FOLIA OECONOMICA CRACOVIENSIA
Vol. XLIX–L (2008–2009) PL ISSN 0071-674X

WYBRANE ZAGADNIENIA KONCEPCJI GŁĘBI DANYCH

DANIEL KOSIOROWSKI
Katedra Statystyki
Uniwersytet Ekonomiczny w Krakowie
PL 31-510 Kraków, ul. Rakowicka 27
e-mail: *daniel.kosiorowski@uek.krakow.pl*

Praca przedstawiona przez Autora na posiedzeniu Komisji Nauk Ekonomicznych Oddziału PAN w Krakowie w dniu 4 grudnia 2007 r.

ABSTRACT

D. Kosiorowski. *Selected Issues of Data Depth Concept.* Folia Oeconomica Cracoviensia 2008–2009, 49–50: 5–30.

In this paper we present selected aspects of data depth concept. We propose several statistical procedures based on data depth concept. We study a performance of the propositions on various multivariate data sets simulated from skewed, fat tailed distributions and mixtures of them.

KEY WORDS — SŁOWA KLUCZOWE

robust statistical procedure, statistical depth function
odporna procedura statystyczna, statystyczna funkcja głębi

1. WPROWADZENIE

Koncepcja głębi danych jest jednym z trzech dominujących nurtów badań nad pojęciem i zastosowaniami wielowymiarowego kwantyla (patrz np. Liu i in., 1999).

Wykazuje szereg związków z tzw. koncepcją głębi regresyjnej (ang.: *regression depth*) (np. Mizera, 2002) oraz staje się coraz atrakcyjniejszą alternatywą dla historycznie pierwszej koncepcji kwantyli przestrzennych (ang. *spatial (geometrical) quantile*) (np. Chaudhuri, 1996). Kwantyle definiowane w obrębie trzech podejść różnią się interpretacją.

Narzędzia oferowane przez koncepcję głębi danych wykorzystywane są w zagadnieniach nieparametrycznego wnioskowania statystycznego wykorzystującego pojęcia rangi, statystyki porządkowej, kwantyla itp. Dodajmy, że w obrębie tzw. klasycznego podejścia do nieparametrycznego wnioskowania w R^d, $d > 1$, wykorzystuje się wektory jednowymiarowych statystyk. Postępowanie takie nie uwzględnia często szczególnie istotnej geometrii wielowymiarowego zbioru danych. Na przykład, może się zdarzyć, że wektor jednowymiarowych median leży poza powłoką wypukłą wielowymiarowego zbioru danych — nie może zatem być dobrą miarą położenia centrum.

Charakterystyczną dla koncepcji miarę centralności punktu $\mathbf{x} \in R^d$, $d > 1$, będącego realizacją pewnego d-wymiarowego wektora losowego $\mathbf{X}$ o rozkładzie prawdopodobieństwa P, wprowadza się za pomocą specjalnej funkcji nazywanej g ł ę b i ą (ang. *depth*) bądź f u n k c j ą g ł ę b i (ang. *depth function*).

Funkcja głębi przyporządkowuje każdemu punktowi liczbę rzeczywistą z przedziału [0, 1], będącą miarą jego centralności, zważywszy na rozkład go generujący.

Wykorzystując funkcję głębi można uporządkować zbiór wielowymiarowych obserwacji na zasadzie odstawania obserwacji od centrum.

Zazwyczaj punkt, dla którego funkcja głębi przyjmuje wartość maksymalną określa się mianem d-w y m i a r o w e j m e d i a n y.

Oznaczamy przez P rodzinę rozkładów prawdopodobieństwa określonych na σ ciele zbiorów borelowskich w R^d oraz przez P_X rozkład danego wektora losowego $\mathbf{X}$.

Każdy element próby $\mathbf{X}^n = \{X_1, ..., X_n\}$ traktujemy jako $d \times 1$ wektor kolumnowy. Zakładamy ponadto, że rozkład P_X jest absolutnie ciągły.

W literaturze przedstawiono kilka układów postulatów, które powinna spełniać funkcja głębi, aby była odpowiednim narzędziem służącym do budowy nieparametrycznych wielowymiarowych procedur statystycznych (zob. Dyckerhoff, 2004; Zuo, i Serfling, 2000. Poniżej przedstawiamy najczęściej stosowany układ postulatów, który spełnia większość znanych funkcji głębi $D(\mathbf{x}, P)$.

1. Niezmienniczość afiniczna — $D(\mathbf{x}, P)$ jest niezależna od wyboru układu współrzędnych.

2. Wartość maksymalna w centrum — jeżeli rozkład P jest symetryczny względem Θ w pewnym sensie, wówczas $D(\mathbf{x}, P)$ przyjmuje w tym punkcie maksimum.

3. Symetria — jeżeli rozkład P jest symetryczny względem Θ w pewnym sensie, wtedy także $D(\mathbf{x}, P)$ jest symetryczna w tym sensie.

4. Zmniejszanie się wartości wzdłuż promieni — wartość funkcji głębi $D(\mathbf{x}, P)$ zmniejsza się wzdłuż promienia mającego początek w punkcie o maksymalnej głębi.

5. Zanikanie w nieskończoności — $D(\mathbf{x}, P) \to 0$ gdy $\|\mathbf{x}\| \to \infty$.

6. Ciągłość $D(\mathbf{x}, P)$ jako funkcji $\mathbf{x}$.

7. Ciągłość $D(\mathbf{x}, P)$ rozpatrywanej jako funkcjonał P.

8. Quasi-wypukłość $D(\mathbf{x}, P)$ rozpatrywanej jako funkcja $\mathbf{x}$ — zbiór $\{x : D(\mathbf{x}, P) \geq \alpha$ jest wypukły dla każdego $\alpha \in [0,1]$.

Zaznaczmy, że przez centrum rozumiemy punkt symetrii, przez symetrię rozumiemy centralną symetrię.

Zbiór

$$\{\mathbf{x} \in R^d : D(\mathbf{x}, P) = \alpha\}, \tag{1}$$

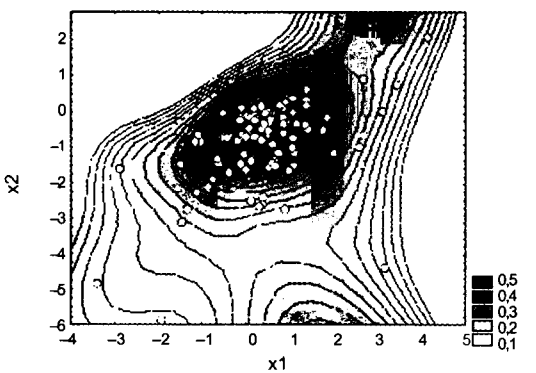
nazywany poziomem bądź konturem głębi α, zbiór ten określany jest mianem d-wymiarowego kwantyla rzędu α, $\alpha \in [0, 1]$.

Zbiór

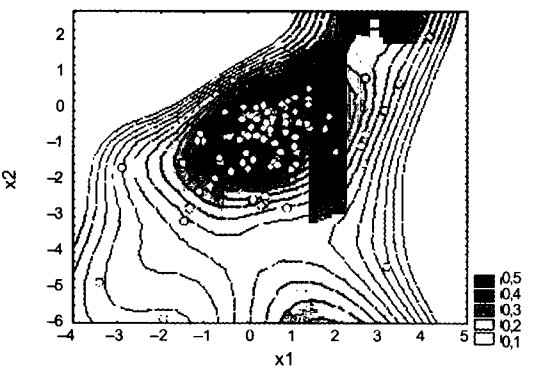
$$D_\alpha(X) = \{\mathbf{x} \in R^d : D(\mathbf{x}, P) \geq \alpha\}, \tag{2}$$

tzn. obszar ograniczony przez kontur głębi α, nazywany jest α przyciętym (centralnym) obszarem $\alpha \in [0,1]$.

Opierając się na danej funkcji głębi, możemy zdefiniować punkt o maksymalnej głębi jako wielowymiarowy analog jednowymiarowej mediany.



Ryc. 1. Projekcyjne obszary centralne ze 100 elementowej próby skośnego rozkładu normalnego z $\Omega = diag(2) \cdot 5$, $\mathbf{m} = (0, 0)$, $\alpha = (2, -5)$



Ryc. 2. Projekcyjne obszary centralne ze 100 elementowej próby z rozkładu Marshalla-Olkina $\lambda = (1, 1, 1)$

Niech będzie funkcją głębi, wtedy medianę indukowana przez definiujemy:

$$M(P) = \operatorname{argsup}_{\mathbf{x}\in\mathbb{R}^d} D(\mathbf{x}, P). \tag{3}$$

Przykłady statystycznych funkcji głębi

Symetryczna głębia projekcyjna punktu $\mathbf{x} \in \mathbb{R}^d$, będącego realizacją pewnego d-wymiarowego wektora losowego $\mathbf{X}$ o rozkładzie prawdopodobieństwa F, $PD(\mathbf{x}, F)$ definiowana jest jako:

$$PD(\mathbf{x}, F) = \left(1 + \sup_{\|\mathbf{u}\|=1} \frac{|Med(\mathbf{u}^t\mathbf{X})|}{MAD(\mathbf{u}^t\mathbf{X})}\right)^{-1}, \tag{4}$$

gdzie $\mathbf{X}$ ma rozkład prawdopodobieństwa F, Med oznacza jednowymiarową medianę, oraz MAD oznacza jednowymiarową medianę odchylenia absolutnego od mediany $MAD(Z) = Med(|Z - Med(Z)|)$.

Głębia projekcyjna oraz 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.

Głębia Tukey'a (głębia domkniętej półprzestrzeni) definiowana jest:

$$D(\mathbf{x}, P) = \inf_H \{P(H) : \mathbf{x} \in H, H \text{ jest domkniętą półprzestrzenią}\}. \tag{5}$$

Głębia punktu $\mathbf{x}$ — najmniejsza masa probabilistyczna znajdująca się na domkniętej półprzestrzeni z punktem $\mathbf{x}$ na brzegu.

Najprostszym przykładem głębi jest tzw. głębia Euklidesa:

$$D_{EUK}(\mathbf{z}\,|\,\mathbf{X}^n) = \frac{1}{1 + \|\mathbf{z} - \bar{\mathbf{x}}\|^2}, \quad \text{gdzie } \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i. \tag{6}$$

Znaną głębię stanowi tzw. głębia Mahalanobisa:

$$D_{MAH}(\mathbf{y}\,|\,\mathbf{X}^n) = \frac{1}{1 + (\mathbf{y} - \bar{\mathbf{x}})^t\mathbf{S}^{-1}(\mathbf{y} - \bar{\mathbf{x}})}, \tag{7}$$

gdzie $\mathbf{S}$ oznacza macierz kowariancji próby $\mathbf{X}^n$.

Pojęcie głębi dopasowania funkcji regresji liniowej zostało wprowadzone przez Rousseeuw i Hubert (Rousseeuw i Hubert, 1998).

I. Mizera (2000) uogólnił ich podejście oraz pokazał je w ramach formalizmu optymalizacji wektorowej uzyskując jednocześnie szereg wyników dotyczących odporności estymatorów maksymalnej głębi regresyjnej.

Przypuśćmy, że zamierzamy dopasować liniową funkcję regresji $y = b_1x + b_2$ do dwuwymiarowego zbioru danych $\mathbf{Z}^n = \{(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\}$.

Oznaczmy współczynniki regresji jako $\mathbf{b} = (b_1, b_2)$. Oznaczmy reszty regresji jako:

$$r_i(\mathbf{b}) = r_i = y_i - b_1 x_1 - b_2.$$

DEFINICJA 1. Powiemy, że dopasowanie $\mathbf{b} = (b_1, b_2)$ nie jest słabo optymalne względem zbioru danych $\mathbf{Z}^n$ jeżeli istnieje liczba rzeczywista $v_b = v$, która nie pokrywa się z żadnym punktem x_i i dla której zachodzi:

$$r_i(\mathbf{b}) < 0 \text{ dla wszystkich } x_i < v \text{ i } r_i(\mathbf{b}) > 0 \text{ dla wszystkich } x_i > v$$

lub:

$$r_i(\mathbf{b}) > 0 \text{ dla wszystkich } x_i < v \text{ i } r_i(\mathbf{b}) < 0 \text{ dla wszystkich } x_i > v$$

(istnienie liczby v odpowiada istnieniu punktu, dookoła którego możemy obracać prostą do pozycji pionowej, nie napotykając żadnej obserwacji).

DEFINICJA 2. Głębia regresyjna $rdepth(\mathbf{b}, \mathbf{Z}^n)$ jest najmniejszą frakcją obserwacji próby $\mathbf{Z}^n$, którą należy usunąć aby dopasowanie $\mathbf{b}$ przestało być słabo optymalne.

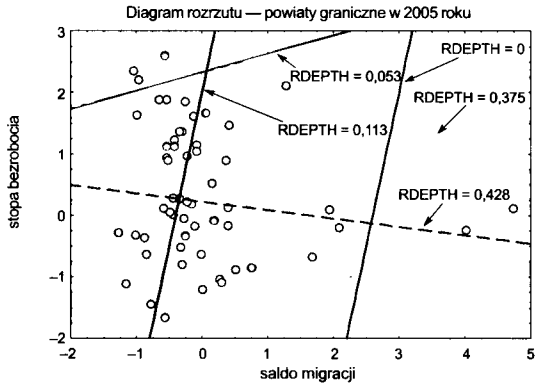
Estymator maksymalnej głębi regresyjnej definiujemy jako:

$$T_r^*(\mathbf{Z}^n) = \arg\max_{\mathbf{b}} rdepth(\mathbf{b}, \mathbf{Z}^n) = \arg\max_{\mathbf{b}^{ij}} rdepth(\mathbf{b}^{ij}, \mathbf{Z}^n). \quad (8)$$

Warto zauważyć, że w przypadku koncepcji głębi regresyjnej nie czyni się jakichkolwiek założeń odnośnie regularności składnika losowego, wpływającego na wartości obserwacji.

Dopasowanie największej głębi jest prostą najlepiej równoważącą chmurę danych.

Zaproponowane przez Mizerę ogólniejsze spojrzenie na koncepcję głębi regresyjnej z jednej strony pokazuje jej miejsce w ramach całej koncepcji głębi danych, z drugiej strony — w wielu przypadkach prowadzi do prostszych metod badania własności estymatorów maksymalnej głębi.



Ryc. 3. Stopa bezrobocia vs. saldo migracji w polskich powiatach granicznych w 2005 roku. Głębie dopasowań funkcji regresji do danych empirycznych. Linia przerywana reprezentuje estymator maksymalnej głębi, linia oznaczona kropkami reprezentuje estymator metody heteroskedastycznej regresji t-Studenta

2. KRZYWA SKALI — ODPORNA I NIEPARAMETRYCZNA METODA BADANIA ROZRZUTU WEKTORA LOSOWEGO I STOPNIA ZALEŻNOŚCI ROZKŁADÓW BRZEGOWYCH

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. Macierz kowariancji z próby ma nieograniczona 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 kowariancji z próby ma nieograniczone maksymalne obciążenie Hubera, czyli nie jest odporna m.in. na błędną specyfikację modelu generującego obserwacje. Praktyczne wykorzystanie macierzy kowariancji z próby wiąże się z istnieniem momentów drugiego stopnia wektora losowego, reprezentującego badane zjawisko. Interpretacja macierzy kowariancji jest utrudniona w przypadku skośnych populacji.

Wykorzystując głębię 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):

$$PC_F(r) = \{\mathbf{x} : PD(\mathbf{x}, F) \geq r\}. \tag{9}$$

W przypadku, gdy rozkład F jest centralnie symetryczny, obszary centralne odznaczają się taką samą własnością.

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

Krzywa skali definiowana jest jako:

$$v_F(r) = \Delta(PC_F(r)),\ 0 \leq r < 1, \tag{10}$$

gdzie: $\Delta(\cdot)$ oznacza miarę Lebesgue'a a $PC_F(\cdot)$ — projekcyjny obszar centralny.

Krzywa skali jest dwuwymiarową metodą opisu rozrzutu wartości wektora losowego wokół mediany projekcyjnej. W związku z faktem, 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.

Niech F_0 będzie „rozkładem niezależności"[1] związanym z danym rozkładem F. Łatwo zauważyć, że krzywa skali F_0 powinna przebiegać powyżej krzy-

[1] Polski odpowiednik niewątpliwie wymaga dopracowania.

wej skali F. Można wykorzystać obszar pomiędzy krzywą skali F_0 i krzywą skali F do pomiaru stopnia zależności rozkładów brzegowych F.

M. Romanazzi (2004) sugeruje, aby w tym celu wykorzystać znormalizowaną wersję odległości Euklidesa pomiędzy krzywymi skali nazywaną krzywą korelacji:

$$\gamma_1(\alpha, F) = \left(\frac{\int_0^{\alpha} (\Delta C(t, F_0) - \Delta C(t, F))^2 dt}{\int_0^{\alpha} (\Delta C(t, F_0) - \Delta C(t, F))^2 dt} \right)^{1/2}, \tag{11}$$

gdzie:
$\Delta(\cdot)$ oznacza miarę Lebesque'a, F_0 jest „rozkładem niezależności".

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

Najczęściej nie jesteśmy w stanie wskazać „rozkładu niezależności", gdyż nie znamy klasy rozkładów, do której należy rozkład generujący dane. Okazuje się, że nawet w takich sytuacjach można z powodzeniem wykorzystywać krzywą korelacji.

Przypuśćmy, że $\mathbf{X}$ jest $n \times p$ macierzą losową, której wiersze $\mathbf{X}_i = (X_{i1}, ..., X_{ip})$ są obserwacjami z n-elementowej próby losowej z p-wymiarowego rozkładu F oraz rozważmy odwzorowanie $\mathbf{X} \to \tau\mathbf{X}$, gdzie τ jest przekształceniem, które zamienia każdą kolumnę $\mathbf{X}^{(j)} = (X_{1j}, ..., X_{nj})^T$ macierzy $\mathbf{X}$ permutacją jej składowych. Zauważmy, że jeśli kolumny mają różne elementy, wtedy jest takich przekształceń.

Niech $\hat{F}_{n\tau}$ będzie rozkładem τX oraz niech X będzie klasą takich rozkładów. Romanazzi (2004) dowodzi twierdzenia głoszącego, że warunkując obserwowaną próbą stosownym „rozkładem niezależności" dla F jest mieszanina:

$$\hat{F}_{n,0} = \frac{1}{(n!)^p} \sum_{\tau \in C} \hat{F}_{n,\tau}. \tag{12}$$

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

Można pokazać, że dla $F \sim N_p(\mathbf{m}, \Sigma)$, gdzie Σ jest dodatnio określoną macierzą wymiaru $p \times p$ dla $0 < \alpha < 1$ zachodzi:

$$\gamma_1(\alpha, F) = (1 - (\det R)^{1/2})/(1 + (\det R)^{1/2}), \tag{13}$$

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

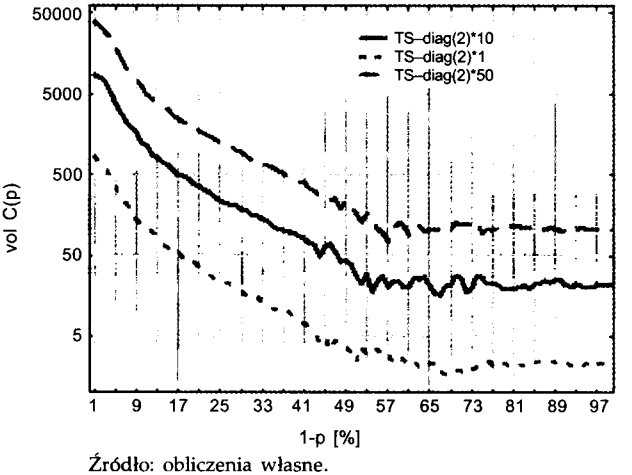
W celu sprawdzenia wybranych statystycznych własności krzywych skali i korelacji z próby przeprowadzono badania symulacyjne. Generowano mianowicie po 500 prób złożonych ze 100 obserwacji pochodzących z dwuwymiarowych rozkładów skośnych normalnych, skośnych T-Studenta, Marshalla-Olkina oraz mieszanin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

Uzyskane wyniki skłaniają do następujących wniosków:

— Krzywe skali dobrze dyskryminują rozkłady skośne normalne i skośne *T*-Studenta różniące się charakterystykami rozrzutu.

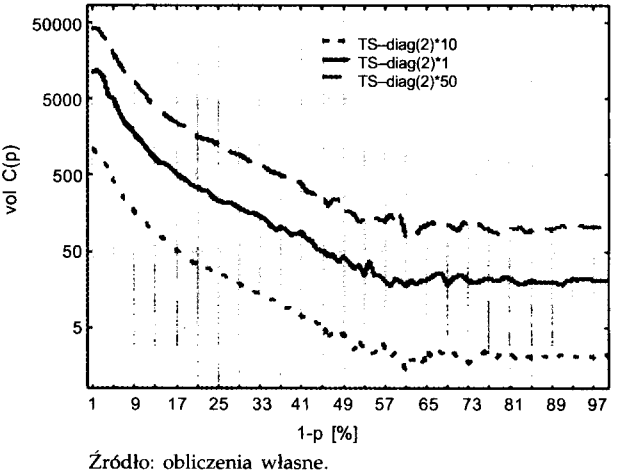
— Krzywe korelacji sporządzone dla izotropowych skośnych rozkładów normalnych i *T* właściwie „wychwytują" wpływ skośności na brak niezależności rozkładów brzegowych.

— Krzywa korelacji dobrze się sprawuje w przypadku rozkładu nie należącego do rodziny wykładniczej. Należy podkreślić, że z krzywej korelacji można odczytać jak przedstawia się struktura zależności rozkładu w zależności od bliskości centrum rozkładu.



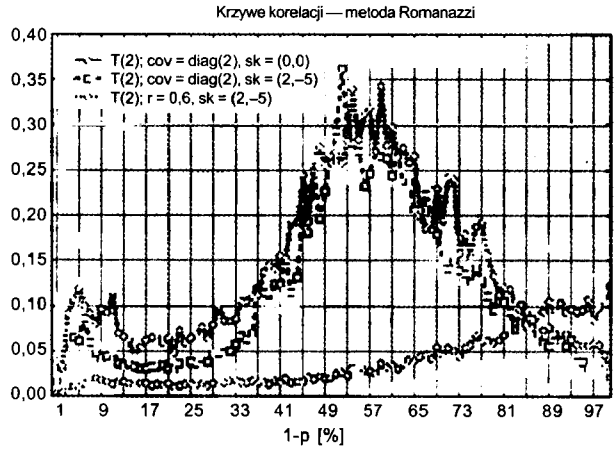
Źródło: obliczenia własne.

Ryc. 4. Krzywe skali — dwuwymiarowe rozkłady normalne



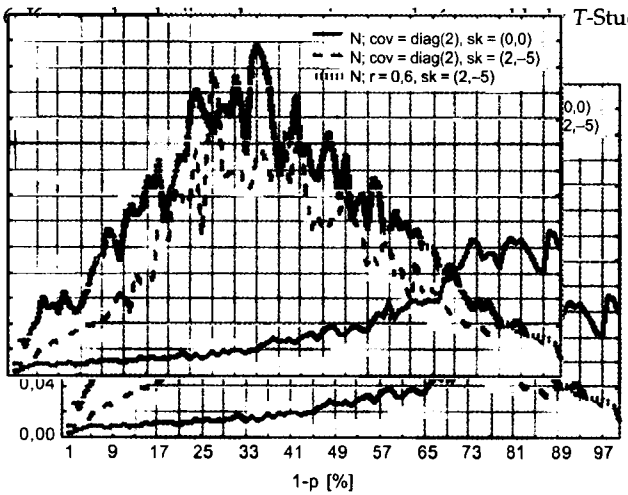
Źródło: obliczenia własne.

Ryc. 5. Krzywe skali — dwuwymiarowe rozkłady *T*-Studenta



Źródło: obliczenia własne.

Ryc. 6. Krzywe korelacji — dwuwymiarowe rozkłady *T*-Studenta



Źródło: obliczenia własne.

Ryc. 7. Krzywe korelacji — dwuwymiarowe skośne rozkłady normalne

— Krzywa korelacji jest wrażliwa na zmieszanie populacji, co w zależności od punktu widzenia można poczytać za jej wadę bądź zaletę.
— Krzywa skali jest względnie niewrażliwa na zmieszania populacji.

3. KWANTYLOWY FUNKCJONAŁ ASYMETRII ROZKŁADU WEKTORA LOSOWEGO

Odstępstwo wielowymiarowego rozkładu prawdopodobieństwo od ustalonego pojęcia symetrii określa się mianem skośności rozkładu. Przez symetrię na ogół rozumie się własność obiektu polegającą na tym, że istnieje pewne,

różne od tożsamościowego przekształcenie, które odwzorowuje dany obiekt na niego samego.

Zaznaczmy, że w przypadku wielowymiarowego rozkładu prawdopodobieństwa wykorzystuje się wiele różniących się wzajemnie pojęć wielowymiarowej symetrii, które sprowadzają się do zwykłego pojęcia symetrii w przypadku jednowymiarowym tzn. symetrii zwierciadlanej.

Mówimy, że wektor losowy $\mathbf{X}$ ma rozkład centralnie symetryczny względem θ, jeżeli:

$$\mathbf{X} - \theta \stackrel{r}{=} \theta - \mathbf{X}.$$

gdzie symbol „$\stackrel{r}{=}$" oznacza równość rozkładów.

Propozycja nawiązująca do koncepcji głębi

Stopień odstępstwa rozkładu prawdopodobieństwa od centralnej symetrii można mierzyć za pomocą funkcjonału asymetrii wykorzystującego stosownie wybraną funkcji głębi, np. głębię projekcyjnej.

Niech PM_F oznacza medianę indukowaną przez głębię projekcyjną. Aby zmierzyć asymetrię rozkładu dla każdego $r \in (0, 1)$, badamy różnicę pomiędzy przeciętną punktów wewnątrz obszaru centralnego rzędu r a medianą PM_F.

Rozważmy mianowicie:

$$\|\tilde{s}_F(p)\| = 2\frac{\left\| \int_{PC_F(p)} W(\mathbf{x}) m_p(d\mathbf{x}) - PM_F \right\|}{\tilde{v}_F(p)^{1/d}}, \quad 0 < p < 1, \tag{14}$$

gdzie: $PC_F(p)$ to projekcyjny obszar centralny rzędu r, $m_p(d\mathbf{x})$ oznacza rozkład, np. jednostajny na $PC_F(p)$, PM_F oznacza indukowaną przez głębię projekcyjną medianę, $W(\cdot)$ właściwą dla zagadnienia funkcję wagową np. $W(\mathbf{x}) = \mathbf{x}$.

Proponowany funkcjonał asymetrii, dzięki własnościom głębi projekcyjnej i własnościom indukowanych przez nią obszarów centralnych, jest afinicznie niezmienniczy, tzn. nie zależy od przyjętego w badaniu układu współrzędnych. Umożliwia nieparametryczny pomiar asymetrii populacji nie posiadającej momentów. Można pokazać, że funkcjonał (14) z próby jest mocno zgodnym w sensie odległości Kołmogorowa estymatorem odpowiednika w populacji.

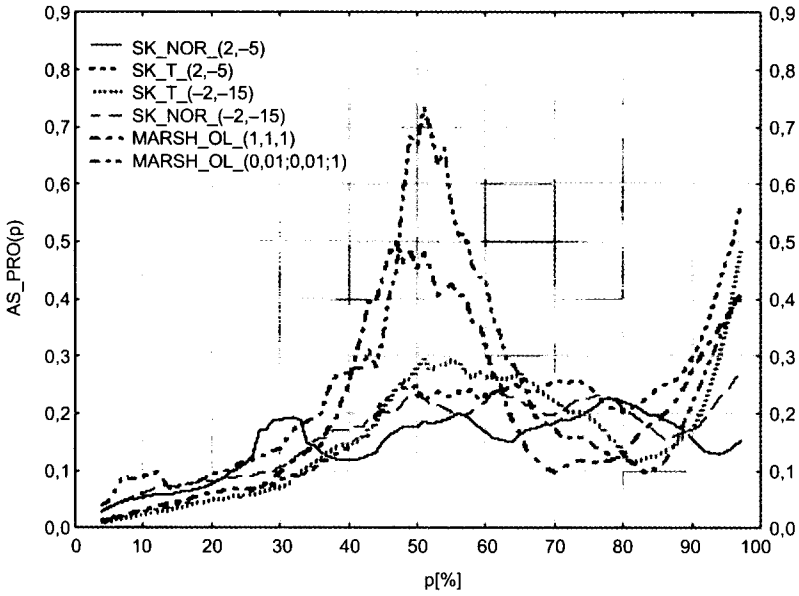
W celu sprawdzenia wybranych własności (14) przeprowadzono badania symulacyjne. Generowano mianowicie po 100 prób 100 elementowych z dwuwymiarowych rozkładów:

a) skośnego normalnego i skośnego T o dwóch stopniach swobody o parametrach: położenia $\mathbf{m} = (0, 0)$, rozrzutu $\Sigma = diag(2) \cdot 5$, kształtu $\Omega = (2, -5)$;

b) skośnego normalnego i skośnego T o dwóch stopniach swobody o parametrach: położenia $\mathbf{m} = (0, 0)$, rozrzutu $\Sigma = diag(2) \cdot 5$, kształtu $\Omega = (-2, -15)$;

c) Marshalla-Olkina o parametrach $\lambda_1 = (1, 1, 1)$ i $\lambda_2 = (0{,}01, 0{,}01, 1)$.

Z ryciny 8 wynika m.in., że proponowany funkcjonał dobrze rozróżnia zarówno wyszczególnione typy rozkładów (skośny normalny, skośny *T*, Marshalla-Olkina), jak i rozkłady należące do tego samego typu a różniące się parametrem asymetrii. Dodajmy, że z prowadzonych wcześniej badań wynika, że nasza propozycja lepiej rozróżnia pomiędzy skośnymi rozkładami normalnym i *T* aniżeli oryginalna propozycja Chaudhuri (1996) oraz propozycje Liu i in (1999).



Ryc. 8. Wyniki badań symulacyjnych proponowanego funkcjonału asymetrii

4. ODPORNOŚĆ METOD KLASYFIKACYJNYCH WYKORZYSTUJĄCYCH FUNKCJE GŁĘBI

Rozważamy k populacji p wymiarowych $C_1, \ldots, C_k$, $k \geq 2$. Przypuśćmy, że z każdą populacją C_j związana jest gęstość prawdopodobieństwa $f_j(\mathbf{z})$ na R^p, w ten sposób, że jeśli obserwacja pochodzi z populacji C_j, to została wygenerowana przez rozkład o gęstości $f_j(\mathbf{z})$.

W tzw. zagadnieniu dyskryminacji obiektów interesuje nas racjonalny sposób przyporządkowywania obserwacji $\mathbf{z} \in R^p$ do jednej ze wspomnianych k populacji (zob. np. Krzyśko, 2006; Jajuga, 1993).

Reguła dyskryminacyjna L odpowiada podziałowi R^p na rozłączne obszary $R_1, ..., R_k$, spełniające warunek $\bigcup R_j = R^p$. Reguła definiowana jest jako:

Przyporządkuj $\mathbf{z}$ do C_j jeżeli $\mathbf{z} \in R_j$, dla $j = 1, ..., k$.

Indeks $i \in \{1, 2, ..., k\} = Y$ wskazujący na rozważaną populację C_i często określa się mianem etykiety populacji. W takim ujęciu zagadnienie dyskryminacji sprowadza się do predykcji etykiety $i \in Y$ na podstawie obserwacji wektora cech $\mathbf{z}$.

Reguła klasyfikacyjna nazywana klasyfikatorem jest zatem funkcją:

$$L : R^p \ni \mathbf{x} \longrightarrow i \in Y.$$

Gdy obserwujemy wektor $\mathbf{x} \in X$, to prognozą jego przynależności jest $L(\mathbf{x}) \in Y$.

Często w praktyce rozważa się sytuację, gdy wprawdzie ogólne postaci gęstości $f_j(\mathbf{z})$ są znane jednak jesteśmy zmuszeni szacować ich parametry. Estymacja w takim przypadku opiera się na tzw. próbie uczącej — macierzy danych $\mathbf{Z}_{n \times p}$, której wiersze są podzielone na k grup odpowiadających k rozpatrywanym populacjom, macierz zostaje podzielona na $k(n_j \times p)$ macierz $\mathbf{Z}_j$ odpowiada próbie n_j obserwacji z populacji C_j.

Klasyczne metody dyskryminacyjne, takie jak liniowe bądź kwadratowe funkcje dyskryminacyjne, zakładają wielowymiarową normalność (szerzej eliptyczność) populacji. Metody te nie sprawują się dobrze w przypadku skośnych populacji. Metody klasyczne zakładają, że rozpatrywane populacje posiadają momenty. Fakt ten stanowi istotne ograniczenie ich stosowalności, np. w przypadku wielowymiarowego rozkładu Cauchy'ego. Metody wykorzystujące wektory przeciętnych, macierze kowariancji bądź kryterium najmniejszych kwadratów są skrajnie nieodporne na jednostki odstające.

Odporność reguły dyskryminacyjnej

DEFINICJA 3 (na podst. Krzyśko, 2006): Rozważamy k populacji p wymiarowych $C_1, ..., C_k$, $k \geq 2$ oraz ustaloną próbę uczącą reprezentującą populacje. Rzeczywisty poziom błędu klasyfikatora L jest równy:

$$Err(L) = P\{L(\mathbf{X}) \neq i \mid \mathbf{X} \in C_i\},$$

gdzie $\mathbf{X}$ oznacza obserwację niezależną od próby uczącej.

$Err(L)$ jest prawdopodobieństwem zdarzenia, że klasyfikator błędnie zakwalifikuje nową obserwację niezależną od próby uczącej pod warunkiem, że próba ucząca jest ustalona.

Propozycja. Rozważamy k populacji p wymiarowych $C_1, ..., C_k$, $k \geq 2$, oraz próbę uczącą $\mathbf{Z}$ reprezentującą populacje. Punkt załamania próby uczącej $\mathbf{Z}$ klasyfikatora L w j-tej klasie C_j określamy jako:

$$BP_j(L, C_j^m) = \inf_{C_j^m}\left(\frac{m}{n_j} : \frac{P\{L(\mathbf{z}) \neq i \mid \mathbf{z} \in C_i\}}{P\{L(\mathbf{z}) = i \mid \mathbf{z} \in C_i\} + P\{L(\mathbf{z}) \neq i \mid \mathbf{z} \in C_i\}} > \frac{1}{2}\right), \tag{15}$$

gdzie C_j^m oznacza $(n_j \times p)$ macierz $\mathbf{Z}_j$ próby uczącej $\mathbf{Z}$ odpowiadającą próbie n_j obserwacji z populacji C_j, w której zastąpiono m wierszy (m obserwacji) dowolnymi wierszami (obserwacjami).

Globalny punkt załamania próby uczącej klasyfikatora określamy jako:

$$BP(L, C_1, ..., C_k) = \min_j BP_j(L, C_j^m). \tag{16}$$

Każda funkcja głębi D indukuje pewną regułę klasyfikacyjną, tzn.:

$$L(\mathbf{z}) = \operatorname{argmax}_j D(\mathbf{z} \mid C_j),\ j = 1, ..., k, \tag{17}$$

która klasyfikuje obserwację do tej klasy C_j, dla której głębia z przyjmuje wartość maksymalną (zob. Hoberg i Mosler, 2006).

Przykłady reguł klasyfikacyjnych indukowanych przez głębie

Niech $C_j = \{\mathbf{x}_{j1}, ..., \mathbf{x}_{jn_j}\}$ oznacza próbę n_j obserwacji z populacji j, $j = 1, ..., k$.

Przykładem reguły klasyfikacyjnej indukowanej przez funkcję głębi jest reguła wykorzystująca tzw. głębię Euklidesa:

$$D_{EUK}(\mathbf{z} \mid C_j) = \frac{1}{1 + \|\mathbf{z} - \bar{\mathbf{x}}_j\|^2}, \text{ gdzie } \bar{\mathbf{x}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{x}_{ji}.$$

Zauważmy, że głębia ta prowadzi do znanej reguły k-centroidów.

Wykorzystując znaną głębię Mahalanobisa:

$$D_{MAH}(\mathbf{z} \mid C_j) = \frac{1}{1 + (\mathbf{y} - \bar{\mathbf{x}}_j)'\mathbf{S}_j^{-1}(\mathbf{y} - \bar{\mathbf{x}}_j)},$$

gdzie $\mathbf{S}_j$ oznacza macierz kowariancji próby C_j, otrzymamy regułę klasyfikacyjną Mahalanobisa.

Propozycja. Rozważmy regułę klasyfikacyjną indukowaną przez tzw. symetryczną głębię projekcyjną (własności tej głębi przedstawia m.in. Zuo, 2003)

$$D_{PRO}(\mathbf{z} \mid C_j) = \left(1 + \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^t\mathbf{z} - med(\mathbf{u}^tC_j)|}{MAD(\mathbf{u}^tC_j)}\right)^{-1},$$

gdzie $\mathbf{u}^tC_j = \{\mathbf{u}^t\mathbf{x}_{j1}, ..., \mathbf{u}^t\mathbf{x}_{jn_j}\}$, $MAD(Y) = med\{|Y - med(Y)|\}$.

Dla ustalonego zbioru danych (rozkładu prawdopodobieństwa) głębia projekcyjna w R^p jest afinicznie niezmiennicza, quasi-wypukła, głębia punktu

zmierza do zera, jeśli norma punktu zmierza do nieskończoności, głębia przyjmuje maksimum w centrum symetrii rozkładu.

Własności klasyfikatora projekcyjnego porównano z liniową oraz kwadratową funkcją dyskryminacyjną oraz z klasyfikatorem głębi Tukey'a na przykładzie znanego zbioru danych (Fisher, 1936; Rao, 1982), dotyczącego 3 gatunków Irysa, rozpatrywanych ze względu na 4 cechy kwiatu. Rozpatrywano próby uczące wielkości 25:25:25 oraz 40:40:40. Na ich podstawie klasyfikowano obserwacje należące do zbioru Irys. Na dobre własności proponowanej reguły wskazują wyniki zawarte w tabeli 1.

Tabela 1

Wyniki badań symulacyjnych proponowanego klasyfikatora projekcyjnego

	3 × 25 obserwacji w próbie uczącej			
	liniowa funkcja dyskryminacyjna	kwadratowa funkcja dyskryminacyjna	klasyfikator głębi projekcyjnej	klasyfikator głębi Tukay'a
Rzeczywisty błąd predykcji	2.6%	4%	3.3%	66%
	3 × 40 obserwacji w próbie uczącej			
Rzeczywisty błąd predykcji	3.8%	2.6%	2.6%	66%

Generowano także 00 razy zbiór 3000 dwuwymiarowych obserwacji : 1000 z rozkładu Marshalla-Olkina, 1000 z izotropowego rozkładu normalnego, 1000 ze skośnego rozkładu Studenta T. Ze zbioru pobierano 100 × 100 × 100 próby uczące. Na jej podstawie klasyfikowano zbiory 3000 obserwacji. Na dobre własności proponowanej reguły wskazują wyniki zawarte w tabeli 2.

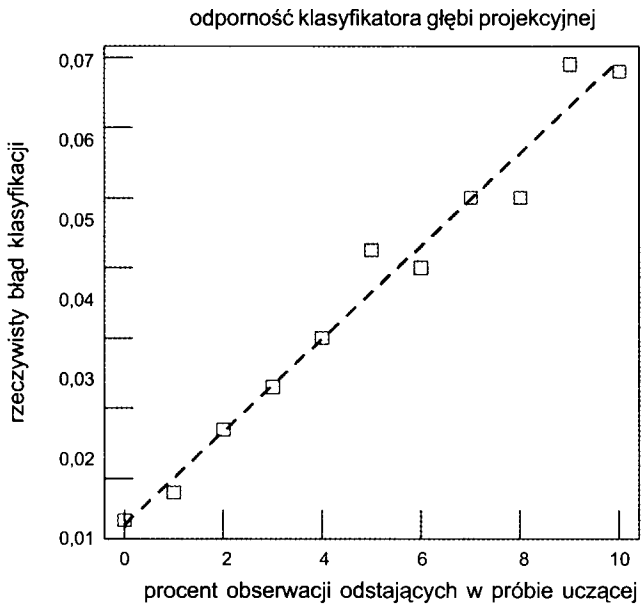
Tabela 2

Wyniki badań symulacyjnych proponowanego klasyfikatora projekcyjnego

	Klasyfikator			
	liniowa funkcja dyskryminacyjna	kwadratowa funkcja dyskryminacyjna	klasyfikator głębi projekcyjnej	klasyfikator głębi Tukey'a
Rzeczywisty błąd predykcji	12.6%	12.6%	0.3%	68%

W celu oszacowania punktu załamania próby skończonej generowano zbiory 3000 dwuwymiarowych obserwacji z populacji będącej mieszaniną trzech sko-

śnych rozkładów T-Studenta o równych udziałach a różniących się parametrami położenia i kształtu. Następnie zastępowano po 0%, 1%, ..., 10% obserwacji w 3 × 100 elementowej próbie uczącej, reprezentującej każde skupisko obserwacjami dalece odstającymi od centrów rozkładów mieszaniny. Obliczano rzeczywisty błąd klasyfikacji po takim zastąpieniu.



Ryc. 9. Punkt załamania proponowanej reguły dyskryminacyjnej

Propozycja reguły grupowania obserwacji na jednorodne skupiska

Przypuśćmy, że dysponujemy n obserwacjami p wymiarowymi $C_0 = \{x_1, ..., x_n\}$. Naszym celem jest wskazanie pewnego optymalnego rozbicia zbioru C_0 na k jednorodnych rozłącznych podzbiorów p wymiarowych $C_1, ..., C_k$, $k \geq 2$, $C_i \cap C_j = \emptyset$, $i \neq j$, $\bigcup C_i = C_0$.

Propozycja reguły. Niech $\tilde{C}_1, ..., \tilde{C}_k$, $k \geq 2$, $\tilde{C}_i \cap \tilde{C}_j = \emptyset$, $i \neq j$, $\bigcup C_i = C_0$, będzie pewnym możliwym rozbiciem zbioru obserwacji C_0. Powiemy, że rozbicie $\tilde{C}_1, ..., \tilde{C}_k$ jest lepsze niż rozbicie trywialne C_0 i $\emptyset$, jeżeli zachodzi:

$$vol(D^{\alpha}_{PRO}(C_0)) > \sum_{i=0}^{k} vol(D^{\alpha}_{PRO}(\tilde{C}_i)), \text{ dla ustalonego } \alpha \in 0, 1, \qquad (18)$$

gdzie $vol(D^{\alpha}_{PRO}(C_0))$ oznacza objętość centralnego obszaru centralnego.

5. ODPORNE WYKRYWANIE PUNKTU ZMIANY TENDENCJI W WYBRANYCH MODELACH REGRESJI

Zagadnienie wskazania punktu zmiany tendencji (ang. *change point*) w modelu regresji było studiowane w ekonomii m.in. w kontekście badań zmian natury zależności pomiędzy zjawiskami powyżej pewnego poziomu jednego z nich. Zagadnienie wiąże się m.in. z badaniem zysku z inwestycji z wielkością poniesionych wydatków, badaniem przyrostu naturalnego i PKB na mieszkańca. Na poziomie technicznym pojawia się w zagadnieniach dyskryminacji i klasyfikacji obiektów, ma związek z jakością tychże metod.

W literaturze statystycznej przedstawiono szereg podejść dotyczących estymacji i weryfikacji hipotez dla modeli regresji w dwóch fazach.

Chen (1998) zaproponował wykorzystanie kryterium informacyjnego Schwarza (SIC), do wskazania punktu zmiany tendencji w modelu liniowym przy założeniu normalności.

Osorio i Galea (2005) rozszerzyli wyniki Chena (1998) na model regresji liniowej z niezależnymi błędami *t*-Studenta.

Sformułowanie problemu

Opierając się na n niezależnych obserwacji (Y_1, x_1^t), (Y_2, x_2^t), ..., (Y_n, x_n^t), zamierzamy zweryfikować hipotezę, że parametry regresji nie zmieniają się:

$$H_0 : Y_i = x_i^t \beta + \varepsilon_i,\ i = 1, 2, ..., n,$$

przeciw hipotezie alternatywnej, że następuje zmiana (parametrów regresji) na nieznanej pozycji k, nazywanej punktem zmiany tendencji:

$$H_1 : Y_i = x_i^t \beta_1 + \varepsilon_i,\ i = 1, 2, ..., k,$$

$$Y_i = x_i^t \beta_2 + \varepsilon_i,\ i = k + 1, ..., n,$$

gdzie $\beta_1 = (\beta_0, \beta_1, ..., \beta_{p-1})^t$, $\beta_2 = (\beta_0^*, \beta_1^*, ..., \beta_{p-1}^*)^t$, i ε oznacza błąd.

W celu rozwiązania powyższego problemu Osorio i Galea (2005), odwołując się do pracy Chena (1998), proponują zamienić proces testowania hipotez na procedurę wyboru modelu z wykorzystaniem kryterium informacyjnego Schwarza (SIC) definiowanego:

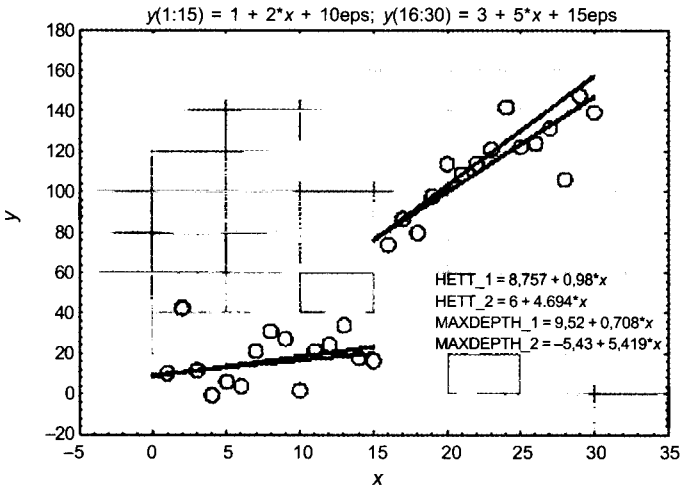
$$SCI = -2L(\hat{\Theta}) + s\log n, \tag{19}$$

gdzie $L(\Theta)$ to logarytm wiarygodności obliczony dla oszacowań parametrów uzyskanych metodą największej wiarygodności, s jest liczbą parametrów w modelu, n to wielkość próby.

Osorio i Galea rozważają model regresji liniowej:

$$Y_i = \mathbf{x}_i^t \beta + \varepsilon_i,\ i = 1, 2, ..., n, \tag{20}$$

gdzie $\mathbf{x}_i$, $i = 1, 2, ..., n$ odpowiada i-temu wierszowi $n \times p$ macierzy $\mathbf{X}$, $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})^t$ jest wektorem nieznanych parametrów, natomiast ε_i są to niezależne błędy losowe o takim samym rozkładzie $t(0, \phi, n)$.



Ryc. 10. Ilustracja procedury detekcji punktu zmiany tendencji

W celu estymacji parametrów i obliczania informacji Schwarza wykorzystują zmodyfikowany algorytm EM.

Przy założeniu H_0, nie występuje zmiana współczynników regresji; przy założeniu H_1 mamy zbiór możliwych modeli ze zmianami w punkcie p or $p + 1$ or...or $n - p$.

Obliczamy wartość SIC przy założeniu H_0 oraz przy założeniu H_1.

Wybieramy model z punktem zmiany na pozycji k, jeżeli dla pewnego k:

$$Sic(n) > SIC(k). \tag{21}$$

W przypadku gdy hipoteza zerowa jest odrzucana, estymator największej wiarygodności punktu zmiany tendencji oznaczany przez $\hat{k}$, powinien spełniać warunek:

$$SIC(\hat{k}) = \min\{SIC(k) : p \le k \le n - p\} = \max\{L_k(\Theta) : p \le k \le n - p\}. \tag{22}$$

w dalszej części porównujemy propozycję z podejściem kryterium informacyjnego Schwarza wykorzystując pakiet autorstwa Juliana Taylora heteroscedastic t-regression package (R Project).

Propozycja

Uporządkujmy obserwacje według wartości zmiennej $\mathbf{x}$:

$$(y, x_{(1)}), ..., (y, x_{(p-1)}), (y, x_{(p)}), (y, x_{(p+1)}), ..., (y, x_{(n-1)}), (y, x_{(n)}),$$

dla $p = \left\lceil \frac{4n}{10} \right\rceil, \ldots, \left\lceil \frac{6n}{10} \right\rceil$ („okno poszukiwań długości 2/10*n").

Oznaczmy:

$$\mathbf{Z}^n = (y, x_{(1)}), (y, x_{(2)}), \ldots, (y, x_{(n-1)}), (y, x_{(n)}),$$

$$\mathbf{Z}_-^n = (y, x_{(1)}), (y, x_{(2)}), \ldots, (y, x_{(p-1)}), (y, x_{(p)}),$$

$$\mathbf{Z}_+^n = (y, x_{(p)}), (y, x_{(p+1)}), \ldots, (y, x_{(n-1)}), (y, x_{(n)}),$$

Następnie obliczmy:

1. Estymator maksymalnej głębi regresyjnej dla wszystkich obserwacji $T_r^*(\mathbf{Z}^n) = (b_0, b_1) = \hat{b}_r$ oraz głębię maksymalnego dopasowania $rdepth(\hat{b}_r)$.

2. Estymator maksymalnej głębi regresyjnej dla $T_r^*(\mathbf{Z}_-^p) = (b_0^-, b_1^-) = \hat{b}_r^-$ oraz głębię tego dopasowania $rdepth(\hat{b}_r^-)$.

3. Estymator maksymalnej głębi regresyjnej dla $T_r^*(\mathbf{Z}_+^p) = (b_0^+, b_1^+) = \hat{b}_r^+$ oraz głębię tego dopasowania $rdepth(\hat{b}_r^+)$.

Jeżeli dla pewnego $p = \lceil 4n/10 \rceil, \ldots, \lceil 6n/10 \rceil$ ma miejsce:

$$rdepth(\hat{\mathbf{b}}_r) < \min\{rdepth(\hat{\mathbf{b}}_r^-), rdepth(\hat{\mathbf{b}}_r^+)\}, \tag{23}$$

wtedy uznajemy, że w punkcie p następuje zmiana parametrów regresji. W takim przypadku za parametry regresji dla obserwacji $i = 1, 2, \ldots, p$ przyjmujemy $\hat{\mathbf{b}}_r^-$ natomiast dla obserwacji $i = p + 1, \ldots, n$ — za parametry przyjmujemy $\hat{\mathbf{b}}_r^+$.

Z przeprowadzonych przez autora badań symulacyjnych wynika, że w przypadku nie występowania punktu zmiany tendencji propozycja spisuje się lepiej niż metoda kryterium Schwarza w przypadkach, gdy błędy mają rozkład normalny bądź Cauchy'ego.

W przypadku, gdy błąd ma rozkład normalny oraz występują dwie obserwacje odstające kryterium SIC sprawuje się nieco lepiej.

W przypadku występowania punktu zmiany tendencji obie rozważane metody zachowują się podobnie, metoda SIC odznacza się nieco mniejszym rozrzutem wskazania.

6. ANALIZA DANYCH PANELOWYCH Z WYKORZYSTANIEM GŁĘBI REGRESYJNEJ

W klasycznej analizie regresji na ogół zakłada się, że obserwacje są pobierane z tej samej populacji, są niezależne i o takim samym rozkładzie. W przypadku analizy regresji prowadzonej z wykorzystaniem modeli mieszanych stosuje się słabsze założenia. Mianowicie, dane mogą tworzyć skupiska, obserwacje pomiędzy skupiskami są niezależne, jednak nie muszą być niezależne wewnątrz skupisk.

Modele mieszane wydają się użyteczne np. w badaniach gmin z uwzględnieniem podziału na województwa, w badaniu wydatków konsumpcyjnych z uwzględnieniem grupy dochodowej itd.

W ekonometrii przyjęto nazywanie obserwacji d a n y m i p a n e l o w y m i wówczas, gdy dotyczą poszczególnych jednostek przekrojowych w dłuższym czasie (więcej niż jednym okresie).

Przypuśćmy, że na dane patrzymy z punktu widzenia liniowego modelu mieszanego, w postaci zaproponowanej przez Lairda i Ware'a w 1982 roku:

$$\mathbf{y}_i = \mathbf{x}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i,\ i = 1, \ldots, N, \tag{24}$$

gdzie:

$\mathbf{y}_i$ — jest $n_i \times 1$ wektorem odpowiedzi i-tego skupiska (n_i odpowiedzi jednostek z i-tego skupiska);

$\mathbf{x}_i$ — jest $n_i \times m$ macierzą ustalonych efektów w i-tym skupisku;

β — uśredniony dla wszystkich skupisk wektor parametrów związanych ze stałymi efektami;

$\mathbf{Z}_i$ — $n_i \times k$ macierz eksperymentu efektów losowych w i-tym skupisku;

ε_i — $n_i \times 1$ wektor błędu dla i-tego skupiska, wektor o niezależnych składowych, każda o przeciętnej zero i wariancji σ^2;

$\mathbf{b}_i$ — jest $k \times 1$ wektorem parametrów związanych z efektami losowymi w i-tym skupisku, wektor o zerowej przeciętnej i macierzy kowariancji $\mathbf{D}* = \sigma^2\mathbf{D}$.

Zakładamy, że macierz $\sum \mathbf{X}_i^t\mathbf{X}_i$ jest nieosobliwa oraz, że $\sum n_i > m$ dla zapewnienia identyfikowalności modelu (24) względem β. Dla zapewnienia identyfikowalności modelu (24) względem σ^2 i D, zakładamy, że przynajmniej jedna macierz $\mathbf{z}_i^t\mathbf{z}_i$ jest dodatnio określona oraz, że $\sum_{i=1}^{N}(n_i - k) > 0$.

Wypada wspomnieć także, że często wykorzystuje się skalowaną macierz kowariancji efektów losowych:

$$\mathrm{D} = \frac{1}{\sigma^2}\mathbf{D}* = \frac{1}{\sigma^2}Cov(\mathbf{b}_i). \tag{25}$$

W celu estymacji parametrów modelu (24) metodą NW zakładamy, że:

$$\varepsilon_i \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}),\ \mathbf{b}_i \sim N(\mathbf{0}, \sigma^2\mathbf{D}). \tag{26}$$

Warto zauważyć, że przy założeniach (3) model (1) można zapisać w następującej postaci brzegowej:

$$\mathbf{y}_i \sim N(\mathbf{X}_i\beta, \sigma^2(\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t)),\ \mathrm{i} = 1, \ldots, \mathrm{N}. \tag{27}$$

Okazuje się, że ustalając macierz $\mathbf{D}$, logarytm wiarygodności dla modelu (24) maksymalizowany jest przez uogólniony estymator najmniejszych kwadratów NK

$$\hat{\beta}_{UNK} = \left[\sum_{i=1}^{N}\mathbf{X}_i^t(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t)\right]^{-1}\left[\sum\mathbf{X}_i^t(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t)^{-1}y_i\right], \tag{25}$$

Zauważmy, że w specjalnym przypadku, gdy $\mathbf{D} = \mathbf{0}$, estymator (36) sprowadza się do zwykłego estymatora NK:

$$\hat{\beta}_{NK} = (\sum X_i^t X_i)^{-1}(\sum X_i^t y_i). \tag{26}$$

W wielu zastosowaniach wykorzystuje się szczególną wersję liniowego modelu mieszanego (24), w którym dopuszcza się jeden efekt losowy dotyczący wyrazu wolnego, a który definiowany jest jako:

$$y_{ij} = a_i + \gamma^t x_{ij} + \varepsilon_{ij}, \; j = 1, \ldots, n_i, \; i = 1, \ldots, N, \tag{27}$$

gdzie y_{ij} interpretowane jest jako j-ta odpowiedź w i-tym skupisku na wartość predyktora x_{ij}, $\gamma = [\alpha \; \beta]$, przy czym to uśredniony dla całej populacji wyraz wolny, β uśredniony dla całej populacji parametr nachylenia.

W modelu (27) wyraz wolny w skupisku i interpretowany jest jako suma uśrednionego dla całej populacji parametru α oraz efektu losowego specyficznego dla i-tego skupiska:

$$a_i = \alpha + b_i. \tag{28}$$

Na ogół zakłada się, że $\varepsilon_{ij} \sim N(0, \sigma^2)$ oraz $b_i \sim N(0, \sigma^2 d)$ są niezależne, gdzie σ^2 jest wariancją błędu wewnątrz skupiska oraz d jest skalowaną wariancją efektu losowego.

Model (27) może pojawić się przykładowo w następującej sytuacji. Rozważamy grupę N województw w ujęciu powiatów. Każde województwo badamy ze względu na stopę bezrobocia, przeciętne wynagrodzenie, saldo migracji w gminach województwa, przyrost naturalny. Przypuśćmy, że interesuje nas zależność przyrostu naturalnego od pozostałych zmiennych. Kluczowym dla stosowalności modelu (27) w przedstawionej sytuacji jest założenie, że struktura zależności nie zmienia się od województwa do województwa, co znaczy, że γ jest ustalonym wektorem.

Model z losowymi wyrazami wolnymi wydaje się bardziej realistyczny niż model klasyczny, gdyż dopuszcza charakterystyczny dla każdego z województw oddzielnie, poziom przyrostu naturalnego.

Uogólniony Estymator NK podobnie jak zwykły estymator NK jest skrajnie nieodporny na jednostki odstające, BP wektora parametrów wynosi 0%.

Eksperymenty symulacyjne wskazują na niską efektywność estymatora UNK w przypadkach, gdy wariancja efektów losowych jest istotnie większa od wariancji błędu oraz gdy wariancja efektów losowych lub wariancji błędu są nieznane.

Propozycja odpornego estymatora głębi regresyjnej

Przypuśćmy, że na obserwacje patrzymy z punktu widzenia modelu (27).

Oznaczmy przez $\mathbf{Z}_i^{n_i}$ zbiór par obserwacji y i x w i-tym skupisku tzn.

$$\mathbf{Z}_i^{n_i} = ((x_{i1}, y_{i1})^t (x_{i2}, y_{i2})^t \ldots (x_{in_i}, y_{in_i})^t)^t, \; i = 1, \ldots, N.$$

Niech $T_i^{rdepth}(\mathbf{Z}_i^{n_i}) = \text{argmax}rdepth(\mathbf{b}, \mathbf{Z}_i^{n_i}) = \hat{\mathbf{b}}_i = (\hat{b}_i^0, \hat{b}_i^1)$ oznacza estymator maksymalnej głębi regresyjnej wektora parametrów $\mathbf{b}_i = (b_i^0, b_i^1)$ liniowej funkcji regresji $y_i = (b_i^0, b_i^1 x)$ w i-tym skupisku.

Niech:

$$T^{rdepth}\left((\mathbf{Z}_i^{n_1})^t ... (\mathbf{Z}_N^{n_N})^t\right)^t = \text{argmax}rdepth\left(\mathbf{b}, \left((\mathbf{Z}_i^{n_1})^t ... (\mathbf{Z}_N^{n_N})^t\right)^t\right) = \hat{\mathbf{b}} = (\hat{b}_0, \hat{b}_1), \quad (29)$$

oznacza estymator maksymalnej głębi regresyjnej wektora parametrów $\hat{\mathbf{b}} = (\hat{b}_0, \hat{b}_1)$ liniowej funkcji regresji $y = b_0 + b_1 x$ dla obserwacji z wszystkich skupisk.

Weźmy za estymatory parametrów modelu zdefiniowanego przez (27) i (28)

$\beta^{rdepth} = \hat{b}_1$ — nachylenie uśrednione dla wszystkich skupisk;

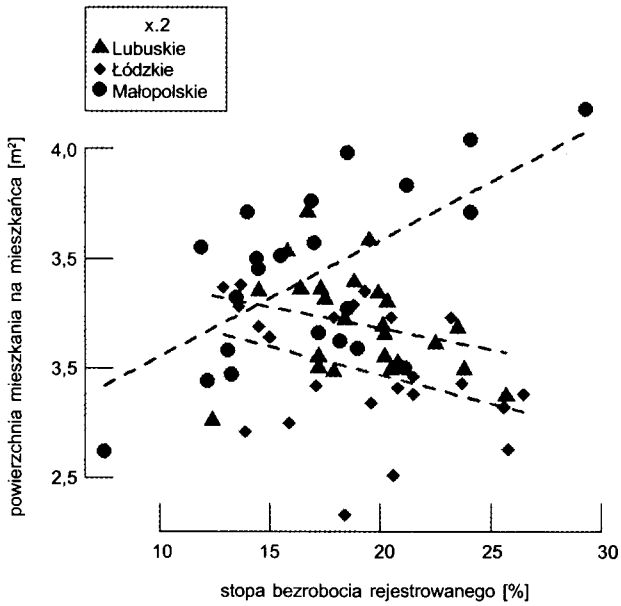
$\beta^{rdepth} = \hat{b}_0$ — wyraz wolny uśredniony dla całej populacji;

$\alpha_i^{rdepth} = \hat{b}_{i0}$ — wyraz wolny specyficzny dla i-tego skupiska, $i = 1, ..., N$;

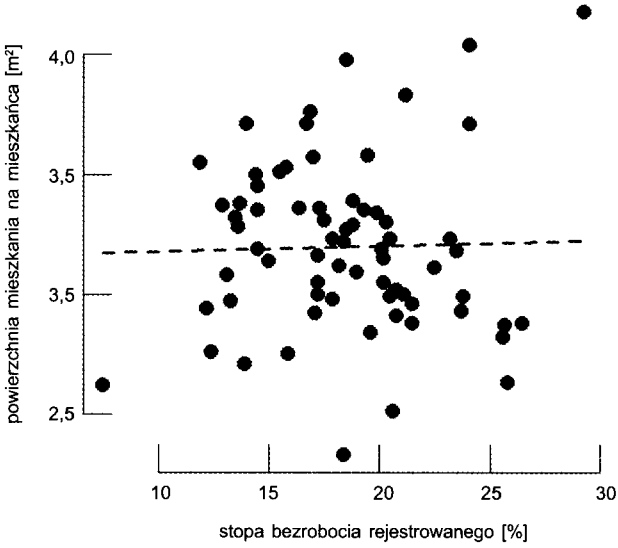
$b_i^{rdepth} = \alpha_i^{rdepth} - \alpha^{rdepth}$ efekt losowy w i-tym skupisku, $i = 1, ..., N$.

Rozpatrzmy zbiór danych złożony z 69 powiatów województw lubelskiego (24), łódzkiego (23) i małopolskiego (22), badanych ze względu na stopę bezrobocia rejestrowanego i powierzchnię mieszkania na 1 mieszkańca (w m^2) w roku 2005.

Zauważmy, że zaletą danych panelowych jest możliwość weryfikacji oraz złagodzenia założeń, które domyślnie są przyjmowane w analizie danych przekrojowych. Przy stosunkowo łagodnych założeniach dotyczących media-

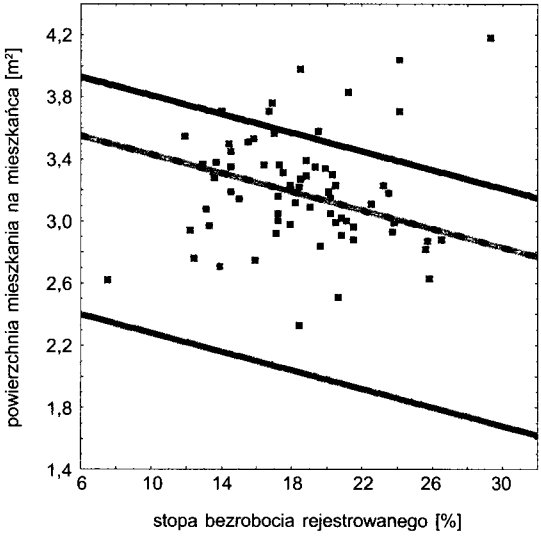


Ryc. 11. Powierzchnia mieszkania vs. stopa bezrobocia — dane dotyczące powiatów trzech województw traktowanych oddzielnie

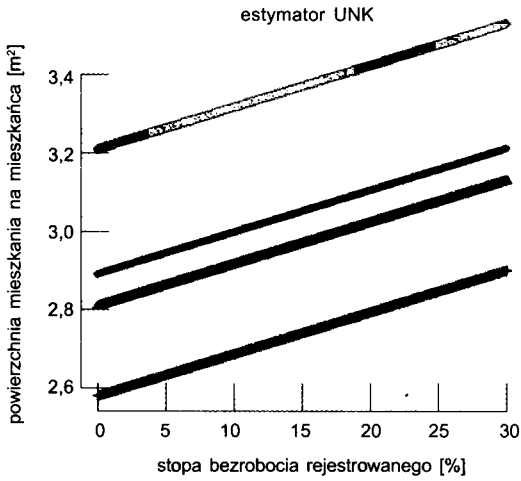


Ryc. 12. Powierzchnia mieszkania vs. stopa bezrobocia — dane dotyczące powiatów trzech województw traktowanych łącznie

ny warunkowego błędu punkty załamania estymatora maksymalnej głębi w każdym ze skupisk wynoszą niezależnie $BP \geq \frac{1}{d+1}$, gdzie d oznacza liczbę zmiennych objaśniających.



Ryc. 13. Powierzchnia mieszkania vs. stopa bezrobocia — dane dotyczące powiatów w trzech województwach — estymator maksymalnej głębi parametrów modelu



Ryc. 14. Powierzchnia mieszkania vs. stopa bezrobocia — dane dotyczące powiatów w trzech województwach — estymator UNK parametrów modelu

Estymator maksymalnej głębi regresyjnej dobrze radzi sobie ze skośnymi oraz heteroskedastycznymi rozkładami błędu i efektów losowych. Wyniki symulacji wskazują na nieobciążoność i niezłą efektywność estymatora maksymalnej głębi regresyjnej w porównaniu z uogólnionym estymatorem NK oraz niezależnymi ocenami parametrów w skupiskach za pomocą zwykłego estymatora NK.

7. PODSUMOWANIE

Poszukiwanie nieparametrycznych i odpornych zarazem procedur statystycznych, adekwatnych dla badania wielowymiarowych układów społeczno-ekonomicznych jest ważne zarówno z teoretycznych, jak i praktycznych względów. Zdaniem autora naszkicowana w pracy perspektywa badań, ma zastosowanie do lepszego zrozumienia struktury współzależności układów ekonomicznych. Przedstawione własności proponowanych metod, odwołujących się do koncepcji głębi danych, wydają się wystarczającym uzasadnieniem dla dalszych studiów nad tym zagadnieniem.

BIBLIOGRAFIA

Azzalini A., Capitanio A. 2003. *Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew t Distribution*, J. Roy. Statist. Soc. B 65, 367–389.

Chaudhuri P. 1996. *On a Geometric Notion of Quantiles for Multivariate Data*, Journal of the American Statistical Association 91, 862–872.

Chen J. 1998. *Testing for a Change Point in Linear Regression Models*, Communications in Statistics — Theory & Methods 27, 2481–2493.

Demidenko E. 2004. *Mixed Models — Theory and Applications*, John Wiley & Sons, Inc., Hoboken–New Jersey.

Dyckerhoff R. 2004. *Data Depths Satisfying the Projection Property*, Allgemeines Statistisches Archiv 88, 163–190.

Hoberg R., Mosler K. 2003. *Classification based on data depth*, Bulletin of the ISI 54th Session.

Hoberg R., Mosler K. 2006. *Data analysis and classification with the zonoid depth*, [w:] *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, R. Liu, R. Serfling, D. Souvaine red., American Mathematical Society, 2006, 49–59.

Koltchinskii V. 1997. *M-estimation, Convexity and Quantiles*, The Annals of Statistics 25, 435–477.

Kosiorowski D. (w druku). *Odporność metod klasyfikacyjnych wykorzystujących funkcje głębi*, Acta Universitatis Lodziensis, Folia Oeconomica, Materiały z Konferencji Multivariate Statistical Analysis 2007.

Kosiorowski D. (w druku). *Analiza danych panelowych z wykorzystaniem głębi regresyjnej*, Acta Universitas Lodziensis, Folia Oeconomica, Materiały z Konferencji SPSG'07.

Kosiorowski D. 2007. *O Kwantylowym funkcjonale asymetrii rozkładu wektora losowego w badaniach szeregów finansowych*, [w:] *Dynamiczne modele ekonometryczne*, Z. Zieliński red., Wydawnictwo UMK w Toruniu, Toruń, 129–136.

Kosiorowski D. 2007. *O odpornej analizie regresji w ekonomii na przykładzie koncepcji głębi regresyjnej*, Przegląd Statystyczny 1, 109–121.

Kosiorowski D. 2007. *Krzywa skali — odporna i nieparametryczna metoda badania rozrzutu wektora losowego i stopnia zależności jego rozkładów brzegowych*, Ryzyko Asekuracja Statystyka nr 44, Raport Techniczny Katedry Statystyki AE we Wrocławiu, 35–36.

Kosiorowski D. 2008. *About Robust Detection of a Change — Point in Selected Linear Regression Models*, Konferencja Multivariate Statistical Analysis 2006, Uniwersytet Łódzki, Acta Universitatis Lodziensis, Folia Oeconomica 216, 109–117.

Kosiorowski D. 2008. *Krzywa skali — odporna i nieparametryczna metoda badania rozrzutu wektora losowego i stopnia zależności jego rozkładów brzegowych*, Konferencja Statystyka Aktuarialna, Teoria i Praktyka, Wrocław, Badania Operacyjne i Decyzje 4, 47–60 .

Krzyśko M. (w druku). *Modele Klasyfikacyjne*, referat plenarny na konferencji Multivariate Statistical Analysis 2006, Łódź.

Laird N., Ware J. 1982. *Random-effects models for longitudinal data*, Biometrics 38, 963–974.

Liu R.Y., Parelius J.M., Singh K. 1999. *Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference (with discussion)*, The Annals of Statistics 27, 783–858.

Maddala G.S. 2006. *Ekonometria*, PWN, Warszawa.

Mizera I. 2002. *On Depth and Deep Points: A Calculus*, Annals of Statistics, 30, 1681–1736.

Mizera I., Muller Ch.H. 2002. *Breakdown Points of Cauchy Regression-Scale Estimators*, Statistics and Probability Letters 57, 79–89.

Mosler K. 2002. *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach*, Springer, New York.

Osorio F., Galea M. 2005. *Detection of a Change — Point in Student-t Linear Regression Models*, Statistical Papers 45, 31–48.

Romanazzi M. 2004. *Data Depth and Correlation*, Allgemeines Statistisches Archiv 88, 191–214.

Rousseeuw P.J., Hubert M. 1998. *Regression Depth*, J. Amer. Statist. Assoc. 94, 388–433.

Rousseeuw P.J., Leroy A.M. 1987. *Robust Regression and Outlier Detection*, Willey, New York.

Serfling R.J. 2004. *Nonparametric Multivariate Descriptive Measures Based on Spatial Quantiles*, Journal of Statistical Planning and Inference 123, 259–278.

Serfling R.J. 2004. *Nonparametric Multivariate Descriptive Measures Based on Spatial Quantiles*, Journal of Statistical Planning and Inference 123, 259–278.

Serfling R.J. 2006. *Multivariate Symmetry and Asymmetry*, [w:] *Encyclopedia of Statistical Sciences*, wyd. 2, S. Kotz, N. Balakrishnan, C.B. Read, B. Vidakovic red., Wiley, 5338–5345.

Van Aelst S., Rousseeuw P.J. 2000. *Robustness Properties of Deepest Regression*, J. Multiv. Analysis 73, 82–106.

Wang J., Serfling R. 2006. *Influence Functions for a General Class of Depth — Based Generalized Quantile Functions*, Journal of Multivariate Analysis 97, 810–826.

Zuo Y. 2003. *Projection Based Depth Functions and Associated Medians*, The Annals of Statistics 31(5), 1460–1490.

Zuo Y., Serfling R. 2000. *General Notions of Statistical Depth Function*, The Annals of Statistics 28, 461–482.

Zuo Y. 2004. *Robustness of Weighted Lp — Depth and Lp — Median*, AStA 88, 215–234.

Zuo Y. 2003. *Projection Based Depth Functions and Associated Medians*, The Annals of Statistics 31(5), 1460–1490.

Zuo Y., Cui H., Young D. 2004. *Influence Function and Maximum Bias of Projection Depth Based Estimators*, The Annals of Statistics 32(1), 189–218.

Zuo Y., Cui H., Young D. 2004. *Influence Function and Maximum Bias of Projection Depth Based Estimators*, The Annals of Statistics 32(1), 189–218.

Zuo Y., Serfling R. 2000. *Nonparametric Notions of Multivariate „Scatter Measure" and „More Scattered" Based on Statistical Depth Function*, Journal of Multivariate Analysis 75, 62–78.