PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Bartosz Kaszuba
Uniwersytet Ekonomiczny we Wrocławiu

ODPORNE MACIERZE KOWARIANCJI W KONSTRUKCJI PORTFELA

Streszczenie: Założenia klasycznych metod statystycznych wykorzystywanych w teorii portfela zazwyczaj nie są spełnione (np. rozkład stóp zwrotu nie jest rozkładem normalnym), dlatego wskazane jest poszukiwanie alternatywnych metod, których założenia będą spełnione w praktyce. Takimi metodami mogą być metody odporne. Celem artykułu jest zaprezentowanie odpornych macierzy kowariancji, opisanie ich podstawowych właściwości oraz metod wyznaczania. Dodatkowo w artykule przeprowadzono analizę możliwości i zasadności zastosowania macierzy odpornych w praktyce.

1. Wstęp

W klasycznej statystyce, nawet w najprostszych sytuacjach, występują założenia (jawne bądź ukryte) m.in. o losowości i niezależności, o rozkładach prawdopodobieństwa danej próby. Założenia te niekoniecznie muszą być rzeczywiste, jednakże matematycznie są one wygodne. W praktyce nawet niewielkie odstępstwa od zbyt wygórowanych założeń mogą spowodować w efekcie niewłaściwe wnioski. W celu uniknięcia błędnych wniosków, nawet w przypadku gdy odstępstwa od założeń są niewielkie, zaproponowano metody odporne. Odporność rozumiana jest jako niewrażliwość na niewielkie odchylenia od założeń [Huber, Ronchetti 2009]. Statystyka odporna koncentruje się głównie na odstępstwach od rozkładu prawdopodobieństwa (zazwyczaj gaussowskiego), gdzie nieznaczne zmiany w rozkładzie (który jest funkcją) przekładają się na niewielkie zmiany w estymatorach (które są liczbą). Sama koncepcja i metody statystyki odpornej zostały zapoczątkowane w latach 50. XX wieku.

Klasyczna statystyka jest również (a nawet nadmiernie często) stosowana w teorii portfela, mimo iż stopy zwrotu nie mają rozkładu normalnego [Jajuga 2000]. Zatem wskazane jest poszukiwanie innych metod niż klasyczne, które będą bardziej dokładne i adekwatne do modelu. Takimi metodami mogą być np.: metody odporne, które nie zakładają określonego rozkładu stóp zwrotu. Istnieje wiele metod odpornych i algorytmów, jednak niedokładna wiedza oraz poziom skomplikowania może doprowadzić do wyboru nieodpowiedniej metody i w efekcie przyczynić się do nieprawdziwych wniosków [Kaszuba 2009].

Celem artykułu jest zaprezentowanie odpornych macierzy kowariancji, zebranie ich podstawowych właściwości oraz analiza możliwości i zasadności zastosowania macierzy odpornych w praktyce. W artykule pokazano również możliwości zastosowania macierzy odpornych w konstrukcji portfela.

2. Odporność statystyczna w teorii portfela

W teorii portfela wielu spółek [Markowitz 1952] wyliczenie optymalnych udziałów w portfelu sprowadza się do odpowiedniego wyestymowania parametrów średniej i kowariancji. W klasycznym przypadku zakłada się, że rozkład stóp zwrotu pochodzi z rozkładu normalnego.

W przypadku odpornym zakłada się, że rozkład stóp zwrotu zawiera obserwacje zagłuszające. Jednym z modeli, który dobrze określa powyższe założenie, może być model, gdzie rozkład stóp zwrotu pochodzący z wielowymiarowego rozkładu normalnego zagłuszany jest innym wielowymiarowym rozkładem:

$$F_\varepsilon(G) = (1-\varepsilon)F + \varepsilon G, \quad 0 < \varepsilon < \frac{1}{2},$$

gdzie F jest wielowymiarowym rozkładem normalnym, natomiast G jest rozkładem zagłuszającym. Przy zastosowaniu takiego modelu do rozkładu stóp zwrotu zakłada się, że danego dnia stopy zwrotu pochodzą albo z głównego rozkładu (normalnego), albo z rozkładu zagłuszającego. Takie założenie jest prawdziwe, gdy istnieją duże zależności pomiędzy stopami zwrotu wszystkich spółek, co w praktyce nie jest spełnione. Powyższy model ε-zaburzony zakładany jest m.in. przy wyznaczeniu estymatorów MCD i MVE.

Innym modelem odpornym, bardziej praktycznym, może być następujący model ε-zaburzony:

$$F_\varepsilon(G) = (1-E)F + EG, \text{ gdzie } E = diag([\varepsilon_1 \quad \varepsilon_2 \quad \dots \quad \varepsilon_n]), \quad 0 < \varepsilon_i < \frac{1}{2}.$$

Jak można zauważyć, nie ma w tym przypadku założenia o zależności rozkładów stóp zwrotu, zatem powyższy model jest lepiej dostosowany do wielowymiarowej analizy odpornej rozkładów stóp zwrotu. Powyższy model wykorzystywany jest w metodzie szacowania poszczególnych elementów macierzy kowariancji dla par zmiennych losowych.

Można zauważyć na powyższych modelach, że przy wyborze odpornych metod niezbędne jest dokładne określenie zakładanego modelu, a następnie wyszukanie najbardziej odpowiedniej metody (o określonej efektywności, punkcie załamania, funkcji wpływu itp.). W przypadku nieodpowiedniego doboru lub niewystarczającej wiedzy efekt zastosowania wybranej metody może być odwrotny do oczekiwanego.

3. MVE – *Minimum Volume Ellipsoid*

Pierwszą omawianą klasą estymatorów macierzy kowariancji będą *S*-estymatory MVE i MCD o wysokim punkcie załamania, dla których funkcje ρ są nieróżniczkowalne, stąd $\rho(d_i)=0$ lub $\rho(d_i)=1$. Metoda najmniejszej objętości elipsoidy (*Minimum Volume Ellipsoid* – MVE) zaproponowana przez Rousseeuw [1984] polega na wybraniu najmniejszej elipsoidy zawierającej co najmniej *h* elementów ze zbioru $X=\{\mathbf{x_1},\ldots,\mathbf{x_n})\subset \mathrm{R}^{\mathrm{p}}$. Macierz kowariancji wyliczona dla zbioru elementów z najmniejszej elipsoidy nazywa się estymatorem MVE, natomiast środek najmniejszej elipsoidy jest estymatorem MVE parametru położenia. W celu wyznaczenia estymatora MVE rozważmy następujący problem:

$$(\hat{\boldsymbol{\mu}},\hat{\mathbf{S}})=\underset{\substack{(\boldsymbol{\mu},\mathbf{S})\in \mathrm{R}^{\mathrm{p}}\times SPD(p)\\ |\mathbf{S}|=1}}{\operatorname{argmin}} d_h^2(\boldsymbol{\mu},\mathbf{S}), \tag{1}$$

gdzie $SPD(p)$ jest zbiorem wszystkich dodatnio określonych, symetrycznych macierzy $\mathbf{S}\in \mathrm{R}^{\mathrm{p}\times\mathrm{p}}$. Macierz $\hat{\mathbf{S}}$ może być nazwana „macierzą kształtu", ponieważ określa ona kształt elipsoidy, a nie jej wielkość, stąd $|\hat{\mathbf{S}}|=1$. Wyrażenie $d_h^2(\boldsymbol{\mu},\mathbf{S})$ określa *h*-tą wartość porządkową odległości kwadratowej pomiędzy $\mathbf{x}_i$ oraz $\boldsymbol{\mu}: d_h^2(\boldsymbol{\mu},\mathbf{S})=\{(\mathbf{x}_i-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x}_i-\boldsymbol{\mu});1\le i\le n\}_h$. Stąd estymator MVE jest określony przez:

$$(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})=(\hat{\boldsymbol{\mu}},c(n,p,h)d_h^2(\hat{\boldsymbol{\mu}},\hat{\mathbf{S}})\hat{\mathbf{S}}),$$

gdzie $c(n,p,h)$ jest współczynnikiem korekcyjnym.

W celu maksymalizacji punktu załamania należy wybrać wartość $h=\left\lfloor\frac{n+p+1}{2}\right\rfloor\approx\frac{n}{2}$ [Lopuha, Rousseeuw 1991]. Metoda MVE jest rekomendowana, gdy co najmniej pięć obserwacji przypada na jeden wymiar, czyli $n/p>5$. Algorytm wyznaczania estymatora MVE można znaleźć w pracach takich autorów, jak Maronna, Martin, Yohai [2006] lub Agullo [1996].

Dla tej macierzy kowariancji zakłada się, że zagłuszenia rozkładu są jak w modelu pierwszym, opisanym w pkt 2.

4. MCD

Celem metody minimalnego wyznacznika macierzy kowariancji (*Minimum Covariance Determinant* – MCD) zaproponowanej przez Rousseeuw [1985] jest znalezienie *h* obserwacji (spośród wszystkich *n* obserwacji), dla których klasyczna ma-

cierz kowariancji ma najmniejszy wyznacznik. Estymatorem MCD parametru położenia jest wtedy średnia tych *h* punktów, natomiast estymatorem MCD parametru rozproszenia (skali) jest macierz kowariancji wyliczona dla otrzymanych *h* punktów. Formalnie estymator MCD można zdefiniować (zachowując wcześniejsze oznaczenia) w następujący sposób:

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \underset{(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}) \in \mathrm{R}^{\mathrm{p}} \times SPD(p)}{\operatorname{argmin}} \sum_{i=1}^{h} d_i^2(\boldsymbol{\mu}_i, \hat{\boldsymbol{\Sigma}}) \ .$$

Punkt załamania estymatora MCD jest taki sam jak punkt załamania estymatora MVE, jednak estymator MCD ma więcej zalet niż MVE. Efektywność estymatora MCD jest lepsza, ponieważ jest asymptotycznie normalny [Butler, Davies, Jhun 1993], natomiast MVE ma w tym przypadku niższy współczynnik zbieżności [Davies 1992], co powoduje dużo niższą asymptotyczną efektywność. Lepsze dopasowanie estymatora MCD sprawia, że jest bardzo użyteczny przy wyliczeniu estymatorów w regresji jednokrokowej. Ponadto estymator MCD jest kluczowym składnikiem estymatorów hybrydowych opisanych przez autorów, takich jak Woodruff i Rocke oraz w liniowej analizie dyskryminacyjnej o wysokim punkcie załamania opisywanej przez Hawkinsa i McLachlana [Rousseeuw 1985].

W celu wyliczenia estymatora MCD niezbędne byłoby porównanie wszystkich $\binom{n}{h}$ zbiorów, co w przypadku dużej liczby danych i wielu wymiarów powodowałoby długi czas obliczenia. Dlatego do obliczenia estymatora MCD można użyć metody FAST-MCD polegającej na próbkowaniu, która została zaproponowana i opisana przez autorów, takich jak Rousseuw i van Driessen [1999].

Wyniki symulacyjne i empiryczne pokazują, że metoda FAST-MCD daje „dobre" wyniki, czas obliczeń zaś w porównaniu z innymi metodami dla estymacji MCD jest dużo krótszy.

Dla tej macierzy kowariancji zakłada się, że zagłuszenia rozkładu są jak w modelu pierwszym, opisanym w pkt 2.

5. Estymator OGK

Zortogonalizowany estymator Gnanadesikana i Ketteringa (*Orthogonalised Gnanadesikan-Kettenring* – OGK) jest oparty na metodzie szacowania poszczególnych elementów macierzy kowariancji dla par zmiennych losowych (*pairwise estimates*). Estymator OGK jest modyfikacją estymatora zaproponowanego przez autorów, takich jak Gnanadesikan i Kettenring [1972] , gdzie dla każdej pary estymatorów $\mathbf{X_i}, \mathbf{X_j}$ wyliczana jest odporna kowariancja:

$$\operatorname{cov}(\mathbf{X_i}, \mathbf{X_j}) = \frac{1}{4}\left(\sigma(\mathbf{X_i} + \mathbf{X_j})^2 - (\mathbf{X_i} - \mathbf{X_j})^2\right) , \qquad (2)$$

gdzie σ jest odpornym estymatorem wariancji (np.: odchylenie medianowe). Niestety, otrzymana w ten sposób macierz nie zawsze jest dodatnio określona i afinicznie ekwiwariantna. Maronna i Zamar [2002] zaproponowali modyfikację macierzy GK opartą na dekompozycji na wektory własne, co pozwoliło na osiągnięcie dodatniej określoności i afinicznej ekwiwariancji. Estymator macierzy kowariancji wyliczany jest w następujący sposób (szczegółowy algorytm znajduje się m.in. w [Maronna, Zamar 2002] lub [Maronna, Martin, Yohai 2006]): najpierw wyliczona jest macierz $\hat{\mathbf{\Sigma}}$ za pomocą estymatora GK, następnie tworzona jest dekompozycja na wektory własne $\hat{\mathbf{\Sigma}} = \mathbf{U}\Lambda\mathbf{U}^{-1}$. W kolejnym kroku tworzone jest odwzorowanie danych na bazie wektorów własnych, następnie estymuje się wariancję w nowym układzie współrzędnych, w ostatnim kroku zaś następuje wyliczenie macierzy OGK według wzoru $\hat{\mathbf{\Sigma}} = \mathbf{U}\Gamma\mathbf{U}^{-1}$, gdzie Γ jest macierzą diagonalną dla wariancji wyestymowanej w kroku 4. Punkt załamania tego estymatora jest nie mniejszy od punktu załamania estymatora wariancji użytego do wyliczenia macierzy GK we wzorze (2).

Dla tej macierzy kowariancji zakłada się, że zagłuszenia rozkładu są zgodne z drugim modelem, opisanym w pkt 2.

6. *CM*-estymatory

Kent i Tyler [1996] zaproponowali ograniczone *M*-estymatory (constrained *M*-estimates, *CM*-estimates), które łączą cechy dobrej lokalnej odporności *M*-estymatorów oraz dobrej globalnej odporności *S*-estymatorów. *GM*-estymatory wylicza się według następującego wzoru:

$$(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}) = \underset{(\boldsymbol{\mu},\mathbf{S})\in \mathrm{R}^{\mathrm{p}}\times SPD(p)}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\rho(d_i^2) + \frac{1}{2}\log|\mathbf{\Sigma}| \right\}$$

przy ograniczeniach:

$$\frac{1}{n}\sum_{i=1}^{n}\rho(d_i^2) \le \varepsilon,$$

gdzie $d_i^2(\boldsymbol{\mu}, \hat{\mathbf{\Sigma}}) = (\mathbf{x}_i - \boldsymbol{\mu})'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$ oraz funkcja ρ jest ograniczona. Dzięki stałym regulującym *CM*-estymatory umożliwiają odpowiednie dobranie funkcji wpływu oraz efektywności estymatora. Dodatkowo modyfikacja efektywności nie ma wpływu na punkt załamania [Kent, Tyler 2001]. Za funkcję ρ można przyjąć np. zmodyfikowaną funkcję podwójnie ważoną (*translated biweight*):

$$\rho'(d;c,M) = \begin{cases} d, & 0 \le d \le m \\ d\left(1 - \left(\dfrac{d-m}{c}\right)^2\right)^2, & M \le d \le M + c \\ 0, & d > M + c \end{cases},$$

która zwiększa efektywność *CM*-estymatora w porównaniu z funkcją podwójnie kwadratową. Więcej na temat właściwości powyższej funkcji można znaleźć np. w pracach autorów, takich jak P. Filzmoser, R. Maronna i M. Werner [2008] lub Rocke [1996].

Dla tej macierzy kowariancji zakłada się, że zagłuszenia rozkładu są jak w modelu pierwszym, opisanym w pkt 2.

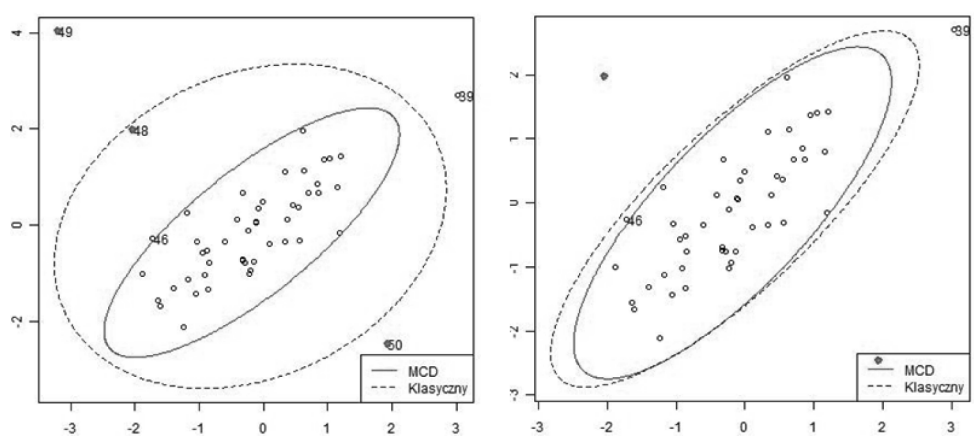
7. Praktyczne zastosowanie

W tej części zostaną opisane możliwości zastosowań wymienionych metod na podstawie danych symulacyjnych oraz historycznych.

W przypadku danych symulacyjnych wygenerowano dane z następującej mieszanki rozkładów: $F = (1-0,06)N(\mu_1,\Sigma_1) + 0,06N(\mu_2,\Sigma_2)$, gdzie $\Sigma_1 = \begin{pmatrix} 1 & 0,8 \\ 0,8 & 1 \end{pmatrix}$, $\mu_1 = (0 \quad 0)$, $\Sigma_2 = \begin{pmatrix} 4 & -3,6 \\ -3,6 & 4 \end{pmatrix}$, $\mu_2 = (0 \quad 0)$. Zatem, jak można zauważyć w części 2 artykułu, najlepszymi estymatorami dla tego modelu są estymatory MCD i MVE.

Na rysunku 1 wyznaczone są obszary eliptyczne dla zasymulowanych danych, dla macierzy MCD i klasycznej macierzy kowariancji (w przypadku pozostałych macierzy wyniki są zbliżone). Obserwacje 46-50 są obserwacjami wygenerowanymi z rozkładu zagłuszającego.

Jak można zauważyć, w tym przypadku obszar eliptyczny klasycznej macierzy kowariancji nie odrzuca nietypowych obserwacji 46 i 48, które odrzucane są przez



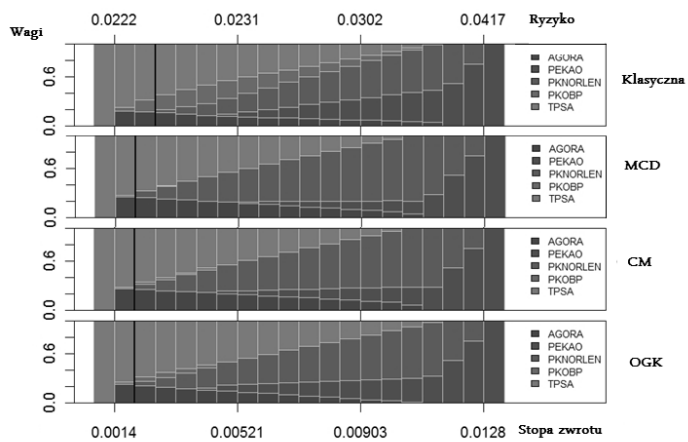
Rys. 1. Obszary eliptyczne estymatora MCD

Źródło: opracowanie własne.

MCD. Dodatkowo, w przypadku wyliczenia obszaru eliptycznego klasycznej macierzy dla obserwacji niezagłuszonych, obszar ten jest zbliżony do obszaru wyznaczonego przez MCD. Zatem gdy dane pochodzą z rozkładu zagłuszonego, lepszym dopasowaniem cechują się estymatory odporne (wniosek ten wypływa z przesłanek teoretycznych).

Dane historyczne, dla których został zobrazowany przykład praktyczny, pochodzą z GPW w Warszawie, zaś liczba spółek i analizowany okres nie mają w tym przypadku wpływu na wnioski z danego przykładu. Analizowane są logarytmiczne stopy zwrotu dla losowo wybranych spółek Agora, PEKAO, PKN Orlen, PKO BP, TPSA z przykładowego okresu 18.02.2009-14.04.2009. Celem przykładu jest jedynie zobrazowanie różnic pomiędzy poszczególnymi estymatorami.

Dla estymatorów MCD, OGK, *CM*-estymatora oraz klasycznego estymatora macierzy kowariancji wyliczono portfele, a następnie policzono ich wagi w zależności od ryzyka i stopy zwrotu. Rysunek 2 przedstawia wyliczone w ten sposób wagi.



Rys. 2. Wagi portfeli

Źródło: opracowanie własne.

Na każdej z 4 części rysunku oś dolna pozioma przedstawia średnią stopę zwrotu, oś górna pozioma przedstawia ryzyko (odchylenie standardowe), natomiast osie pionowe przedstawiają udziały poszczególnych spółek w portfelu. Jak można zaobserwować, portfele wykorzystujące klasyczną macierz kowariancji (portfele Markowitza) mają istotnie różne wagi od portfeli odpornych. Na przykład udział spółki PKO BP jest największy w portfelu wyznaczonym metodą klasyczną, natomiast w portfelach z użyciem metod MCD i CM spółka nie jest uwzględniana. W metodzie OGK udział spółki jest większy, co może wynikać z założonego modelu dla tego estymatora.

8. Podsumowanie

Założenia klasycznych metod statystycznych wykorzystywanych w teorii portfela zazwyczaj nie są spełnione (np. rozkład stóp zwrotu nie jest rozkładem normalnym), dlatego wskazane jest poszukiwanie alternatywnych metod, których założenia będą spełnione w praktyce. W artykule opisano *S*-estymatory – MCD i MVE, *CM*-estymatory oraz estymator OGK. Estymator MCD ma lepsze właściwości niż estymator MVE, jednak oba oparte są na modelu, który zakłada silną korelację pomiędzy stopami zwrotu wszystkich spółek. *CM*-estymatory mają bardzo dobre właściwości teoretyczne i umożliwiają dokładne dopasowanie efektywności i punktu dopasowania, jednak również w tym przypadku istnieje założenie o zależności rozkładów stóp zwrotu. Ostatni z estymatorów, estymator OGK, jest najbardziej adekwatny do zastosowania w praktyce, jednak w jego przypadku punkt załamania jest jedynie określony przez nierówność.

Jak można zauważyć na przykładach, wszystkie wyliczone estymatory cechuje duża odporność na obserwacje odstające. Dodatkowo w przypadku analizy danych historycznych można zaobserwować różnice w udziałach spółek wchodzących w skład danego portfela. W zależności od wybranego modelu efekt inwestowania może być istotnie różny.

Literatura

Agullo J., *Exact iterative computation of the multivariate minimum volume ellipsoid estimator with a branch and bound algorithm*, Proceedings of the 12th Symposium in Computational Statistics (COMPSTAT 12), 1996.

Butler R.W., Davies P.L., Jhun M., *Asymptotics for the minimum covariance determinant estimator*, „The Annals of Statistics" 1993, 21.

Davies P.L., *The asymptotics of Rousseeuw's minimum volume ellipsoid estimator,* „The Annals of Statistics" 1992, 20.

Fabozzi F.J., Kolm P.N., Pachamanova D., Focardi S., *Robust Portfolio Optimization and Management*, Hoboken, John Wiley & Sons, NJ 2007.

Filzmoser P., Maronna R., Werner M., *Outlier identification in high dimensions*, „Computational Statistics and Data Analysis" 2008 vol. 52.

Gnanadesikan R., Kettenring J.R., *Robust estimates, residuals, and outlier detection with multiresponse data*, Biometrics 1972 28.

Huber P.J., Ronchetti E. M., *Robust Statistics*, Wiley Series in Probability and Statistics. 2. Edition, 2009.

Jajuga K., *Metody ekonometryczne i statystyczne w analizie rynku kapitałowego*, AE, Wrocław 2000.

Kaszuba B., *Odporne metody konstrukcji portfela wieloskładnikowego*, Taksonomia 16, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, UE, Wrocław 2009.

Kent J.T., Tyler D.E., *Regularity and uniqueness for constrained m-estimates and redescending m-estimates*, „The Annals of Statistics" 2001 vol. 29, no 1.

Kent J.T., Tyler D.E., *Constrained M-estimation for multivariate location and scatter*, „The Annals of Statistics" 1996 vol. 24, no 3.

Lopuha H.P., Rousseeuw P.J., *Breakdown points of affine equivariant estimators of multivariate location and covariance matrices*, „The Annals of Statistics" 1991 no 19.

Markowitz H.M., *Mean-variance analysis in portfolio choice and capital markets*, „Journal of Finance" 1952 no 7.
Maronna R., Martin R., Yohai V., *Robust Statistics: Theory and Methods*, John Wiley 2006.
Maronna R., Zamar R., *Robust estimates of location and dispersion for high-dimensional data sets*, „Technometrics" 2002 no 44 (4).
Rocke D.M., *Robustness properties of S-estimators of multivariate location and shape in high dimension*, „Annals of Statistics" 1996 no 24.
Rousseeuw P.J., *Multivariate estimation with high breakdown point*, „Mathematical Statistics and Applications" 1985 vol. B.
Rousseeuw P.J., van Driessen K., *A fast algorithm for the minimum covariance determinant estimator*, Technometrics 1999 vol. 41, no 3.
Rousseeuw P.J., van Zomeren B.C., *Unmasking multivariate outliers and leverage points*, „Journal of the American Statistical Association" 1990 no 85.
Rousseeuw P.J., *Least median of squares regression*, „Journal of the American Statistical Association" 1984 no 79.

ROBUST COVARIANCE MATRIX ESTIMATION IN PORTFOLIO STRUCTURE

Summary: The assumptions of the classical statistical methods used in the portfolio theory are usually not met (e.g. the distribution of rates of return is not a normal distribution), thus it is advisable to seek alternative methods, the assumptions of which would always be met. Robust methods can be such an alternative. The purpose of the article is to present robust covariance matrixes, describe their basic properties and methods of defining. Additionally, the article features an analysis of possibility and legitimacy of applying robust matrixes in practice.