Three applications of the Method of the Least Trimmed Squares

Tsvetan B. Georgiev
New Bulgarian University, Department of Natural Science
21 Montevideo Str., BG-1618, Sofia

ABSTRACT

The ordinary Method of the least squares (MLS) minimizes the scatter of all squares of deviations from the demanded estimation. However, only one large outlier (large error in the data) causes strong change of the result. The more sofisyical Method of the least trimmed squares (MLTS), introduced by Rousseeuw in 1984, minimizes the left part of the ranged squares of the deviations, including at least the half of the data. Therefore, numerous large squares of deviations may be present in the right part of the ranged squares of deviations, up to about 40 % of the data, but MLTS ignores them. However, while the MLS gives formulas for calculations of the statistical parameters, the MLTS tests and qualifies each available pattern of possible solution: every data point in 1D case, each line through pair of points in 2D case, each plane through triad of points in 3D case, etc. The solution is the pattern with the shortest MTLS deviation. Three chosen applications of the MLTS are presented here: mode estimation in an assymetric 1D random distribution, model of the "main sequence" in a pistol-like diagram and reveal of the short-time outbursts in a light curve of an active star. Some recommendations about the application of the MLTS are given too.

Keywords: *data analysis – models; statistical - methods*

Три приложения на Метода на отбраните най-малки квадрати

Цветан Б. Георгиев
Нов български университет Ддепартамент „Природни науки"

РЕЗЮМЕ

Ординарният Метод на най-малките квадрати (МНК) минимизира разсейването на всички квадрати на отклонения от търсената оценка. Обаче, дори едно силно отклонение (голяма грешка в данните) предизвиква силно изменение на резултата. По-софистичният Метод на отбраните най-малки квадрати (МОНК) минимизира разсейването на лявата част на подредените квадрати на отклонения, включваща поне на половината данни. Ето защо множество големи квадрати на отклонения могат да присъстват в дясната част на реда на квадратите на отклоненията, до около 40 % от данните, но МОНК ги игнорира. Обаче, докато МНК дава формули за изчисляване на статистически параметри, МОНК тества и окачествява всеки от достъпните образци на възможни решения – всяка отделна точка в

едномерния случай, права през всяка двойка точки в двумерния случай, равнина през всяка тройка точки в тримерния случай и т.н. Решението е образецът, който дава най-малко МОНК разсейване. В тази работа са представени три избрани приложения на МОНК: оценяване на мода на асиметрично едномерно случайно разпределение, моделиране на „главна последователност" в пистолетовидна диаграма и изявяване на кратковременни избухвания в крива на блясъка на активна звезда. Дадени са и някои препоръки за прилагането на МОНК.

Ключови думи: *анализ на данни – модели; статистически методи*

1.Методът на най-малките квадрати (МНК)

Фитирането на емпирични зависимости чрез регресионни линии или повърхнини, построени по Метода на най-малките квадрати (МНК; Method of the Least Squares, MLS), е широко разпространено в практиката на научните изследвания. При изглаждане на времеви редове се използва методът на движещия се прозорец данни, като оценката на стойността на централната точка в прозореца данни обикновено се дава чрез полином от ниска степен, построен пак по МНК. Всъщност МНК се базира на Принципа на най-малките квадрати, въведен от Льожандр и Гаус в началото на XIX век: Най-добрата оценка на статистическия параметър (напр. средна стойност на данни, наклон на регресионна права) е тази, която минимизира квадратите на отклоненията на данните от оценката на параметъра. Основното достойнство на МНК е, че той дава формули за оценяването на параметрите. А понеже формулата за средно-аритметичната стойност като оценка на средната стойност (математическото очакване) при едномерна случайна величина може да се изведе чрез МНК, то всички оценки на параметри, получени по МНК имат смисъл на средно-аритметични оценки.

Според теорията МНК е коректно приложим при изпълнени множество статистически допускания: Стойностите на аргумента (независимата променлива) x_j са точно известни, стойностите на функцията (зависимата променлива) y_j съдържат адитимна нормална грешка (случайна величина с нулево математическо очакване и крайна ненулева дисперсия) и др. Обикновено се смята, че статистическите допускания са изпълнени и удобният за практиката МНК се прилага повсеместно. Обаче, големият недостатък на МНК е, че дори една груба грешка (*y*-данна, която се отклонява силно от очакваната стойност) води до силно изменение в оценката на параметъра. А при изглаждането на редове от данни всеки импулс (груба грешка в данните) се „размива" върху околните данни и ги изкривява. Затова се говори, че МНК има асимптотично нулева робастост (якост, устойчивост) спрямо силни отклонения в данните.

Обикновено тази слабост на МНК не е сериозен проблем, понеже изследователят познава точките си и изхвърля предварително грубите грешки. Обаче, проблемът с сериозен, (i) когато фракцията на данните с груби грешки е голяма, напр. 30-40 % , (ii)

когато аргументите са повече от един, т.е. визуалният контрол е труден и (iii) когато МНК се прилага многократно и „сляпо", напр. при обработка на редове от данни или на числени изображения. В такива случаи следва да се прилага метод, който да е робаст, т.е. да е устойчив спрямо силно отклоняващи се данни. Следва да се отбележи, че думата робаст (англ. – robust) е възприета напр. в руската литература по математическа статистика и няма причина да бъде заменяна с точния неин български превод „як".

2. Методът на отбраните най-малки квадрати (МОНК)

Забележителен и прост по същността си робаст метод е Методът на отбраните най-малки квадрати (МОНК; Method of the Least Trimmed Squares, MLTS) е въведен от Peter Rousseeuw (1984). Анализ и примерни приложения са дадени от Rousseeuw & Leroy (1987), и Georgiev (2008). Докато МНК минимизира *квадратите на всички n отклонения*, МОНК минимизира *само лявата част на квадратите на отклоненията, наредени по нарастване, включвайки поне* $n_h = n/2+1$ *данни*. В дясната част на наредените квадрати на отклонения може да присъстват произволно големи отклонения, но МОНК ги игнорира. Теоретично МОНК се характеризира с максимално възможната асимптотична робастост 50% и затова се определя като „екстремално робаст". От тази гледна точка робастостта на ординарния МНК е асимптотично 0 %. На практика МОНК игнорира успешно до около 40% от данните, чиито y-стойности са силно отклонени, независимо колко силно. Така за разлика от МНК, който е средно-аритметичен оценител, МОНК е модален оценител. Всъщност при асиметрично разпределение оценката на модата на разпределението, получена чрез МОНК, остава малко отместена към по-тежката опашка на разпределението, т.е. МОНК е квази-модален оценител.

Обаче, докато коефициентите на полиномите, построени по МНК, се изчисляват чрез аналитично изведени формули, при МОНК това става чрез последователно тестване на достъпни образци на възможни решения. За решение се избира образецът, който има най-малко МОНК-разсейване. Изискването за употреба на най-малко 50% +1 от квадратите на отклоненията, т.е. за „просто болшинство", следва от необходимостта да се изявява главната (най-населената) мода при мулти-модално разпределение на отклоненията. При n данни броят на данните за просто болшинство е $n_h = n/2+1$, но изследователят може да избере и „болшинство данни", по-голямо от n_h. При това, когато n_h клони към n, тогава резултатът от МОНК клони към резултата от МНК.

Ето алгоритъмът на МОНК в най-простия случай, като модален оценител на средната стойност μ чрез емпиричната модална стойност m_O за n данни y_j (фиг.1).

1. За всяка точка (данна) y_j се изчисляват всичките n квадрати на отклонения от нея $\Delta y_{ij} = (y_i - y_j)^2$, $i=1,\ldots, n$;

2. За всяка точка се отбират (измъкват и подреждат възходящо) n_h квадрати на отклонения Δy_{ij}, като няма нужда да се подреждат и останалите отклонения.

3. Пресмята се сумата S_j на отбраните квадрати на отклонения, която е основната характеристика на точката y_j като потенциална оценка на модата;

4. Извежда се като решение (в случая – мода на 1D данни) $m_O = y_k$, където данната y_k е тази, за която сумата S_k на отбраните n_h квадрати на отклонения спрямо нея е най-малка.

5. Извежда се като оценката s_O на стандартното отклонение ϭ на математическото очакване μ във вида $s_O = 2.[S_k/(n_h-1)]^{1/2}$. Тук умножаването по 2 компенсира определянето на стандартно отклонение s по само ½ от данните. Така резултатът от МОНК става сравним с резултата, който би се получил по всички данни чрез МНК, ако техните отклонения имаха нормално разпределение.

Когато се търси център на 2D, 3D и т.н. дискретно разпределение МОНК проверява всяка точка (вектор) r_j като потенциално решение. За целта се използват пространствените квадрати на отклонения $\Delta r_{ij} = |r_i - r|^2$, $i=1,\dots, n$. Решението е векторът r_k ,спрямо който сумата на отбраните квадрати е най-малка. В разглежданите случаи броят на тестваните образци на решения, каквито тук са отделните точки, е $N = n$.

Ето как работи МОНК и в най-разпространения 2D случай, като оценител на модална регресионна права от вида $<y> = a + b.x$, по n данни (x_j, y_j).

1. За всяка двойка точки $[(x_i, y_i)(x_j, y_j)]$, разглеждана като потенциален носител на търсената права и номерирана като k,, се определят параметрите a_k and b_k на нейната права $<y> = a_k + b_k .x$. След това се изчисляват всички квадрати на отклоненията на точките от тази права $\Delta y_{ik} = (y_i - a_k - b_k.x_i)^2$, i=1,… , n

По-нататък се изпълняват стъпки 2 - 5 от предния случай. Броят на тестваните образци на решения – правите през всички двойки точки – е равен на броя комбинациите от 2 елемента сред n елемента, т.е. $N = n.(n-1)/2$.

При търсенето на модална квадратична регресия зависимост от вида $<y> = a + b_1.x + b_2.x^2$ (фиг.3) или на модална регресионна равнина от вида $<z> = a.x + b.y + c$ се оценяват 3 параметъра. Аалгоритъмът е подобен на горните, като образците на решения се задават като квадрагични функции през всички тройки точки, а броят на тестваните образци е $N = n.(n-1).(n-2)/6$. При търсенето на модална кубична регресия от вида $<y> = a + b_1.x + b_2.x^2 + b_3.x^3$ или на модална регресионна хипер-равнина от вида $<t> = a.x + b.y + c.z + d$ се използват по аталогичен начин четворки точки, оценяват се 4 параметъра и броят на тестваните образци става $N = n.(n-1).(n-2).(n-3)//24$.

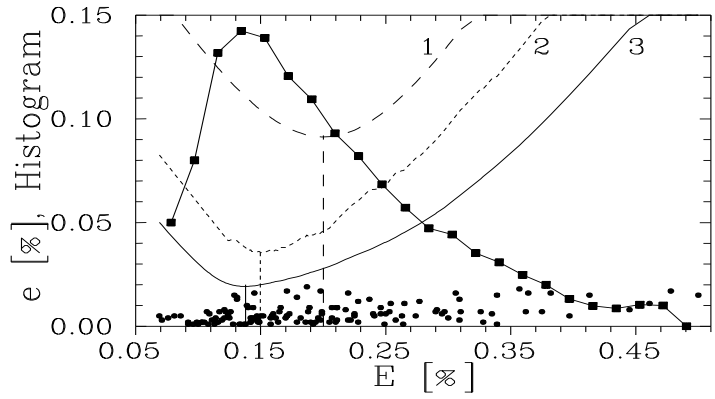
Трябва да се отбележи, че при оценяване на средната стойност в едномерния случай, както и при изглаждането на редове от данни, вместо средно-аритметична стойност може да се използва медианна стойност, която също има асимптотична робастост 50%. Обаче, не е ясно дали и как може да се построи медианна права, медианна равнина и изобщо – медианна регресия.

За прилагането и пропагандирането на МОНК авторът е разработил система от компютърни С-програми. В тази статия са представени три приложения на МОНК в случаи когато МНК е безполезен – за оценяване на средна стойност (математическо

очакване) при силно асимеметрично разпределание чрез модата на данните (Раздел 3), за построяване на робаста регресионна крива при множество (≈40 %) силно отклоняващи се данни (Раздел 4) и за изявяване на съществени детайли в сложен ред от данни (Раздел 5).

3.Атмосферното поглъщане над Националната астрономическа обсерватория Рожен

Фигура 1 представя приложение на МОНК като модален оценител на средната стойност μ чрез емпиричната модална стойност m_O. Там са показани 146 електро-фтометрични измервания на атмосферно поглъщане E [%] във визуалната област от спектъра в 2003-2007 г, с относителна точност $e \approx 1\%$ (по данни, любезно предоставени от Dimitrov (2007). Разпределението е силно асиметрично, с тежка дясна опашка. Пресмятането показва, че средно-аритметичната стойност на поглъщането над Обсерваторията (на 1750 м над морското ниво) е $E = (19 \pm 9)$ %. Обаче, при привидно чисто небе е възможно поглъщане и $E \approx 40$ %, което пък е характерно за чисто небе над морското ниво. В този случай модалната стойност, изчислена по МОНК и характеризираща поглъщането при незадимена атмосфера е $E=(14 \pm 6)$ %. Това е и стойността, която следва да характеризира астроклимата на Обсерваторията при сравнение с други обсерватории.



Фиг.1. Разпределение на данните за атмосферното поглъщане E и техните стандартни отклонения e (точки) заедно с хистограмата на разпределението (квадратчета). Кривите 1, 2 и 3 са криви на грешките за средно-аритметичната стойност, съответна на МНК, медианната стойност и модалната стойност (получена чрез МОНК). Стойностите, съответни на минимумите на кривите, са маркирани с вертикални отсечки (по Georgiev, 2008).

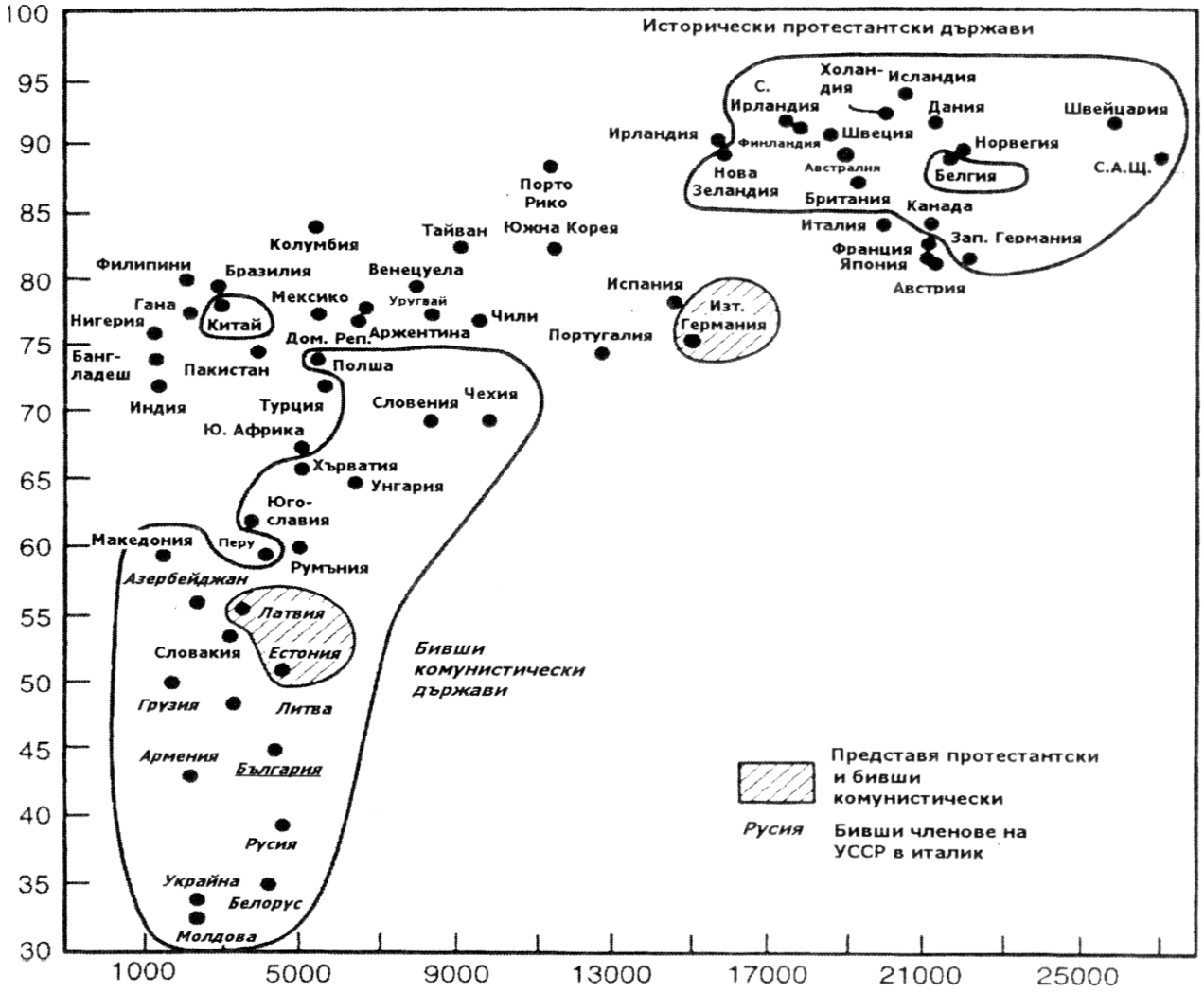
Модата на разпределението на атмосферното поглъщане, дадено на фиг.1, може да се определи и чрез хистограма. Но това значи изследователят предварително да избере (субективно) нул-пункт и стъпка на хистограмата. Чрез прилагането на МОНК като модален оценител субективността и „ръкопашността" се избягват (Georgiev, 2008).

4.Зависимост „доход – самочувствие“ за гражданите от различни държави

Фигура 2 представя резултати от едно социологично проучване на тема доход – самочувствие на населението в 65 държави през 1999 г (Inglehart & Klingemann, 2000). Фракцията от населението с положително самочувствие („субективно психично благополучие“) е определена като средно-аритметичен процент на фракциите на „щастливите в живота“ и „удовлетворените от живота“. За доход на населението е използван брутният вътрешен продукт на глава от населението за 1995 г, в щатски долари.

По данните от това проучване Baychinska & Bakracheva (2907) изявяват две специални групи държави, оградени на диаграмата с криви. В горния десен ъгъл на диаграмата стоят най-високо развитите държави, характеризирани като „исторически протестантски държави“. Тяхното население има най-висок доход и най-високо самочувствие. В долния ляв ъгъл на диаграмата стоят бедни и не съвсем бедни държави, характеризирани като „бивши комунистически държави“. Тяхното население има нисък доход и ниско самочувствие. Все пак, в Китай, Полша, Словения,Чехия и Източна Германия самочувствието на населението е по-скоро високо.

На фиг.2 явно се вижда и трета, най-многобройна група, от бедни до средно богати (нормални?) държави, разположени в горе вляво и в средата на диаграмата Те имат главно католически и мюсюлмански религии, а населението им има високо самочувствие. Любопитно е, че най-долу, вляво, като под-група, се разполагат 8 държави, сред които и България, които очевидно се отделят и като исторически ортодоксални (източно-православни) християнски държави. Населението на тези държави изглежда най-бедно и с най-ниско самочувствие. (За Гърция няма данни.)

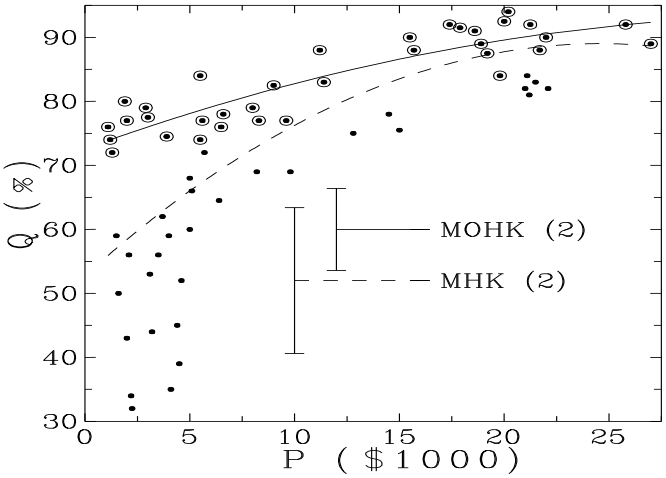


Фиг.2. Съпоставяне на средния годишен доход на населението в щатски долари за 1995 г (абсциса) и процента от населението с високо самочувствие през 1999 г (ордината) за 65 държави (Inglehart & Klingemann, 2000; Baychinska & Bakracheva, 2907).

Дали от данните, представени на фиг. 2, може да бъде извлечена статистическа зависимост? Тези данни, образуват облак от точки формата на пистолет, но в горната част на диаграмата се вижда широка ивица на повишена населеност, с положителен наклон. Тази „главна последователност" е изявена математически чрез МОНК на фиг. 3.

Както се вижда на фиг. 3, квадратичната регресия по МНК, със стандартна грешка s = 11.4 %, се влияе силно от данните в долната лява част на диаграмата и не изявява видимата „главна последователност". Напротив, квадратичната регресия със стандартна грешка s_O = 6.4 %, построена по МОНК, игнорира данните в долния ляв ъгъл на диаграмата като нетипични и изявява максималната населеност като „главна последователност".

В подобни случаи изследователят отстранява лошите точки и фитира предварително заподозряната зависимост чрез МНК. Обаче, в този случай не е лесно да се разпознаят и отделят всички лоши точки. Прилагането на МОНК решава задачата като изявява главната последователност и я описва математически (по Georgiev, 2014).



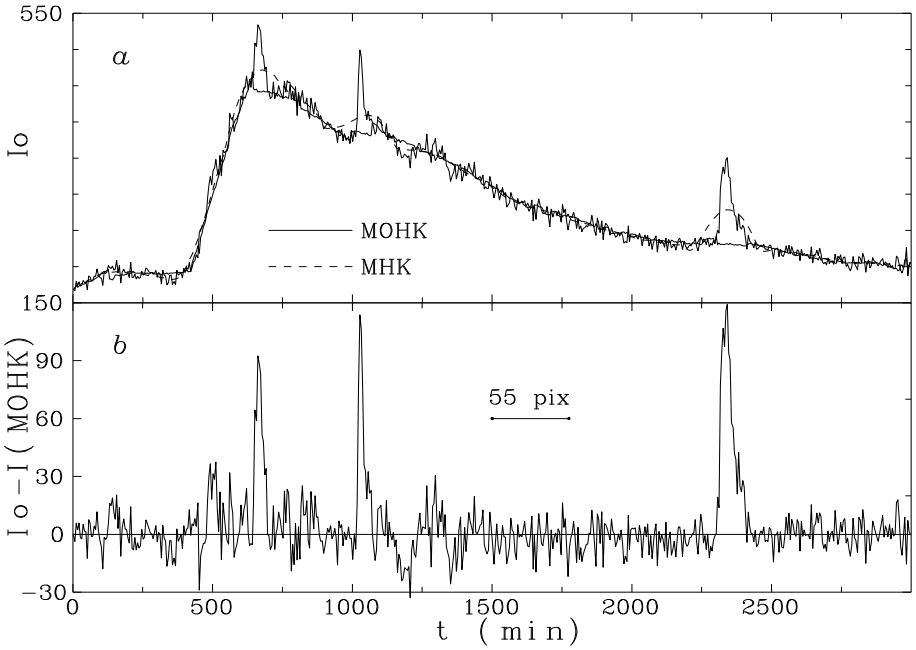
Фиг.3. Фитиране на пистолетовидната диаграма „доход *P* – самочувствие *Q*" от фиг.2 с квадратичен полином чрез МНК (прекъсната линия) и МОНК (плътна линия). Точките с елипси са „добрите" 33 точки, използвани от МОНК. Точките без елипси са „лошите" 32 точки, игнорирани от МОНК. Вертикалните отсечки представят стандартните грешки на двата вида фитирания (по Georgiev, 2014).

5.Голямо избухване и продължаващ фликеринг при активна звезда-джудже

В някои активни тесни звездни двойки (симбиотични и катаклизмични), както и в активни галактични ядра (квазари) стават мощни избухвания, проявяващи се като усимания на общия блясък до секетки пъти. Наблюдава се и т.н. фликеринг - перманентни усилвания и отслабвания на блясъка с до десетки проценти за времена от минути до часове. Избухванията и фликеритгът пораждат сложни фотометрични криви на блясъка, в които едро-мащабните и дребно-мащабните вариации на блясъка представляват значителен астрофизичен интерес. В такива случаи изглаждането на редове от фотометрични данни чрез МОНК се оказва особено ефективно.

Фигура 4*a* представя 600 измервания на едно забележително избухване на активната звезда от тип червено джудже EV Lac в ултравиолетови лъчи. Данните са от мониторинг в Националната астрономическа обсерватория Рожен, предоставени любезно от Konstantinova-Antova & Bogdanovski (2013). Главното избухване има амплитуда около 250 %. „Гладкото" поведение на избухването е представено след изглаждане на реда от данни с ширина на прозореца 55 точки, съответни на ~275 минути наблюдателно време.

Използвано е изглаждане с пълзящa парабола от вида $<z> = b_0 + b_2.t^2$, построена чрез МНК и МОНК Забележете, че МНК изглажда и размива импулсите на фликеринга (късите избухвания), докато МОНК игнорира данните във фликеринга и ги замества с по-вероятни стойности, съответстващи на гладкото поведение на кривата на блясъка.



Фиг.4. *a:* Крива на блясъка на избухване на звездата EV Lac и резултатите от нейното изглаждане с пълзяща парабола по 55 точки – чрез МНК (прекъсната крива) и чрез МОНК (плътна крива). *b*: Остатъчна крива на блясъка спрямо изгладеното чрез МОНК крива.

Фигура 4*b* показва остатъчната крива на блясъка, в която изпъкват поне 3 добре изразени къси избухвания с продължителност 100 – 150 минути и относителна амплитуда около 50 %. Тук прилагането на МОНК позволява да се изявят и фотометрират съпътстващите избухвания и да се оценят уверено техните енергии. (по Georgiev, 2013).

Трябва да се отбележи, че при изглаждане на редове от данни МОНК, за разлика от МНК, произвежда назъбена крива. За получаване на гладка крива се препоръчва след изглаждането с МОНК, с което се игнорира импулсния шум, да се приложи изглаждане с МНК, с което се постига гладка крива. В примера на фиг.4*a*. това не е направено с цел да се види доколко резултатът от МОНК е назъбен.

6.Заключение

МОНК не е широко разпространен поради големия обем изчисления при тестването на всички възможни образци. Например, за полином от 3та степен образците (полиноми през четворки точки) са всички комбинации от 4ти клас сред *n* елемента, с

брой $N = n(n\text{-}1)(n\text{-}2)(n\text{-}3)/24$. Така при n =101 данни броят образци е N = 4 082 925. А при всяка комбинация трябва да се подреждат по големина поне $n_h = n/2+1$ данни. Затова компютърното време на МОНК може да трае милиони пъти повече отколкото при МНК.

Най-простият начин за намаляване на изчислителното време при МОНК е да се използват не всички комбинации, а примерно 1/100 – 1/1000 от тях, подходящо подбрани. Един алгоритъм с прогресивно спрямо n намаляване на броя на използваните комбинации, който дава възможност за редуциране на изчислителното време от множество часове до няколко минути, е даден от Georgiev (2013).

ЛИТЕРАТУРА

Baychinska K., Bakracheva M ., 2007, Filosofski Alternativi, No.. 2/3, 32-46 (in Bulgarian)
Dimitrov, D., 2007, Private communication.
Georgiev, Ts. B. 2008. Bulg. Astron. J. 10, 93-117
Georgiev, Ts. B. 2013. Aerospace Research in Bulgaria, in print
Georgiev, Ts. B. 2014. Publ. Astron. Soc..Bulg., in print (in Bulgarian)
Inglehart R., Klingemann H.D., 2000, Genes, Culture, Democracy and Happiness. Cambridge MA, MIT Press.
Konstantinova-Antova, R., Bogdanovski. R., 2013, Private communication.
Rousseeuw P. J., 1984, J. Am. Stat. Assoc. 79, 871
Rousseeuw P. J., Leroy A. M., 1987, Robust Regression and Outlier Detection, John Willy & Sons