COMPUTATIONAL SCIENCE AND TECHNIQUES
Special Issue for Research Innovations Fundamentals 2011
Volume 1, Number 1, 2013, 30-35

Publisher: Klaipėda University
http://journals.ku.lt/index.php/CST
Online ISSN: 2029-9966

PUSIAU EMPIRINIŲ ROBASTINIŲ ALGORITMŲ TAIKYMAS INVESTICINIO PORTFELIO SUDARYMUI

Natalja Kaskevič[1], Šarūnas Raudys[2], Vitalij Denisov[1]
[1]Informatikos katedra, Klaipėdos universitetas, Herkaus Manto g. 84, LT-92294, Klaipėda
natalja.kaskevic@gmail.com, vitalij@ik.ku.lt
[2]Informatikos katedra, Vilniaus universitetas, Naugarduko g. 24, LT-03225, Vilnius
raudys@ktl.mii.lt

Anotacija. Analizuojant finansų rinkos duomenis dažnai susiduriama su stebėjimais, kurie yra nukrypę nuo stebimo dėsningumo. Tai vadinama didelių nuokrypių arba robastiškumo problema. Šiame straipsnyje analizuojami būdai mažinti didelių nuokrypių įtaką ir pateikiama pusiau empirinė nuokrypių šalinimo procedūra, skirta naudoti vertybinių popierių prekybinių sistemų optimizacijos procese.

Pagrindiniai žodžiai: duomenų išskirtis, robastiniai įverčiai, suskirstymo vertė, vertybinių popierių investicinis portfelis.

Įvadas

Sprendžiant duomenų apdorojimo problemas susiduriama su modelio parametrų vertinimo užduotimi. Parametrų vertinimui didelės įtakos turi duomenų kokybė, tačiau dažnai turimi duomenų masyvai turi duomenų, nepriklausančių duotai imčiai, talpina tuščius įrašus ir pan. Tokie nukrypimai (triukšmai) atsiranda dėl įvairių priežasčių, pavyzdžiui dėl pačios duomenų struktūros ypatumų, dėl klaidingo duomenų įvedimo ir pan. Jeigu sudarytas prognostinis modelis yra paremtas triukšmingais duomenimis, jis bus nekorektiškas ir teiks klaidingus rezultatus. Triukšmingų duomenų problemos sprendimą apsunkina tai, kad ne visada įmanoma apibrėžti kokie duomenys yra pašaliniai, o kokie tinkami. Jeigu būtų įmanoma tai nustatyti, tuomet modelį pavyktų sudaryti remiantis vien tik patikimais, neturinčiais triukšmo duomenimis, ignoruojant nukrypius. Šiame straipsnyje duomenų apdorojimo metodų tikslumas didinamas panaudojus atsparius triukšmams robastinius įverčius (*angl.* robust estimators).

1. Robastinių įverčių apžvalga

Dažniausiai susiduriama su duomenų imtimis, kurių dydis yra n, o x_i – yra nepriklausomas stebėjimas iš duotos duomenų imties su pasiskirstymo funkcija $F[(x-\mu)/\sigma]$; F yra žinoma (paprastai – standartinio normalaus (Gauso) pasiskirstymo funkcija). Tuomet reikia nustatyti centro parametrą μ, ir pasiskirstymo parametrą σ. Klasikiniu centro parametru yra skaitomas aritmetinis vidurkis. Tačiau, atsiradus nors vienai išskirčiai, šis parametras itin pasikeičia. Viena iš priemonių apibūdinti įverčio robastiškumą yra suskirstymo vertė (*angl.* breakdown value) (Hubert and Debruyne, 2009). Suskirstymo vertė nurodo mažiausią dalį išskirčių (triukšmų) imtyje, kuri sukelia įverčio iškraipymą, t. y.

įvertis įgauna netikslias, klaidingas tiriamojo modelio atžvilgiu, reikšmes. Tokiu atveju, aritmetinio vidurkio suskirstymo vertė yra lygi *1/n*, t. y. didelės apimties imtims lygi 0%.

Vietoj aritmetinio vidurkio, galima pasinaudoti kitais centro parametrais, pavyzdžiui mediana. Mediana yra centrinis imties požymio stebėjimas, kuris lygus *x[(n+1)/2]*, jei *n* yra nelyginis skaičius, ir yra *[x(n/2) + x(n/2+1)]/2*, jei *n* yra lyginis. Jos suskirstymo vertė yra apie 50% (Rousseeuw and Hubert, 2011), t. y. toks įvertis yra atsparus penkiasdešimčiai procentų triukšmų. Klasikinis pasiskirstymo parametras σ yra standartinis nuokrypis, išreiškiamas formule (1):

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\,. \tag{1}$$

Pažymėtina, kad netgi viena išskirtis gali stipriai iškraipyti jo reikšmę. Atsparus triukšmams pasiskirstymo įvertis yra *MAD* – visų absoliučių nuokrypių nuo medianų mediana (*angl.* **M**edian of all **A**bsolute **D**eviations from the median) (Rousseeuw and Hubert, 2011):

$$MAD = 1.483\ median_{i=1,\dots,n}\left|x_i - median_{j=1,\dots,n}(x_j)\right|, \tag{2}$$

kur 1,483 yra koregavimo koeficientas, kuris daro MAD nešališku normaliuoju pasiskirstymu. Panašius atsparumo rezultatus galima gauti pasinaudojus Q_n įverčiu: (Nunkesser et al., 2009)

$$Q_n = 2.2219\left\{\left|x_i - x_j\right|; i < j\right\}_{(k)}, \tag{3}$$

kur $k = \binom{n}{2}/4$. Tokiu būdu, pasiskirstymo parametru tampa visų porinių skirtumų, tarp dviejų duomenų taškų, pirmasis kvartilis. *MAD* ir Q_n įverčio suskirstymo vertė yra apie 50% (Brandt, 2010).

Atsparus triukšmams yra interkvartilinis diapazonas (*angl.* interquartile range, IQR), išreikštas kaip skirtumas tarp duomenų trečiojo ir pirmojo kvartilio: $IQR = x_{\lceil 3n/4 \rceil} - x_{\lfloor n/4 \rfloor}$ (Kang and Lee, 2005). *IQR* suskirstymo vertė yra apie 25%, tačiau šis įvertis pasižymi paprastumu.

Naudoti aukščiau išvardintus įverčius yra prasminga tik tada, kai yra žinoma, kad duomenyse yra daug triukšmų, tačiau naudojimasis jais aprašant normalius modelius teikia mažiau efektyvius rezultatus, negu taikant klasikinius metodus. Norint gauti didesnį įverčių veiksmingumą, galima pasinaudoti ir kitomis robastinėmis procedūromis (Bickel and Fruehwirthe, 2006), pvz. realizuojančiomis M-įverčius (Ozturk, 2005)(Raudys, 2008).

2. Robastinių įverčių panaudojimas investicinio portfelio sudaryme

Vertybinių popierių prekybiniuose sistemose duomenų nukrypimai atsiranda pagrindinai dėl duomenų struktūros ypatumų. Sistemos, naudojamos optimaliam akcijų

prekybos portfelio sukūrimui, sudaro įrašus tokiu būdu: teigiamas skaičius, kai buvo gautas pelnas; neigiamas – nuostolis, nulis – kai investuoti į rinką nevertėjo, t. y. sistema nebuvo panaudota. Tokių duomenų fragmentas yra pateiktas 1 paveiksle.

pnls <142x2078 double>

	135	136	137	138	139	140	141
16	0	0	0	0	0	0	0
17	0	0	0	0	160.2800	-14.7200	0
18	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0
20	80.1400	20.1400	0	0	0	155.1400	-4.8600
21	155.1400	362.5000	75	155.1400	0	-32.3600	-342.3600
22	-19.8600	362.5000	75	87.5000	12.5000	-77.2200	-217.3600
23	0	0	0	705.5600	0	0	0

1 pav. Duomenų fragmentas (MatLab 2010 aplinka).

Pateiktas pavyzdys yra iškarpa iš masyvo, sudaryto iš 142 prekybinių sistemų, 2078 dienų laikotarpio duomenų, naudotų akcijų prekybos portfelio sudarymui. Jei kiekviena iš 142 sistemų būtų panaudota kiekvieną dieną, įrašai sudarytų tos dienos gauto pelno arba nuostolio, o nuliniai įrašai būtų tie, kurie nedavė prieaugio arba praradimų.

Sudarant investicinį portfelį naudojami įvairūs optimizacijos metodai, populiariausi yra *James–Stein*, *Black–Litterman*, *Markowitz* ir kt. (Rousseeuw and Hubert, 2011) Formuojant portfelį optimizacijos procedūros tikslas yra sumodeliuoti turto grąžą kaip normalinio skirstinio funkciją, kurioje standartinis nuokrypis aprašo riziką, o vidurkis – tikėtiną grąžą, ir nustatyti strategijai atitinkamą svorį (koeficientą). Tokiu būdu optimizacijos kriterijus formuluojamas kaip investavimo strategijų svorinė kombinacija (tikėtina investicijų grąža):

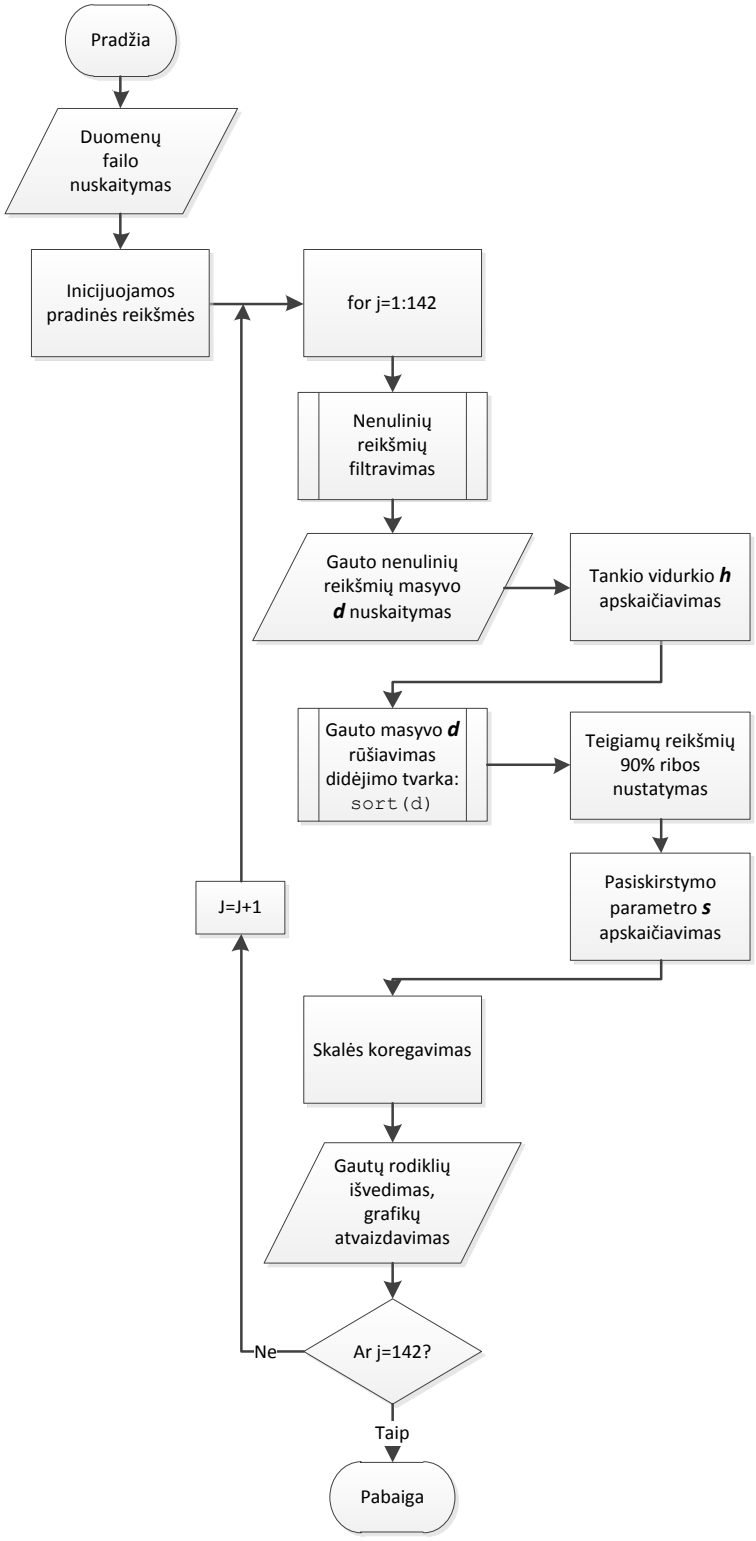
$$E(G_p) = \sum_{i=1}^{n} \omega_i E(G_t), \tag{4}$$

kur G_p – portfelio grąža, G_t – turto grąža, ω_i – strategijos i svorinis koeficientas, nagrinėjamu atveju $i = 142$. Norint pasinaudoti pasirinktu optimizacijos metodu, reikia atlikti duomenų normalizavimą, t. y. priartinti pradinius duomenis prie pasiskirsčiusių pagal normalųjį (Gauso) skirstinį:

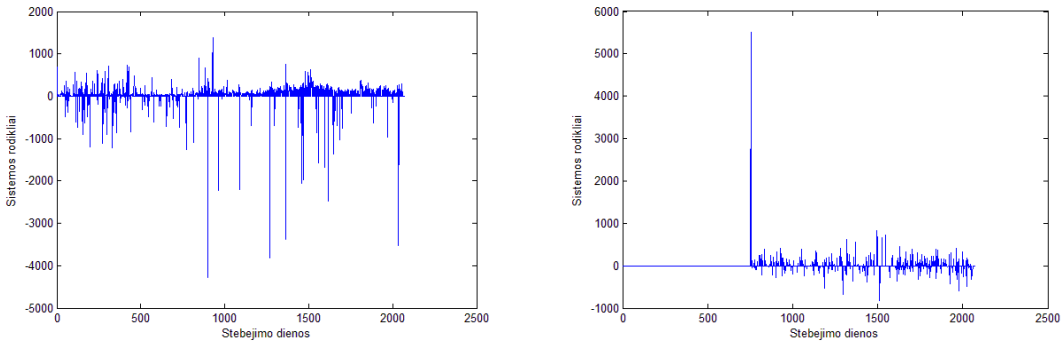
$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}, \tag{5}$$

kur μ – kiekvieno požymio (prekybinės sistemos) aritmetinis vidurkis, o σ – standartinis nuokrypis. Dėl labai didelio „nulinių“ duomenų kiekio vertybinių popierių prekybiniuose sistemose standartiniai metodai tampa neefektyvūs. Norint tiksliau įvertinti didelius nukrypimus ir atsižvelgti į tą faktą, kad paprastai daug dienų prekybos sistemos nebuvo panaudotos, galima aproksimuoti duomenys normaliniu pasiskirstymu. 2 pav. pateikiami siūlomos pradinių duomenų normalizavimo procedūros žingsniai, pritaikyti realizacijai MatLab programinio paketo aplinkoje. 2 paveiksle parodytoje procedūroje kiekvienai iš 142 sistemų pritaikytas binarinis filtras, kuris leidžia atsikratyti nulinių stebėjimų. Taip gautas naujas masyvas d tik su nenuliniais įrašais. Analizuojant turimus įrašus, galima pastebėti, kad triukšmai atsiranda ne tik dėl didelio skaičiaus „nulinių reikšmių“, bet ir dėl

pasitaikančių klaidingų įrašų, kurie aiškiai išsiskiria nuo kitų (žr. 3 pav.). Tokių nukrypimų pašalinimui binarinis filtras nepadės, todėl reikia dar „nukirpti uodegas“.



2 pav. Siūloma pradinių duomenų normalizavimo procedūra

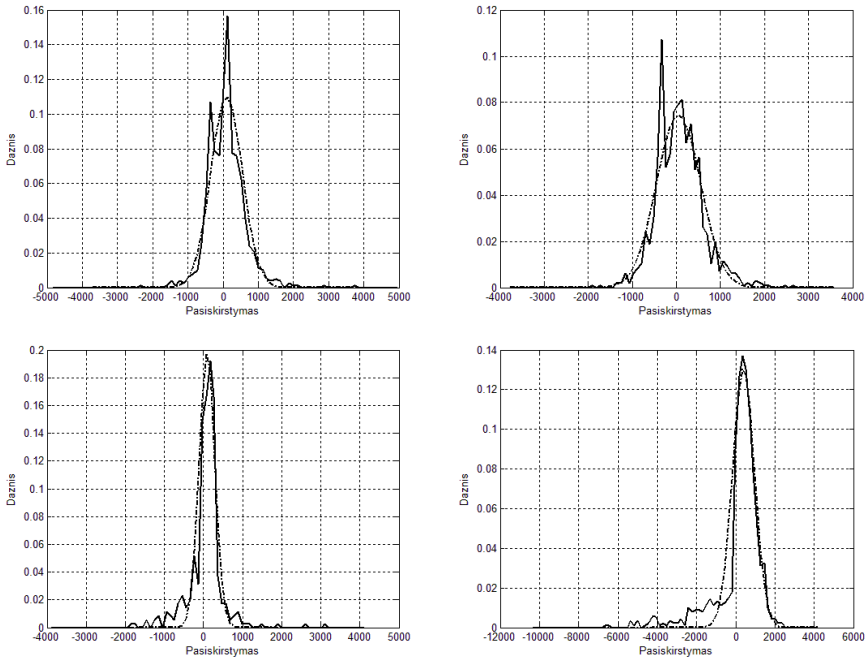


3 pav. 130-os ir 142-os sistemos duotų rezultatų grafikai su aiškiai matomais nuokrypiais (grafikai sugeneruoti MatLab 2010)

Šiuo atveju, Gauso pasiskirstymo aproksimacijai gauti, vietoj vidurkio μ ir standartinio nuokrypio σ pasinaudosime kitais įverčiais. Pirma, panaudojus MatLab funkcija *„interp1“*, apskaičiuojamas aproksimuojančio normalinio pasiskirstymo tankio vidurkis – h“. Antra, atmetant 10% didžiausių teigiamų reikšmių panaikinama „uodega“ (procentas buvo parinktas empiriniu būdu; galima taikyti ir kitokią reikšmę). Reikšmė, kuri panaudota vietoj standartinio nuokrypio, yra randama tokiu būdu:

$$s = t/(1.28155 + h), \qquad (6)$$

kur t – yra dviejų ribinių, esančių ant 10% ribos, stebėjimo vidurkis. Tokiu būdu atsikratoma neįprastiniais stebėjimais, ir lieka duomenys tarp jau nustatytų mažiausių ir didžiausių reikšmių, kurie užimtų 1.28155 (viršuje - iki 10% ribos) plius h (apačioje - iki mažesniu už nulį reikšmių) kvadratiniu nukrypimu s. Turėdami gautus s ir h parametrus, galime aproksimuoti duomenis ir palyginti gautą aproksimaciją su normaliniu pasiskirstymu (žr. 4 pav.):



4 pav. Gautų duomenų (ištisinė linija) ir normalinio pasiskirstymo (dryžuota linija) grafikai suvienodintoje skalėje (grafikai sugeneruoti MatLab 2010).

Išvados

- Statistinės analizės metu naudojamos klasikinės duomenų imties charakteristikos, tokios kaip centro ir pasiskirstymo parametrai, yra labai jautrūs triukšmams, tad viena išskirtis labai iškreipia šių parametrų reikšmes.
- Jeigu tiriamoje duomenų imtyje yra daug nuokrypių, tuomet efektyviau yra pasinaudoti atspariais (robastiniais) įverčiais, kurių atsparumo triukšmams vertė, išreiškiama suskirstymo verte, yra didesnė, nei klasikinių metodų.
- Analizuojant vertybinių popierių prekybinių sistemų duomenis buvo pasiūlytas ir realizuotas pusiau empirinis metodas, kuris pašalina išskirtis ir paruošia duomenis tolimesnei parametrų vertinimo procedūrai.

Literatūra

Bickel, D. R., Fruehwirth, R. (2006). On a Fast, Robust Estimator of the Mode: Comparisons to Other Robust Estimators with Applications, Computational Statistics and Data Analysis 50, 3500-3530.

Brandt, M. W. (2010). Portfolio Choice Problems, Handbook of Financial Econometrics: Tools and Techniques, 269-336.

Hubert, M., Debruyne, M. (2009). Breakdown Value, Computational Statistics. Volume 1, Issue 3, November/December 2009, 296–302.

Kang, S. J., Lee, M. (2005). Q-convergence with interquartile ranges, Journal of Economic Dynamics and Control, vol. 29, issue 10, 1785-1806.

Nunkesser, R., Fried, R., Schettlinger, K., Gather, U. (2009). Online analysis of time series by the Qn estimator, Computational Statistics & Data Analysis, Issue (Month): 6 (April), , volume 5, 2354-2362.

Ozturk, O. (2005). Robust joint estimation of location and scale parameters in ranked set samples, Journal of Statistical Planning and Inference, Volume 127, Issues 1-2, 1, January 2005, 295-308.

Raudys, Š. (2008). Žinių išgavimas iš duomenų. Klaipėdos Universiteto leidykla. Pusl.: 32-34. Klaipėda.

Rousseeuw, P. J., Hubert, M. (2011). Robust statistics for outlier detection, Data Mining and Knowledge Discovery, Volume 1, Issue 1, January/February 2011, 73–79.

SEMIEMPIRICAL ROBUST ALGORITHM FOR INVESTMENT PORTFOLIO FORMATION

Natalja Kaskevič, Šarūnas Raudys, Vitalij Denisov

Summary

When analyzing stock market data, it is common to encounter observations that differ from the overall pattern. It is known as the problem of robustness. Presence of outlying observations in different data sets may strongly influence the result of classical (mean and standard deviation based) analysis methods or models based on this data. The problem of outliers can be handled by using robust estimators, therefore making aberrations less influential or ignoring them completely. An example of applying such procedures for outlier elimination in stock trading system optimization process is presented.

Key words: data outliers, robust estimators, breakdown value, stock portfolio.