغر من [n – [n+ p + 1] /2] ،عدا ذلك فان المشاهدات التي عددها [n- [n + p +1]/2] والتي لها اكبر قيم الى ri فهي تعتبر شواذ وتزال من البيانات، وبعدها يتم حساب مسافة مهالوبس الحصينة (V_i) بالاعتماد على البيانات الخالية من القيم الشاذة وحسب المعادلة (٢٧-٢) فاذا كانت مسافة مهالوبس أكبر من قيمة القطع $\chi^2_{p,1-.99}$ تصنف على انها قيم شاذة . وفي حالة $\tilde{S}_z^{-1}$ غير موجودة سوف نستعمل المعكوس العام لها.وتكرر هذه الخطوات حتى لاتبقى قيم شاذة ونستطيع ان نستخدم هذه الطريقة في حالة عدد المتغيرات أكبر من المشاهدات أيضاً بأستخدام خوارزمية SVD.

٣-١ الجانب التجريبي [15,5,8]

من اجل الوصول الى الهدف تم استعمال اسلوب المحاكاة للمقارنة بين تقدير معلمات الانحدارلطريقة المربعات الصغرى الجزئية بأستعمال خوارزمية SIMPLS وطريقة المربعات الصغرى الجزئية الحصينة لخوارزميتي MPLSKURSD, PLSKURSD لعدة انواع من الشواذ وبنسب التلوث المختلفة ولابعاد متغيرات وحجوم عينات مختلفة.

٣-٢ وصف تجربة المحاكاة

في هذه التجربة تم توليد العينات بحجوم (200,150,100,50,30) وعدد المتغيرات التوضيحية(p=5) في حالة n>P اما في حالة p>n تم توليد العينات بالحجم n=20 وعدد المتغيرات التوضيحية (p=40,60,80,100,150)

وكان عدد المركبات (2) في الحالتين اما مصفوفة التباين الى t (مصفوفة القياس) فهي قطرية diag_ [6,4] وتم استعمال نسب تلوث ϵ مساوية الى (0.1, 0.2 ,0.3) وقد كررت التجربة 1000 مرة للحصول على الدقة العالية. في هذة التجربة سنقارن بين نوعين من الشواذ بالاضافة الى المشاهدات الاعتيادية وقد تمت المقارنة بين متجه معلمات الانموذج للطرائق المختلفة ومتجه المعلمات المقدر $\hat{\beta}_a^{(l)}$ حيث تم تكرارالتجربة لـ 1000 بمعيار

Norm(β) فقد تم حساب المقياس كالاتي :

$$MSE_a(\hat{\beta}) = 1/m \sum_{l=1}^{m} \left\| \hat{\beta}_a^{(l)} - \beta \right\|^2 = Norm(\beta) \qquad \cdots(3-1)$$

حيث || || يشيرالى اقليديس الطبيعي و متجه مقدرات الانحدار $\hat{\beta}_a^{(l)}$ في التكرر l و m يشير الى عدد التكرارات.

بالاعتماد على النموذج الابتدائي(٢-٢)و(٣-٢) وعلى فرض ان الوسط الحسابي يساوي صفر حيث النموذج يتبع التوزيع الطبيعي وغير المرتبط تم توليد البيانات التالية:

١-المشاهدات الاعتيادية (*Regular observation*)وهي المشاهدات التي تكمن من ضمن الغالبية من المشاهدات التي تقع على خط الانحداركالاتي:

$$t \propto N_a(0_a, \sum_t)$$
$$x = I_{p,a} t + N_p(0_p, 0.1\, I_p) \ , \quad p > a \qquad \cdots(3-2)$$
$$y = q^T t + N(0,1)$$

حيث $I_{p,a}$ = 1 عندما $i=j$ و $I_{p,a}$=0 ماعدا ذلك .

q متجه من الواحدات ذو بعد $ax1$

I_pمصفوفة الوحدة ذو بعد $p\ x\ p$

اما بالنسبة للبيانات الملوثة فقد تم توليد مجموعة تتكون من [100 *(1- ϵ)] من المشاهدات وفقا للنموذج اعلاه وتم اضافة [100 *ϵ] من المشاهدات الملوثة التي تولد من خلال نماذج التلوث ادناه لتكون انواع التلوث التالية

٢- الشواذ العمودية *vertical outliers* وهي النقاط التي تكون بعيدة عن خط الانحداربينما اسقاطاتها تكون ضمن المشاهدات الجيدة .

$$y_\varepsilon = q't + N(30, 0.3) \qquad \cdots(3-3)$$

٣- نقاط الانعطاف جيدة *(good leverage point)* وهي النقاط التي تقع على خط الانحدار لكنها بعيدة عن تجمع الغالبية من المشاهدات التي تحدد خط الانحدار .

$$x_\varepsilon = t_\varepsilon\, I_{a,p} + N_p((0_a, 3_{p-a}), 0.01 I_p)$$
$$t_\varepsilon \propto N_a(3_a, \sum_t)$$
$$y_\varepsilon = q't + N(30, 0.3) \qquad \cdots(3-4)$$

نتائج المحاكاة

تم تقدير المعلمات للطريقتين المذكورة في الجانب النظري وتم حساب ($\hat{\beta}$) MSE_a عند جميع حجوم العينات ولمختلف اعداد المتغيرات التوضيحية وبنسب تلوث مختلفة والتي تم ذكرها وكانت النتائج كما في الجداول ادناه:

جدول رقم (١) يوضح قيم MSE لمتجه المعلمات المقدرة عند (n=30) و (p=5) عند نسب التلوث (0.3,0.2,0.1)

الطرائق / الشواذ	PLS	PLSKUR	MPLSKUR
No contamination لاتوجد شواذ	3.93198	6.705675	6.29413
10%Vertical outliers الشواذ العمودية	17.85582	6.683227	7.938724
10%tGood leverage point نقاط الرفع الجيدة	26.62007	6.15471	6.179438
20%Vertical outliers الشواذ العمودية	30.9934	11.32486	9.957953
20%Good leverage point نقاط الرفع الجيدة	28.49926	6.120336	5.882657
30%Vertical outliers الشواذ العمودية	52.44824	28.51931	26.12804
30%Good leverage point نقاط الرفع الجيدة	29.05623	6.703405	7.145314

جدول رقم (٢) يوضح قيم MSE امتجه المعلمات المقدرة عند (n=50) و (p=5) عند نسب التلوث = (0.3,0.2,0.1)

الطرائق / الشواذ	PLS	PLSKUR	MPLSKUR
No contamination لاتوجد شواذ	3.143267	4.455928	4.145776
10%Vertical outliers الشواذ العمودية	12.74727	4.238654	4.396087
10% Good leverage point نقاط الرفع الجيدة	27.27451	4.212243	4.217679
20% Vertical outliers الشواذ العمودية	22.78322	4.551482	4.438952
20% Good leverage point نقاط الرفع الجيدة	29.07668	4.707521	4.964694
30% Vertical outliers الشواذ العمودية	30.0315	10.47879	11.1865
30% Good leverage point نقاط الرفع الجيدة	29.26588	4.551511	4.448104

جدول رقم (٣) يوضح قيم MSE لمتجه المعلمات المقدرة عند (n=100) و (p=5) عند نسب التلوث = (0.3,0.2,0.1)

الطرائق / الشواذ	PLS	PLSKUR	MPLSKUR
No contamination لاتوجد شواذ	2.784376	2.944539	2.955385
10% Vertical outliers الشواذ العمودية	7.436278	3.101319	3.100473
10% Good leverage point نقاط الرفع الجيدة	27.10441	2.959698	2.945401
20% Vertical outliers الشواذ العمودية	12.25034	3.202687	3.128556
20% Good leverage point نقاط الرفع الجيدة	29.06709	3.078286	3.070211
30% Vertical outliers الشواذ العمودية	18.85743	4.1549	3.713399
30% Good leverage point نقاط الرفع الجيدة	29.88383	3.466043	3.560048

جدول رقم (٤) يوضح قيم MSE لمتجه المعلمات المقدرة عند (n=150) و (p=5) عند نسب التلوث = (0.3,0.2,0.1)

الطرائق / الشواذ	PLS	PLSKUR	MPLSKUR
No contamination لاتوجد شواذ	2.564526	2.619275	2.616338
10% Vertical outliers الشواذ العمودية	5.421788	2.669105	2.666211
10% Good leverage point نقاط الرفع الجيدة	27.49836	2.646156	2.651284
20% Vertical outliers الشواذ العمودية	9.084631	2.828726	2.82872
20% Good leverage point نقاط الرفع الجيدة	28.98738	2.711855	2.704045
30% Vertical outliers الشواذ العمودية	11.94638	3.17712	2.992124
30% Good leverage point نقاط الرفع الجيدة	29.57321	3.069245	3.128813

جدول رقم (٥) يوضح قيم MSE لمتجه المعلمات المقدرة عند (n=200) و (p=5) عند نسب التلوث = (0.3,0.2,0.1)

الطرائق / الشواذ	PLS	PLSKUR	MPLSKUR
No contamination لاتوجد شواذ	2.468142	2.498064	2.504395
10% Vertical outliers الشواذ العمودية	4.565388	2.555479	2.553579
10% Good leverage point نقاط الرفع الجيدة	27.33878	2.570506	2.567473
20% Vertical outliers الشواذ العمودية	7.58146	2.602076	2.605444
20% Good leverage point نقاط الرفع الجيدة	29.00699	2.623537	2.621256
30% Vertical outliers الشواذ العمودية	9.250026	2.688	2.748775
30% Good leverage point نقاط الرفع الجيدة	29.69553	3.007342	3.29134

جدول رقم (٦) يوضح قيم MSE لمتجه المعلمات المقدره عند (n=20) و (p=40) عند نسب التلوث = (0.3 ,0.2,0.1)

الطرائق / الشواذ	PLS	PLSKUR	MPLSKUR
No contamination لاتوجد شواذ	90.76676	83.67835	83.62133
10 % Vertical outliers الشواذ العمودية	203.7097	130.7868	129.3845
10% Good leverage point نقاط الرفع الجيدة	76.85447	76.39718	76.36648
20% Vertical outliers الشواذ العمودية	255.2324	121.5665	119.8704
20% Good leverage point نقاط الرفع الجيدة	43.28528	42.79096	42.77652
30% Vertical outliers الشواذ العمودية	399.5773	163.3382	159.6765
30% Good leverage point نقاط الرفع الجيدة	51.60886	51.29083	51.2872

جدول رقم (٧) يوضح قيم MSE لمتجه المعلمات المقدرة عند (n=20) و (p=60) عند نسب التلوث = (0.1,0.2, 0.3)

الطرائق / الشواذ	PLS	PLSKUR	MPLSKUR
No contamination لاتوجد شواذ	66.45336	58.18977	58.18797
10% Vertical outliers الشواذ العمودية	174.4606	95.6495	95.23426
10% Good leverage point نقاط الرفع الجيدة	58.34411	57.55186	57.5446
20% Vertical outliers الشواذ العمودية	281.1992	123.3543	122.2643
20% Good leverage point نقاط الرفع الجيدة	52.75271	51.99841	51.98885
30% Vertical outliers الشواذ العمودية	421.7863	175.031	173.1078
30% Good leverage point نقاط الرفع الجيدة	74.58289	74.13526	74.10261

جدول رقم (٨) يوضح قيم MSE لمتجه المعلمات المقدرة عند (n=20) و (p=80) عند نسب التلوث = (0.1,0.2, 0.3)

الطرائق / الشواذ	PLS	PLSKUR	MPLSKUR
No contamination لاتوجد شواذ	70.61393	62.59702	62.54708
10% Vertical outliers الشواذ العمودية	177.7266	98.43487	97.7031
10% Good leverage point نقاط الرفع الجيدة	57.87697	56.80069	56.78465
20% Vertical outliers الشواذ العمودية	320.3175	168.6173	167.5111
20% Good leverage point نقاط الرفع الجيدة	99.34089	98.55477	98.51187
30% Vertical outliers الشواذ العمودية	397.5902	166.802	165.8302
30% Good leverage point نقاط الرفع الجيدة	79.74979	79.03737	78.97919

جدول رقم (٩) يوضح قيم MSE لمتجه المعلمات المقدرة عند (n=20) و (p=100) عند نسب التلوث = (0.1,0.2, 0.3)

الطرائق / الشواذ	PLS	PLSKUR	MPLSKUR
No contamination لاتوجد شواذ	74.42864	66.54818	66.44204
10% Vertical outliers الشواذ العمودية	173.8288	100.4384	100.4113
10% Good leverage point نقاط الرفع الجيدة	63.48287	62.27119	62.24431
20% Vertical outliers الشواذ العمودية	303.9499	155.7178	155.3580
20% Good leverage point نقاط الرفع الجيدة	97.46692	96.38357	96.34031
30% Vertical outliers الشواذ العمودية	417.1883	188.8668	188.0394
30% Good leverage point نقاط الرفع الجيدة	109.4832	108.5655	108.5351

جدول رقم (١٠) يوضح قيم MSE لمتجه المعلمات المقدرة عند(n=20) و(p=150) عند نسب التلوث = (0.1,0.2, 0.3)

الطرائق / الشواذ	PLS	PLSKUR	MPLSKUR
No contamination لاتوجد شواذ	147.6072	141.1587	141.1408
10% Vertical outliers الشواذ العمودية	230.5816	169.0253	168.9759
10% Good leverage point نقاط الرفع الجيدة	143.02	141.3591	141.3234
20% Vertical outliers الشواذ العمودية	339.5476	215.9492	215.4868
20% Good leverage point نقاط الرفع الجيدة	167.7448	166.3475	166.3294
30% Bad leverage point نقاط الرفع السيئة	158.4073	156.3259	156.265
30% Good leverage point نقاط الرفع الجيدة	155.5631	154.501	154.4384

تحليل النتائج والاستنتاجات

في حالة عدد المشاهدات اكبر من المتغيرات (n>p) بتضح مايلي :-

١- يتضح من الجدول رقم (١) والذي يتضمن (n=30 p=5) مايلي

- عند عدم وجود الشو اذ كانت الطريقة PLS هي الافضل .
- عند الشواذ العمودية كانت الطريقة plskursd هي الافضل عند نسب التلوث 0.1 والطريقة mplskursdهي الافضل عند نسب التلوث 0.3,0.2
- عند نقاط الرفع الجيدة كانت الطريقة plskursd هي الافضل عند نسب التلوث 0.3,0.1 و mplskursd هي الافضل عند نسب التلوث 0.2

٢- يتضح من الجدول الجدول رقم (٢) والذي يتضمن (n=50 p=5) مايلي :

- عند عدم وجود الشو اذ كانت الطريقة PLS هي الافضل .
- عند الشواذ العمودية كانت الطريقة plskursd هي الافضل عند نسب التلوث 0.3,0.1, والطريقة mplskursdهي الافضل عند نسب التلوث 0.2
- عند نقاط الرفع الجيدة كانت الطريقة plskursd هي الافضل عند نسب التلوث 0.2,0.1 و mplskursd هي الافضل عند نسب التلوث 0.3

٣- يتضح من الجدول رقم (٣) والذي يتضمن (n=100 p=5) مايلي :

- عند عدم وجود الشو اذ كانت الطريقة PLS هي الافضل .
- عند الشواذ العمودية كانت الطريقة mplskursd هي الافضل في كل الحالات نسب التلوث
- عند نقاط الرفع الجيدة كانت الطريقة mplskursd هي الافضل عند نسب التلوث 0.1,0.2 وان plskursd هي الافضل عند نسب التلوث 0.3

٤- يتضح من الجدول رقم (٤) والذي يتضمن (n=150 p=5) مايلي :

- عند عدم وجود الشو اذ كانت الطريقة PLS هي الافضل.
- عند الشواذ العمودية كانت الطريقة mplskursd هي الافضل في كل الحالات نسب التلوث.
- عند نقاط الرفع الجيدة كانت الطريقة plskursd هي الافضل عند نسب التلوث 0.1,0.3 وان mplskursd هي الافضل عند نسب التلوث 0.2 .

٥- يتضح من الجدول رقم (٥) والذي يتضمن (n=200 p=5) مايلي :

- عند عدم وجود الشو اذ كانت الطريقة PLS هي الافضل.
- عند الشواذ العمودية كانت الطريقة mplskursd هي الافضل عند نسب التلوث 0.1 و plskursd الافضل في الحالات الاخرى
- عند نقاط الرفع الجيدة كانت الطريقة mplskursd هي الافضل عند نسب التلوث و plskursd الافضل عند نسب التلوث 0.1,0.2 0.3

٦ - يتضح من الجدول رقم (٦و٧و٨و٩و١٠)نجد مايلي:-

- عند عدم وجود الشواذ والشواذ العمودية ونقاط الرفع الجيدة كانت الطريقة mplskursd هي الافضل في جميع الجداول.

ويمكن ان نلخص النتائج بما يلي:

١- ان طريقة PLS هي الافضل في تقدير المعلمات في حالة عدم وجود الشواذ عندما عدد المشاهدات اكبر من المتغيرات وذلك لامتلاكها اقل متوسط مربعات الخطأ (MSE).

٢- ان طريقة PLSKURSD الافضل تقريبا بالمقارنة مع MPLSKUR في تقدير المعلمات عند أحتواء البيانات على القيم الشاذة عندما عدد المشاهدات اكبر من المتغيرات

٣- ان طريقة MPLSKURSD الافضل في تقدير المعلمات في حالة عدد المتغيرات اكبر من المشاهدات وعند عدم وجود الشواذ وفي الشواذ العمودية ونقاط الانعطاف(الرفع) الجيدة . .

٤ - اشارت النتائج الى ان قيم متوسط مربعات الخطأ تتناقص بزيادة حجم العينة ولكلا الطرائق في التقدير.

التوصيات

نوصي باستخدام الخوارزمية الحصينة المقترحة MPLSKURSD في حالة عدد المتغيرات اكبر من المشاهدات لكل نسب التلوث واستعمال PLSKURSD في حالة المشاهدات اكبر من المتغيرات لبعض نسب التلوث

المصادر

1- Andersson, M., (2009) "Acomarison of nine PLS1 algorithms'' J. chemometrics; 23:518-529
2- Collins,B., 2010 " partial least squares regression "october 13
3-De Jong, S., 1993. SIMPLS: an alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 18: 251–26
4- Fritz,H., Filzmoser,P. , Croux,C., 2010'' A comparison of algorithms for the multivariate l1_median''
http://www.statistik.tuwien .ac.at
5-Gonzales, J., & et al. (2009) "A robust partial least squares regression method with application" J . Chemometrics; 23:78-90
6- Hossjer and Croux (1995) "Generalizing Univariate Signed Rank Statistics for Testing and Estimating a Multivariate Location Parameter", Non-parametric Statistics, 4, 293
7-Hubert, M., Rousseeuw, P.J., Verboven,S (2002), ''A fast robust method for principal components with applications to chemometrics" J . Chemometrics and Intelligent Laboratory Systems, 60,101-111.
8- Hubert, M., & Vanden , K., (2003) ''Robust Methods for Partial Least Squares Regression "9th October J. chemometrics p 537-549.
9- - Nguyen,D.V. and Rocke ,D.M.,(2002) ''Tumor classification by partial least squres using microarray gene expression data'' Bioinformayics vol.18 no .1 pages 39-50.
10-Pena,D. and Prieto,F.J., (1997) "Robust Covariance Matrix Estimation and Multivariate Outlier Detection" working paper 97-08, statistics and Econometrics series 04
11- Pena,D. and Prieto,F.J.,(2001) "Multivariate Outlier Detection and Robust Covariance Matrix Estimation" Technometrics,vol.43,no.3
12- Pena,D., and Prieto,F.J., (2007)"Combining Random and Specific Direction for Outlier Detection and Robust Estimation in High-Dimensional Multivariate Data" journal of computational and graphical statistics, ,vol.16, number 1,pages 228-254
13-Rosipal ,R., Kr¨amer ,N.,(2006)'' Overview and Recent Advances in Partial Least Squares'' Springer-Verlag Berlin Heidelberg pp. 34–
14-Vanden Branden. K. & Hubert. M,(2003) ''Robustness Properties of a robust PLS regression method'' December 10. J.Analytica Chimica p229-241.
15-verardi .v and croux .V(2009) ''Robust Regression in stata'' K.u.leuven Facutly of Business and Economices.
16- Wakeling .I.N and Macfie. H.J.H ''A Robust PLS Procedure'' (1992) journal of chemometrics, vol.6,189-198.
17- Yuditskaya,s., (2010)" An Overview of Methods in Linear Least-Squares Regression"pattern recognition and Analysis ,November 4

Comparison of some robust methods to estimate parameters of partial least squares regression (PLSR)

Abstract

The technology of reducing dimensions and choosing variables are very important topics in statistical analysis to multivariate. When two or more of the predictor variables are linked in the complete or incomplete regression relationships, a problem of multicollinearity are occurred which consist of the breach of one basic assumptions of the ordinary least squares method with incorrect estimates results.

There are several methods proposed to address this problem, including the partial least squares (PLS), used to reduce dimensional regression analysis. By using linear transformations that convert a set of variables associated with a high link to a set of new independent variables and unrelated with each other, which are called, the components. These components are orthogonal and independent from each other.

The method of partial least squares PLS is failed in dealing with data that consist of the presence of Outliers values and hence the success of this method depends on the absence of such outliers values that have undesirable effect on the results. In order to reduce the presence of these values, we resorted to use the robust methods.

In this research a method of PLSKURSD that applied SIMPLS algorithms on variance-covariance robust matrix. Also the proposed method MPLSKURSD are used which is a modified method to the PLSKURSD method. parameters linear regression model by partial least squares(PLS) is compared with modalities robust partial least squares through the simulation experiments depends on the presence of several types of outlier values of data for different rates of pollution, volumes of samples, and variables dimensions

Keyword: Partial Least Squares; Robust Covariance Matrix; kurtosis; Projection