

Daniel KOSIOROWSKI\*

## KRZYWA SKALI – ODPORNA I NIEPARAMETRYCZNA METODA BADANIA ROZRZUTU WEKTORA LOSOWEGO I STOPNIA ZALEŻNOŚCI ROZKŁADÓW BRZEGOWYCH

W pracy pokazano wybrane teoretyczne aspekty budowy i interpretacji krzywej skali oraz pewne propozycje jej modyfikacji dostosowujące ją do zagadnień ekonomicznych. Rozważania teoretyczne zilustrowano za pomocą symulacji obserwacji z wielowymiarowego skośnego rozkładu  $T$  i mieszanin rozkładów oraz za pomocą przykładu empirycznego wielowymiarowego szeregu finansowego

Słowa kluczowe: *statystyczna funkcja głębi, odporność, wielowymiarowa mediana*

### 1. Wprowadzenie

Wskazanie odpornej i zarazem efektywnej alternatywy dla wektora przeciętnych oraz macierzy kowariancji jako estymatorów odpowiednio położenia centrum i rozrzutu wektora losowego należy do najważniejszych celów współczesnej wielowymiarowej analizy statystycznej<sup>1</sup>. Powszechnie wiadomo, że macierz kowariancji z próby jest optymalnym estymatorem rozrzutu w przypadku tzw. wielowymiarowych modeli eliptycznych. Wiadomo również, że jest to estymator skrajnie wrażliwy na obserwacje nietypowe, pojawiające się w próbie np. wskutek zmieszania populacji z innymi zakłócającymi analizę rozkładami. Macierz kowariancji z próby ma nieograniczoną funkcję wpływu Hampela co znaczy, że nie jest odporna na lokalne punktowe zmieszania. Punkt załamania (BP) próby skończonej Donoho i Hubera macierzy kowariancji z  $n$ -elementowej próby wynosi  $1/n$  – zaledwie jedna obserwacja odstająca jest w stanie istotnie zniekształcić ocenę rozrzutu rozpatrywanego wektora. Macierz ko-

---

\* Katedra Statystyki, Uniwersytet Ekonomiczny w Krakowie, ul. Rakowicka 27, 31-510 Kraków, e-mail: dkosioro@ae.krakow.pl

<sup>1</sup> Patrz np. Zuo [15].

wariancji z próby ma nieograniczone maksymalne obciążenie Hubera<sup>2</sup>, czyli nie jest odporna m.in. na błędną specyfikację modelu generującego obserwacje.

Praktyczne wykorzystanie macierzy kowariancji z próby wiąże się z istnieniem momentów drugiego rzędu wektora losowego, reprezentującego badane zjawisko (nie posiada ich np. wielowymiarowy rozkład Cauchy’ego). Interpretacja macierzy kowariancji jest utrudniona w przypadku skośnych rozkładów (np. wielowymiarowego skośnego rozkładu normalnego bądź skośnego  $T$ -Studenta czy Marshalla–Olkina).

W artykule przedstawiono oraz wskazano na użyteczność w ekonomii nieparametrycznej miary rozrzutu wartości wektora losowego wokół jego wielowymiarowej mediany, jaką jest tzw. krzywa skali. Pokazano także blisko związane z krzywą skali narzędzie badania stopnia zależności rozkładów brzegowych wektora, jakim jest tzw. krzywa korelacji. Sprawdzamy wybrane statystyczne własności krzywych skali i korelacji za pomocą symulacji oraz na przykładzie dwuwymiarowego szeregu finansowego, złożonego z procentowych zmian cen akcji spółek Duda i Lotos notowanych na GPW w Warszawie. Zdaniem autora prezentowane w pracy pojęcia oraz wnioski mogą znaleźć zastosowanie m.in. w analizie ryzyka portfeli ubezpieczyciela.

W pracy  $\Delta(S)$  oznacza miarę Lebesgue’a zbioru mierzalnego  $S \subset \mathbb{R}^d$ ,  $\mathbf{X}^n = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$  oznacza  $n$ -elementową próbę z  $\mathbf{X}$ ,  $BP(T, \mathbf{X}^n)$  to punkt załamania Hubera estymatora  $T$  z  $n$ -elementowej próby,  $[x]$  oznacza największą liczbę całkowitą nie większą od  $x$ , przez  $\mathcal{F}$  oznaczamy klasę rozkładów na borelowskim  $\sigma$ -ciele  $\mathbb{R}^d$ .

## 2. Wybrane wiadomości na temat wielowymiarowych kwantyli

Korzystne pod względem odporności własności jednowymiarowych metod statystycznych wykorzystujących statystyki porządkowe sprawiają, że studia nad pojęciem wielowymiarowego kwantyla od wielu lat cieszą się dużym zainteresowaniem wśród osób zajmujących się ekonomią. Zaznaczmy, że trudność definiowania wielowymiarowego kwantyla wiąże się m.in. z nieistnieniem naturalnego porządku w  $\mathbb{R}^d$ ,  $d \geq 2$ , który to porządek stanowi podstawę definicji w przypadku jednowymiarowym.

W obrębie wspomnianych badań można wyróżnić co najmniej trzy dominujące nurty, tzw. **koncepcję głębi danych**<sup>3</sup> (ang. *data depth concept*), **koncepcję głębi**

<sup>2</sup> Rozważamy możliwe rozkłady  $G$  i model mieszaniny  $(1 - \varepsilon)F + \varepsilon G$  dla ustalonego  $\varepsilon > 0$ . W takim przypadku maksymalne obciążenie  $T(\cdot)$  w punkcie  $F$  definiowane jest jako  $B(\varepsilon, T, F) = \sup_G \|T((1 - \varepsilon)F + \varepsilon G) - T(F)\|$ .

$B(\varepsilon, T, F)$  wskazuje najgorsze obciążenie, spowodowane zmieszanym wielkości  $\varepsilon$  założonego rozkładu. Uważa się, że  $T(\cdot)$  jest odporny, jeżeli dla niewielkich  $\varepsilon$  krzywa maksymalnego obciążenia przyjmuje umiarkowane wartości.

<sup>3</sup> Wprowadzenie może stanowić praca Liu i in. [3].

**regresyjnej**<sup>4</sup> (ang. *regression depth*) oraz **koncepcję kwantyli przestrzennych**<sup>5</sup> (ang. *spatial (geometrical) quantile*). W ramach każdego z podejść definiowana jest tzw. wielowymiarowa mediana. Mediany te wprowadzane są na bazie różnych podejść analitycznych, różnią się interpretacją, a jednak wykazują pomiędzy sobą wiele związków. W kontekście dalszych rozważań przedstawimy jedynie kilka faktów dotyczących koncepcji głębi danych, odsyłając zainteresowanego szczegółami czytelnika np. do prac [1], [2], [7].

W ramach koncepcji rozważa się specjalne funkcje, nazywane **statystycznymi funkcjami głębi** (głębiami), które służą porządkowaniu obserwacji generowanych przez wielowymiarowe rozkłady na zasadzie odstawiania od centrów tychże rozkładów. Punkt, dla którego funkcja głębi przyjmuje wartość maksymalną określa się **medianą indukowaną przez** stosownie wybraną **funkcję głębi**.

W niniejszej pracy wykorzystujemy tzw. **funkcję głębi projekcyjnej**<sup>6</sup>, której własności badali m. in. Zuo [14] oraz Zuo i in. [15].

Otóż głębia projekcyjna<sup>7</sup> punktu  $\mathbf{x} \in \mathbb{R}^d$  definiowana jest jako

$$PD(\mathbf{x}, \mathbf{X}^n) = \left[ 1 + \sup_{\|\mathbf{u}\|=1} |\mathbf{u}^T \mathbf{x} - m(\mathbf{u}^T \mathbf{X}^n)| / \sigma(\mathbf{u}^T \mathbf{X}^n) \right]^{-1}, \quad (1)$$

gdzie  $m$  i  $\sigma$ , to miary położenia i rozrzutu w  $\mathbb{R}$ ,  $\mathbf{u}^T \mathbf{X}^n = \{\mathbf{u}^T \mathbf{X}_1, \dots, \mathbf{u}^T \mathbf{X}_n\}$ .

Głębia projekcyjna i indukowane przez nią estymatory położenia centrum oraz rozrzutu wektora losowego odznaczają się bardzo dobrymi własnościami w kategoriach odporności oraz efektywności dla szerokiej klasy populacji. Głębia ta jest afinicznie niezmiennicza.

Najczęściej wykorzystuje się odporne  $m$  i  $\sigma$ , takie jak mediana (Med) bądź zmodyfikowana mediana odchylenia absolutnego od mediany  $MAD_d$ :

$$MAD_d(\mathbf{x}^n) = \text{Med}_d \{|\mathbf{x}_i - \text{Med}(\mathbf{x}^n)|\}, \quad (2)$$

gdzie  $\text{Med}_d = (x_{(\lfloor (n+d)/2 \rfloor)} + x_{(\lfloor (n+d+1)/2 \rfloor)})/2$

oraz  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  oznacza uporządkowany zbiór obserwacji.

Wykorzystując głębnię projekcyjną  $PD(\mathbf{x}, F)$ , definiuje się tzw. **projekcyjne obszary centralne** rzędu  $r$  (w obrębie koncepcji głębi danych ich brzegi określa się mianem  $d$ -wymiarowych kwantyli)

<sup>4</sup> Zagadnienie przedstawiono w pracy Mizera [7].

<sup>5</sup> Patrz: Chaudhuri [1].

<sup>6</sup> Zuo [14] wykazał, że gdy punkty  $\mathbf{X}^n$  są w ogólnej pozycji, tzn. nie więcej niż  $d$  punktów z  $\mathbf{X}^n$  leży w  $d - 1$  wymiarowej hiperpłaszczyźnie, to wtedy  $BP(PM_F, \mathbf{X}^n) = \lfloor (n - d + 1)/2 \rfloor / n$ , gdzie  $PM_F$  oznacza medianę projekcyjną.

<sup>7</sup> W dodatku zamieszczonym na końcu artykułu można znaleźć funkcję środowiska  $R$ , umożliwiającą przybliżone obliczanie głębi projekcyjnej za pomocą algorytmu Dyckerhoffa [2].

$$PC_F(r) = \{\mathbf{x} : PD(\mathbf{x}, F) \geq r\}. \quad (3)$$

Jeśli rozkład  $F$  jest centralnie symetryczny, to obszary centralne odznaczają się taką samą własnością.

### 3. Krzywa skali

Istotne geometryczne własności wielowymiarowego rozkładu prawdopodobieństwa można wyrazić za pomocą stosownie definiowanych konturów. Kontury rozumiane jako poziomy równej gęstości prawdopodobieństwa wykorzystuje się m.in. do przedstawienia lokalnej struktury rozkładu. Jeśli interesuje nas raczej globalna niż lokalna struktura rozkładu, to kontury można zdefiniować jako pewną funkcję, która mierzy odstawanie punktu od stosownie wybranego centrum rozkładu. W takiej sytuacji kontury mogą posłużyć do nieparametrycznego opisu rozrzutu rozkładu poprzez wykorzystanie objętości obszarów ograniczonych przez kontur.

Nawiązując do wprowadzonej powyżej głębi projekcyjnej zauważmy, że rozkład prawdopodobieństwa w  $d$ -wymiarowej przestrzeni można opisać za pomocą projekcyjnych obszarów centralnych. Obszary te tworzą zagnieżdżoną rodzinę zbiorów, które zawierają właściwie zdefiniowane centrum i których wielkość i kształt odzwierciedla rozrzut i stopień zależności rozkładu. Wykorzystując obszary centralne, możemy zaproponować częściowy porządek rodziny rozkładów prawdopodobieństwa  $\mathcal{F}$  na  $\mathbb{R}^d$ ,  $d \geq 2$ , ze względu na stopień rozrzutu wokół mediany projekcyjnej. Rozważmy mianowicie:

**Definicja.** Niech  $PD(\cdot, \cdot)$  będzie funkcjonałem głębi projekcyjnej na  $\mathbb{R}^d \times \mathcal{F}$ . Powiemy, że rozkład  $P$  należący do  $\mathcal{F}$  odznacza się większym rozrzutem niż rozkład  $Q$  w  $\mathcal{F}$ , co oznaczamy  $P \geq_{PD} Q$ , jeżeli

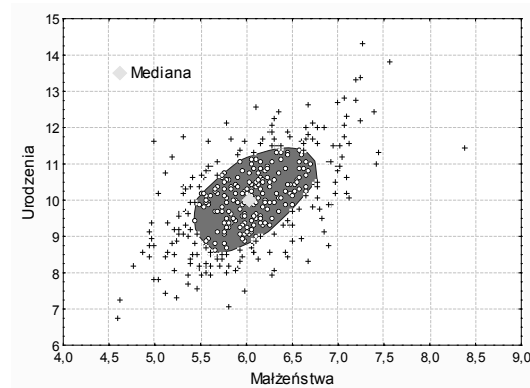
$$\Delta(PC_P(\alpha)) \geq \Delta(PC_Q(\alpha)), \quad \text{dla każdego } \alpha > 0. \quad (4)$$

Innymi słowy rozkład  $P$  odznacza się większym rozrzutem niż  $Q$ , jeżeli dla każdego  $\alpha > 0$ ,  $\alpha$  – centralny obszar projekcyjny opierający się o  $PD(\cdot, P)$  ma większą objętość niż opierający się o  $PD(\cdot, Q)$ .

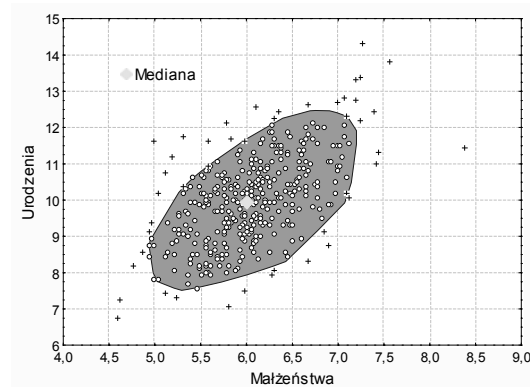
Wykorzystując centralne obszary projekcyjne, możemy zdefiniować tzw. **krzywą skali** będącą rzeczywistym funkcjonałem objętości obszarów centralnych, a służącą do nieparametrycznego opisu rozrzutu wektora losowego wokół wielowymiarowej mediany. **Krzywa skali** jest definiowana jako

$$v_F(r) = \Delta(PC_F(r)), \quad 0 \leq r < 1. \quad (5)$$

Krzywa skali jest dwuwymiarową metodą opisu rozrzutu wartości wektora losowego wokół mediany projekcyjnej. W związku z tym, że projekcyjne obszary centralne stanowią zagnieżdżoną rodzinę zbiorów, krzywa skali służy do pomiaru stopnia ekspansji obszarów centralnych wraz ze wzrastającą masą probabilistyczną w nich zawartą (patrz rys. 1 i 2).



**Rys. 1.** Liczba urodzeń i liczba małżeństw w powiatach RP w roku 2006  
– 50% obszar centralny sporządzony z wykorzystaniem głębi Tukeya  
Źródło: Obliczenia własne, dane GUS.



**Rys. 2.** Liczba urodzeń i liczba małżeństw w powiatach RP w roku 2006  
– 90% obszar centralny sporządzony z wykorzystaniem głębi Tukeya  
Źródło: Obliczenia własne, dane GUS.

Krzywa skali została wprowadzona przez Liu i in. [3]), Zuo i Serfling [13] oraz Serfling [10] przedstawili jej teorię asymptotyczną. Autorzy pokazali jednostajną mocną zbieżność krzywej skali z próby do jej odpowiedniczki w populacji. Serfling

wykazał, że krzywa skali ma strukturę uogólnionej funkcji kwantylowej. Wang i Serfling [11] opisali funkcję wpływu funkcjonału objętości zbiorów indukowanych przez uogólnione funkcje kwantylowe (takimi funkcjami w szczególności są funkcje głębi). Z ich pracy wynika, że funkcja wpływu krzywej skali jest funkcją o dwóch wartościach ze skokiem na brzegu  $p$ -tego obszaru centralnego. Funkcja wpływu przyjmuje ujemne wartości wewnątrz oraz dodatnie na zewnątrz  $p$ -tego obszaru centralnego. Wynika stąd, że krzywa skali odznacza się skończoną czułością na błędy grube i nieskończoną czułością na lokalne przesunięcia. Punktowe zmieszanie populacji powoduje niedoszacowanie bądź przeszacowanie w zależności od tego, czy zmieszanie ma miejsce wewnątrz czy na zewnątrz  $p$ -tego obszaru centralnego. Krzywa skali jako estymator rozrzutu jest afinicznie ekwiwariantna. Właśność ta wynika z własności głębi projekcyjnej.

#### 4. Krzywa korelacji

Krzywą skali można także wykorzystać do opisu stopnia zależności pomiędzy rozkładami brzegowymi wielowymiarowego rozkładu. Pomysł na takie jej zastosowanie opiera się na znanej obserwacji, że rozkład należący do pewnej klasy rozkładów i odznaczający się brakiem zależności pomiędzy rozkładami brzegowymi ma większy rozrzut wokół centrum niż rozkład tej klasy odznaczający się zależnością pomiędzy rozkładami brzegowymi. Dla przykładu, gdy  $F \sim N_p(\mathbf{m}, \Sigma)$ , wtedy odpowiednim rozkładem „braku zależności” jest  $F_0 \sim N_p(\mathbf{m}, \text{diag}\Sigma)$ . Myśl tę wykorzystuje się porządkując rozkłady ze względu na siłę liniowej zależności rozkładów brzegowych z zastosowaniem pewnej funkcji macierzy kowariancji np. wyznacznika.

Nawiązując do wprowadzonych wcześniej pojęć nasuwa się pomysł, aby pomiar zależności pomiędzy rozkładami brzegowymi pewnego rozkładu wykonać na bazie odległości pomiędzy krzywymi skali sporządzonymi dla tegoż rozkładu i dla rozkładu reprezentującego brak zależności. Pomysł ten rozwijali m.in. Mosler [8] i Romanazzi [9].

Niech  $F_0$  będzie „rozkładem niezależności” związanym z danym rozkładem  $F$ . Łatwo zauważyć, że krzywa skali  $F_0$  powinna przebiegać powyżej krzywej skali  $F$ . Można wykorzystać obszar pomiędzy krzywą skali  $F_0$  i krzywą skali  $F$  do pomiaru stopnia zależności rozkładów brzegowych  $F$ .

Romanazzi sugeruje, aby w tym celu wykorzystać znormalizowaną wersję odległości Euklidesa pomiędzy krzywymi skali, nazywaną **krzywą korelacji**:

$$\gamma_1(\alpha, F) = \left( \frac{\int_0^\alpha (\Delta C(t, F_0) - \Delta C(t, F))^2 dt}{\int_0^\alpha (\Delta C(t, F_0) + \Delta C(t, F))^2 dt} \right)^{1/2}, \quad (6)$$

gdzie  $\Delta$  oznacza miarę Lebesgue'a,  $F_0$  jest „rozkładem niezależności”.

Krzywa korelacji wyraża odległość  $F$  od rozkładu niezależności  $F_0$  dla  $0 \leq \alpha \leq 1$ . Aby przedstawić ją graficznie, sporządzamy diagram rozrzutu  $\gamma_1(\alpha, F)$  vs.  $\alpha$ .

Mosler proponuje natomiast wykorzystać do pomiaru stopnia zależności rozkładów brzegowych iloraz

$$GS_\alpha^*(F) = \frac{\Delta D_\alpha^*(F_{ind})}{\Delta D_\alpha^*(F)} \quad (7)$$

dla pewnego ustalonego  $\alpha^*$  oraz zagregowaną wersję miary (7):

$$\int_0^1 \alpha^d \Delta D_\alpha(F) d\alpha, \quad (8)$$

gdzie każdy  $\alpha$ -obszar centralny  $D\alpha$  otrzymuje wagę  $\alpha^d$ .

Zasadniczy problem związany z wykorzystaniem krzywej korelacji w praktyce wiąże się z tym, że na ogół nie jesteśmy w stanie wskazać „rozkładu niezależności”, gdyż nie znamy klasy rozkładów, do której należy interesujący nas rozkład. Jedyną informacją jaką dysponujemy stanowi próba. Okazuje się, że także w takiej sytuacji możemy z powodzeniem wykorzystywać krzywą korelacji. Romanazzi [9] przedstawia interesujący sposób postępowania w takich sytuacjach, wywodzący się z zasady randomizacji.

Przypuśćmy, że  $\mathbf{X}$  jest  $n \times p$  macierzą losową, której wiersze  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$  są obserwacjami z  $n$ -elementowej próby losowej z  $p$ -wymiarowego rozkładu  $F$  oraz rozważmy odwzorowanie  $\mathbf{X} \rightarrow \tau \mathbf{X}$ , gdzie  $\tau$  jest przekształceniem, które zamienia każdą kolumnę  $\mathbf{X}^{(j)} = (X_{1j}, \dots, X_{nj})^T$  macierzy  $\mathbf{X}$  permutacją jej składowych. Zauważmy, że jeśli kolumny  $\mathbf{X}$  mają różne elementy, to wtedy jest  $(n!)^p$  takich przekształceń. Niech  $\hat{F}_{n,\tau}$  będzie rozkładem  $\tau X$  oraz niech  $\chi$  będzie klasą takich rozkładów. Romanazzi dowodzi twierdzenia głoszącego, że warunkując obserwowaną próbą, stosownym „rozkładem niezależności” dla  $F$  jest mieszanina

$$\hat{F}_{n,0} = \frac{1}{(n!)^p} \sum_{\tau \in \chi} \hat{F}_{n,\tau}. \quad (9)$$

Aproksymację  $\tilde{F}_{n,0}$  uzyskuje się biorąc losową próbę  $m$  rozkładów z  $\chi$ .

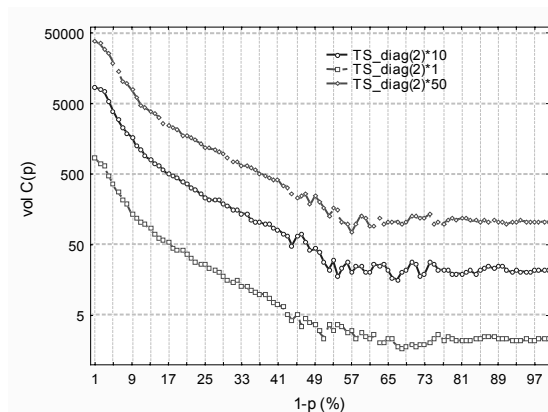
Krzywa korelacji sporządzona z wykorzystaniem głębi projekcyjnej jest afinicznie ekwiwariantna. Romanazzi podkreśla, że krzywa korelacji, podobnie jak krzywa skali, musi być uważnie definiowana w przypadkach, gdy  $F$  jest skoncentrowane z prawdopodobieństwem 1 na pewnej  $q$ -wymiarowej podprzestrzeni  $\mathbb{R}^p$ , z  $q < p$ . Gdy  $F$  jest wielowymiarowym rozkładem normalnym, wtedy krzywa korelacji jest funkcją stałą. Można pokazać, że dla  $F \sim N_p(\mathbf{m}, \Sigma)$ , gdzie  $\Sigma$  jest dodatnio określoną macierzą wymiaru  $p \times p$  dla  $0 < \alpha < 1$  zachodzi

$$\gamma_1(\alpha, F) = (1 - (\det R)^{1/2}) / (1 + (\det R)^{1/2}), \quad (10)$$

gdzie  $R = (\text{diag}\Sigma)^{-1/2}\Sigma(\text{diag}\Sigma)^{-1/2}$  jest macierzą korelacji  $F$ .

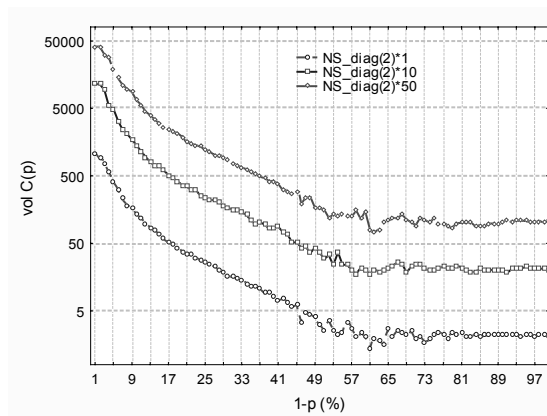
## 5. Wyniki badań symulacyjnych

W celu sprawdzenia wybranych statystycznych własności krzywych skali i korelacji z próby przeprowadzono badania symulacyjne. Generowano mianowicie po 500 prób, złożonych ze 100 obserwacji pochodzących z dwuwymiarowych rozkładów skośnych normalnych, skośnych  $T$ -Studenta, Marshalla–Olkina oraz mieszanin tychże rozkładów. Eksperymenty powtarzano kilkadziesiąt razy dla sprawdzenia stabilności oszacowań.



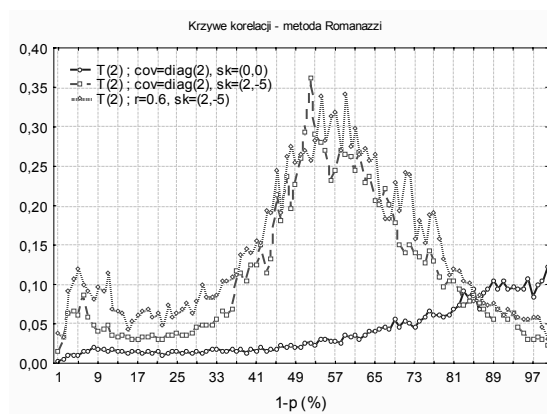
**Rys. 3.** Krzywe skali – dwuwymiarowe skośne rozkłady  $T$ -Studenta o dwóch stopniach swobody  
Źródło: Obliczenia własne.





**Rys. 4.** Krzywe skali – dwuwymiarowe skośne rozkłady normalne  
 Źródło: Obliczenia własne.

Na rysunkach 3 i 4 zamieszczono oszacowania krzywych skali odpowiednio dla prób pochodzących ze skośnego rozkładu  $T$ -Studenta o dwóch stopniach swobody i skośnego rozkładu normalnego, o takich samych parametrach: położenia  $\mathbf{m} = (0,0)$ , rozrzutu  $\mathbf{\Omega}_1 = \text{diag}(2) \cdot 1$ ,  $\mathbf{\Omega}_2 = \text{diag}(2) \cdot 5$ ,  $\mathbf{\Omega}_3 = \text{diag}(2) \cdot 50$ , kształtu  $\boldsymbol{\alpha} = (2, -5)$ .



**Rys. 5.** Krzywe korelacji – dwuwymiarowe skośne rozkłady  $T$ -Studenta o dwóch stopniach swobody  
 Źródło: Obliczenia własne.

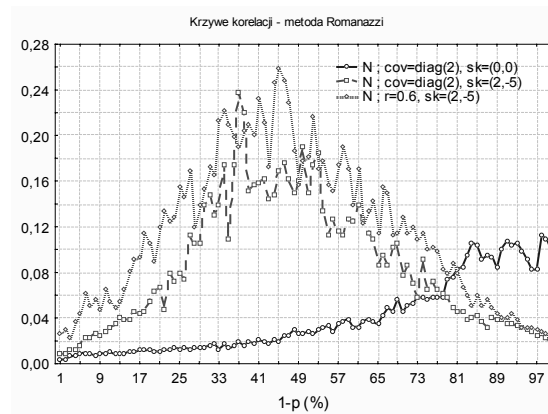
Na rysunkach 5 i 6 przedstawiono oszacowania krzywych korelacji proponowanych przez Romanazzi’ego dla prób pochodzących z rozkładów  $T$ -Studenta o dwóch stopniach swobody i skośnego rozkładu normalnego o parametrach:

- a) położenia  $\mathbf{m} = (0,0)$ , rozrzutu  $\mathbf{\Omega} = \text{diag}(2) \cdot 1$ , kształtu  $\boldsymbol{\alpha} = (0, 0)$ ,

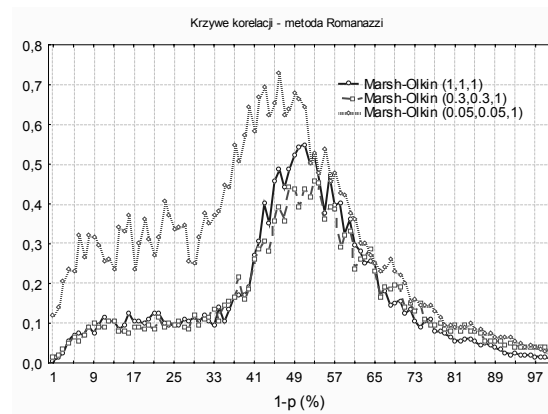
b) położenia  $\mathbf{m} = (0,0)$ , rozrzutu  $\mathbf{\Omega} = \text{diag}(2) \cdot 1$ , kształtu  $\boldsymbol{\alpha} = (2, -5)$ ,

c) położenia  $\mathbf{m} = (0,0)$ , rozrzutu  $\mathbf{\Omega}[2,1] = \mathbf{\Omega}[1,2] = 0,6$ ;  $\mathbf{\Omega}[1,1] = \mathbf{\Omega}[2,2] = 1$ , kształtu  $\boldsymbol{\alpha} = (2, -5)$

Na rysunku 7 przedstawiono krzywe korelacji dla prób pochodzących z dwuwymiarowych rozkładów Marshalla–Olkina o parametrach  $\boldsymbol{\lambda}_1 = (1,1,1)$  ( $\rho \approx 0.33$ ),  $\boldsymbol{\lambda}_1 = (0.3,0.3,1)$  ( $\rho \approx 0.62$ ),  $\boldsymbol{\lambda}_1 = (0.05,0.05,1)$  ( $\rho \approx 0.9$ ).



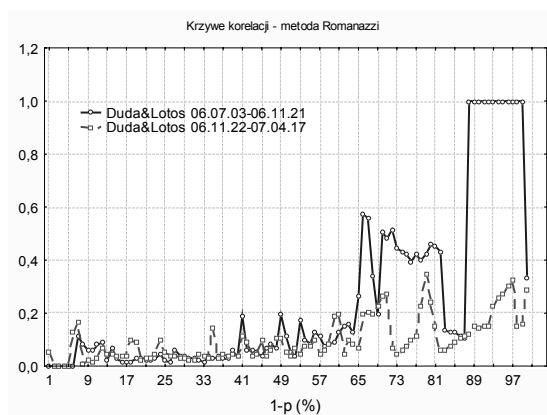
**Rys. 6.** Krzywe korelacji – dwuwymiarowe skośne rozkłady normalne  
Źródło: Obliczenia własne.



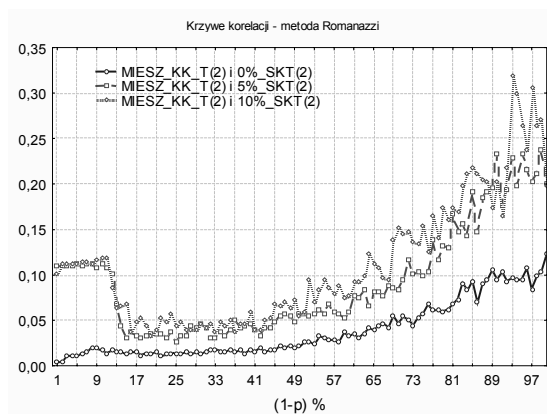
**Rys. 7.** Krzywe korelacji – dwuwymiarowe rozkłady Marshalla–Olkina,  
 $r = 0,33$ ;  $r = 0,62$ ;  $r = 0,9$   
Źródło: Obliczenia własne.

W celu sprawdzenia własności proponowanego funkcjonału na danych empirycznych analizowano dwuwymiarowy szereg finansowy, złożony z obserwacji procento-

wych przyrostów cen akcji spółek Duda i Lotos, notowanych na GPW w Warszawie w okresie 2006.07.03–2007.04.17 (rys. 8). Rozpatrywany okres umownie podzielono na dwa podokresy, obejmujące po sto notowań spółek. Na rysunku 9 zamieszczono krzywe korelacji obliczone dla dwóch następujących po sobie podokresów. Rysunek 9 przedstawia oszacowania krzywych korelacji dwuwymiarowych 0%, 5%, 10% mieszanin rozkładów  $T(df=2)$  i skośnego  $T(df=2)$  o parametrach położenia odpowiednio  $\mathbf{c} = (0,0)$  i  $\mathbf{c} = (100,100)$ , kształtu  $\boldsymbol{\alpha} = (0,0)$  i  $\boldsymbol{\alpha} = (2,-5)$  oraz identycznej charakterystyce rozrzutu  $\boldsymbol{\Omega}_2 = \text{diag}(2) \cdot \mathbf{1}$ . Na rysunku 10 przedstawiono krzywe skali dla tych samych co na rysunku 9 populacji.

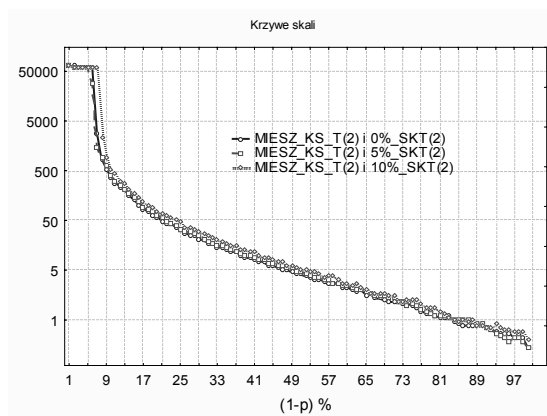


**Rys. 8.** Krzywe korelacji – Duda & Lotos, przyrosty cen akcji (%) w okresie 2006.07.03–2007.04.17  
 Źródło: Obliczenia własne.



**Rys. 9.** Krzywe korelacji – 0%, 5%, 10% mieszaniny rozkładów  $T(df=2)$  i skośnego  $T(df=2)$   
 Źródło: Obliczenia własne.

Z rysunków 3 i 4 wynika, że krzywe skali dobrze dyskryminują rozkłady skośne normalne i skośne  $T$ -Studenta, różniące się charakterystykami rozrzutu. Zaznaczmy, że dla wykorzystywanego w symulacjach skośnego rozkładu  $T$  o dwóch stopniach swobody nie istnieje macierz kowariancji. Rysunki 5 i 6 pokazują, że krzywe korelacji sporządzone dla izotropowych skośnych rozkładów normalnych i  $T$  właściwie „wychwytują” wpływ skośności na brak niezależności rozkładów brzegowych.



**Rys. 10.** Krzywe skali – 0%, 5%, 10% mieszaniny rozkładów  $T(df=2)$  i skośnego  $T(df=2)$

Źródło: Obliczenia własne.

Rysunek 7 pokazuje, że krzywa korelacji dobrze się sprawuje w przypadku rozkładu nienależącego do rodziny wykładniczej. Należy podkreślić, że z krzywej korelacji można odczytać, jak przedstawia się struktura zależności rozkładu w zależności od bliskości centrum rozkładu. Przykładowo z rysunku 8 odczytujemy, że zależność pomiędzy przyrostami cen akcji spółek Duda & Lotos manifestuje się mocniej bliżej centrum, słabiej w obszarach od centrum znacznie oddalonych, inaczej niż w przypadku rozkładu Marshalla–Olkina. Z rysunku 9 wynika, że krzywa korelacji jest wrażliwa na zmieszanie populacji, co w zależności od punktu widzenia można po czytać za jej wadę bądź zaletę. Z rysunku 10 wynika, że krzywa skali jest względnie niewrażliwa na zmieszania populacji.

## Podsumowanie

Analiza rozrzutu wektora losowego reprezentującego zjawisko ekonomiczne oraz pomiar stopnia zależności jego rozkładów brzegowych stanowią istotne części procesu weryfikacji teorii ekonomicznej, modelowania ryzyka aktuarialnego bądź finansowego.

Przedstawiane w pracy krzywe skali oraz korelacji stanowią, zdaniem autora, interesujące alternatywy dla klasycznych metod analizy rozrzutu i stopnia zależności rozkładów brzegowych, zwłaszcza w przypadku niewielkiej wiedzy *a priori* na temat badanego zjawiska. Na uwagę zasługują własności metod polegające na tym, że krzywe pokazują struktury rozrzutu i korelacji rozkładu w powiązaniu z miarą odstawania od centrum rozkładu. Wypada podkreślić, że zarówno krzywa skali, jak i krzywa korelacji mogą zostać sporządzone z wykorzystaniem innych funkcji głębi (np. głębi Tukey'a czy symplecjoidalnej) bądź z wykorzystaniem kwantyli przestrzennych. Z badań autora wynika jednak, że właśnie głębia projekcyjna odznacza się najlepszymi w sensie odporności i efektywności własnościami.

### Bibliografia

- [1] CHAUDHURI P., *On a Geometric Notion of Quantiles for Multivariate Data*, Journal of the American Statistical Association, 1996, 91, 862–872.
- [2] DYCKERHOFF R., *Data Depths Satisfying the Projection Property*, Allgemeines Statistisches Archiv., 2004, 88, 163–190.
- [3] LIU R.Y., PARELIUS J.M., SINGH K., *Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference (with discussion)*, The Annals of Statistics, 1999, 27, 783–858.
- [4] KOSIOROWSKI D., *Nonparametric Equity of Two Shapes Test Based on Multivariate Quantile Functional*, Bulletin of the International Statistical Institute, 56th Session of the ISI, 2007.
- [5] KOSIOROWSKI D., *O kwantylowym funkcjonałe asymetrii rozkładu wektora losowego w badaniach szeregów finansowych [w:] Dynamiczne modele ekonometryczne*, Z. Zieliński (red.), Wydawnictwo UMK w Toruniu, Toruń, 2007, 129–136.
- [6] KRZYŚKO M., *Statystyka matematyczna*, Wydawnictwo Naukowe UAM, Poznań 2004.
- [7] MIZERA I., *On Depth and Deep Points: A Calculus*, The Annals of Statistics, 2002, 30, 1681–1736.
- [8] MOSLER K., *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach*, Springer, New York 2002.
- [9] ROMANAZZI M., *Data Depth and Correlation*, Allgemeines Statistisches Archiv., 2004, 88, 191–214.
- [10] SERFLING R.J., *Nonparametric Multivariate Descriptive Measures Based on Spatial Quantiles*, Journal of Statistical Planning and Inference, 2004, 123, 259–278.
- [11] WANG J., SERFLING R., *Influence Functions for a General Class of Depth – Based Generalized Quantile Functions*, Journal of Multivariate Analysis, 2006, 97, 810–826.
- [12] ZUO Y., CUI H., YOUNG D., *Influence Function and Maximum Bias of Projection Depth Based Estimators*, The Annals of Statistics, 2004, 32(1), 189–218.
- [13] ZUO Y., SERFLING R., *Nonparametric Notions of Multivariate “Scatter Measure” and “More Scattered” Based on Statistical Depth Function*, Journal of Multivariate Analysis, 2000, 75, 62–78.
- [14] ZUO Y., *Projection Based Depth Functions and Associated Medians*, The Annals of Statistics, 2003, 31 (5), 1460–1490.
- [15] ZUO Y., *Robust Location and Scatter Estimators in Multivariate Analysis* (Invited book chapter to honor Peter Bickel on his 65th Birthday), *The Frontiers in Statistics*, Imperial College Press 2005.

## Dodatek

### (przybliżony algorytm umożliwiający obliczanie funkcji głębi projekcyjnej z próby)

Należy wgrać pakiet {MASS}.

## argumentem funkcji jest macierz danych wymiaru  $n*k$ , funkcja zwraca wartości głębi dla każdej obserwacji – wiersza macierzy danych##

```
PROJEKCYJNAT<-function(A){mi<-c(seq(length = ncol(A),from=0,to=0))
sigma<-diag(ncol(A))
proj<-mvrnorm(n=1000,mi,sigma)
u<-t(proj)
f<-c(length=1000)
for (i in 1: 1000) f[i]<-sqrt(sum(u[,i]^2))
f
us<-matrix(nrow=nrow(u),ncol=ncol(u))
for (i in 1: 1000) us[,i]<-u[,i]/f[i]
us
xut<-A%*%us
MED<-c(length=1000)
for(i in 1: 1000) MED[i]<-median(xut[,i])
MAD<-c(length=1000)
for(i in 1:1000) MAD[i]<-median(abs(xut[,i]-MED[i]))
OD<-matrix(nrow=nrow(xut),ncol=ncol(xut))
for(i in 1:1000) OD[,i]<-(1/(1+(abs(xut[,i]-MED[i])/MAD[i])))
PD<-c(length=nrow(OD))
for(i in 1:nrow(OD)) PD[i]<-min(OD[i,])
PD
}
```

### Scale curve – a robust and nonparametric approach to study a dispersion and interdependence of multivariate distributions

A scale curve is a nonparametric approach to study a dispersion of a random vector around a multivariate median. The scale curve is a volume functional based on probabilities allocated on the so-called central regions induced by a given statistical depth function. The curve expresses a degree of dispersion of random vector in a central regions expansion categories. We can also use the curve to display a degree of interdependence of marginal components of a specific distribution.

In this paper we discuss selected theoretical aspects of the scale curve induced by a projection depth function. We study the performance of the propositions on various multivariate data sets simulated from skewed, fat tailed and including outliers distributions.

Keywords: *statistical depth function, robustness, multivariate median*