طريقة مقترحة لتعديل معلمة الموقع

أ. م. د. لقاء علي العلوي
كلية الادارة والاقتصاد- جامعة بغداد
قسم الاحصاء

الملخص

تقدير الموقع والتشتت في متعدد المتغيرات مع خاصيتي تساوي التغاير والانهيار الموجب دائمـا يكون صعب والمقدر الذي يحقق هذه الخواص هو مقدر القطع البيضوي الأصغر Minimum Volume Ellipsoid (MVE) Estimator . حساب (MVE) المضبوط غالبا يكون غير مقنع لذلك يتم إعادة ترتيب خوارزمية تقريبية لحسابه. في الانحدار هناك خوارزميات عديـدة لحـساب مقـدرات الانهيـار الموجب مثل اصغر وسيط للمربعات الذي يعيد حساب التقاطع في كل خطـوة لتحقيـق النتيجـة. هـذا الأسلوب يدعى بـ (تعديل الموقع). في هذا البحث اشرنا إلى تقنيات مشابهة تدعى تعديل الموقع، ممكن وضعها في حساب (MVE). لهذا الغرض تم استخدام اصغر قطع كروي Minimum Volume Ball (MVB) لغرض تقليل دالة الهدف لـ (MVE). تم اعتبار خوارزمية مـضبوطة لحـساب (MVE). وكبديل لتعديل الموقع (MVB) تم استخدام تعديل الموقع (L_1) في مرحلة ثانية، والـذي لـيس مـن الضروري ان يقلل دالة الهدف لـ (MVE) ولكنه يفرز مقدرات أكثر كفاءة لجزء الموقع. والمحاكـاة تمت لمقارنة هذه الأنواع في تعديل الموقع.

Abstract

Estimating multivariate location and scatter with both affine equivariance and positive break down has always been difficult. Awell-known estimator which satisfies both properties is the Minimum volume Ellipsoid Estimator (MVE) Computing the exact (MVE) is often not feasible, so one usually resorts to an approximate Algorithm. In the regression setup, algorithm for positive-break down estimators like Least Median of squares typically recomputed the intercept at each step, to improve the result. This approach is called intercept adjustment. In this paper we show that a similar technique, called location adjustment, Can be applied to the (MVE). For this purpose we use the Minimum Volume Ball (MVB). In order to lower the (MVE) objective function. An exact algorithm for calculating the (MVB) is presented. As an alternative to (MVB) location adjustment we propose (L_1) location adjustment, which does not necessarily lower the (MVE) objective function but yields more efficient estimates for the location part. Simulations Compare the two type of location adjustment.

1. المقدمة Introduction

ان وجود قيم شاذة outliers ضمن مجموعة البيانات تؤثر بـشكل كبيـر فـي نتـائج التحليـل الاحصائي لتلك البيانات ويظهر هذا في حقيقة ان اغلب الطرائق الإحصائية تؤكـد علـى افتـراض ان البيانات المستخدمة في التحليل هي بيانات متجانسة homogeneous data أي ان كل نقاط البيانـات الموجودة في مجموعة معينة تتبع نفس التوزيع المفترض لذلك طرائق بديلة طورت للتعامل مع الأخطاء في النموذج والتلويث في البيانات.

في التحليل الحصين مقدر اصغر قطع بيضوي Minimum Volume Ellipsoid Estimotor (MVE) غالبا مايستخدم في تقدير الموقع والتشتت، حيث يعرف (MVE) نسبة الى (Rousseeuw 1985) كاصـغر قطـع بيـضوي منـتظم يحـوي علـى الاقـل h مـن عناصـر مجموعـة البيانات $X=\{X_1,X_2,\dots,X_n\}\subset R^p$، حيث ان مقـدر الموقـع Locayion Estimator للـ (MVE) هو مركز ذلك القطع، بينما مقدر التشتت Scatter Estimator في الـ (MVE) يعود الى شكل المصفوفة، وبذلك يمكن تقليل مقدرات (MVE) بالصورة ادناه :

$$\left(\hat{\mu},\hat{\Sigma}\right)=\underset{\substack{(\mu,\Sigma)\in R^p\, XSPD(p)\\ |\Sigma|=1}}{arq\min}\; d_h^2(\mu,S) \qquad \dots\dots\dots (1)$$

حيث SPD(P) مجموعة كل المصفوفات المتماثلة الموجبة قطعـا Symmetric Positive definit matrix ($S\in R^{P\times P}$) . يطلق على $\hat{S}$ مصفوفة الشكل Shape matrix لان $\hat{S}$ تحدد شكل القطـع البيضوي وليس الكمية ، طالما ضروريا $|\hat{S}|=1$.

بواسطة $d_h^2(\mu,S)$ تعرف المسافة التربيعية h^{th} بين X_i و µ بواسطة المصفوفة S كما يلي

$$d_h^2(\mu,S)=\left\{(X_i-\mu)' S^{-1}(X_i-\mu);\le i\le n\right\}_{(h)}=\left\{\|X_i-\mu\|s^2;1\le i\le n\right\}_{(h)}$$
$$=median_i\|X_i-\mu\|_s^2 \qquad \dots\dots\dots\dots(2)$$

حيث الوسيط للاحصاءات المرتبة (h^{th}) Order Statistic . ولهذا مقدر(MVE) يكون بالصيغة

$$\left(\hat{\mu},\hat{\Sigma}\right)=\left(\hat{\mu},c(n,p,h)d_h^2\left(\hat{\mu},\hat{S}\right)\hat{S}\right) \qquad \dots\dots\dots (3)$$

حيث c(n,p,h) هو عامل التصحيح Correction Factor لتكون $\hat{\Sigma}$ متسقة Consistent لـ Σ في النموذج الطبيعي وقيمته مساوية الى $c=1+\dfrac{15}{n-p}$.

اذا اردنا تقليل نقطة الانهيار للمقدر، القيمة لـ h في المعادلة (1) و (3) تكون $h=\dfrac{n+p+1}{2}\approx\dfrac{n}{2}$.

لكن اذا (هذا ما يحدث غالبا) عرفنا كسر الشواذ على الاكثر α حيث $0 < \alpha < 1/2$. نستطيع العمل مع مقدر(α) MVE حيث $h = [n(1-\alpha)]$. اختيار $\alpha = 1/4$ يكـون جيد وبهذا نستطيع وضع خوارزمية ايجاد مقدر (MVE) كما يلي :

1. بعد تحديد حجم العينة (n) يتم تحديد العينات الجزئية بحجم (p+1) في المشاهدات ، أي نختار C^{n}_{p+1} من العينات .
2. لكل عينة جزئية يتم حساب متجه الاوساط ومصفوفة التشتت ليجري بعدها استخراج المسافات التربيعية .
3. ترتب قيم المسافات تصاعديا ويتم اختيار d_h^2 لكل عينة جزئية ، ليتم بعدها اختيار افضل عينة جزئية وهي العينة التي تمتلك اصغر دالة هدف (1).
4. بعد ايجاد مصفوفة التشتت المصححة يتم استخراج المسافات التربيعية وفقا لكافة المـشاهدات ليتم بعدها الاسلوب التعاقبي في الحساب والذي يتم بزيادة مشاهدة واحدة الى حجم المجموعة الجزئية ليصبح (p+2) ثم لتصل بعدها الى (h) من المشاهدات .

اذا وجدنا نقاط h نستطيع تغطيتها بشكل بيضوي باصغر قطع عـشوائي . فـي المعادلـة (1) اصـغر $d_h^2(\mu, S)$ تصبح صفر ولكنها لاتمتد لكافة المشاهدات الشرط هذا كافي لايجاد مقدرات(MVE) طالما اصغر قطع هو الاصغر للعدد المنتهي $\binom{n}{h}$ للقطوع الموجبة بالضبط .

2. حساب المقدرات الحصينة للانحدار

Computation of the Robust Estimators in Regression

في نموذج الانحدار$\gamma_i = \beta' X_i + \alpha + \in_i (i = 1,.......n)$ معلمة الميل **Slope Parameter** هي $\beta \in R^{p-1}$ ومعلمة التقاطع **intercept parameter** هي $\alpha \in R$ لذلك فان مقدر اصغر وسيط للمربعات **Least Median of Square (LMS) Estimator** لـ **(Rousseeuw 1985)** بالنـسبة لـ (α،β) يعرف بواسطة :

$$(\hat{\alpha}, \hat{\beta}) = \underset{(\alpha,\beta)\in R\times R^{p-1}}{\arg\min}\, median(y_i - \beta' X_i - \alpha)2 \qquad \text{.......... (4)}$$

عادة، تقدير التقاطع $\hat{\alpha}$ يحسب مشروطا بقيمة تقدير الميل $\hat{\beta}$ لغرض تخفيض قيمة دالة الهدف. هـذا الاسلوب في تعديل التقاطع **intercept adjustment** يتضمن تشريح مشكلة التصغير (4) الى جزئين:

$$\hat{\beta} = \underset{\beta\in R^{p-1}}{\arg\min}\, median(y_i - \beta' X_i - \hat{\alpha}(\beta))2 \qquad \text{......... (5)}$$

$$\hat{\alpha}(\hat{\beta}) = \underset{\alpha\in R}{\arg\min}\, median(y_i - \beta' X_i - \alpha)2 \qquad \text{.......... (6)}$$

وهذا يعود لقولنا ان $\hat{\alpha}(\beta)$ هي تقدير الموقع لـ (LMS) لحادي التغيـر لـ ـ ـ (n) مـن العناصـر $\gamma_i - \beta' X_i \ , \ (i = 1,2,.......n)$. وهناك خوارزمية مضبوطة لتقدير موقع (LMS) لحادي المتغيـر، طالما انها نقطة الوسط لاصغر فترة تحتوي h من المشاهدات .
بصورة ملائمة لما ذكر اعلاه، نستطيع كتابة (1) بالشكل :

$$\hat{S} = \underset{\substack{S \in SPD(P) \\ |S|=1}}{\arg\min} \ d_h^2\left(\hat{\mu}(s), S\right) \qquad .... (7)$$

حيث لـ (s) المعطاة مع $|s| = 1$ نضع :

$$\hat{\mu}(s) = \underset{\mu \in R^p}{\arg\min} \ d_h^2\left(\mu, S\right) \qquad .... (8)$$

$$\hat{\mu}(s) = \underset{\mu \in R^p}{\arg\min} \, median \left\| x_i - \mu \right\| S$$

$$\therefore \ \hat{\mu}(s) = \underset{\mu \in R^p}{\arg\min} \, median \left\| S^{-1/2} x_i - S^{-1/2} \mu \right\| \qquad .... (9)$$

حيث $S^{1/2}$ تعرف بالجذر المتماثل لـ ـ S أي ان ($S = S^{1/2} S^{1/2}$) مع $S^{1/2}$ المتماثلة . باسـتخدام مجموعة بيانات التحويلات $\mathrm{Y} = \left\{ y_i = S^{-1/2} X_i \ , \ i = 1,........n \right\}$ نحصل على:

$$\hat{\mu}(s) = S^{1/2} \ \arg \underset{\theta}{\min} \, median \left\| y_i - \theta \right\| \qquad .... (10)$$

نعرف ان $\hat{\mu}(\hat{S})$ مع $\hat{S}$ المعرفة بـ ـ (7) تـساوي $\hat{\mu}$ ولـذلك تبقـى متـساوية التغـاير affine equivariant . قيمة θ التي تصغر وسيط $\left\| y_i - \theta \right\|$ هي مركـز الكـرة مـع مقـدر اصـغر قطـع (غير صفري) الذي يحوي على الاقل h من مجموعة البيانات Y. وهذا يقود لمقدر اصغر قطع كـروي Minimum Volume Ball Estimator (MVB) معرف من قبل (Rousseeuw 1984,PP877) علما ان هذا المقدر يتمتع بخصائص الحـصانة (متـساوي التغـاير ومتعامد وذا نقطة انهيار مساوية الى 50%) .

كذلك نحن نفترض مقدرات الانحدار الحصينة والمعرف مقياس الحـصانة لهـا بتقليـل انتـشار البواقي. كمثال عندما نـستخدم الانحـراف القياسـي المـشذب Trimmed standard deviation فنحصل على مقدر المربعات المشذبة الـصغرى Least Trimmed squares estimator (LTS) لـ ـ ـ (Rousseeuw &Leroy 1987) للعلم فان الغرض من اجراء تعديل الموقع في (MVE) هو لتخفيض دالة الهدف (1). هذا مشابه لتقنية تعديل التقاطع في الانحدار الحصين والذي يكون ذو فائدة.

3. حساب اصغر قطع كروي Computation of Minimum Volume Ball

لناخذ $Y = \{\gamma_1, \gamma_2, ..., \gamma_n\} \subset R^p$ ، مقدر موقع (MVB) يعرف بـ

$$MVB(Y) = \underset{\mu \in R^p}{\arg\min}\, median \|\gamma_i - \mu\| \quad .... (11)$$

وثانية الوسيط يمثل الاحصاءة المرتبة h^{th}

الخطوات ادناه تعطي الخوارزمية المضبوطة لمقدر (MVB) .

1. نضع R_{best} الابتدائية مساوية الى ∞ +

2. لاي عـــدد $2 \le k \le P+1$ ولاي مجموعــة جزئيــة k وان $J = \{i_1, i_2, ..., i_k\} \subset \{1,2,......., n\}$ يكون:

   a. نضع $A_J = Affinespan\{\gamma_i : j \in J\}$ واذا كان $\dim(A_J) \prec k-1$ نرجع للخطـوة 2 (أي نحذف هذه المجموعة الجزئية J)

   b. لذلك يكون $\dim(A_J) = k-1$. نحدد النقطة الوحيدة μ_J في A_J التي تقع في نفس البعد الاقليدي Euclidean distance لكل y_j لكل $j \in J$ وذلك بحل kxk من نظام المعـادلات الخطية .

   c. نحسب $R_J = median_i \|\gamma_i - \mu_J\|$ واذا كان $R_J \ge R_{best}$ نرجع للخطوة 2

   d. نسجل $R_{best} = R_J$ وكذلك $\mu_{best} = \mu_J$

3. نسجل $MVB(Y) = \mu_{best}$

المشكلة هي في ايجاد k والمجموعة الجزئية $\{i_1, i_2, ..., i_k\}$ حيث نبحث خلال كل المجموعـات الجزئية J الممكنة لكل الاحجام الممكنة ونحسب R_J .

طالما ان خوارزمية (MVB) المضبوطة تاخذ وقت طويل (وقت تكرار العملية يعتمد علـى n) لـذلك وجدت خوارزمية تقريبية بديلة ، حيث الخطوة 2 تتعامل فقط مع المجموعة الجزئية N_{samp} ذات الحجم (p+1) (بواسطة المحاكاة نجد ان المجموعة الجزئية J التي تنتج R_{best} و μ_{best} ذات حجـم (p+1) باحتمال عالي).

بديل اخر يكون قريب جدا لخوارزمية المجموعة الجزئية (p+1) التقريبية يكون :

1. لكل j=1,2,…….. n نحسب $R_j = median_i \|\gamma_i - \gamma_j\|$.

2. نضع MVB(y) مساوية الى المشاهدة y_j لنحصل على اوطا R_j .

هذه الخوارزمية اقل كلفة ولكنها تبقى تحسب مقدرات متـساوية التغـاير متعامـدة وذات 50% نقطـة انهيار.

4. تعديل الموقع بواسطة Location Adjustment by MVB (MVB)

ايجاد الحل المضبوط لـ (MVB) الذي يقال المشكلة (1) يكون غالبا غير مقنع . لذلك تتم اعـادة ترتيب خوارزمية المجموعة الجزئية (p+1) التقريبية. بسهولة يمكن تكييف هـذه الخوارزميـة بعـدم اشتراك تعديل الموقع باستخدام مقدر اصغر قطع كروي. وهذا يقود الى الخوارزمية الاتية:

1. نحدد القيمة الابتدائية لـ R_{best} بـ ـ ∞ +
2. لكل مجموعة جزئية (p+1) حيث $J \subset \{1,2,........,n\}$ نعمل مايلي:
   a. نحـــسب $\mu_j = \frac{1}{p+1}\sum_{i\in J} \mathrm{X}_i$ وكـــذلك $C_J = \frac{1}{p}\sum_{i\in J}(\chi_i - \mu_J)(\chi_i - \mu_J)'$ واذا كـــان $|C_J| = 0$ نرجع للخطوة 2.
   b. نحسب $S_J = |C_J|^{-1/p} C_J$ لتكون $|S_J| = 1$.
   c. تحول مصفوفة البيانات X الى Y حيث $\mathrm{Y} = \{S_J^{-1/2}\mathrm{X}_i \ ; \ i = 1,2,.......n\}$
   d. نحسب تقدير $\theta_J = MVB(Y)$ ونضع $R_J = median\|\gamma_i - \theta_J\|$
   e. اذا كان $R_J = R_{best}$ نرجع للخطوة 2
   f. نسجل التقدير الاخير $(\hat{\mu}, \hat{\Sigma})$ حيث $\hat{\mu} = S_{best}^{1/2}\theta_{best}$ وكذلك $\hat{\Sigma} = C(n,p,h)R_{best}^2 S_{best}$ نعرف ان اوطا قيمة لدالة الهدف (1) مساوية الى R_{best}^2 .

بزيادة الابعاد P، يصبح من غير الممكن اعتماد كل المجموعات الجزئية $\binom{n}{p+1}$ لذلك نـستطيع ان نبقى نبحث باسلوب الاختيار العشوائي للمجموعات الجزئية N_{samp} ذات الحجم (p+1) . نستطيع كذلك وضع تعديل (MVB) مرة واحدة كتحسين اخير لخوارزمية المجموعة الجزئية (P+1) العادية لمقـدر (MVE) . من الضروري الاشارة الى ان كل هذه النسخ لـ (MVE) المقارنة مـع تعـديل الموقـع (MVB) هي نظريات متساوية التغاير . لاننا نضع (MVB) المتساوية التغاير والمتعامد للبيانات فـي مصفوفة (MVE) .

5. المحاكاة Simulation:

خوارزمية المجموعة الجزئية (p+1) لـ (MVB) وضعت من قبل (Rousseeuw & Leroy 1987 pp 259-260) ووضعت ضمن برامجيات SAS,S-plus . هذا المبحث ضم دراستين للمحاكاة ، الاولى ضمن متعدد المتغيرات والثانية في تحليل الانحدار . دراسة المحاكاة الاولى قارنت بين المقدرات الاتية من حيث الموقع والتشتت: مقدر المجموعة الجزئية (p+1) الاصلي $(\hat{\mu},\hat{\Sigma})$، مقدر المجموعة الجزئية (p+1) ذا $(\tilde{\mu},\tilde{\Sigma})$ مع تعديل (MVB) في كل خطوة ، وفي النهاية مقدر المجموعة الجزئية (p+1) ذا $\left(\tilde{\tilde{\mu}},\tilde{\tilde{\Sigma}}\right)$ مع تعديل (MVB) مرة واحدة في النهاية. تم توليد نوعين من البيانات الاول هو حالة البيانات الطبيعية حيث $X_i \approx N(0,I_p)$ لكل $i=1,2,......n$. وفي الحالة الثانية هناك 20% من المشاهدات الملوثة (بابدالها بـ $100e_1$ حيث e_1 هو المتجهة الاحادي الاول first unit vector) . هذا ينتج عنقود من الشواذ المتطرفة . تم اولا اعتماد , n=30, p=2 مع $N_{samp}=400$ وفي الحالة الثانية p=3, n=400 و $N_{samp}=500$. القيم الملخصة لـ m=500 من التكرارات تم تسجيلها اضافة الى التحيز Bias ومتوسط مربعات الخطا mean squared error لمقدرات الموقع . علما ان:

$$Bias(\hat{\mu}) = \|\bar{\mu}-\mu\| = \left\|(\frac{1}{m}\sum_{k=1}^{m}\hat{\mu}^k)-\mu\right\| \qquad .... (12)$$

$$MSE = \frac{1}{m}\sum_{k=1}^{m}\left\|\hat{\mu}^k-\mu\right\|^2 \qquad .... (13)$$

حيث $\hat{\mu}^k$ هو تقدير الموقع لعينة المحاكاة k^{th} والمعلمة الحقيقية هي $\mu=0$.

لقياس انحراف التسطح deviation from sphericity لمصفوفة التشتت المقدرة $\hat{\Sigma}^{th}$ للعينة k^{th} ، يتم حساب

$$\phi_k = \frac{trac\left(\hat{\Sigma}^k/p\right)^p}{\det\left(\hat{\Sigma}^k\right)} \qquad .... (14)$$

والجدول (1) ادناه يسجل قيم $\underset{k=1,2,.....,m}{median\ In\ \phi_k}$ نعرف ان المصفوفة $\hat{\Sigma},\tilde{\tilde{\Sigma}}$ تختلف بعامل واحد فقط ، لذلك لها نفس انحراف التسطح .

اخيرا، القيمة المتوسطة لدالة الهدف (1) لـ (m) من التكرارات تم ادراجها ايضا ضمن الجدول (1) والذي نلاحظ من خلاله ان تعديل (MVB) يخفض دالة الهدف لـ (MVE) مقارنة بخوارزمية (MVE) ذات المجموعة الجزئية (p+1) الاصلية وخاصة عندما يتم التعديل في كل خطوة. في الحقيقة، بواسطة الخوارزمية نعرف ان $(\tilde{\mu},\tilde{\Sigma})$ دائما تنتج قيم اوطا لدالة الهدف عن $\left(\tilde{\tilde{\mu}},\tilde{\tilde{\Sigma}}\right)$ والتي بتكرارها تنتج قيم اوطا عن $(\tilde{\mu},\tilde{\Sigma})$. نلاحظ من جدول (1) ايضا ان التحيز ومتوسط مربعات الخطا اضافة الى $median_k\ In\phi_k$ يبقى نفسه. حتى عندما يتم التعديل في كل خطوة فاننا نحصل على الدقة بصعوبة. بينما وقت الحسابات يزداد فجاة.

جدول رقم (1)

التحيز ومتوسط مربعات الخطا لمقدرات الموقع مع وسيط $In\phi_k$ لمقدرات التشتت اضافة الى متوسط قيمة دالة الهدف (MVE) المحسوبة باستخدام تعديل الموقع (MVB) .

		البيانات الطبيعية			البيانات الملوثة 20%		
		$(\hat{\mu},\hat{\Sigma})$	$(\tilde{\mu},\tilde{\Sigma})$	$(\tilde{\tilde{\mu}},\tilde{\tilde{\Sigma}})$	$(\hat{\mu},\hat{\Sigma})$	$(\tilde{\mu},\tilde{\Sigma})$	$(\tilde{\tilde{\mu}},\tilde{\tilde{\Sigma}})$
p=2	$Bias(\hat{\mu})$	**0.0079**	**0.0079**	**0.0077**	**0.028**	**0.015**	**0.018**
n=30	$MSE(\hat{\mu})$	**0.234**	**0.229**	**0.231**	**0.259**	**0.228**	**0.252**
	$med_k\ In\phi_k$	**0.594**	**0.586**	**0.594**	**0.408**	**0.364**	**0.408**
	$Ave_k\ obj_k$	**1.016**	**0.878**	**0.985**	**1.508**	**1.293**	**1.453**
p=3	$Bias(\hat{\mu})$	**0.018**	**0.023**	**0.023**	**0.037**	**0.035**	**0.028**
n=40	$MSE(\hat{\mu})$	**0.295**	**0.266**	**0.319**	**0.310**	**0.301**	**0.353**
	$med_k\ In\phi_k$	**0.921**	**0.786**	**0.921**	**0.702**	**0.719**	**0.702**
	$Ave_k\ obj_k$	**2.150**	**1.897**	**2.108**	**2.991**	**2.616**	**2.885**

دراسة المحاكاة الثانية تمت لمشابهة النتائج اعلاه لشكل او اطار الانحدار ، حيث تـم مقارنـة مقـدر المربعات المشذبة الصغرى (LTS) مقارنة مع او بدون تعديل التقاطع. في حالة وحيد المتغير (LTS) يحسب بالضبط من خلال خوارزمية

(Rousseeuw & Leroy 1987) . تم توليد ثلاث حالات مختلفة ، في الاولى النموذج معطى بالشكل:

$$\gamma_i = \beta_1\chi_{i1} + \beta_2\chi_{i2} + \beta_3\chi_{i3} + \alpha + e_i$$

حيث i =1,2,…….,40 مع $\beta_1 = \beta_2 = \beta_3 = \alpha = 1$ حيث $e_i \approx N(0,1)$ والمتغيرات التوضـيحية مولدة باستقلال (غير معتمدة) بالشكل $X_{ij} \approx N(0,10)$ لكل j =1,2,3 . في الحالة الثانيـة تبـدل اول (8) نقاط بالشواذ في الاتجاه y (20% من التلوث) اي ان $e_i \approx N(10,1)$ لكـل i =1,2,………8 . وفي الحالة الثالثة تبدل بالشواذ في اتجـاه X اي ان $X_{ij} \approx N(100,10)$ مـع $e_i \approx N(0,1)$ لكـل i =1,2,………8 . التحيز ومتوسط مربعات الخطا المعرفة في (12) و (13) سجلت لمعـالم الانحـدار ومعالم التقاطع باستخدام خوارزمية المجموعة الجزئية (p) بدون تعديل التقاطع ، ومع تعديل التقـاطع في كل خطوة . واخيرا باستخدام تعديل التقاطع فقط في الخطوة النهائية.

من جدول (2) نرى ان تعديل التقاطع في الحقيقة يخفض دالة الهدف لـ (LTS) خـصوصا اذا اجري التعديل في كل خطوة ، حيث التحيز و (MSE) للمعاملات لا تتغير معنويـا. هـذا يؤيـد نتـائج المحاكاة للموقع والتشتت في التجربة الاولى الخاصة بمتعدد المتغيرات (جدول (1)).

جدول رقم (2)

التحيز ومتوسط مربعات الخطا لمقدرات الميل والتقاطع مع متوسط قيمة دالة الهدف لـ (LTS) المحسوبة بتعديل التقاطع لانحدار (LTS).

	البيانات القياسية			20% شواذ عمودية			20% شواذ افقية		
	$(\hat{\beta},\hat{\alpha})$	$(\tilde{\beta},\tilde{\alpha})$	$(\tilde{\tilde{\beta}},\tilde{\tilde{\alpha}})$	$(\hat{\beta},\hat{\alpha})$	$(\tilde{\beta},\tilde{\alpha})$	$(\tilde{\tilde{\beta}},\tilde{\tilde{\alpha}})$	$(\hat{\beta},\hat{\alpha})$	$(\tilde{\beta},\tilde{\alpha})$	$(\tilde{\tilde{\beta}},\tilde{\tilde{\alpha}})$
$10_2^2 \text{X}\, Bias(\hat{\beta})$ $10^2 MSE(\hat{\beta})$	**0.142 0.079** **0.142** **0.463 0.460 0.463**			**0.331 0.262 0.331** **0.422 0.430 0.422**			**0.125 0.172 0.125** **0.414 0.385 0.414**		
$10^2\ Bias(\hat{\alpha})$ $MSE(\hat{\alpha})$	**1.677 1.749 1.70** **0.129 0.124 0.126**			**0.227 0.275 0.786** **0.134 0.123 0.122**			**2.010 1.428 2.720** **0.020 0.014 0.027**		
$AVe_k\ obj_k$	**0.810**	**0.781**	**0.799**	**1.195**	**1.148**	**1.174**	**1.194**	**1.14 7**	**1.174**

6. **تعديل** L_1 L_1 Adjustment

مما تقدم نرى ان تعديل (MVB) يخفض قيمة دالة الهدف لـ (MVE) لكن هذا التعـديل لـيس من الضروري ان يزيد من كفاءة المقدرة للعينة النهائية لكنه يبقى يستلزم وقت اطول فـي الحـساب ، لهذا يتم ابدال تعديل (MVB) بتعيل اخر يستلزم وقت اقصر في الحساب . مقدر الموقع L_1 **يكون ملائم طالما له نفس خصائص مقدر (MVB) بالنسبة لخاصية تساوي التغاير ونقطة الانهيار العالية اضـافة الى كفاءته الاحصائية المحسوبة بوقـت اقـصر.لمجموعـة البيانـات المعطـاة ذات p مـن الابعـاد** $\text{Y} = \{y_1, y_2, ..., y_n\}$ **فان مقدر** L_1 **أي** $\mu_L(\text{Y})$ **يكون حل لمشكلة التقليل:**

$$\mu_L = \arg\min_{\mu \in R^p} \sum_{i=1}^{n} \| y_i - \mu \| \qquad \text{..........} \ (16)$$

علما ان الخوارزمية الاسرع لمقدر L_1 **مقدمة من قبـل (Hossjer & Croux 1995) . الان يمكـن عمل تعديل الموقع من خلال مقدر** L_1 **ولاجل ذلك نطبق الخوارزمية المقدمة في المبحث الثالث وتبـديل الخطوة d بالاتي :**

d. نحسب تقدير L_1 **،** $\theta_J = \mu_L(\text{Y})$ **ونحسب كذلك** $R_J = median \| y_i - \theta_J \|$ **. أي شيء غيـر ذلـك يبقى نفسه. بالتاكيد ممكن حساب المقدرات سوية في حالة البحث في كـل المجموعـات الجزئيـة ذات الحجم (p+1) من خلال** N_{samp} **وتسحب عشوائيا المجموعات الجزئية (p+1) . كذلك ممكـن وضـع تعديل الموقع** L_1 **مرة واحدة في النهاية . هذا يقود الى مقدر الخطـوتين two – stage estimator المعرف بـ :**

نحسب مقدرات (MVE) أي $\left(\hat{\mu},\hat{\Sigma}\right)$ وتبدل بعدها $\hat{\mu}$ بـ :

$$\tilde{\tilde{\mu}} = \arg\min_{\mu \in R^p} \sum_{i=1}^{n} \left\| x_i - \mu \right\| \qquad .... (17)$$

نقطة الانهيار والخصائص التقريبية لهذا المقدر مقدمة من قبل (Hossjer & Croux 1995) علما ان كل هذه النسخ لـ (MVE) مع تعديل الموقع L_1 تبقى متساوية التغاير . ولغرض تطبيق المقدر تم اعادة دراسة المحاكاة الاولى المقدمة في المبحث الخامس لمقارنة خوارزمية المجموعة الجزئية (p+1) العشوائية لـ (MVE) محسوبة مع او بدون تعديل الموقع L_1 . الجـدول (3) ادناه يوضح النتائج التي تم التوصل اليها، وكما هو متوقع فان هذا النوع من تعديل الموقع لايقلـل دالة الهدف لـ (MVE) كذلك فان تاثير تعديل L_1 على التحيز $Bias(\hat{\mu})$ ومتوسط مربعـات الخطـا $MSE(\hat{\mu})$ وكذلك $median_k\ In\phi_k$ تبقى نفسها عندما يجري التعديل في كل خطوة او فقط في النهاية. النسخة الاخيرة تعرف بـ $\left(\tilde{\tilde{\mu}},\tilde{\tilde{\Sigma}}\right)$ وهي الاسرع في الحساب.

جدول رقم (3)

التحيز ومتوسط مربعات الخطا لمقدرات الموقع مع وسيط $In\phi_k$ لمقدرات التشتت ومتوسط قيمة دالة الهدف لـ (MVE) المحسوبة باستخدام تعديل L_1

		البيانات الطبيعية			20% البيانات الملوثة		
		$\left(\hat{\mu},\hat{\Sigma}\right)$	$\left(\tilde{\mu},\tilde{\Sigma}\right)$	$\left(\tilde{\tilde{\mu}},\tilde{\tilde{\Sigma}}\right)$	$\left(\hat{\mu},\hat{\Sigma}\right)$	$\left(\tilde{\mu},\tilde{\Sigma}\right)$	$\left(\tilde{\tilde{\mu}},\tilde{\tilde{\Sigma}}\right)$
p=2 n=30	$Bias(\hat{\mu})$	0.0079	0.0062	0.0034	0.0211	0.4082	0.4015
	$MSE(\hat{\mu})$	0.234	0.093	0.094	0.259	0.277	0.272
	$med_k\ In\phi_k$	0.594	0.531	0.594	0.408	0.404	0.406
	$Ave_k\ obj_k$	1.016	1.055	1.332	1.508	1.677	2.178
p=3 n=40	$Bias(\hat{\mu})$	0.0178	0.0036	0.0057	0.0369	0.4605	0.4601
	$MSE(\hat{\mu})$	0.295	0.092	0.093	0.310	0.328	0.328
	$med_k\ In\phi_k$	0.921	0.845	0.921	0.702	0.693	0.702
	$Ave_k\ obj_k$	2.150	2.093	2.461	2.991	3.109	3.695

من الجدول اعلاه نرى ان $median_k\ In\phi_k$ لمقدر مصفوفة التشتت يبقـى نفـسه فـي جميـع الحالات. اما بالنسبة لمقدرات الموقع، فهي تعتمد على فيما اذا كانت البيانات طبيعية او ملوثة. للبيانات الطبيعية (غير الملوثة) نرى ان تعديل L_1 يقدم تحيز قليل وتخفيض في متوسط مربعات الخطا طالما ان L_1 له افضل كفاءة احصائية ، لكن للبيانات الملوثة فان (MSE) تبقى نفسه بينما التحيز يصبح عـالي جدا.

7. مقارنة مع خوارزمية الحل المقنع

Acomparison with the feasible Solution Algorithm

مما تقدم نرى ان الحسابات المضبوطة لـ (MVE) تكون كالاتي : اعتبر كل المجموعات الجزئية الممكنة ذات الحجم (h) وتدعى بالعينات النصفية half samples ويحسب القطع لاصغر شكل بيضوي يحوي كل النقاط المحسوبة لـ (MVE) هو الشكل البيضوي الخاص بتلك العينة النصفية المثلى: طالما العدد الكلي للعينات النصفية الممكنة كبير جدا فالحسابات المضبوطة تكون غير مقنعة في التطبيق العملي، ماعدا حجم العينة الصغير جدا لذلك عادة ما يتم اعادة ترتيب الخوارزميات التقريبية.

في هذا المبحث سوف نركز على خوارزمية Feasible Solution Algorithm (FSA) المعروفة والمستخدمة في حل مقدر الانحدار اصغر وسيط للمربعات Least Median of square regression . (Hawkins 1993 a)

FSA تبدا بالاختيار العشوائي للعينة النصفية ليتم بعد ذلك ابدال النقاط في العينة النصفية مع النقاط غير العائدة لها، وهذا يقلل قيمة دالة الهدف، وهذا معناه ان القطع لاصغر شكل بيضوي يحوي كل النقاط في العينة النصفية. اذا لم نلاحظ انخفاض أكثر فالعينة النصفية الحاصل عليها يطلق عليها الحل المقنع Feasible Solution . الخوارزمية تعتمد على العدد الكلي N_{fsa} للبدايات العشوائية. هذا العدد يكون كبير بدرجة كافية خاصة عند تعامل FSA مع مجموعة بيانات قلقة Nosily data set (يبدا البرنامج مع $N_{fsa}=50$).

وقت الحسابات الذي يستلزمه FSA اكثر كثيرا من خوارزمية المجموعة الجزئية (p+1) . كلفة المقارنة لوقت الحسابات صعب، طالما معالم التذبذب Tuning parameters تختار من قبل المستخدم (N_{samp}, N_{fsa}, h) ، واذا كان حجم العينة والابعاد تزداد فان هذا الاختلاف في وقت الحسابات يصبح كبير جدا (علما اننا لا نتعامل مع دالة التعديل في كل خطوة).

من ملاحظة الجدول (4) نرى ان FSA نجحت في ايجاد اوطا قيمة لدالة الهدف خلال المقدرات المعتمدة . حيث تم هنا اعادة دراسة المحاكاة للمبحث الخامس ومنه يمكن القول ان FSA حققت هدفها (حققت قيمة واطئة لدالة الهدف (1)).

على كل حال الفوائد الاحصائية بقيت محددة، حيث لجزء التشتت لا يوجد اختلاف معنوي مع $(\hat{\mu}, \hat{\Sigma})$ او مع $\left(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}}\right)$ اما لجزء الموقع فهناك تحسن البسيط في (MSE) عندما p=3 فان هناك تحيز عالي وايضا MSE وهذا ضمن حالة التعامل بالبيانات الملوثة، وذا بسبب ان الانهيار يحصل 8 مرات في 500 تكرار. زيادة قيمة N_{fsa} الى 100 يلغي هذا الانهيار، ولكن يضاعف وقت الحسابات.

عند الحصول على حل FSA لمشكلة (MVE) يمكن ملاحظة التحسن في تقدير الموقع باضافة (MVB) او تعديل (L_1) في النهاية. الجدول (4) يسجل نتائج المحاكاة لـ FSA مع تعديل الموقع باستخدام (MVB) او L_1 .

جدول رقم (4)

استخدام تعديل الموقع لـ FSA لكل الحالات محسوبة ومتوسط مربعات الخطا لمقدر الموقع ووسيط $In\phi_k$ لمقدرات التشتت اضافة الى متوسط قيمة دالة الهدف لـ (MVE)

		البيانات الطبيعية			20% البيانات الملوثة		
		FSA	+MVB	$+L_1$	FSA	+MVB	$+L_1$
p=2	$Bias(\hat{\mu})$	0.028	0.028	0.022	0.027	0.026	0.388
n=30	$MSE(\hat{\mu})$	0.242	0.242	0.092	0.228	0.227	0.261
	$med_k\ In\phi_k$	0.606	0.606	0.606	0.385	0.385	0.385
	$Ave_k\ obj_k$	0.823	0.823	1.293	1.257	1.257	2.119
p=3	$Bias(\hat{\mu})$	0.006	0.006	0.011	0.411	0.411	0.784
n=40	$MSE(\hat{\mu})$	0.248	0.248	0.098	10.09	10.14	7.580
	$med_k\ In\phi_k$	0.973	0.973	0.973	0.682	0.682	0.682
	$Ave_k\ obj_k$	1.519	1.519	2.241	2.365	2.365	3.752

المدهش انه لا يوجد اختلاف نشاهده بين حلول FSA و FSA+MVB . في الحقيقـة كـلا الاسـلوبين غالبـا (ولكن ليس دائما) نفس النتيجة . اذا FSA وجدت MVE المضبوطة، عندها تعديل (MVB) لايغيـر التقـدير. لاختبار فيما اذا تعديل (MVB) يغير مقدر الموقع يكون لذلك غير ضروري لكن غير كافي وهـو شـرط ايجـاد (MVE) المضبوط. استخدام تعديل (L_1) يفرز فيما مقاربة الى $\left(\tilde{\tilde{\mu}},\tilde{\tilde{\Sigma}}\right)$ في جدول (3): زيـادة قيمـة دالـة الهدف، لكن الكفاءة الاحصائية الافضل مقاسة بواسطة MSE.

8. النتائج

في هذا البحث لاحظنا ان مقدر اصغر كروي (MVB) ممكن استخدامه لتعـديل الموقـع لمقـدر اصغر بيضوي (MVE) . هذا التعديل غالبا ما يقلل دالة الهدف في (MVE) ولكن له تاثير قليل علـى التحيز Bias ومتوسط مربعات الخطا MSE للمتوسط $\hat{\mu}$. من جانب اخر فان تعديل الموقع بالاعتمـاد على (L_1) ليس ضروريا ان يقلل دالة الهدف لـ (MVE) ولكنه يبـرهن علـى كفـاءة $\hat{\mu}$ للبيانـات الطبيعية Normal data.

لغرض تقليل تاثير الشواذ المتطرفة على تعديل (L_1) ممكن استخدام مقدر (L_1) معـاد الـوزن Reweighted (L_1) estimator بدلا عنه وكما في حالة مقدر (L_1) غير معاد الـوزن، فـان هـذا النوع من التعديل لا يخفض دالة الهدف لـ (MVE) بينما التحيز للمقدر الناتج يكون مشابه لما هو في مقدر (MVE)، بينما MSE له يكون افضل حتى عند التعامل بالبيانات الملوثة.

في المبحث الخامس لاحظنا ان تعديل الموقع بـ (LTS) يخفض قيمة دالة الهدف ولكن ليس له تاثير كبير على متوسط مربعات الخطا للمعاملات لكنه يفرز مقدر تقاطع بتحيز واطئ وكفاءة اكثر. الغاية اذا من استخدام تعديل (L_1) هي في انه لا يكلف وقت حسابات عالي، يملك منحني تحيز واطئ ويعطـي مقدرات اكثر كفاءة لمعلمة الموقع في حالة البيانات الطبيعية متعددة المتغيرات.

استخدام تعديل (MVB) او ما يسمى بـ Feasible Solution Algorithm الموضحة في المبحث الثامن يعطي قيم اوطا لدالة الهدف المتعلقة بمقدرات اصغر قطع بيضوي (MVE) ولكن له حسابات عالية، علما ان اوطا القيم لدالة الهدف لا تولد بالضرورة دقة احصائية عالية.

اخيرا من الضروري الاشارة الى ان هناك خوارزميات عديدة وجديدة لحسابات مقدرات عالية الانهيار، كالخوارزمية المعتمدة لحساب مقدر اصغر محدد تباين مشترك Minimum Covariance determinant (MCP) ولكن حساب مقدر (MVE) يبقى هو الاصعب.

المصادر

1. العلوي، لقاء علي، (مقارنة مقدرات التباين المشترك الحصينة في تحليل المركبات الرئيسية) رسالة دكتوراه- كلية الادارة والاقتصاد- جامعة بغداد.
2. Croux,C.,& Haesbroeck, (2002), "Finite – sample efficiencies of Estimators for the Minimum volume Ellipsoid", Journal of statistical computing and simulation, 72,585-596.
3. Croux, C., Van Aelst, S., & Dehon. C., (2003), " Bounded Influence Regression using High Breakdown Scatter Matrix", Annals of the Institute of statistical Mathematics, 55,256-285.
4. Hawkins, D.M.(1993a), (the Feasbile set Algorithm for Least Median of squares Regression", Computational statistics and data Analysis, 16,81-101.
5. Hossjer, O., & Corux, C.(1995), "Generalizing univariate signed Rank statistics for Testing and estimating a multivariate Location parameter", Nan parametric statistics, 4,293-308.
6. Rousseeuw, P.J.(1985), "Multivariate Estimation with High Breakdown point", in Mathematical statistics and Application, VolB, eds.W.Grossmann, G. Pflug, I. Vincze and w.werts, (Dordrecht:Reidel), 283-297.
7. Rousseeuw, P.J. & Leroy, A.M.(1987), "Robust Regression and outlier Detection", New york: John wiley.