Problema stabilității estimărilor econometrice

Date anormale(outliers), metode robuste de regresie ortogonală

Corina Sâman[1]

Introducere

Datele anormale(outliers), prin definiție, pun sub semnul întrebării orice analiză şi inferență pe care dorim s-o obținem din structura statistică a oricărui set de date. Sursa acestei anormalități (erori de măsurare, erori datorate abaterii față de structura statistică considerată a datelor) poate fi de interes, dar mai important este ca aceste date să fie ponderate corespunzător sau excluse din setul de date pentru a preveni perturbarea parametrilor estimați şi a permite inferențe statistice.

Acestea vor fi prezente mai ales în seturi de date cu multe variabile şi/sau multe observații.

Cu privire la o distribuție multivariată a datelor, distanța statistică a unei observații prezintă proprietatea că o creştere a distanței de la medie reflectă o descreştere în probabilitate. Datele anormale (outliers) vor fi deci acele observații care sunt anormale în raport cu distribuția lor de probabilitate, având o probabilitate semnificativ mică de apariție. Datele anormale sunt deci un tip de contaminare. Pot exista astfel de date rătăcite în setul de date sau pot exista clusteri cu astfel de puncte. Prezența anomaliilor în date (outliers) face ca încrederea în inferențele statistice bazate pe modelul asociat datelor să fie mică.

Tehnicile de analiză multidimensională (regresii multivariate, componente principale, analiza factorilor, clasificarea, analiza clusterilor, etc.) se bazează pe statistici empirice (medie, covarianță, corelație) şi minimizarea unor funcții de reziduri care pot fi afectate sever de existența unor date anormale, fie ele cât de puține. Doua lucruri se întâmplă:

- estimarile diferă substanțial de cele obținute dacă nu ar exista date anormale;
- modelul rezultat nu permite detectarea datelor anormale pe baza rezidualelor, distanțelor Mahalanobis sau a altor teste diagnostic.

Două proceduri pot fi folosite:

- estimari robuste: găsirea unui model robust care este similar cu cel ce ar rezulta dint-un set de date fără anomalii;
- detectarea datelor anormale.

[1] Institutul de Prognoză Economică, Centrul de Modelare Macroeconomică, Academia Română

Ambele răspund problemei, dar pornesc din direcţii diferite. In primul caz se gaseşte modelul potrivit cu datele fără anomalii, ceea ce permite identificarea datelor anormale pornind de la reziduale, iar în cel de-al doilea caz detectarea anomaliilor permit să se trateze acestea prin înlăturare sau orice altă procedură de diminuare a efectului lor asupra estimării.

Ne vom ocupa în primul rând de detectarea anomaliilor în date (puncte numite outliers).

Media unui eşantion şi matricea de covarianță empirică stau la baza analizei clasice a datelor, fiind estimatori optimi ai parametrilor de localizare şi împrăştiere, dar foarte sensibili la date anormale. Aceşti estimatori sunt necesari pentru multe tehnici de analiză ca regresia, analiza componentelor principale, analiza factorilor, clasificarea, etc.

In prima parte a studiului am discutat tehnici robuste de estimare a localizării şi împrăştierii datelor, iar în partea a doua am aplicat aceste metode pe un caz real cu trei variabile.

În partea a treia am descris o metodă robustă pentru estimarea unui model de regresie ortogonală, iar in secțiunea a patra am făcut simulări Monte-Carlo pentru a compara performanța diferiților estimatori ai modelului de regresie ortogonală robustă.

1. Tehnici de estimare robustă

1.1 Distanțe statistice

Dacă avem un set de date de dimensiune d şi presupunem că ele reprezintă obsevații ale unui vector de variabile aleatoare $\mathbf{X}^T=(\mathbf{X}_1, ..., \mathbf{X}_d)$, iar o realizare a variabilei aleatoare este $x^T=(x_1, ..., x_d)$.

Pentru a analiza acest set de date şi a vedea dacă există date anormale, adică date care nu aparțin modelului statistic considerat, este necesar să avem definită o distanță. Doi parametrii esențiali pentru definirea distanței sunt media μ şi matricea de covarianță Σ a variabilei aleatoare **X**.

Putem porni în analiza datelor de la distanța Euclidiană, D, față de un estimator al mediei $\hat{\mu}$:

$D^2 = ((x_i - T)'(x_i - T))$, dar atunci nu am luat in considerare variația setului de date de-a lungul axelor care este dată de un estimator al împrăştierii datelor. O distanță mare fată de centru este mai semnificativă pentru un set de date care prezintă o variație mică decât pentru un altul cu variație mare. Distanța trebuie deci să fie determinată în mod invers proporțional de o măsură a împrăştierii datelor. Astfel de distanțe sunt numite distanțe statistice.

Cu privire la o distribuție multivariată a datelor, distanța statistică a unei observații prezintă proprietatea că o creştere a distanței de la medie reflectă o descreştere în probabilitate.

Datele anormale (outliers) vor fi deci acele observații care sunt anormale în raport cu distribuția lor de probabilitate, având o probabilitate semnificativ mică de apariție.

O distibuție multivariat normală de dimensiune d este definită de densitatea de probabilitate:

$$f(x) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}} \exp(-1/2(x-\mu)'\Sigma^{-1}(x-\mu)) \qquad (1)$$

unde Σ este matricea de covarianță şi $\frac{1}{|\Sigma|^{1/2}}$ este Iacobianul transformării.

Exponentul din această expresie corespunde cu unul dintre exemplele de distantă statistică foarte cunoscute, distanța Mahalanobis definită de acesta în 1930 numită şi distanța T^2 a lui Hotelling.

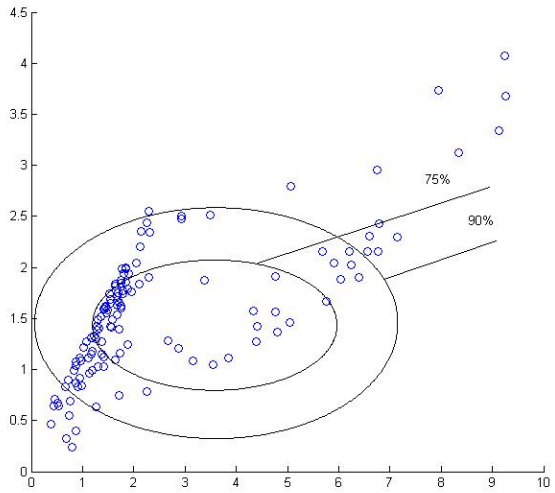
Dacă **X** este d-multivariat normală atunci orice punct cu distanță echivalentă față de centrul datelor se găseşte pe un elipsoid care descrie mulțimile de nivel ale funcției de densitate de probabilitate.

Pentru date astfel distribuite este cunoscut că pătratul distanței Mahalanobis urmează o distribuție chi-pătrat cu d grade de libertate:

$$M_i = ((x_i - \mu)'\Sigma^{-1}(x_i - \mu))^{1/2} \sim \chi^2_d, \qquad (2)$$

unde $P(M_i \le \chi^2_{0.90,d}) = 0.90$. Astfel încât, dacă d = 2, $c^2 = \chi^2_{0.90,2} = 4.605$ şi este 90% şansă ca variabila aleatoare x_i să se găsească în interiorul elipsei descrise de curba $M^2 = c^2$.

O ilustrare a unei astfel de situații pentru un set de date particular este dat în figura următoare. O valoare mai mică pentru c > 0 înseamnă că mai puține date vor fi conținute în elipsa corespunzătoare.



Când identificăm datele anormale (outliers) folosind metode bazate pe distanțe localizăm acele obsrvații care sunt în exteriorul unei astfel de elipse. Determinarea frontierei acestei elipse care separă cele două regiuni se face pornind de la estimarea robustă a a localizării (centrului) şi împrăştierii (covarianței).

Seturile de date pot conține date anormale (outliers), puncte care deviază de la modelul sugerat de majoritatea datelor, aşa că o analiză statistică a acestora trebuie să înceapă în mod necesar cu estimarea lor.

1.2 Invarianța Afină şi Estimarea de Verosimilitate Maximă

In cazul particular când **X** are o densitate de probabilitate multivariat normală (1), estimatorii de verosimilitate maximă pentru eşantionul $x_1,..., x_n$ pentru medie şi împrăştiere sunt dați de acei parametrii $\hat{\mu}$, $\hat{\Sigma}$ care satisfac condiția:

$$L_n(\hat{\mu},\hat{\Sigma}) = \max_{\mu,\Sigma} \quad L_n(\mu,\Sigma), \tag{3}$$

unde $L_n(\mu,\Sigma)$ este probabilitatea dată de modelul asociat datelor

$$L_n(\mu,\Sigma) = \frac{1}{|\Sigma|^{1/2}} \exp(-\frac{1}{2}\sum_{i-1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)). \tag{4}$$

Estimatorii sunt:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \ \Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'$$

Invarianța estimatorilori la transformările afine este o proprietate naturală făcând analiza independentă de unitatea de măsură şi orice translație sau rotație a datelor. Spunem că un estimator T(X) al localizării şi un estimator C(X) al covarianței este invariant în sensul precizat dacă îndeplinesc condițiile:

$$\begin{aligned} T(XA+1_n v') &= T(X)A+v \\ C(XA+1_n v') &= A'C(X)A \end{aligned} \tag{5}$$

pentru orice vector v de dimensiune *dx1* şi orice matrice nesingulară A de dimensiune *dxd,* unde *d* este dimensiunea spațiului de date.

Estimatorii de verosimilitate maximă pentru modelul multivariat normal al datelor sunt afin-invarianți.

Totuşi aceşti estimatori pot fi afectați de prezența datelor anormale pe care vrem să le detectăm, diminuând şansele de a le identifica. De aceea este necesar să considerăm variante robuste ale localizării şi împrăştierii datelor.

1.3 Funcția de influență

O altă discuție legată de estimatorii robuşti priveşte funcția de influență (Hampel, 1974).

O metodă de a măsura robustețea unui estimator T(X), dacă se presupune că distribuția lui X este F, este de a evalua efectul asupra estimatorului când distribuția reală a datelor este perturbată intr-un punct x din R^d. Fie δ_x măsura de probabilitate exprimată de funcția Dirac care atribuie 1 punctului x. Hampel (1971) a definit funcția de influență pentru T(X) dat de distribuția F într-un punct x:

$$IF(x;T,F) = \lim_{0<\varepsilon\to 0} \frac{T[(1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon} \qquad (6)$$

dacă limita există.

Funcția de influență măsoară efectul asupra lui T a unei conatminări infinitezimale a distribuției F.

Estimatorul T care are o funcție de influentă mică sau cel puțin mărginită este considerat robust.

Dacă înlocuim ε cu 1/n în expresia funcției de influență putem măsura efectul unui outlier pe poziția x.

Funcția de influență are scăderea că măsoară efectul unei singure date anormale (Lopuhaa, 1989).

Pentru media şi covarianța clasică avem:

$IF(x;\mu,F) = x - \mu(F)$

$IF(x;\Sigma,F) = (x-\mu)(x-\mu)'-\Sigma(F)$ ’

ceea ce dovedeşte că sunt nemărginite şi deci că estimatorii sunt nerobuşti.

1.4 Estimatori robuşti pentru localizare şi împrăştiere

Bickel (1964) pare să fie primul care a considerat alternative robuste pentru media unui vector: mediana pe fiecare direcție şi estimatorul Hodge-Lehman. Aceşti estimatori fiind mai robuşti decât media empirică, totusi nu respectă proprietate de invarianță afină.

Hampel (1973) a fost primul care a propus o procedură iterativă pentru un estimator afin-invariant pentru matricea de împrăştiere (scatter), care s-a dovedit a fi un M-estimator.

Inspirat de Huber (1964), Maronna (1976) a definit şi studiat pentru prima oară M-estimator pentru localizare şi împrăştiere.

O metoda clasică pentru detectarea datelor anormale (outliers) este sa se calculeze distanța Mahalanobis față de centrul spațiului datelor (norul de puncte) pentru fiecare punct:

$MD_i = ((x_i - T)'C^{-1}(x_i - T))^{1/2}$,

unde T şi C sunt respective media şi matricea de covarianță a eşantionului.

MD ar trebui să măsoare depărtarea de centrul norului de date, ținându-se cont în acelaş timp de forma acestuia.

Numai că aceasta distanță este influențată datele anormale(outliers) care pot exista care denaturează măsurile *T* şi *C* făcând să apară efectul de mascare prin care datele anormale nu primesc neapărat o măsură *MD* mare.

Robustificarea acestei măsuri se poate face folosind alte marimi pentru *T* şi *C*, care să aprecieze correct centrul (locația) şi împrăştierea. Dacă se consideră estimatori robuşti pentru *T* şi *C* atunci distanța Mahalanobis va fi mare pentru datele anormale si va deveni:

$RD_i = ((x_i - T_R)' C_R^{-1} (x_i - T_R))^{1/2}$, unde indicele R semnalează o măsură robustă pentru estimator.

Estimatorii robuşti pentru localizare şi împraştiere pot fi calculați deodată ca în cazul M-estimatorilor, sau separat.

Invarianța estimatorilori la transformările afine este o proprietate naturală făcând analiza independentă de unitatea de măsură şi orice translație sau rotație a datelor:

$T(XA + 1_n v') = T(X)A + v$

$C(XA + 1_n v') = A'C(X)A$,

pentru orice vector v de dimensiune *dx1* şi orice matrice nesingulară A de dimensiune *dxd,* unde *d* este dimensiunea spațiului de date.

O măsură globală de robustețe a estimatorilor este punctul de rupere (breakdown point).

Donoho şi Huber (1983) au considerat definirea acesteia pentru eşantioane finite, iar Hampel (1971) sensul asimtotic.

Acest punct de rupere se poate defini informal ca proporția minimă de contaminare a datelor din mulțimea X care să facă estimatorul T să devieze oricât de mult de la adevărata valoare:

$$BP(T, X) = \min\{\frac{m}{n} : \sup_{X_m} |T(X_m) - T(X)| = \infty\} . \qquad (6)$$

Unde X_m este o mulțime de date contaminată rezultând din înlocuirea a m puncte din *X* cu puncte arbitrare din R^d .

Pentru matricea de covarianță se inlocuieşte în definiție T cu vectorul valorilor proprii logaritmate ale lui *C*. Estimatorul lui *C* ar deveni nefolosior dacă valorile proprii ar fi 0 sau infinit aşa cum pentru localizare ar fi infinit de mare.

1.4.1 M-estimatori

M-estimatorii nu sunt influențați de perturbații mici în date şi au o eficiență relativ bună pentru multe modele ale unei populații, dar nu sunt destul de robuşti ca măsură globală (în termini de punct de rupere = breakdown point) pentru spațile cu multe dimensiuni.

M-estimatorii pentru localizare şi împrăştierte sunt soluțiile T (în R^d) şi C (o matrice simetrică pozitiv-definită) ale ecuațiilor următoare:

$$\frac{1}{n}\sum_{i=1}^{n} u_1(((x_i - T)'C^{-1}(x_i - T))^{1/2})(x_i - T) = 0 \quad (7)$$

$$\frac{1}{n}\sum_{i=1}^{n} u_2(((x_i - T)'C^{-1}(x_i - T))^{1/2})(x_i - T)(x_i - T)' = C \quad (8)$$

unde u_i, i=1,2, sunt funcții care satisfac nişte condiții:

$u:\mathbf{R} \to [0,\infty)$ este simetrică, $u(0)=0$ şi $u(y) \to \infty$ când $y \to \infty$.

Sunt generalizări ai estimatorilor de verosimilitudine maximă şi pot fi priviți ca media şi matricea de covarianță ponderate.

Verosimilitatea maximă pentru distribuțiile normale este dată de alegerea:

$u_1(s) = -\frac{1}{s}\frac{d(\log f(s))}{ds}$ şi $u_2(s^2) = u_1(s)$ pentru s > 0.

Ei au funcție de influență mărginită pentru anumiți u_i, dar un punct de rupere relativ mic (<1/(d+1)) (Maronna, 1976) şi de aceea nu sunt global robuşti în spațiile de date cu multe dimensiuni.

1.4.2 Estimatori Stahel-Donoho

Stahel (1981) şi Donoho (1982) au introdus ca măsuri pentru localizare şi împrăştiere o medie şi matrice de covarianță ponderate bazate pe proiecții care au în acelaş timp proprietatea de afin-invarianță şi o robustețe globală înaltă dată de punctul de rupere (breakdown point).

Detectarea punctelor anormale se face după o măsură care se bazează pe proiecția unui punct x din mulțimea datelor $X \subset R^d$ (d ≥ 1 este dimensiunea datelor) în R:

$$O(x, X) = \sup_{\{u \in R^d, \|u\|=1\}} \frac{|u'x - \mu(uX)|}{\sigma(uX)} \quad (9)$$

unde $X = \{u'X_1, ..., u'X_n\}, X = \{X_1, ..., X_n\}$.

Atunci locația şi împrăştierea se calculează astfel:

$$T_{SD}(X) = \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i , \quad (10)$$

$$C_{SD}(X) = \sum_{i=1}^{n} w_i (x_i - T_{SD}(x))(x_i - T_{SD}(x))' / \sum_{i=1}^{n} w_i , \quad (11)$$

unde $w_i = w(O(x_i, x))$ şi w este o funcție pondere care penalizează punctele anormale.

Donoho a demonstrat robusteţea acestor estimatori pentru $(\mu,\sigma)=(Med,MAD)$, mediana şi median absolute deviation şi *w* funcţia pondere cu proprietăţi asemănătoare funcţiilor *u* de la M-estimatori.

1.4.3 Estimatori MVE şi MCD

Rousseeuw (1985) a introdus ca estimatori afin-invarianţi pentru localizare şi împrăştiere cei ce se bazează pe MVE (minimum volume ellipsoid) elipsoidul de volum minim şi cei ce se bazează pe MCD (minimum covariance determinant).

Estimatorii MVE pentru localizare şi împrăştiere sunt respectiv centrul şi elipsoidul de volum minim care cuprind cel puţin h puncte din X. S-a dovedit de către Davies (1987) că aceşti estimatori au un punct de rupere foarte înalt pentru $h=\lfloor(n+d+1)/2\rfloor$, dar nu sunt asimtotic normali nici $\sqrt{n}$ consistenţi (Davies, 1992)

Estimatorii MCD sunt media empirică şi un multiplu al matricea de covarianţă a h puncte din X pentru care determinantul matricei de covarianţă este minim.

TMCD şi CMCD au punctul de rupere aproximativ (n-h)/n, adică ½ pentru $h=\lfloor(n+d+1)/2\rfloor$(Lopuhaa and Rousseeuw, 1991) valoarea maximă pe care o poate atinge orice estimator invariant la transformări afine.

Aceşti estimatori MCD sunt $\sqrt{n}$ consistenţi (Davies and Jhun, 1993) şi de asemenea este demonstrată Rousseeuw normalitatea asimtotică pentru localizare dar nu şi pentru matricea de covarianţă şi nu au o eficienţă mare pentru modelele multivariat normale (Croux and Haesbroeck, 1999), dar sunt foarte des folosite datorită algoritmilor rapizi de calcul existenţi.

Un algoritm eficient de calcul pentru estimatorilor MCD pentru localizare şi covarianţă este FAST-MCD dezvoltat de Rousseeuw şi Van Driessen (1999) care calculează estimatori ponderaţi (one step reweighted):

$T_{RMCD}(X)=\sum_{i=1}^{n}w_i x_i/\sum_{i=1}^{n}w_i$,

$C_{RMCD}(X)=D_{h,n}\sum_{i=1}^{n}w_i(x_i-T_{RMCD}(x))(x_i-T_{RMCD}(x))'/\sum_{i=1}^{n}w_i$,

unde $w_i=\begin{cases}1 & daca \quad RD_{MCD}(X_i)\le\sqrt{\chi^2_{d,0.975}}\\ 0 & altfel\end{cases}$, iar $D_{h,n}$ este un factor de corecţie-

1.4.4 S-estimatori

O metodă de a îmbunătăţi eficienţa estimatorilor MVE şi MCD este să se considere o funcţie obiectiv netedă. Astfel o clasă importantă de esimatori sunt S-esimatorii

(Rousseeuw and Leroy , 1987, Davies, 1987) pentru localizare şi covarianță care minimizează funcția:

$$\min \det(C)$$

$$\frac{1}{n}\sum_{i=1}^{n}\rho(\sqrt{(x_i - T)'C^{-1}(x_i - T)} \le b, \tag{12}$$

pentru o funcție ρ care de multe ori se alege a fi funcția lui Tukey:

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x4}{2c^2} + \frac{x^6}{6c^4}, & daca \quad |x| \ge c \\ \frac{c^2}{6}, & daca \quad |x| \ge c \end{cases} \tag{13}$$

iar punctul de rupere este dat de valoarea lui c , BP=6b/c^2.

Dacă considerăm $\rho(x) = x^2$ şi b=1 obținem cazul metodei celor mi mici pătrate.

S-estimatorii au un punct de rupere înalt şi proprietatea de normalitate asimtotică (Rousseeuw şi Zohai 1984) dar metodele de calcul sunt intensive, iar când datele sunt foarte contaminate nu sunt unici (Woodruff şi Rocke, 1994).

1.4.5 Detectarea datelor anormale (outliers)

Considerăm un model standard localizare-covarianță, adică un set de date d-dimensionale x_1, x_2,...,x_n care sunt realizări ai unor variabile aleatoare X_1, X_2,..., X_n cu o distribuție de probabilitate eliptică $P_{\mu,\Sigma}$ cu densitatea de probabilitate:

$$f(x) = |\Sigma|^{-1/2} g(\|x - \mu\|^2_{\Sigma^{-1}}) \ (*)$$

cu $\mu \in R^d, \Sigma \in PDS(d)$, clasa matricilor de dimensiune dxd simetrice şi pozitiv definite şi $g : [0,\infty) \to [0,\infty)$.

O dată ce am obținut estimări robuste ale localizării şi împrăştierii putem aborda problema căutării datelor care au o deviație anormală față de media estimată şi conform cu matricea de covarianță.

Există trei tipuri de contaminare care cer diferite tehnici de detecție.

Prima categorie sunt acele date anormale pentru care distanțele robuste RD sunt mai mari decât un prag fix (Rousseeuw şi van Zomeren 1990, Rocke şi Woodruff 1996) sau decât un prag adaptativ care depinde nu numai de structura datelor ci şi de mărimea eşantionului (Gervini 2003).

Dacă f este multivariat normală atunci $RD(X_i) > \sqrt{\chi^2_{d,0.975}}$ determină aceste anormalități cu prag de semnificație 0.975.

Al doilea tip este cel al datelor anormale (shift outliers) care corespund unor clusteri care urmează aceiaşi structură ca a majorității datelor, dar sunt deplasați față de media

majorității populației. De exemplu pentru o populație distribuită $N(\mu,\Sigma)$ un nor de date anormale deplasate pot fi distribuite $N(\mu+s,\Sigma)$ şi astfel corup metrica bazată pe distanța Mahalanobis (Rocke şi Woodruff 1999) făcând ca diferența dintre datele normale ți acestea să dispară dacă se consideră această distanță.

Al treilea tip de anormalitate este cel al datelor anormale localizate (point mass outliers) dat de cel rezultat dintr-o concentrare a contaminării într-o regiune redusă (Pena şi Prieto 2001)

2. Un exemplu de estimare a datelor anormale folosind estimatorii MCD

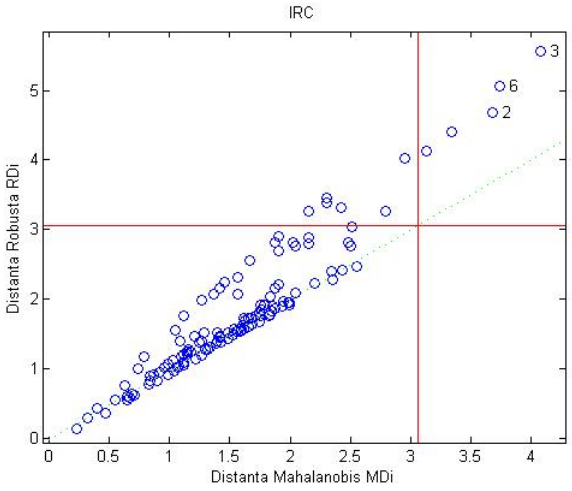
Considrăm în spațiul R^3, variabilele Inflație, Rata dobânzii active a clienților nebancari, Cursul Euro-Ron în perioada februarie 2000 - iunie2010 (n=125 de puncte) care urmează un model multivariat normal şi calculăm distanțele Mahalanobis MD_i şi distanțele robuste RD_i cu pragul de semnificație 97,5% şi h=0.9n ceea ce asigură un punct de rupere BP=10%.

Din graficul urmator (figura 1) se observă efectul de mascare al distanțelor nerobuste MD_i, care în comparație cu RD_i sunt mai mici.

Punctele desemnate ca outliers sunt cu indexii în intervalul 1-12 şi 26. Iată că deviația față de modelul normal se întimplă in primul interval (an), apoi se poate considera că datele urmează acest model (graficul 3 în care se raportează RD_i la rădăcin pătrată a cuantilelor distribuției chi-pătrat). De fiecare dată, pentru orice h în intervalul (n/2, n), datele anormale se găsesc în prima perioadă. Pentru h=n obținem puncte anormale 1-4, 6, 12, 26, care apar ca anormale în toate estimările aşa că putem cochide că aceste puncte sunt outliers.

In figura 1, se poate observa efectul de mascare asupra distanțelor Mahalanobis, care sunt subevaluate datorită măsurilor nerobuste ale mediei şi covarianței.

Figura 1



In figura 2 se pot observa datele anormale comparându-se distanțele robuste cu valoare chi-pătrat, iar în figura 3 se raportează cuantilele distribuției empirice a distanțelor robuste cu cuantilele distribuției chi-pătrat cu d grage de libertate şi prag de semnificație 0.975.

Figura 2

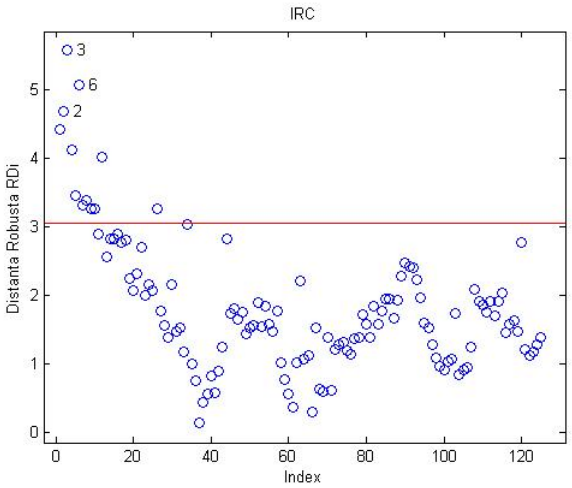
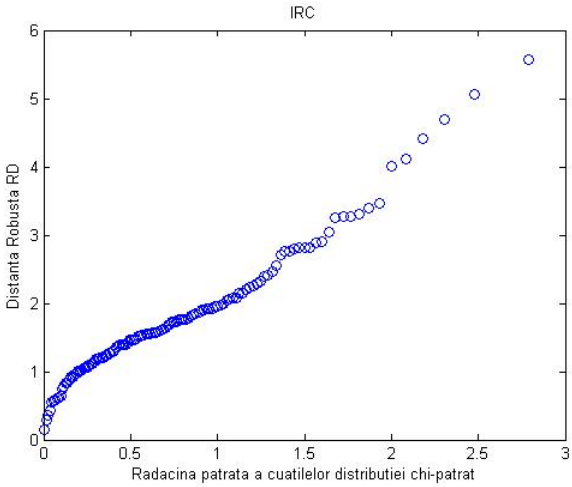


Figura 3



3. Regresie ortogonală robustă

Fekri şi Ruiz-Gazen (2004) au definit regresie ortogonală ponderată robustă.

Fie $X \in R^{dxn}$ matricea conținând pe coloane cele n date observate $X_i \in R^d$, i=1,..,n şi fie x_i aparținând direcției afine $D_{A,b} = \{x \in R^d, A'x = b; A \in R^d, b \in R\}$, $X_i = x_i + e_i$ şi e_i distribuit normal cu medie 0 şi matrice de covarianță diagonală Σ .

Modelul regresiei ortogonale estimează A şi b minimizând criteriul:

$$\sum_{i=1}^{n}\left\|X_i - x_i\right\|^2 = \sum_{i=1}^{n}\left\|X_i - \Pi_{D_{A,b}}(X_i)\right\|^2 + \sum_{i=1}^{n}\left\|x_i - \Pi_{D_{A,b}}(X_i)\right\|^2$$

Minimizarea acestui criteriu duce la $x_i = \Pi_{D_{A,b}}(X_i)$ ţi minimizarea :

$$\sum_{i=1}^{n}\left\|X_i - \Pi_{D_{A,b}}(X_i)\right\|^2 = \sum_{i=1}^{n}\left\|A'X_i - b\right\|^2 = \left\|(A'X - bu'\right\|_F^2$$

cu $u \in R^n = [1,1,..,1]$ şi $\|M\|_F^2 = Tr(M'M) = Tr(MM')$ reprezentând norma Frobenius.

Regresia ortogonală robustă schimbă criteriul de minimizare intr-unul ponderat:

$$\sum_{i=1}^{n} w(\left\|X_i - \mu\right\|_{\Sigma^{-1}}^2)\left\|X_i - x_i\right\|^2$$

care conduce la $\sum_{i=1}^{n} w(\left\|X_i - \mu\right\|_{\Sigma^{-1}}^2)\left\|A'X_i - b\right\|^2 = \left\|(A'X - bu')W\right\|_F^2$,

unde $W = [w_1,...,w_n], w_i = w(\left\|X_i - \mu\right\|_{\Sigma^{-1}}^2)$.

Soluţia optimă (A*,b*) este dată de:

A*=vectorul propriu de normă unitate asociat matricii pozitiv semidefinite $H = X(W^2 - \frac{1}{\nu}W^2uu'W^2)X'$ şi $b^* = \frac{1}{\nu}A^{*\prime}XW^2u$.

Fie :

$$\mu_{R,n} = \frac{\sum_{i=1}^{n} w(\left\|X_i - \mu_n\right\|_{\Sigma_n^{-1}}{}^2)X_i}{\sum_{i=1}^{n} w(\left\|X_i - \mu_n\right\|_{\Sigma_n^{-1}}{}^2)} \text{ şi } \Sigma_{R,n} = \frac{\sum_{i=1}^{n} w(\left\|X_i - \mu_n\right\|_{\Sigma_n^{-1}}{}^2)(X_i - \mu_{R,n})(X_i - \mu_{R,n})'}{\sum_{i=1}^{n} w(\left\|X_i - \mu_n\right\|_{\Sigma_n^{-1}}{}^2)},$$

unde μ_n şi Σ_n sunt estimatori robuşti ai localizării şi covarianţei.

Minimul este atins de direcţia afină care conţine $\mu_{R,n}$ şi este paralelă direcţiei dată de vectorul propriu asociat celei mai mici valori proprii a matricei $\Sigma_{R,n}$.

Funcţia w este o funcţie descrescătoare care penalizează observaţiile cu distanţa $\left\|X_i - \mu_n\right\|_{\Sigma_n^{-1}}{}^2$ mare. Vom folosi în exemple funcţia pondere $1I_{[0,c]}$ unde c este cuantila 97,5% a distribuţiei chi-pătrat.

Tabelul 1

	Coeficienţii hiperplanului		
	TLS	RS	RMCD
Inflaţie	-0.4239	-0.4875	-0.4870
Rata dobânzii	0.9057	0.8732	0.8734

Curs	0.0041	-0.0028	0.0027
Termen liber	-0.0249	-0.0195	-0.0191

4. Rezultate de simulare

In această secțiune am făcut simulări pentru a compara estimatorii coeficienților hiperplanului obținuți prin metoda robustă şi cea standard. Parametrii populației sunt cei obținuți cu metoda RMCD pentru exemplul prezentat.

Am considerat datele necontaminate care sunt conforme cu hiperplanul generat de ecuația ortogonală robustă obținută folosind covarianța RMCD şi urmează modelul regresiei otogonale cu erori de tip multivariat normal. Contaminarea s-a realizat considerând că erorile urmează alta distribuție de tip eliptic, sau combinații de distribuții normale.

Astfel, rezultatele obținute arată ce s-ar întâmpla dacă ar exista date anormale modelului considerat şi am aplica una dintre metodele robuste sau nerobuste. Aceste simulîri constituie o măsurare a robuteții globale a metodelor.

Simulările constau în 1000 de eşantioane cu dimensiunea 125 pentru mai multe distribuții:

1. distribuția normală $N(\mu,\Sigma)$ (NOR)

2. distribuția normală simetric contaminată (SCN), care este un mixaj de 90% $N(\mu,\Sigma)$ şi 10% $N(\mu,10\Sigma)$

3. distribuția multivariată Cauchy (CAU)

4. distribuția normală asimetric contaminată (ACN), care este o mixtură de 90% $N(\mu,\Sigma)$ şi 10% $N(\mu_1,10\Sigma)$ cu $\mu_1 = 2*\mu$

Tabelul 2

	Media empirică			MSE(media pătratelor erorilor)		
NOR	TLS	RS	RMCD	TLS	RS	RMCD
	-0.4870	-0.4870	-0.4870	4.36e-13	0.25e-13	4.36e-13
	0.8734	0.8734	0.8734	1.41e-11	0.54e-11	1.45e-11
	0.0027	0.0027	0.0027	1.37e-16	0.212-14	2.75e-14
	-0.0190	-0.0191	-0.0191	0.14e-15	0.3e-15	0.1e-15
SCN	-0.4874	-0.4870	-0.4870	2.58e-10	0.31e-11	2.6e-13
	0.8744	0.8734	0.8734	2.35e-12	2.2e-12	2.3e-12
	0.0026	0.0027	0.0027	2.8e-14	3.2e-17	2.7e-17
	-0.0191	-0.0191	-0.0191	1.4e-3	2.3e-4	3.7e-4
CAU	0.7135	-0.4934	-0.4865	0.0514	8.9e-4	2.26e-4

	0.7004	0.8761	0.8735	0.0299	8.4e-4	7.02e-5
	0.0133	0.0023	0.0027	1.132e-4	4.3e-6	8.31e-8
	0.0647	-0.0798	-0.0989	0.0021	0.8260	1.8918
ACN	0.4224	-0.4794	-0.4824	0.0265	5.6e-6	0.4e-6
	-0.49118	0.8700	0.8703	0.0129	3.4e-4	1.69e-4
	6.7188e-04	0.0025	0.0026	2.9e-6	4.6e-7	2.07e-7
	0.0019	-0.0189	-0.0204	0.0004	0.1e-5	0.0003

In cazul contaminării simetrice rezultatele pentru regresia TLS sunt foarte bune deoarece datele (inflația, rata dobânzii, cursul de schimb) sunt puternic corelate aşa modificarea considerată matricei de covarianță însemnă o corelare mai puternică a datelor.

În ce priveşte modelele de regresie ortogonală robuste, ele par echivalente şi deci poate fi folosită oricare dintre ele şi cum algoritmul pentru metoda RMCD este mai eficient este de preferat utilizarea acestei metode.

Ca măsură globală performanței estimatorilor ecuației regresiei ortogonale am considerat valoarea absolută a cosinusului unghiului dintre parametrii estimați şi cei ai populației. Această măsură poate fi interpretată ca mărime a influeței datelor anormale (outliers) asupra estimatorilor.

Figura 4 şi 5 prezintă funcția de distribuție cumulată empirică pentru cele 1000 de eşantioane ale simulărilor, pentru distribuțiile Cauchy şi distribuția normală asimetric contaminată. Distribuția cumulativă empirică ar trebui să fie masată în apropierea lui 1 dacă parametrii estimați au valori aprope de valorile reale deoarece atunci cosinusul unghiului dintre parametrii ar fi aproape de 1.

In cazul distribuției normale toate cele trei curbe au vârful de maxim către 1. Insă nu aceeiaşi este situația pentru estimarea TLS (regresia ortogonală nerobustă) în timp ce metodele robuste dau rezultate bune.

Figura 4

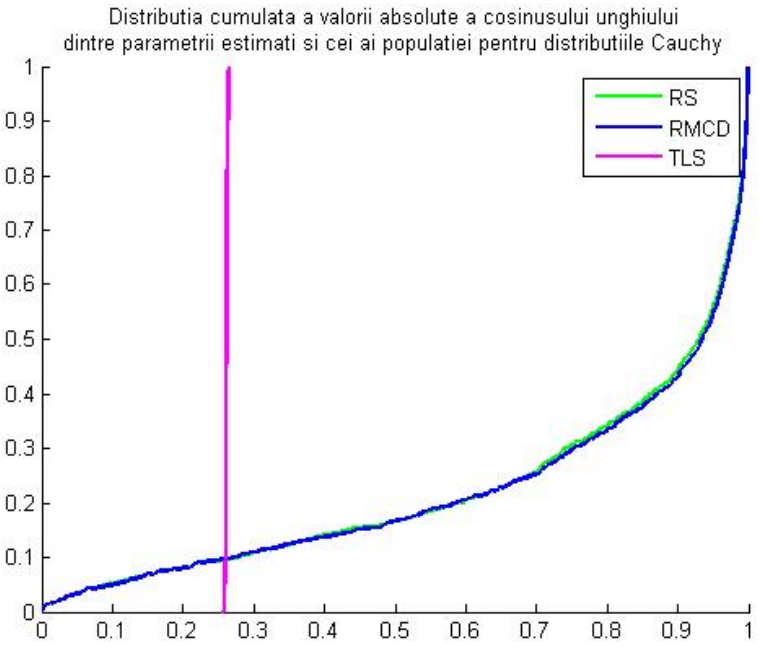
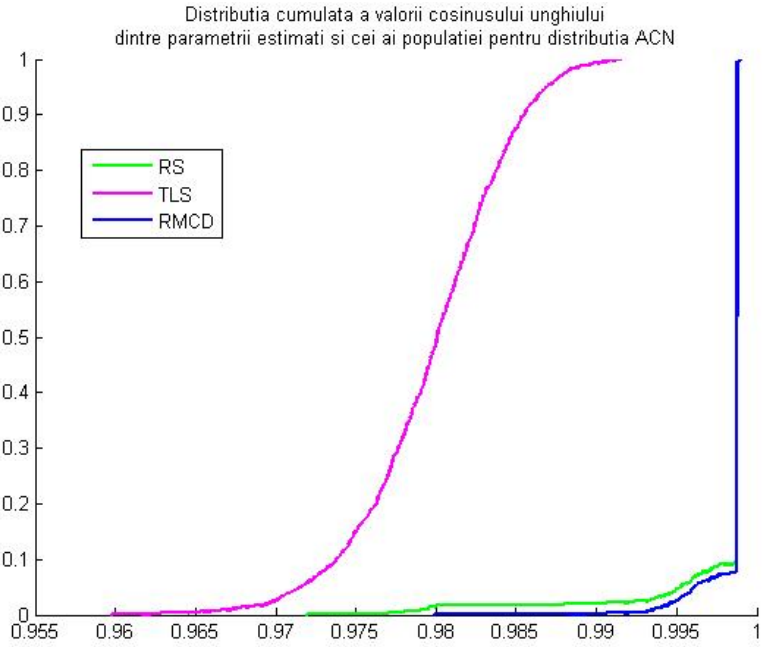


Figura 5



Concluzii

In acest studiu am analizat influența datelor anormale asupra estimatorilor în cazul modelului Eroare-in-Variabile cu distribuția erorilor multivariat eliptice pentru regresia ortogonală.

Estimatorii robuşti sunt definiți ca estimatori ponderați de tip cele-mai-mici-pătrate fiind derivați din estimatori robuşti ai localizării (mediei) şi împrăştierii (covarianței) datelor. In acest studiu am folosit estimatori S şi RMCD, dar pot fi utilizați orice alți estimatori robuşti ai primelor două momente ale distribuțrei empirice a datelor cum sunt estimatorii de tip proiecție Stahel-Donoho sau M-estimatori.

Am văzut că estimatorii clasici ai hiperplanului care se află la distanță minimă de date (distanța Euclidiană deoarece este analizat cazul regresiei ortogonale) se îndepărtează de valorile reale dacă datele sunt contaminate de erori care urmează alte distribuții decât cea gaussiană. Acest caz este des întâlnit în situațiile reale când datele deviază față de distribuția normală prezentând de multe ori distribuții asimetrice sau cu valori extreme.

Folosirea unor metode nerobuste în aceste cazuri duce la estimări incorecte şi implicit la inferențe greşite.

Rezultatele obținute din simulări şi prezentate în tabel sunt similare celor din Fekri şi Ruiz-Gazen (2004) dovedind că pentru modelul necontaminat cei mai eficinți estimatori ai hiperplanului ce reprezintă regresia liniară ortogonală fără ca celelalte metode robuste să fie ineficinte, iar pentru modelele contaminate regresiile robuste sunt eficiente dar regresia ortogonală clasică este complet inadecvată in cazurile distribuției Cauchy şi celei normal asimetric contaminate.

Programele utilizate pentru calculul estimatorilor şi pentru simulări au fost scrise în limbaj MathLab. Estimatorii RMCD au fost calculați folosind rutinele LIBRA.

Referințe

Bickel, P.J. (1964), On some alternative estimates for shift in the p-variate one sample problem, *Ann. Math. Statist.*, 35: 1079-1090.

Brown, M. (1982), Robust line estimation with errors in both variables, *Journal of the American Statistical Association*, 77: 71-79.

Croux, C. and Haesbroeck, G. (1999), Influence and Efficiencz of the Minimum Covariance Determinant Scatter Matrix Estimator, *Journal of Multivariate Analysis*, 71: 161-190.

Davies. P.L. (1987), Asymptotic behaviour of S-estimators of multivariate location parameters and dispersion matrices, Annals of Statistics, 15: 1269-1292.

Fekri, M. and Ruiz-Gazen, A. (2004), Robust weighted orthogonal regression in the errors-in-variables model, *Journal of Multivariate Analaysis,* 88: 89–108.

Gervini, D. (2002), A robust and efficient adaptative reweighted estimator of multivariate location and scatter, *Journal of Multivariate Analaysis*, 84:116-144.

Hampel, F.R. (1974), The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, 69: 383-393.

Hampel, F.R. (1971), A general qualitative definition of robustness, *Ann. Math. Statis.*, 42:1887-1896.

Huber, P.J. (1972), Robust statistics: A review, *Ann. Math. Statis.*, 35:73-101.

Hubert, M. Rousseeuw, P. J. And Van Aelst, S. (2008), High-Breakdown Robust Multivariate Methods, *Statistical Science*, 23(1): 92-119.

Lopuhaä, H.P. (1989), On the relation between S-estimators and M-estimators of multivariate location and covariance, *Ann. Statist*., 17(4):1662-1683.

Lopuhaä, H.P. (1999), Asymptotics of reweighted estimators of multivariate location and scatter, *Ann. Statist*., 27: 1638–1665.

Lopuhaä, H.P. and Rousseeuw, P.J. (1991), Breakdown points of affine equivariant estimators of multivariate location and covariance matrices, *Ann. Statist.*, 19: 229–248.

Maronna, R.A. (1976), RobustM-estimates of multivariate location and scatter, *Ann. Statist.*, 4: 51–67.

R.A. Maronna and S. Morgenthaler, (1986), Robust regression through robust covariances, *Comm. Statist.—Theory Methods,* 15: 1347–1365.

Maronna, R.A. Stahel, W.A. and Zohai, V.J. (1992), Bias-robust estimates of multivariate scatter based projections, *Journal of Multivariate Analysis*, 42:141-161.

Rocke, D.M. and Woodruff, D.L. (1999),

Rousseeuw, P. J. and Van Zomeren, B.C. (1990), Unmasking multivariate outliers and leverage points, *J. Amer. Statist. Assoc*., 85: 633–651.

Rousseeuw, P. J. And Van Driessen, K. (1999), A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics*, 41: 212-223.

Rubinstein, R.Z, and Kroese, D.P. (2008), Simulation and the Monte Carlo Method, John Wilez / Sons, Hoboken, New Jersey.

Woodruff, D.L. and Rocke, D.M. (1994), Computable robust estimation of multivariate location and shape in high dimension using compound estimators, *J. Amer. Statist. Assoc.,* 89: 888-896.

Anexa.

In această anexă se prezintă programul de simulare pentru datele contaminate cu distribuția Cauchy.

```
function res=Multivar_tTLSsim(IRCplan,Mer,Ser,trial,p,multsigma,multm)
% Simulare Monte Carlo pentru setul de date IRCdata contaminat cu distributie Cauchy
n = size(IRCplan(:,1)); % The number of function evaluations
 % --- Generate vectors of random inputs
% r ~ distributie Normala N(medie=m,sd=SIGMA)
% rc ~ distributie Normala N(medie=multm*m,sd=10*SIGMA)
% re ~ distributie Cauchy CAU(sd=Ser)
% rce ~ distributie Normala N(medie=Mer,sd=Ser)

res.TLS=[];
res.TLSm=[];
res.TLSmse=[];
res.RMCDmse=[];
res.sz=[];
res.multsigma=multsigma;
res.multm=multm;
for i=1:4
res.TLSmse(i)=0;
res.RMCDmse(i)=0;
end
j=0;
SIGMA=cov(IRCplan);
m=mean(IRCplan);
while j<trial
r=mvnrnd(m,SIGMA,n(1)-floor(n(1)*multsigma));
rc=mvnrnd(multm*m,10*SIGMA,floor(n(1)*multsigma));
IRCrand=[r; rc];
re=mvtrnd(Ser,1,25); rce=mvnrnd(Mer,Ser,100);
IRCrand=IRCrand+[re;rce];
m=mean(IRCplan);
COVrand=cov(IRCrand); mrand=mean(IRCrand);
   j=j+1;
   [v,d]=eig(COVrand);
   k=1;
   for i=2:3
         if d(i,i) < d(k,k)
         k=i;
         end
   end
   res.TLS(4,j)=0;
   for i=1:3
   res.TLS(i,j)=v(i,k);
   res.TLSmse(i)=res.TLSmse(i)+(abs(v(i,k))-abs(p(i)))*(abs(v(i,k))-abs(p(i)));
   res.TLS(4,j)=res.TLS(4,j)+v(i,k)*mrand(i);
   end
```

```
    res.TLSmse(4)=res.TLSmse(4)+(res.TLS(4,j)-abs(p(4)))*(res.TLS(4,j)-abs(p(4)));
    [rew,raw]=mcdcov(IRCrand,'plots',0, 'alpha',0.75) ;    plan=size(rew.plane);
    res.sz(j)=plan(1);
    [v,d]=eig(rew.cov);
    k=1;
    for i=2:3
          if d(i,i) < d(k,k)
          k=i;
          end
    end
    res.RMCD(4,j)=0;
    for i=1:3
    res.RMCD(i,j)=v(i,k);
    res.RMCDmse(i)=res.RMCDmse(i)+(abs(v(i,k))-abs(p(i)))*(abs(v(i,k))-abs(p(i)));
    res.RMCD(4,j)=res.RMCD(4,j)+v(i,k)*mrand(i);
    end
    res.RMCDmse(4)=res.RMCDmse(4)+(res.RMCD(4,j)-abs(p(4)))*(res.RMCD(4,j)-abs(p(4)));
end
res.TLSm(1)=mean(res.TLS(1,:)); res.TLSm(2)=mean(res.TLS(2,:)); res.TLSm(3)=mean(res.TLS(3,:));
res.TLSm(4)=mean(res.TLS(4,:)); res.RMCDm(1)=mean(abs(res.RMCD(1,:)));
res.RMCDm(2)=mean(abs(res.RMCD(2,:))); res.RMCDm(3)=mean(abs(res.RMCD(3,:)));
res.RMCDm(4)=mean(abs(res.RMCD(4,:)));
end
```