Statystyczne funkcje głębi w odpornej analizie statystycznej strumieni danych ekonomicznych

Daniel Kosiorowski, Katedra Statystyki, UEK w Krakowie

UW, Warszawa 15.01.2013

Dziękujemy za wsparcie finansowe w postaci grantu NCN DEC-011/03/B/HS4/01138

“Whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways.”

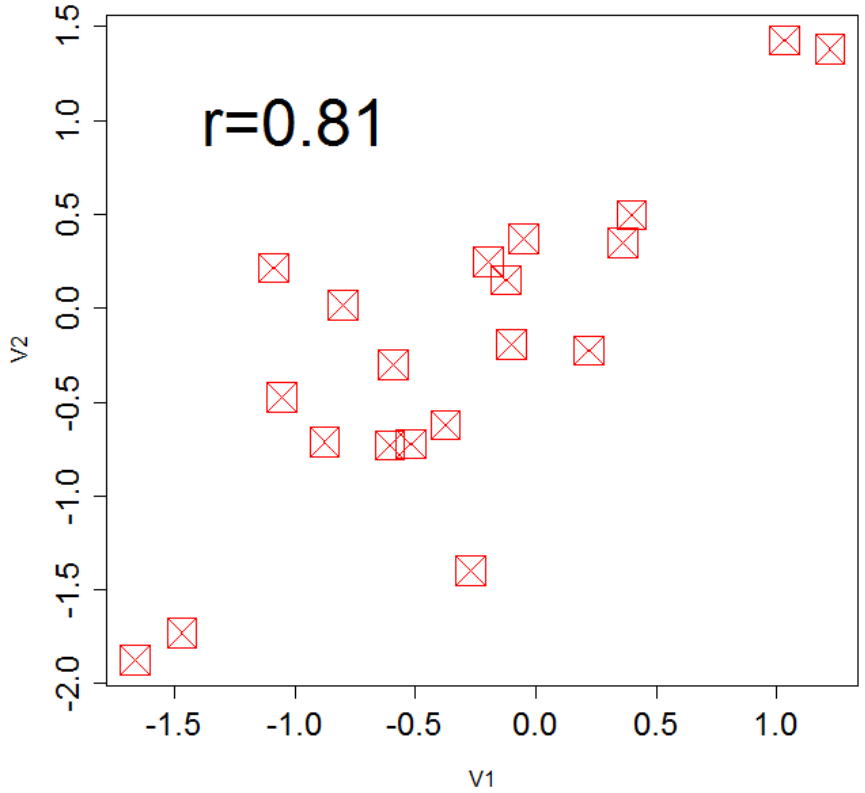
Sir Fransis Bacon (1620)

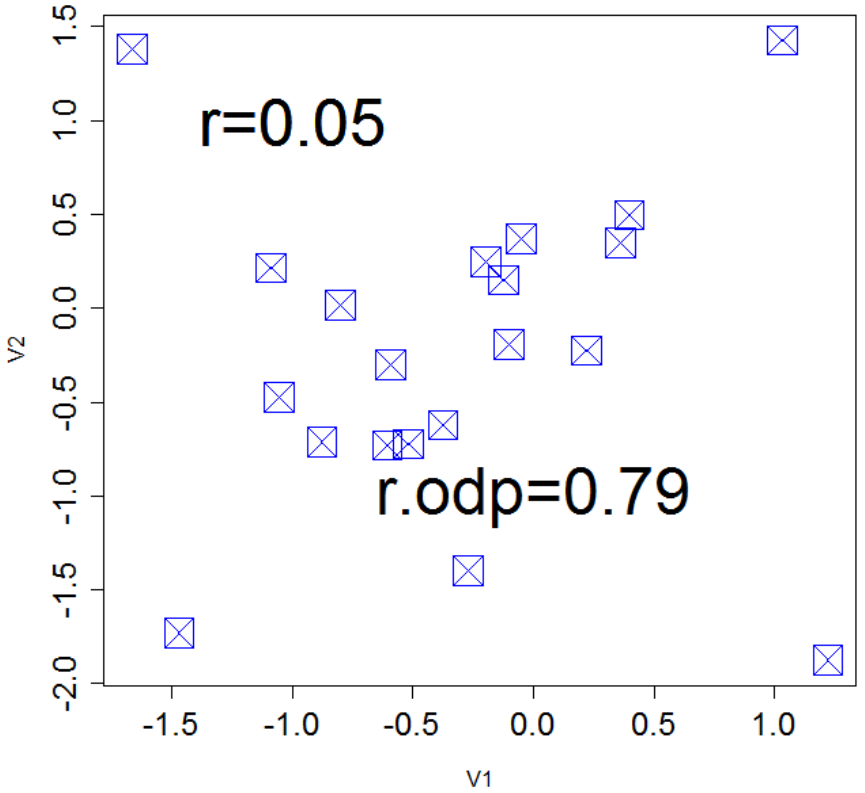
The method of the least squares is seen to be our best course when we have thrown overboard a certain portion of our data – a sort of sacrifice which has often to be made by those who sail the stormy seas of Probability.

Francis Ysidoro Edgeworth (1887)

A tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which were optimal under strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians.

John W. Tukey (1960)





Dwadzieścia obserwacji z dwuwymiarowego rozkładu normalnego współczynnik korelacji wynosi 0.8. Dwie obserwacje (10% danych) zastąpiono obserwacjami odstającymi – współczynnik korelacji wynosi teraz 0.05. Jeżeli zastosujemy odporny estymator wsp. korelacji, to otrzymamy r.odp=0.87 w pierwszym przypadku i r.odp= 0.79 w drugim przypadku (z 10% obserwacji odstających).

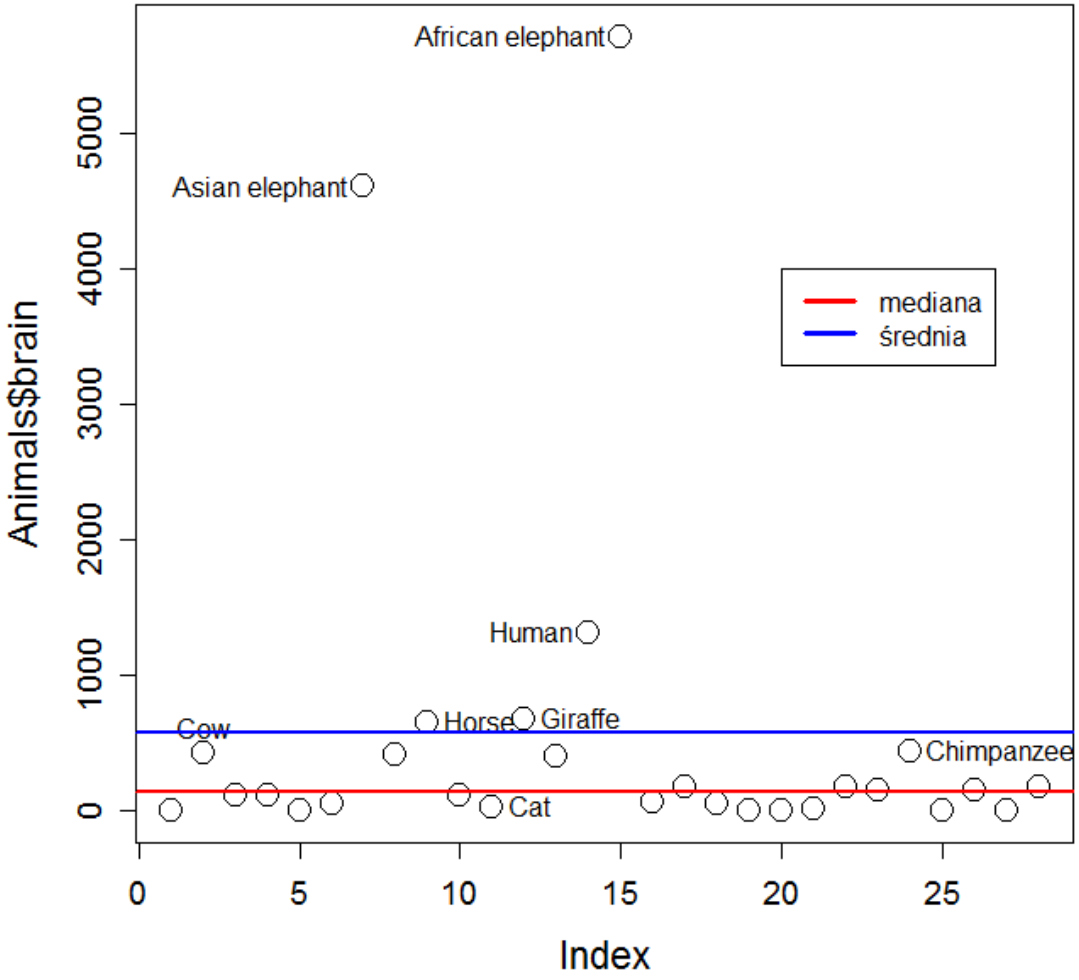
Przykład R

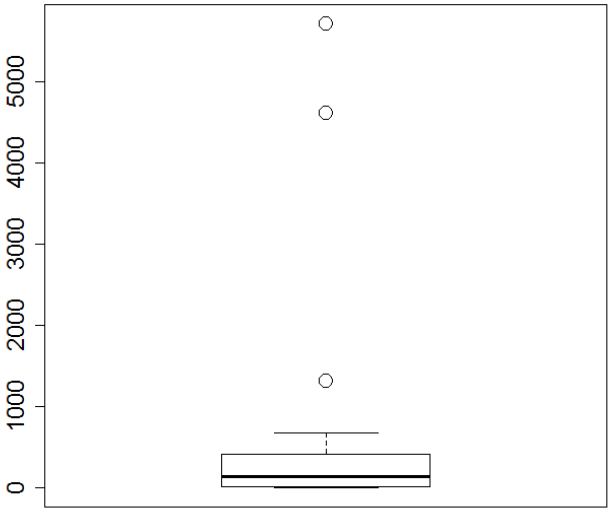
```
library(MASS)
data(Animals, package="MASS")
```

Przeciętne wagi mózgu i ciała dla 28 gatunków zwierząt lądowych

Źródło: P. J. Rousseeuw and A. M. Leroy (1987) *Robust Regression and Outlier Detection.* Wiley,

```
plot(Animals$body, type="p", cex=2,cex.lab=1.4,cex.axis=1.2)
identify(Animals$body, labels=rownames(Animals))
mean(Animals$body)
median(Animals$body)
abline(h=mean(Animals$body), col="blue",lwd=2)
abline(h=median(Animals$body), col="red",lwd=2)
legend(20,40000,c("mediana","średnia"),col=c("red","blue"),lwd=3)
```



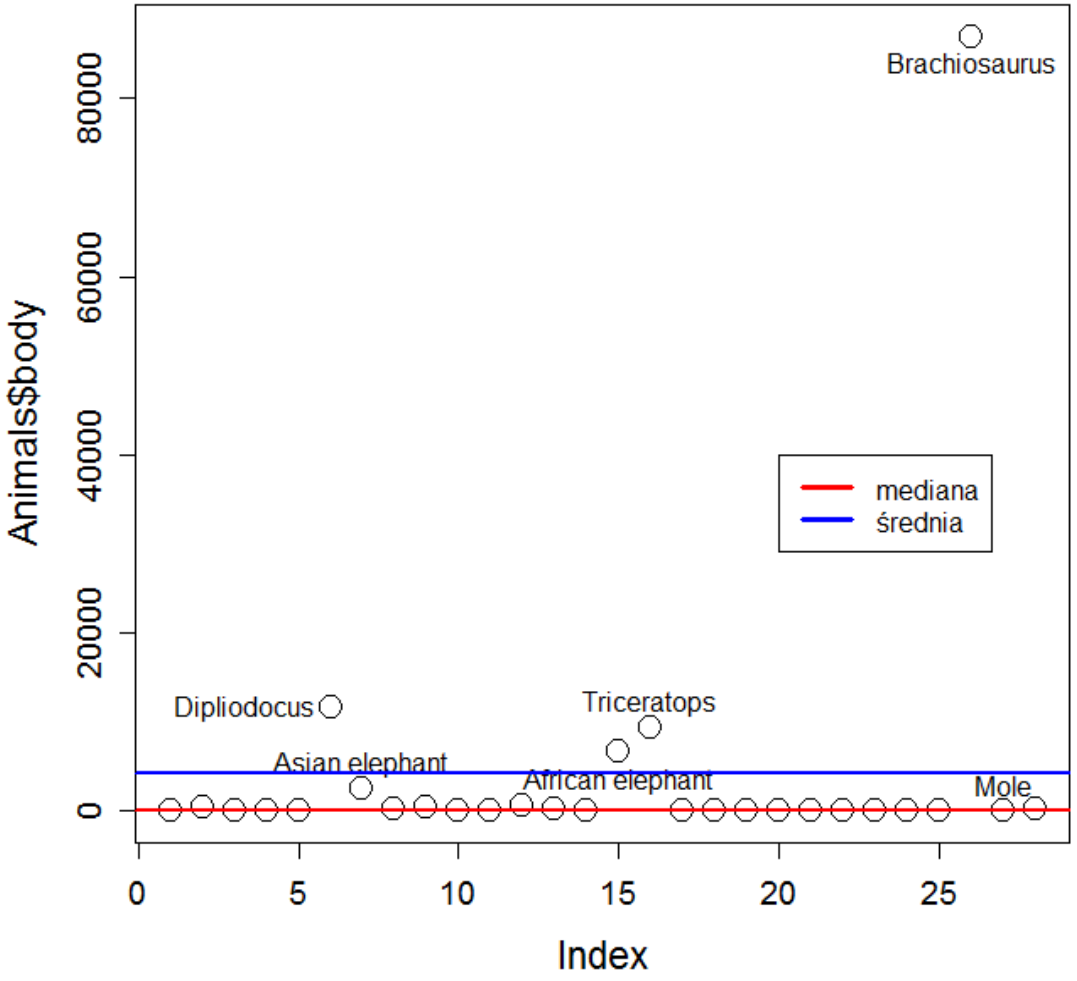


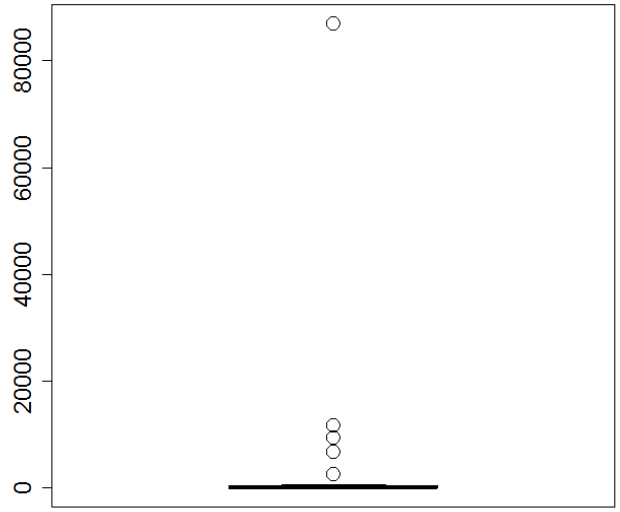
Średnia= 574.5214

Mediana= 137

SD= 1334.929

MAD= 193.0345





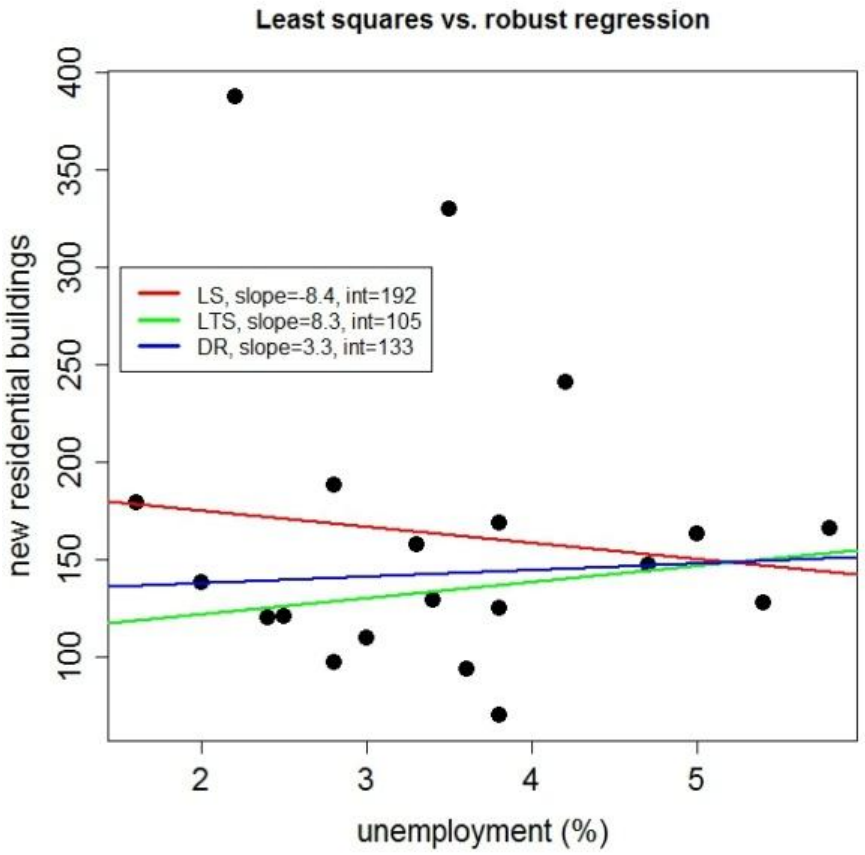
Średnia= 4278.44

Mediana=53.83

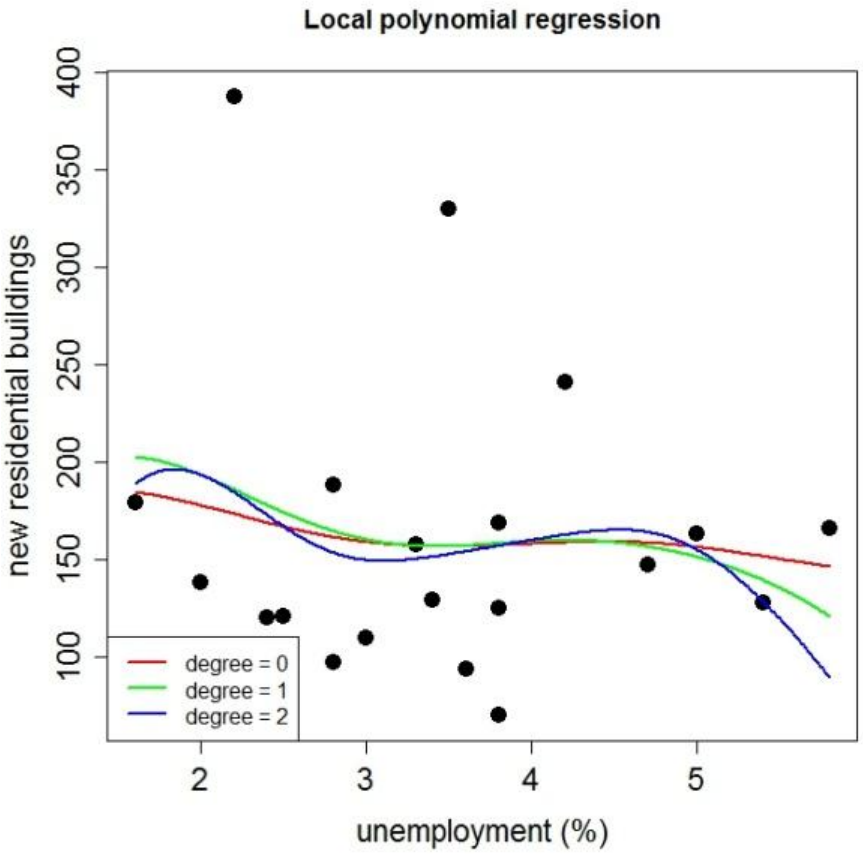
SD= 16480.49

MAD= 79.516

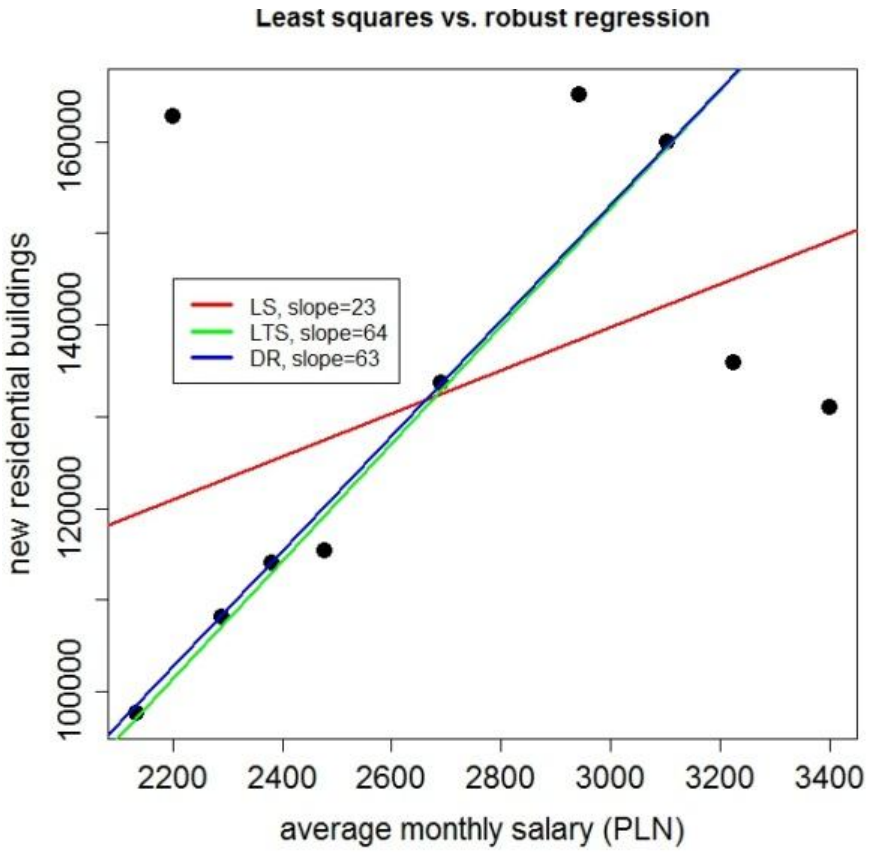
New residential buildings vs. unemployment rate in “near Cracow” districts of Poland in 2007 – 2011.



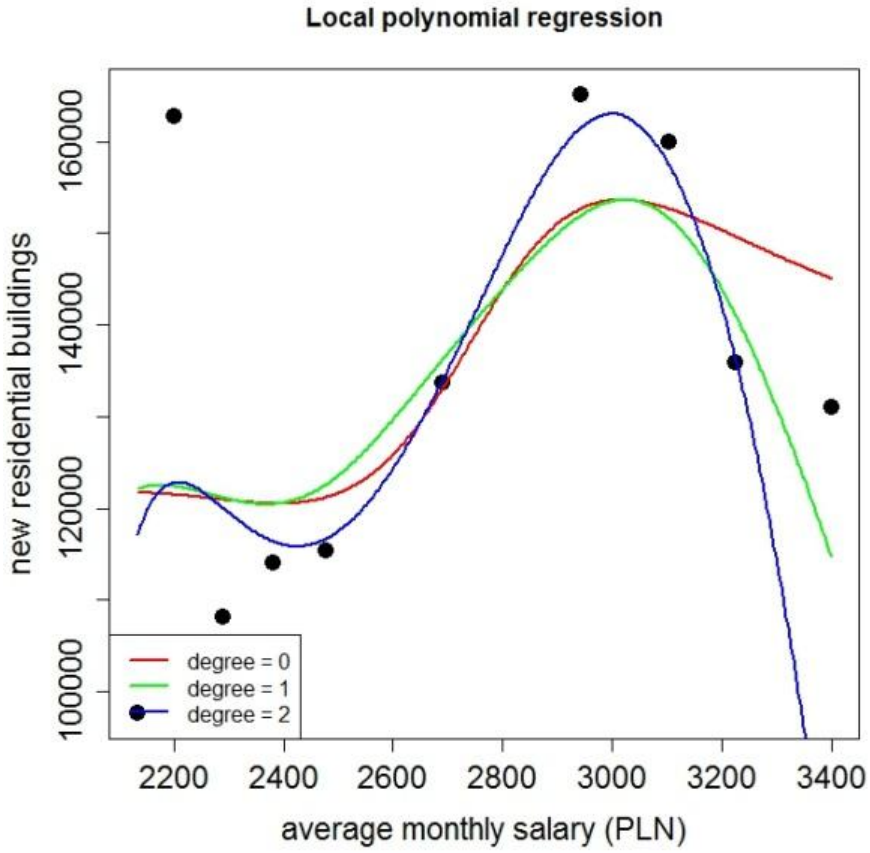
New residential buildings vs. unemployment rate in “near Cracow” districts of Poland in 2007 – 2011.



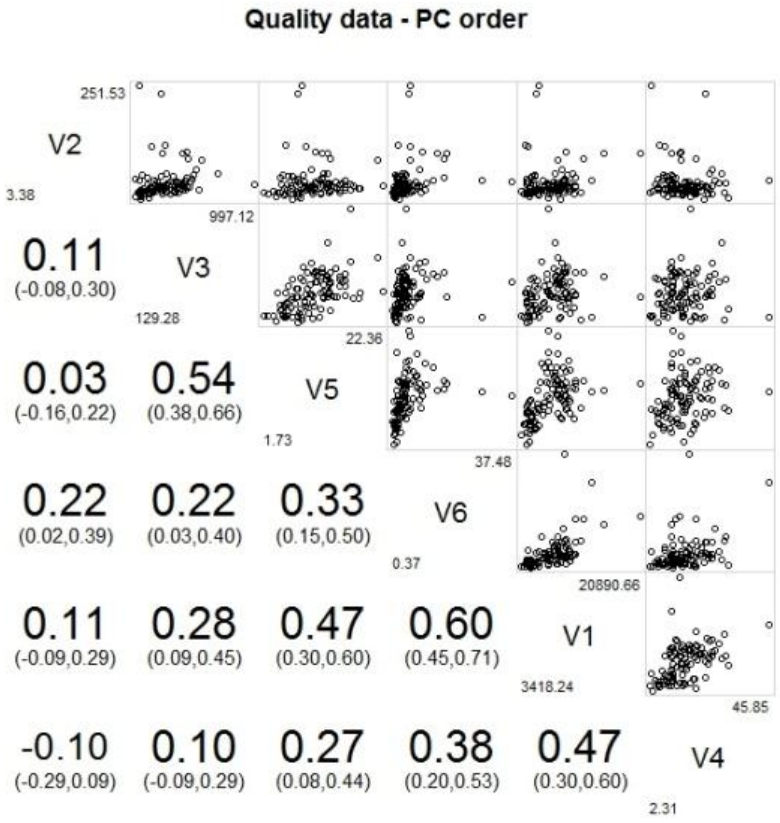
New residential buildings vs. average monthly salary in polish voividships in 2011.



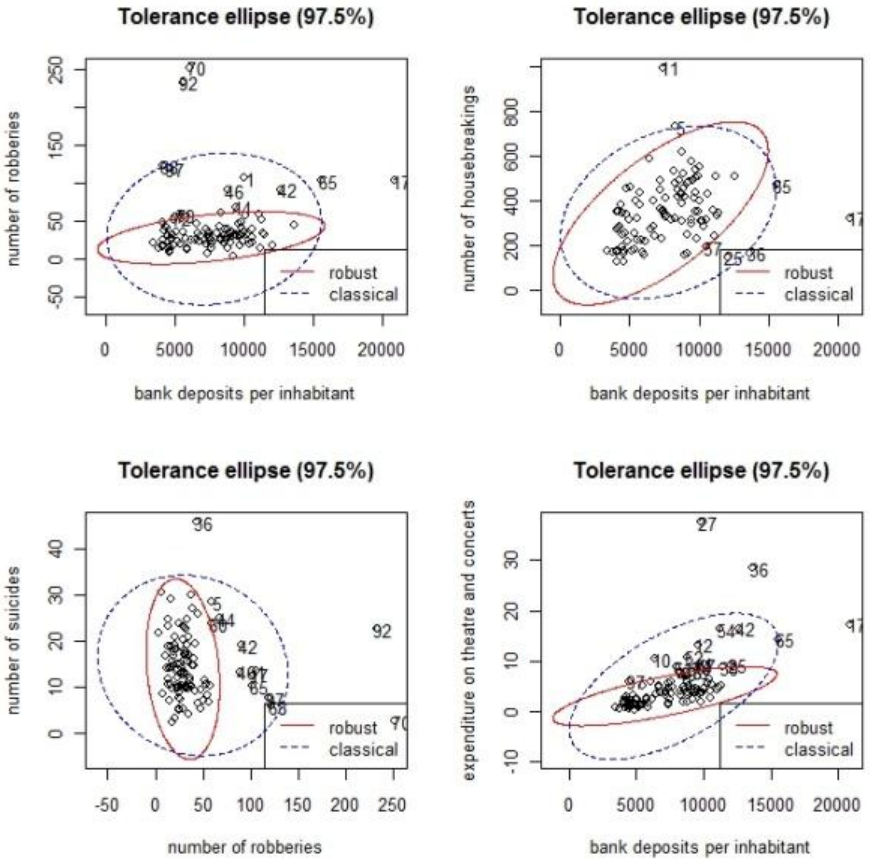
New residential buildings vs. average monthly salary in polish voividships in 2011.



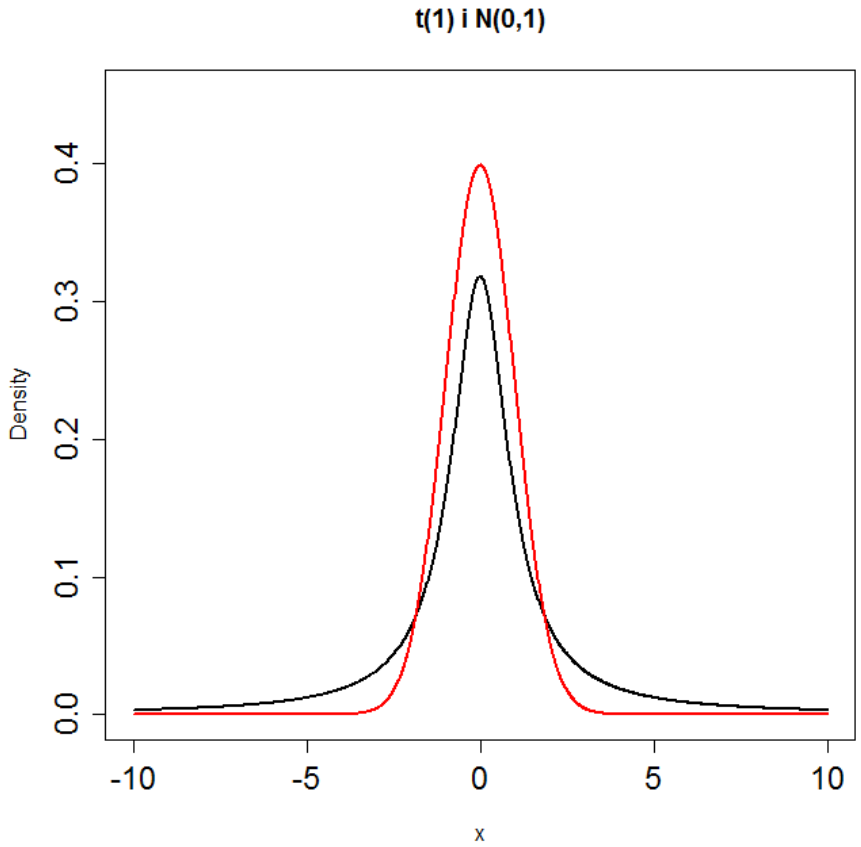
Classical estimate of the correlation matrix for the quality of life data.



Robust (MCD) estimate of the correlation matrix for the quality of life data.



PIERWSZE STARCIE ZE SZCZEGÓŁAMI STATYSTYKI ODPORNEJ



W przypadku symetrycznych rozkładów 1d mediana równa się wartości oczekiwanej

Dla rozkładu normalnego $N(m,\sigma^2)$

Średnia z próby n – elementowej ma rozkład $\bar{X}_n \sim N\left(m,\frac{\sigma^2}{n}\right)$

Mediana z próby n – elementowej $Med \sim N\left(m,\frac{\sigma^2}{n}\frac{\pi}{2}\right)$

Asymptotyczna efektywność względna mediany do średniej

$$ARE(Med,\bar{X}) = \lim_{n\to\infty}\frac{Var(\bar{X}_n)}{Var(Med)} = \frac{2}{\pi} = 0.6366$$

Dla rozkładu Cauchy'ego $C(m,\sigma^2)$

$$f(x;m,\sigma)=\frac{1}{\pi\sigma}\left(1+\left(\frac{x-m}{\sigma}\right)^2\right)^{-1/2}$$

(m – parametr położenia, σ – parametr rozrzutu)

Rozkład średniej z próby $\bar{X}\sim C(m,\sigma^2)$

Rozkład mediany z próby $Med\sim N\left(m,\frac{\pi^2\sigma^2}{4n}\right)$

$$ARE(Med,\bar{X})=\infty,\ ARE(\bar{X},Med)=0$$

t – Student(n)	$n\le 2$	**3**	**4**	**5**
$ARE(Med,\bar{X})$	∞	1.621	1.125	0.96
$ARE(\bar{X},Med)$	0	0.617	0.888	1.041

Dla mieszaniny dwóch rozkładów normalnych

$$X \sim \begin{cases} N(m,\sigma^2) & prawd = 1-\varepsilon \\ N(m,(3\sigma)^2) & prawd = \varepsilon \end{cases}$$

co oznacza, że np. nie wszystkie pomiary odznaczają się tą samą precyzją

$$X \sim (1-\varepsilon)N(m,\sigma^2) + \varepsilon N(m,(3\sigma)^2)$$

- Dla $\varepsilon > 0.10$ $ARE(Med, \bar{X}) > 1$

- MAD jest bardziej efektywna aniżeli odchylenie standardowe dla $\varepsilon > 0.01$
- Odchylenie absolutne jest dwa razy bardziej efektywne niż odchylenie standardowe dla $\varepsilon = 0.05$

JAK ROZUMIEĆ ODPORNOŚĆ PROCEDURY STATYSTYCZNEJ

Procedura statystyczna to pewien algorytm, dla którego wejście mogą stanowić dane, wyjściem może być szerokie spektrum obiektów takich jak liczby, wykresy, obrazy.

DANE = SYGNAŁ + SZUM

Analiza statystyczna może obejmować wiele procedur statystycznych.

ODPORNOŚĆ: Niewielkie zmiany danych powinny skutkować niewielkimi zmianami wyników działania procedury. Niewielkie zmiany rozumiemy, jako małe zmiany wartości obserwacji bądź duże zmiany wartości niewielkiej frakcji danych.

PROBLEMY KTÓRE NAPOTYKAMY W BADANIACH STATYSTYCZNYCH

- Staramy się dotrzeć do systematycznego składnika danych (sygnału), oddzielić ten składnik od losowego zaburzenia, błędu itd.
- Staramy się właściwie wybrać model generujący dane – uniknąć tzw. błędu specyfikacji

Celem statystyki odpornej przedstawienie zgodnych, efektywnych estymatorów, testów statystycznych o stabilnych poziomach błędów, w przypadku, gdy mamy do czynienia z niewielkim błędem specyfikacji.

Przez **niewielki błąd specyfikacji** rozumiemy, że mechanizm generujący dane leży w sąsiedztwie prawdziwego (postulowanego) modelu, takiego modelu, który wydaje się nam użyteczny w danym badaniu.

JAK MOŻEMY ROZUMIEĆ SĄSIEDZTWO ZAKŁADANEGO MODELU?

Dla przykładu *sąsiedztwo modelu* możemy ująć za pomocą mieszaniny rozkładów

$$F_\varepsilon = (1-\varepsilon)F_\theta + \varepsilon G$$

- F_θ – to postulowany model
- θ – zbiór interesujących nas parametrów
- G – dowolny rozkład „zaburzenie”
- $0 \le \varepsilon \le 1$ „rozmiar błędu specyfikacji”

***Zagadnienie trudniejsze**: sąsiedztwo modelu procesu stochastycznego?*

CELE STATYSTYKI ODPORNEJ – RAZ JESZCZE…

- Przedstawić opis zasadniczej części danych
- Zidentyfikować obserwacje odbiegające od zasadniczego wzorca danych reprezentowanego przez ich większość
- Zidentyfikować i ostrzec przez wysoce wpływowymi obserwacjami
- Zaproponować metody radzące sobie z autokorelacją, heteroskedastycznością, skośnością błędu itd.

POMIAR ODPORNOŚCI PROCEDURY – PIERWSZE STARCIE

Zbiór danych $x_1,\ldots,x_{n-1}$

Statystyka $T_{n-1} = T(x_1,\ldots,x_{n-1})$

Zanieczyszczony zbiór danych $x_1,\ldots,x_{n-1},\mathbf{x}$

Statystyka dla tego zbioru $T_n = T(x_1,\ldots,x_{n-1},\mathbf{x})$

KRZYWA WRAŻLIWOŚCI (wprowadzona przez Tukey'a)

$$SC_n(\mathbf{x}) = n(T_n - T_{n-1})$$

Zauważmy $T_n = T_{n-1} + \frac{1}{n} SC_n(\mathbf{x})$

FUNKCJA WPŁYWU - WERSJA KRZYWEJ WRAŻLIWOŚCI DLA POPULACJI

FUNKCJA WPŁYWU HAMPELA

Rozważmy rozkład $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_{\mathbf{x}}$,

gdzie $\delta_{\mathbf{x}}$ oznacza rozkład skoncentrowany w punkcie $\mathbf{x}$

- Możemy porównać $T(F)$ vs. $T(F_\varepsilon)$
- Możemy zdefiniować jakościową odporność (ciągłość) statystyki T

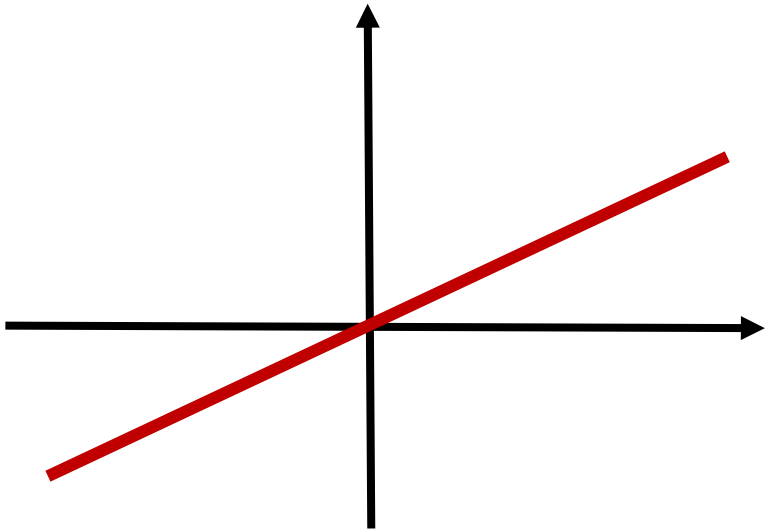
$$T(F_\varepsilon) \to T(F) \text{ gdy } \varepsilon \to 0$$

FUNKCJĘ WPŁYWU DEFINIUJEMY JAKO

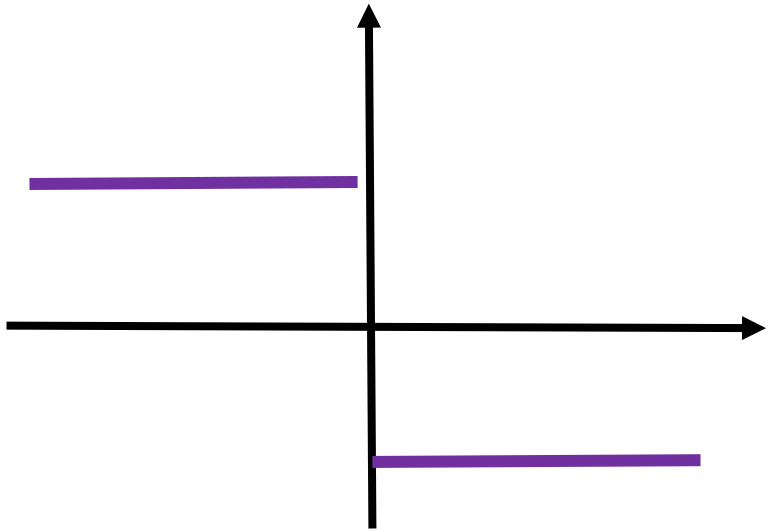
$$IF(\mathbf{x};T,F) = \lim_{\varepsilon\to\infty} \frac{T(F_\varepsilon) - T(F)}{\varepsilon}$$

Okazuje się, że ma miejsce $T_n \approx T_{n-1} + \frac{1}{n} IF(\mathbf{x};T,F)$

- Funkcja wpływu jest lokalną miarą odporności
  - Funkcja wpływu powinna być ograniczona



Funkcja wpływu średniej



Funkcja wpływu mediany

Przykłady odpornych estymatorów położenia

$\alpha-$przycięta średnia: odrzuć $\alpha/2$ największych i $\alpha/2$ najmniejszych obserwacji – policz zwykłą średnią dla pozostałych obserwacji

$\alpha-$ średnia Windsora: zastąp $\alpha/2$ największych i $\alpha/2$ najmniejszych obserwacji odpowiednio najbliższą mniejszą i najbliższą większą obserwacją – policz zwykłą średnią

Przykład: 1,2,3,4,100

20% przycięta średnia = (2+3+4)/3

20% średnia Windsora = (2+2+3+4+4)/3

PUNKT ZAŁAMANIA PRÓBY SKONCZONEJ ESTYMATORA (BP) – najmniejsza frakcja złych obserwacji w próbie, która sprawia, że estymator staje się bezużyteczny – np. jego obciążenie staje się zbyt wysokie.

PUNKT ZAŁAMANIA PRÓBY SKONCZONEJ ESTYMATORA ROZRZUTU – najmniejsza frakcja złych obserwacji w próbie sprawiających, że estymator wskazuje zero bądź nieskończoność.

BP dla odchylenia standardowego wynosi $1 / n \approx 0$

BP dla rozstępu międzykwartylowego $IQR = 0.74 \left| x_{(\lfloor 0.75*n \rfloor)} - x_{(\lfloor 0.25*n \rfloor)} \right|$

wynosi $\approx 25\%$

BP dla mediany odchyleń absolutnych od mediany $MAD = 1.48 * med_i \left| x_j - med_j x_j \right|$

wynosi $\approx 50\%$

(stałe 0.74 i 1.48 gwarantują, że w przypadku danych generowanych przez rozkład normalny estymatory SD, IQR i MAD szacują to samo)

STATYSTYKA ODPORNA I ŚRODOWISKO R

Pakiety {robustbase}, {rrcov}, {MASS},…,{depth}, {depthproc}

Uwaga: w tzw. komercyjnych pakietach statystycznych na ogół odporny estymator = tzw. M estymator (uogólnienie metody największej wiarygodności)

ODPORNA ANALIZA STRUMIENI DANYCH

Współczesna gospodarka w sposób ciągły generuje gigantyczne zbiory danych (truizm?)

Analiza strumienia danych (strumieniowe przetwarzanie danych). Analiza taka przykładowo może obejmować monitorowanie setek tysięcy finansowych szeregów czasowych w celu znalezienia użytecznych inwestycyjnie zależności pomiędzy nimi, analizę danych generowanych przez stacje pogodowe w pewnym obszarze oceanu, monitorowanie centrum miasta za pomocą systemu kamer, decydowanie co do podjęcia interwencji na rynku zbóż w oparciu o dane dostarczane przez giełdy towarowe.

Ujmując zagadnienie nieprecyzyjnie możemy określić **strumień danych** jako „***nieokreślonej długości ciąg z reguły wielowymiarowych obserwacji***” (por. Szewczyk 2010).

W przypadku tradycyjnie rozumianej **analizy procesu stochastycznego**, powiedzmy $\{X_t\}$, zakładamy ustalony przedział czasowy, powiedzmy $[0,T]$. Nasze obliczenia dotyczą tego przedziału a więc wnioskujemy na podstawie informacji uzyskanej do chwili T.

W przypadku analizy strumienia danych nie ustalamy przedziału badania $[0,T]$. **Każda kolejna chwila oznacza nową analizę procesu stochastycznego**.

Strumieniowe przetwarzanie danych, analizę strumienia danych można określić, jako sekwencję analiz procesu stochastycznego.

Terminologia wywodzi się z informatyki, gdzie tego typu zagadnienia były rozważane po raz pierwszy. Statystycy zajmują się strumieniami danych od niedawna (por. Huber 2011)

W literaturze dotyczącej analizy strumieni danych, strumieniowego przetwarzania danych w zasadzie nie podaje się wprost odwołania do probabilistycznego modelu danych. Jednakże wczytując się w tę literaturę można pokusić się o stwierdzenie, że analiza taka jest w istocie rodziną analiz procesu stochastycznego odznaczających się następującymi cechami:

1. Obserwacje generowane są przez proces, w którym ma miejsce nieliniowa zależność teraźniejszości od przeszłości.

2. Obserwacje modeluje się na ogół przez proces niestacjonarny, którego nie da się sprowadzić do procesu stacjonarnego za pomocą różnicowania, usunięcia deterministycznego trendu. Proces na ogół odznacza się występowaniem pewnej ilości reżimów. Typ niestacjonarności, liczba i charakterystyki reżimów mogą zmieniać się w czasie.

3. Analizę strumienia prowadzimy opierając się na stale aktualizowanej próbie – na podstawie ustalonej długości ruchomego okna (można rozważać okna różnej długości dla różnych skal czasu – sekund, minut, dni itd.). Na podstawie takiej stale aktualizowanej próby podejmujemy decyzje, na jej podstawie monitorujemy położenie, rozrzut strumienia.

4. Strumienie na ogół liczą setki tysięcy wielowymiarowych obserwacji. Z reguły dane z racji swej wielkości nie są magazynowane w pamięci komputera – muszą być przetwarzane na bieżąco (ang. *on–line processiong)*.

5. Dane napływają do obserwatora z reguły w nierównych odstępach czasu, w pakietach nierównej wielkości. Można założyć, że modelem strumienia jest proces stochastyczny z czasem ciągłym. Wówczas mamy na uwadze sytuację, gdy częstość próbkowania obserwacji ze strumienia jest zmienną losową. Można założyć stosownie skonstruowany proces dyskretny odwołując się np. do teorii procesów podporządkowanych, warunkowych procesów trwania, bądź tak jak w niniejszej pracy wyjść od takiego procesu, który losowo generuje sygnał (odpowiednio zdefiniowany) w chwilach równo od siebie odległych.

6. Do analizy strumieni stosuje się są na ogół procedury nieparametryczne, które muszą spełniać wysokie wymagania w zakresie złożoności obliczeniowej, które muszą radzić sobie z problemem „***rzadkości danych***” (ang. sparsity of the data) w wielu wymiarach (por. Hastie i in. 2009). (*nie mylić z ang. sparse data albo z ang. sparse method*)

W dalszej części zakładamy, że strumień generowany jest przez pewną konkretną postać ogólnego modelu określanego mianem CHARME (por. Stockis i in. 2010).

Rozważamy tym samym **proces stochastyczny** z czasem dyskretnym o ustalonej liczbie reżimów.

Zakładamy, że w obserwowanych przez nas danych występują obserwacje odstające. Mamy tutaj na uwadze sytuację, gdy na badany proces działa tzw. ***addytywny proces odstawania*** **(AO)** (ang. additive outliers process) – przyjmujemy ramy pojęciowe zaproponowane w klasycznym podręczniku Marona i in. (2006).

Niech x_t oznacza proces warunkowo stacjonarny[1], niech v_t oznacza stacjonarny proces odstawania. Niech $P(v_t = 0) = 1 - \varepsilon$, co oznacza, że „niezerowa" część procesu v_t pojawia się z prawdopodobieństwem ε.

W modelu AO, zamiast x_t obserwujemy $y_t = x_t + v_t$ przy czym zakłada się, że procesy x_t i v_t są wzajemnie niezależne.

AO można określić, jako *proces błędów grubych*, obserwacje odstające na ogół są izolowane.

[1] Mówimy, że jednowymiarowy proces jest warunkowo stacjonarny, jeżeli jego rozkładu warunkowe są niezmiennicze względem przesunięć w czasie (por. Shalizi, Kantorovich, 2007 def. 51 str. 35).

W referacie skupiamy naszą uwagę na procesie podejmowania decyzji na podstawie stale uaktualnianej niewielkiej próby ze strumienia.

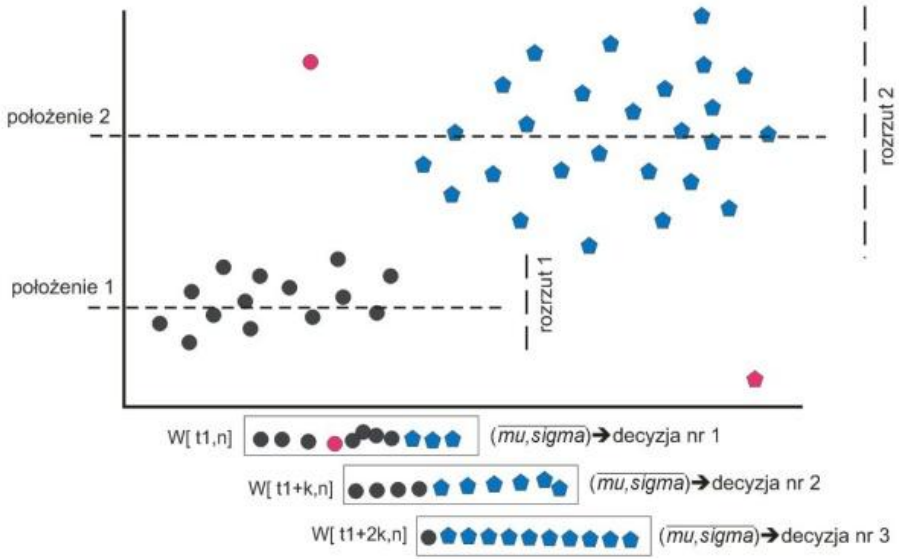
Decyzje dotyczą m.in. prognozowania kolejnych wartości strumienia, prognozowania i monitorowania charakterystyk rozrzutu, położenia i skośności, (bezwarunkowych i warunkowych względem obserwowanej próby w przeszłości), monitorowania zależności pomiędzy teraźniejszością i przeszłością strumienia.

Naszym zadaniem jest stworzenie stosownych narzędzi umożliwiających nam odczytanie sygnału zawartego w strumieniu w sytuacji występowania obserwacji odstających. Należy jednakże podkreślić, że w przeciwieństwie do nauk inżynieryjnych (dane = deterministyczny sygnał + losowy szum) przez sygnał rozumiemy relację pomiędzy charakterystykami liczbowymi probabilistycznego modelu danych[2].

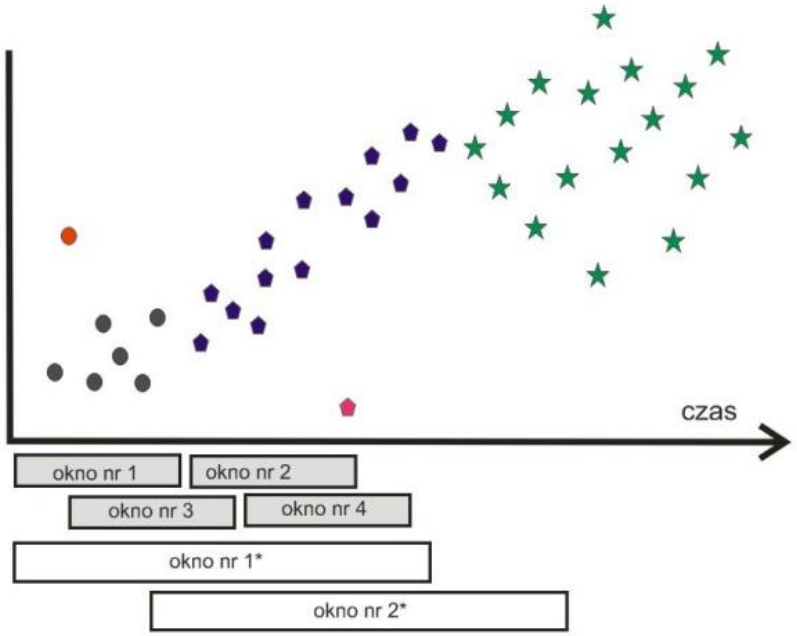
[2] W niniejszym opracowaniu zakładamy, że dane generuje pewien niestacjonarny proces stochastyczny. Sygnał utożsamiamy z charakterystykami liczbowymi jego modelu(li). Jednakże o ile zmienimy rozumienie sygnału – można rozważać strumienie generowane przez procesy stacjonarne bądź układy stricte deterministyczne. W kontekście zastosowań tematyki w ekonomii – przyjęte ramy wydają się być najwłaściwsze.

W zasadzie w przyjętych przez nas dalej ramach pojęciowych odczytanie sygnału wiążemy ze wskazaniem reżimu procesu generującego strumień.

Ilustracja zagadnienia decydowania co do zmiany położenia – rozrzutu na podstawie ruchomego okna.



Trzy reżimy strumienia danych. Dane zawierają obserwacje odstające, rozważamy ruchome okna różnej długości



Odporność naszych propozycji rozumiemy w duchu jednolitego i ogólnego podejścia Gentona i Lucasa (2003) jako odporność reguły decyzyjnej określonej na stale uaktualnianej próbie ze strumienia (za punkt odniesienie bierzemy np. medianę w przestrzeni decyzji, rozważamy różne funkcje straty np. LINEX).

Według Gentona i Lucasa (2003) **krytyczna cecha estymatora sprowadza się do tego, że ten przyjmuje różne wartości dla różnych realizacji próby. Jeżeli możliwe jest kontinuum prób a estymator jest ciągły, to oczekujemy kontinuum jego wartości**.

Załamanie estymatora polega na tym, że ta jego własność zanika, estymator przyjmuje jedynie skończoną liczbę różnych wartości pomimo kontinuum możliwych prób.

Można umownie wyróżnić **dwa nurty podejść do analizy strumieni danych** – nurt związany z metodami eksploracyjnej analizy bardzo wielkich zbiorów danych (ang. very big high-dimensional data mining) oraz nurt związany z klasyczną **nieparametryczną analizą szeregów czasowych** (por. Fan, Yao, 2005).

W obrębie pierwszego nurtu (por. Aggerwal, 2007) wyróżnić można m. in.: dynamiczną redukcję wymiaru zagadnienia za pomocą tzw. **mikro-skupisk, badanie dynamicznych klasyfikacji, stosowanie adaptacyjnej metody najbliższych sąsiadów, wykorzystywanie drzew regresyjnych i klasyfikacyjnych, wykorzystanie sieci neuronowych, sieci bayesowskich.**

Drugi nurt wiąże się z adaptacjami **metod nieparametrycznej analizy szeregów czasowych**. Mamy tutaj na uwadze adaptacje lokalnej liniowej, lokalnej wielomianowej regresji w tym szereg wariantów nieparametrycznej *regresji Nadaraya-Watsona* (patrz Hall i in., 1999), metody wykorzystujące wielomiany ortogonalne, regresję nieliniową z ograniczeniami (np. metody LOESS, LASSO por. Hastie i in., 2009), sklejki itd.

Należy podkreślić, że w przypadku analizy strumieni danych na ogół wielowymiarowych niezmiernie istotne jest, aby procedura radziła sobie z tzw. *„przekleństwem wielowymiarowości" – rzadkością danych* (ang. sparsity ot the data) *w wielu wymiarach*. Owo przekleństwo sprawia m.in., że dla przykładu dobre statystyczne własności jednowymiarowej regresji Nadaraya-Watsona zanikają w wielu wymiarach, istotność statystyczna oszacowań wielowymiarowych modeli stosowanych w empirycznych finansach budzi poważne wątpliwości (por. Kosiorowski, Snarska, 2012).

W literaturze jak dotychczas nie jest znanych wiele odpornych metod analizy strumieni danych. Wiąże się to między innymi z trudnościami z rozumieniem odstawania w przypadku strumieni generowanych przez model o wielu reżimach.

Pojawia się dla przykładu pytanie *czy rozumienie odstawania powinno się w takim przypadku wiązać z konkretnym reżimem procesu*? Co ciekawe w przypadku analizy strumienia z jednostkami odstającymi stosowana procedura powinna być odporna, jednak nie bardzo odporna (tzn. jej punkt załamania nie powinien osiągać maksymalnej możliwej wartości) – tak, aby pomijała wpływ obserwacji odstających, lecz jednocześnie była wrażliwa na zmianę reżimu modelu.

MODEL STRUMIENIA DANYCH EKONOMICZNYCH ORAZ PROBLEMY ZWIĄZANE Z ANALIZĄ STRUMIENI DANYCH

W literaturze nie jest znanych wiele modeli strumienia danych, do nielicznych należy zaliczyć propozycję Hahsler i Dunhamr (2010), w której rozważa się zmienny w czasie łańcuch Markowa dla mikro skupisk.

Wydaje się jednak, że model strumienia danych można skonstruować na podstawie jednego z wykorzystywanych w ekonometrii modeli dla zjawisk o zmiennym reżimie np. model VTAR (ang. vector treshold auroregressive model) bądź jego nieliniową wersję VFAR (ang. functional vector autoregressive model) (por. Fan, Yao, 2005).

Niech $\mathbf{X}_1 = (X_{11},...,X_{1d})$, $\mathbf{X}_2 = (X_{21},...,X_{2d})$,..., oznacza d-wymiarowy strumień danych $d \geq 2$ oraz niech $\mathbf{x}_1 = (x_{11},...,x_{1d})$, $\mathbf{x}_2 = (x_{21},...,x_{2d})$,..., $\mathbf{x}_T = (x_{T1},...,x_{Td})$ oznacza zaobserwowane wartości strumienia w punktach 1,...,T. Zdecydowaliśmy się modelować strumień danych za pomocą modelu **VCHARN** (ang. *vector conditional heteroscedastic autoregressive nonparametric*) definiowanego

$$\mathbf{X}_t = m(\mathbf{X}_{t-1},.,\mathbf{X}_{t-p}) + \sigma(\mathbf{X}_{t-1},.,\mathbf{X}_{t-p})\epsilon_t, \qquad (1)$$

gdzie $m(\cdot)$ i $\sigma(\cdot)$ oznaczają dowolne ale ustalone funkcje (np. $m(\mathbf{x}) = E(\mathbf{X}_t \,|\, [\mathbf{X}_{t-1},.,\mathbf{X}_{t-p}] = \mathbf{x})$, $\sigma^2(\mathbf{x}) = diag\{Var(\mathbf{X}_t \,|\, [\mathbf{X}_{t-1},.,\mathbf{X}_{t-p}] = \mathbf{x})\}$, gdzie $\mathbf{x} = (\mathbf{x}_{t-1},.,\mathbf{x}_{t-p})$, $diag\{Var(\cdot)\}$ oznacza wektor elementów diagonalnych macierzy wariancji-kowariancji oraz ϵ_t oznacza niezależne innowacje o tym samym rozkładzie i wartości oczekiwanej zero (por. Hall i in., 1999).

W kontekście analizy strumieni danych nie zakładamy, że obserwowany proces ma tę samą funkcję trendu m i tę samą funkcję zmienności σ w każdej chwili.

Nie zakładamy, że te funkcje zmieniają się powoli w czasie.

Rozważamy klasę nieparametrycznych modeli szeregów czasowych zawierających obserwacje odstające. Pomiędzy chwilami losowej zmiany obserwowany proces jest względnie stabilny.

W naszym modelu liczba jednostek odstających w danej chwili jest losowa i jest jedynie ograniczona wg. prawdopodobieństwa. Obserwacje odstające niekoniecznie pojawiają się niezależnie, ich wielkość też może być losowa.

Skupiamy naszą uwagę na modelu **CVHARME** (ang. *Conditional Vector Heteroscedastic Autoregressive Mixture of Experts*) (por. Stockis i in., 2010). CVHARME jest ogólną metodą modelowania szeregów czasowych o wielu reżimach. CVHARME obejmuje m. in. wiele znanych liniowych i nieliniowych modeli np. modele

VAR, VTAR (ang. vector threshold models), wielowymiarowe modele GARCH. Dynamiką modelu **CVHARME** $\{\mathbf{X}_t\}$ rządzi **ukryty łańcuch** Markowa $\{Q_t\}$ na skończonym zbiorze stanów $\{1,2,\ldots,K\}$ w następujący sposób:

$$\mathbf{X}_t = \sum_{k=1}^{K} S_{tk}(m_k(\mathbf{X}_{t-1},..,\mathbf{X}_{t-p}) + \sigma_k(\mathbf{X}_{t-1},..,\mathbf{X}_{t-p})\epsilon_t) + b_t\Theta_t, \qquad (2)$$

gdzie $S_{tk} = 1$ dla $Q_t = k$ oraz $S_{tk} = 0$ w przeciwnym wypadku, m_k, σ_k, $k = 1,\ldots,K$, oznaczają dowolne lecz ustalone funkcje, ϵ_t oznaczają niezależne zmienne losowe o tym samym rozkładzie i wartości oczekiwanej zero, człon $b_t\Theta_t$ oznacza składnik związany z obserwacjami odstającymi, b_t jest nieobserwalną binarną zmienną losową wskazującą pojawienie się obserwacji odstającej w chwili t, a Θ_t oznacza wartość obserwacji odstającej, $\boldsymbol{\Theta}_t \sim N_d(\mathbf{m},\boldsymbol{\Sigma})$. Dla uniknięcia tzw. przekleństwa wielowymiarowości proponujemy przyjąć $p = 1$ albo $p = 2$.

W naszych rozważaniach, Q_t zmienia się rzadko, tzn. obserwujemy ten sam reżim przez względnie długi okres czasu zanim nastąpi jego zmiana.

Niech $\mathbf{x}_1, \mathbf{x}_2, \ldots$ oznacza d-wymiarowy strumień obserwacji generowany przez model (2). Okno $\mathbf{W}_{i,n}$ oznacza ciąg obserwacji kończący się na obserwacji $\mathbf{x}_i$ o wielkości n, tzn. $\mathbf{W}_{i,n} = (\mathbf{x}_{i-n+1}, \ldots, \mathbf{x}_i)$. Czasem wygodnie jest rozważać okna postaci $\mathbf{W}_{[i,j]}$ - podciągi strumienia danych pomiędzy obserwacją i a j. Duża część technik monitorowania strumieni danych opiera się na analizie odległości pomiędzy rozkładami wyznaczonymi na podstawie punktów znajdujących się w dwóch bądź większej ilości oknach $\mathbf{W}_t$ i $\mathbf{W}_{t'}$.

PROBLEM 1: Rozważamy sytuację gdy w oparciu o uaktualnianą próbę (ruchome okno) ze strumienia $\mathbf{W}_{i,n}, \mathbf{W}_{i+1,n}, \ldots$, przewidujemy kolejne wartości strumienia $\hat{\mathbf{x}}_{i+1}, \hat{\mathbf{x}}_{i+2}, \ldots$, albo kolejne okna $\hat{\mathbf{W}}_{i+1,k}, \hat{\mathbf{W}}_{i+2,k}, \ldots$, $k \ll n$. Naszym celem jest wskazanie optymalnej procedury w sytuacji gdy dane zawierają obserwacje odstające.

PROBLEM 2: Na podstawie monitorowania ruchomego okna $\mathbf{W}_{i,n}$, $i = 1,2,\ldots$ zamierzamy wykryć bezwarunkowe zmiany w modelu generującym dane. Zakładając model postaci (2), naszym celem jest wykryć reżim Q_k (stan ukrytego łańcucha Markowa), i w konsekwencji funkcji m_k i σ_k pojawiających się w (2).

PROBLEM 3: Monitorujemy d-wymiarowy strumień $\mathbf{X}_1,\mathbf{X}_2,\ldots$, a naszym celem jest wykrycie wykrycie zmian rozkładu warunkowego okna $\mathbf{W}_{i+1,n}$, warunkowanego zaobserwowanym oknem $\mathbf{W}_{i,n}$, $i=1,2,\ldots$, tzn. zmian $P(\mathbf{W}_{i+1,n}\in A\,|\,\mathbf{W}_{i,n}=\mathbf{x})$, $A\subset\mathbb{R}^d$, $i=1,2,\ldots$. Zakładając model (2) naszym celem jest wykrycie zmian w macierzy przejścia związanej z ukrytym łańcuchem Q_k, i/lub zmian postaci funkcji m_k i σ_k.

PROBLEM 4: Monitorujemy d-wymiarowy strumień $\mathbf{X}_1,\mathbf{X}_2,\ldots$, a naszym celem jest wykrycie zmian w łącznym (warunkowym) rozkładzie (conditional) distribution of $\mathbf{X}_i$ na podstawie $\mathbf{W}_{i-1,n}$, $i=1,2,\ldots$. W szczególności chcemy wykryć zmiany liniowego związku pomiędzy współrzędnymi $\mathbf{X}_i$.

PROPOZYCJE NAWIĄZUJĄCE DO WIELOWYMIAROWYCH TESTÓW RANGOWYCH

Rozważamy wielowymiarowy strumień danych zawierający obserwacje odstające, który analizujemy w oparciu o ruchome okno obserwacji z tegoż strumienia,

Proponowane metody analizy, monitorowania wielowymiarowego strumienia powinny radzić sobie z tzw. przekleństwem wielowymiarowości (rzadkością danych w wielu wymiarach, por. Hastie i in. 2009) oraz powinny odznaczać się zadowalającą złożonością obliczeniową.

Procedury powinny być odporne jednak co może wydawać się zaskakujące, nie powinny być bardzo odporne.

Oznacza, to że powinny być odporne na występowanie obserwacji odstających będąc jednocześnie wrażliwe na zmiany reżimu strumienia danych.

Procedury statystyczne wywodzące się z tzw. **nieparametrycznej analizy danych** ogólnie rzecz biorąc zakładają gładkość zmian charakterystyk strumienia, nie są odporne i w związku z tzw. przekleństwem wielowymiarowości – rzadkością danych w wielu wymiarach – odznaczają się słabymi własnościami w przypadku wielowymiarowego strumienia danych (por. Hall i in., 1999, Hardle i Simar, 2012).

Procedury wywodzące się z tzw. **machine learning methodology** (por Hastie i in., 2009) jak np. wykorzystywanie wielomianów ortogonalnych, różnego rodzaju sklejki (np. B-sklejki), regresje z rozmaitymi ograniczeniami (np. LASSO, LOESS) nie są odporne, są niezmiernie złożone pod względem obliczeniowym oraz tracą swe dobre własności w przypadku wielowymiarowym.

Procedury obejmujące klasyczną estymację modelu parametrycznego dla każdego okna w celu jego wykorzystania do predykcji bądź monitorowania wymagają pełnej specyfikacji modelu.

KONCEPCJA GŁĘBI DANYCH

Głębia danych to sposób pomiaru głębi bądź odstawania danego punktu względem wielowymiarowego zbioru danych bądź rozkładu prawdopodobieństwa, który wygenerował ten punkt. Zakładając pewien rozkład prawdopodobieństwa F na $\mathbb{R}^d$, funkcja głębi $D(\mathbf{x},F)$ umożliwia porządkownie punktów $\mathbf{x}$ w $\mathbb{R}^d$ na zasadzie odstawania od centrum rozkładu. Dla próby $\mathbf{X}^n = \{\mathbf{x}_1,...,\mathbf{x}_n\}$, wyrażenie $D(\mathbf{x},\mathbf{X}^n)$ oznacza głębię z próby wygenerowanej z rozkładu F, F_n oznacza rozkład empiryczny wyznaczony na podstawie $\mathbf{X}^n$.

Statystyczna funkcja głębi kompensuje brak porządku liniowego w $\mathbb{R}^d$, $d \geq 2$, poprzez orientowanie punktów względem centrum "centrum". Punkty o wyższej wartości głębi reprezentują wyższy stopień centralności. Takie porządkowanie umożliwia pomiar wielu złożonych własności wielowymiarowego rozkładu prawdopodobieństwa – położenia, kwantyli, rozrzutu, skośności i kurtozy.

Klasyczna statystyka nieparametryczna wykorzystuje statystyki porzadkowe, kwantyle i rangi (patrz Hajek i Sidak, 1967).

W przypadku jednowymiarowym dysponujemy liniowym porządkiem obserwacji, które przyjmują wartości w zbiorze liczb rzeczywistych. Posługując się tym porządkiem definiujemy statystyki porządkowe oraz rangi obserwacji. Zaznaczmy, że ów porządek obserwacji indukowany przez zbiór liczb rzeczywistych nie ma naturalnego uogólnienia na $\mathbb{R}^d$, $d \geq 2$.

Rozsadnym rozwiazaniem w takiej sytuacji wydaje się być porządkowanie obserwacji względem pewnego pojęcia centrum.

W tzw. **koncepcji głębi danych** w jednolity sposób rozszerza sie na przypadek wielowymiarowy jednowymiarowe metody statystyczne wykorzystujące statystyki porządkowe, kwantyle, rangowanie i miary odstawania (patrz Zuo i Serfling, 2000).

Podkreślmy, że mamy tu na uwadze wielowymiarowe statystyki porządkowe (aby uwzględnić wielowymiarową geometrię zbioru danych rozważa się wielowymiarową medianę zamiast wektora jednowymiarowych median).

Za źródło inspiracji prowadzących do powstania koncepcji głębi danych można przyjąć prace Hotellinga dotyczące ekonomicznej teorii gier oraz test znaków zaproponowany przez Hodgesa (patrz Serfling 2006). Upowszechnienie na szerszą skalę zagadnień z nią związanych wiąże się z nazwiskiem wybitnego matematyka i statystyka amerykańskiego Johna Tukey.

Tukey (por. Tukey 1975) zaproponował ***głębię domkniętej półprzestrzeni*** *do generowania konturów w wielu wymiarach, które spełniałyby rolę analogiczną do rang i statystyk porządkowych w przypadku jednowymiarowym*. Jego propozycje zapoczątkowały szereg prac dotyczących wielowymiarowych uogólnień jednowymiarowych statystyk porządkowych, kwantyli i rang.

- Statystyczna funkcja głębi związana z danym rozkładem prawdopodobieństwa P na $\mathbb{R}^d$ umożliwia porządkowanie punktów $\mathbf{x} \in \mathbb{R}^d$ na zasadzie odstawania od rozkładu P.

- W przypadku punktów położonych w pobliżu centrum rozkładu funkcja głębi przyjmuje wartości bliskie jedności, wartości funkcji głębi bliskie zeru odpowiadają peryferiom rozkładu.

- Wykorzystując funkcję głębi można określić centrum rozkładu jako zbiór punktów, które globalnie maksymalizują wartość funkcji głębi. Wielomodalność rozkładu P na ogół jest pomijana.

- Punkt (zbiór punktów), w którym funkcja głębi przyjmuje wartość maksymalną określa się mianem **wielowymiarowej mediany** indukowanej przez tę funkcję głębi.

- Funkcje głębi różnią się pod względem sposobu ujmowania centrum bądź peryferii rozkładu. Różnią się pod względem oferowanej badaczowi informacji gdyż opierają się o odmienne własności rozkładu.
- Niektóre, jak głębia Mahalanobisa bądź projekcyjna uwzględniając geometrię rozkładu, czerpią z własności metrycznych *d*–wymiarowej przestrzeni rzeczywistej, inne jak głębia domkniętej półprzestrzeni bądź głębia symplicjalna mogą zostać zdefiniowane w przestrzeni topologicznej bądź wektorowej.
- Głębie różnią się pod względem niezmienniczości afinicznej, szybkości zbieżności z próby, odporności indukowanych przez nie procedur statystycznych.

Najprostszym przykładem funkcji głębi jest tzw. **głębia Euklidesa**:

$$D_{EUK}(\mathbf{y}, \mathbb{X}^n) = \frac{1}{1 + \left\| \mathbf{y} - \bar{\mathbf{x}} \right\|^2},$$

gdzie $\bar{\mathbf{x}}$ oznacza wektor średnich z n – elementowej próby $\mathbb{X}^n$.

Kolejny popularny przykład głębi stanowi tzw. **głębia Mahalanobisa**

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

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

Zwróćmy uwagę na fakt, że w przypadku głębi Euklidesa zbiory punktów przyjmujących tę samą wartość głębi tworzą rodzinę współśrodkowych sfer, w przypadku głębi Mahalanobisa rodzinę współśrodkowych elipsoid. Mediany indukowane przez te głębie nie są odporne na występowanie obserwacji odstających.

Od funkcji głębi $D(\mathbf{x},P)$ na ogół wymaga się spełnienia następujących postulatów:

Niezmienniczość afiniczna: $D(\mathbf{x},P)$ powinna być niezależna od wyboru układu współrzędnych. Własność z jednej strony ułatwia interpretację, z drugiej ułatwia badania własności funkcji za pomocą symulacji. Własność ta jest ważna, gdy dane w każdym wymiarze są zadane na skali przedziałowej i zamierzamy interpretować liniowe kombinacje zmiennych np. w analizie czynnikowej, składowych niezależnych.

Wartość maksymalna w centrum: Jeżeli rozkład P jest symetryczny względem θ zważywszy na pewne rozumienie symetrii wówczas $D(\mathbf{x},P)$ przyjmuje w tym punkcie maksimum.

Symetria: Jeżeli rozkład P jest symetryczny względem θ w pewnym sensie, wtedy także $D(\mathbf{x},P)$ jest symetryczna w tym sensie.

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

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

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

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

Quasi – wypukłość: $D(\mathbf{x},P)$ rozpatrywanej jako funkcja $\mathbf{x}$.

Zbiór $\{\mathbf{x}: D(\mathbf{x},P)\geq\alpha\}$ jest wypukły dla każdego $\alpha\in[0,1]$.

DEFINICJA (Zuo i Serfling, 2000): Odwzorowanie $D(\cdot,\cdot):\mathbb{R}^d \times \mathcal{P} \longrightarrow [0,1]$, które jest ograniczone, nieujemne oraz spełnia poniższe warunki ZS1 – ZS4:

ZS1: $D(\mathbf{Ax}+\mathbf{b}, P_{\mathbf{AX}+\mathbf{b}}) = D(\mathbf{x}, P_{\mathbf{X}})$ dla dowolnego wektora losowego $\mathbf{X}$ o wartościach w $\mathbb{R}^d$, dowolnych $d \times d$ nieosobliwej macierzy $\mathbf{A}$, i d wektora $\mathbf{b}$.

ZS2: $D(\mathbf{m}, P) = \sup_{\mathbf{x}\in\mathbb{R}^d} D(\mathbf{x}, P)$ zachodzi dla każdego $P \in \mathcal{P}$ mającego centrum $\mathbf{m}$.

ZS3: Dla dowolnego $P \in \mathcal{P}$ mającego punkt o największej głębi $\mathbf{m}$, zachodzi

$D(\mathbf{x}, P) \le D(\mathbf{m} + \alpha(\mathbf{x}-\mathbf{m}), P)$, $\alpha \in [0,1]$.

ZS4: $D(\mathbf{x}, P) \to 0$, gdy $\|\mathbf{x}\| \to \infty$ dla każdego $\mathbf{x} \in \mathbb{R}^d$.

nazywamy **statystyczną funkcją głębi**.

Symetryczną głębię projekcyjną punktu $\mathbf{x} \in \mathbb{R}^d$ będącego realizacją pewnego d wymiarowego wektora losowego $\mathbf{X}$ o rozkładzie F, $D(\mathbf{x}, F)$ definiujemy pomocą

$$D(\mathbf{x}, F) = \left[1 + sup_{\|\mathbf{u}\|=1} \frac{\left| \mathbf{u}^T \mathbf{x} - Med\left(\mathbf{u}^T \mathbf{X} \right) \right|}{MAD\left(\mathbf{u}^T \mathbf{X} \right)} \right]^{-1}, \tag{3}$$

gdzie Med oznacza jednowymiarową medianę, $MAD(Z) = Med(|Z - Med(Z)|)$, wersję z próby tej głębi oznaczaną za pomocą $D(\mathbf{x}, F_n)$ bądź $PD(\mathbf{x}, \mathbf{X}^n)$ uzyskujemy poprzez zastąpienie rozkładu F jego empirycznym odpowiednikiem F_n obliczonym na podstawie próby $\mathbf{X}^n$.

Głębia projekcyjna jest niezmiennicza afinicznie, indukowane przez nią wielowymiarowe charakterystyki położenia i rozrzutu odznaczają się wysokim punktem załamania próby skończonej i ograniczoną funkcją wpływu.

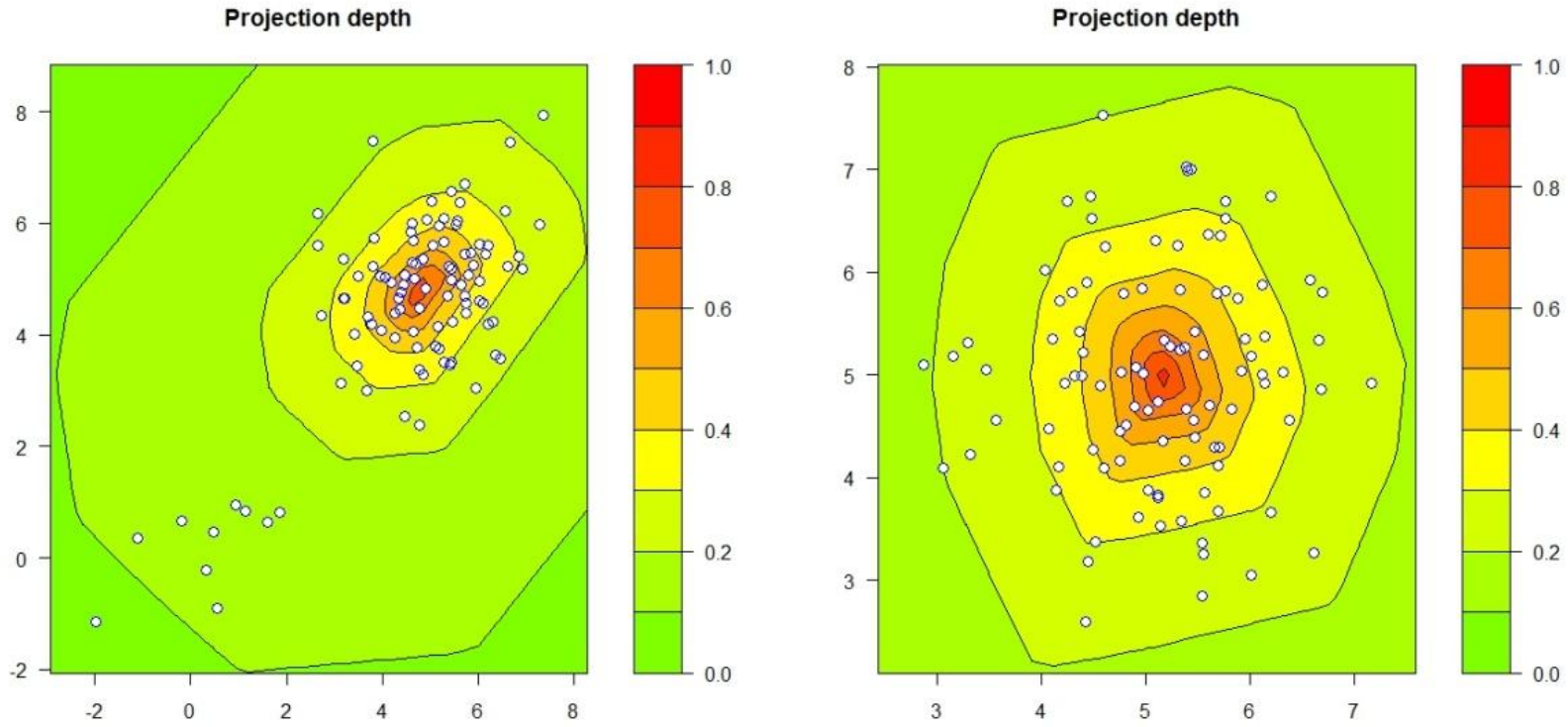
Dla próby $\mathbf{X}^n = \{\mathbf{x}_1,.,\mathbf{x}_n\}$ zbiór punktów

$D_\alpha(\mathbf{X}^n) = \{\mathbf{x} \in \mathbb{R}^d : D(\mathbf{x},\mathbf{X}^n) \geq \alpha\}$ nazywany jest $\alpha -$**obszarem centralnym.**

Jego brzeg możemy traktować jako **wielowymiarowy odpowiednik jednowymiarowego kwantyla**.

Poniższy rysunek przedstawia głębię projekcyjną z próby policzoną dla prób 100 obserwacji pobranych ze dwuwymiarowego rozkładu normalnego i mieszaniny dwuwymiarowych rozkładów normalnych.

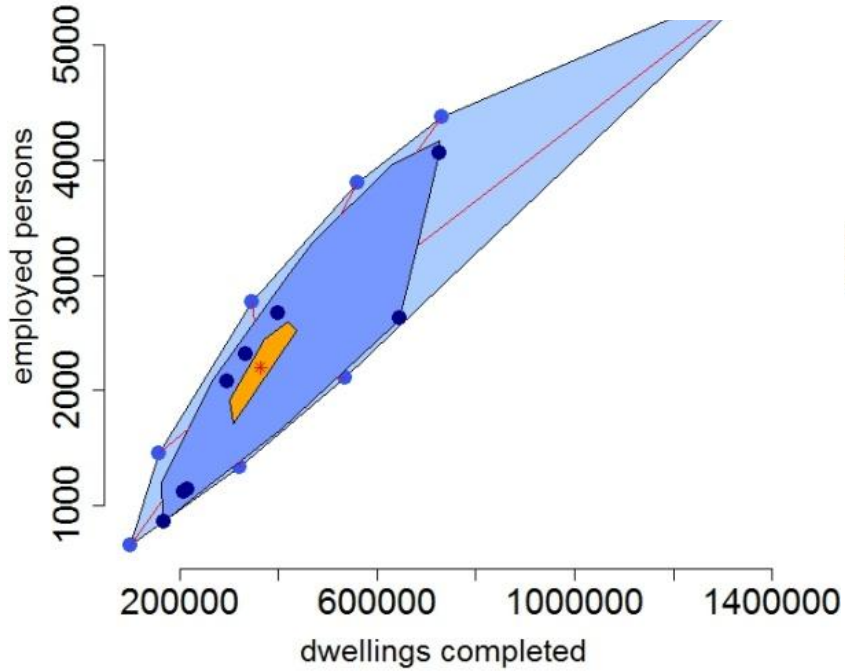
Rysunek został sporządzony z zastosowaniem przybliżonego algorytmu zapropnowanego przez Dyckerhoffa (2004) za pomocą pakietu {depthproc} dostępnego via serwer R-Forge.



Wykres Konturowy głębi projekcyjnej z próby dla mieszaniny dwuwymiarowych rozkładów normalnych (po prawej) i rozkładu normalnego (po prawej).

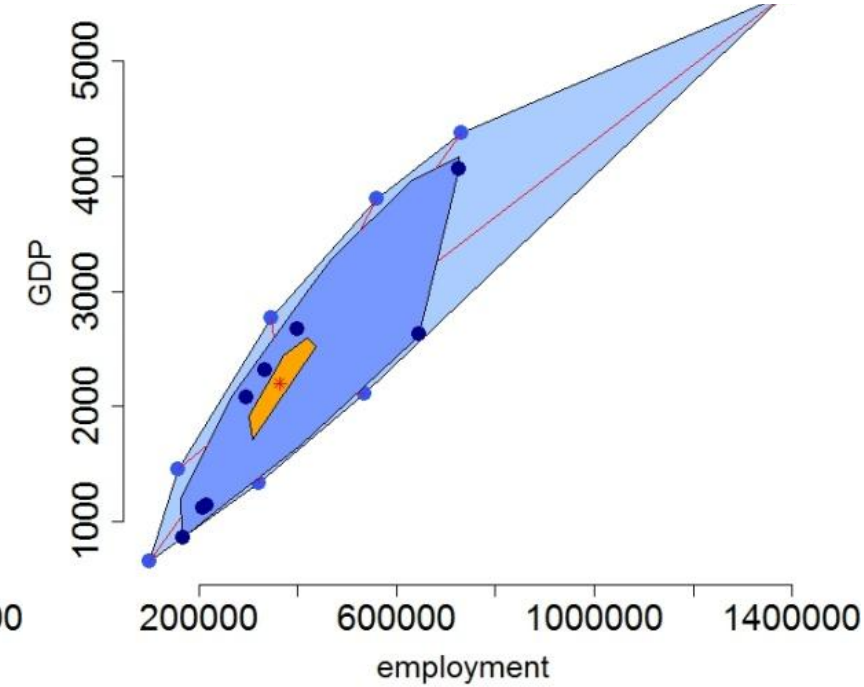
Funkcje głębi indukują zagnieżdżone kontury równego odstawanie. Dla szerokiej klasy rozkładów funkcje głębi określają rozkład jednoznacznie (jak dystrybuanta) (por. Kong and Zuo 2010)

Dwellings completed vs. the number of employed persons in polish voividships in 2011 – 2D boxplot – Tukey depth.



Source: Our own calculations – {aplpack} R package, data GUS

Employment in thousands vs. GDP in polish voivodships – 2D boxplot – Tukey depth.



Source: Our own calculations – {aplpack} R package, data GUS

WYKRESY PUDEŁKOWE 2D! – tu wykorzystujemy głębię Tukey'a

Co możemy policzyć na podstawie funkcji głębi- np. stosując pakiet {depthproc}?

1. Kontury
2. Indukowane przez glebie statystyki porządkowe. Możemy porządkować dane ze względu na wartość głębi - względem centrum.
3. Ważone głębią funkcjonały położenia.
4. Ważone głębią macierze rozrzutu.
5. Krzywe skali - objętość obszaru centralnego względem α (parametr odstawania).
6. Funkcjonały skośności. Skalowana różnica pomiędzy dwoma funkcjonałami położenia.
7. Funkcjonały kurtozy.

Procedury statystyczne indukowane przez głębie

1. Uogólnienia wykresów ramka wąsy.
2. Porównanie dwóch prób za pomocą prostego rysunku 2D - głębia punktu zważywszy na pierwszą próbę vs. głębia punktu zważywszy na druga próbę.
3. Nieparametryczny opis wielowymiarowego rozkładu.
4. Testy wielowymiarowej symetrii.
5. Procedury statystycznej kontroli jakości.
6. Estymacja wielowymiarowej gęstości.
7. Klasyfikacja za pomocą głębi.
8. Analiza skupisk za pomocą głębi.

Bezpośrednie obliczanie wartości funkcji głębi jest w wielu przypadkach zadaniem niezmiernie złożonym pod względem obliczeniowym. Wobec zapotrzebowania na procedury statystyczne indukowane przez funkcje głębi szczególnego znaczenia nabierają propozycje przybliżonego obliczania wartości funkcji głębi z próby.

Na szczególną uwagę zasługują prace Dyckerhoffa (Dyckerhoff (2004)) a dotyczące przybliżonego obliczania pewnej klasy funkcji głębi.

Podejście Dyckerhoffa można wyrazić w uproszczeniu w następujący sposób:

jeżeli przyjmiemy, że punkt jest centralny względem pewnego wielowymiarowego rozkładu prawdopodobieństwa, gdy wszystkie jego jednowymiarowe projekcje są centralne względem jednowymiarowej projekcji rozkładu to możemy definiować wielowymiarową głębię punktu jako minimum jednowymiarowych głębi wszystkich takich jednowymiarowych projekcji.

Głębia regresyjna Rousseeuw i Hubert

Niech $\mathbf{Z}^n = \{(\mathbf{x}_1, y_1),...,(\mathbf{x}_n, y_n)\} \subset \mathbb{R}^d$ oznacza próbę rozpatrywaną z punktu widzenia następującego modelu semiparametrycznego:

$$y_l = a_0 + a_1 x_{1l} + . + a_{(d-1)l} x_{(d-1)l} + \varepsilon_l \text{, l=1,...,n,}$$

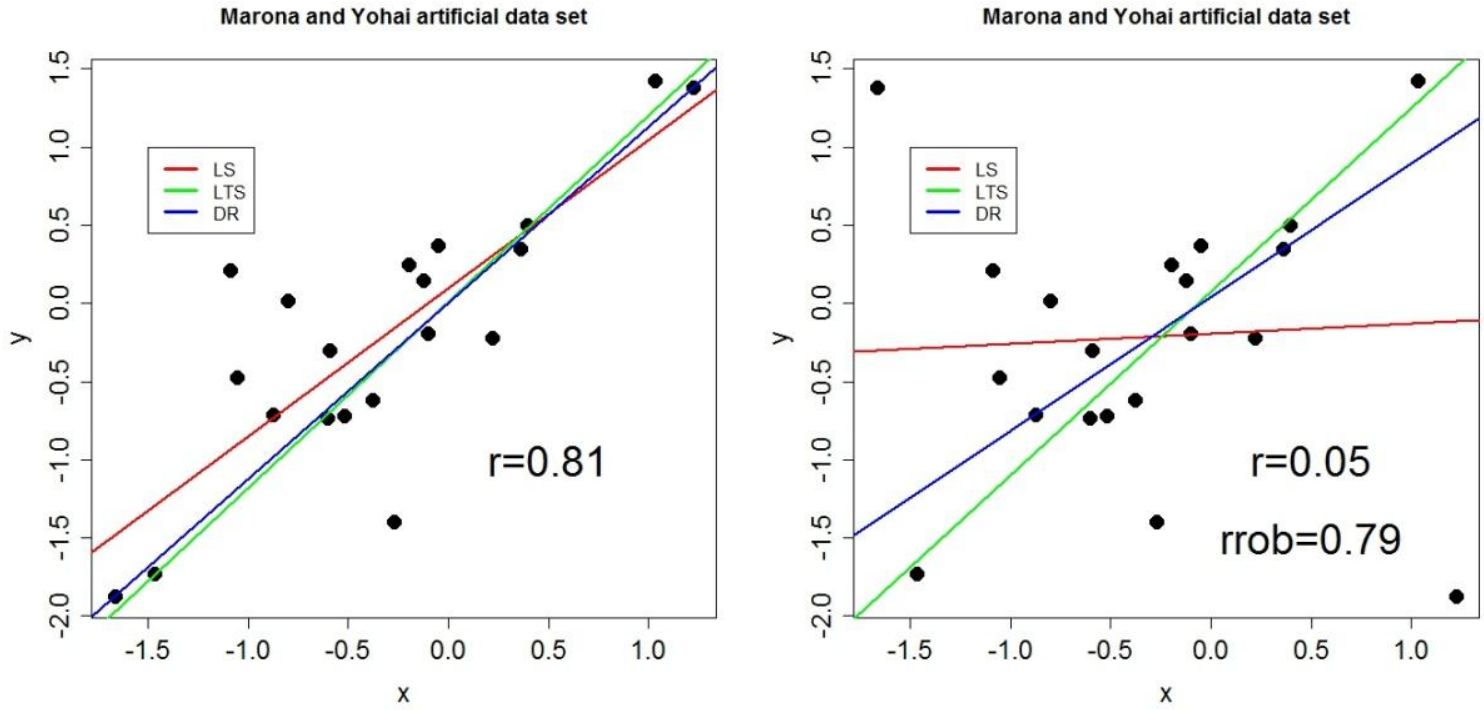
obliczamy **głębię konkretnego dopasowania** $\alpha = (a_0,...,a_{d-1})$ za pomocą

$$RD(\alpha, Z^n) = \min_{\mathbf{u} \neq 0} \#\left\{ \frac{r_l(\alpha)}{\mathbf{u}^T \mathbf{x}_l} < 0, l = 1,...,n \right\},$$

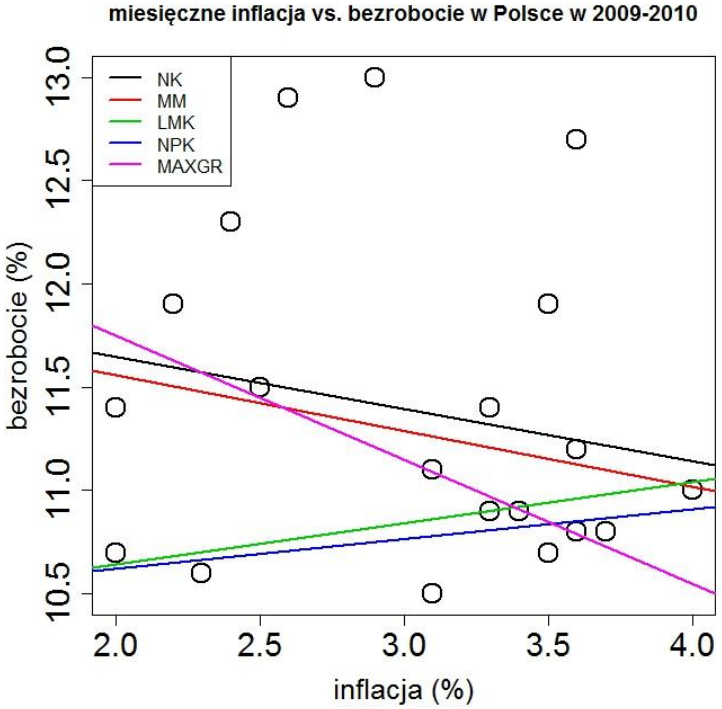
gdzie $r(\cdot)$ oznacza resztę regresji, $\alpha = (a_0,...,a_{d-1})$, $\mathbf{u}^T \mathbf{x}_l \neq 0$.

Estymator maksymalnej głębi regresyjnej (ang. deepest regression) $DR(\alpha,\mathbf{Z}^n)$ definiujemy

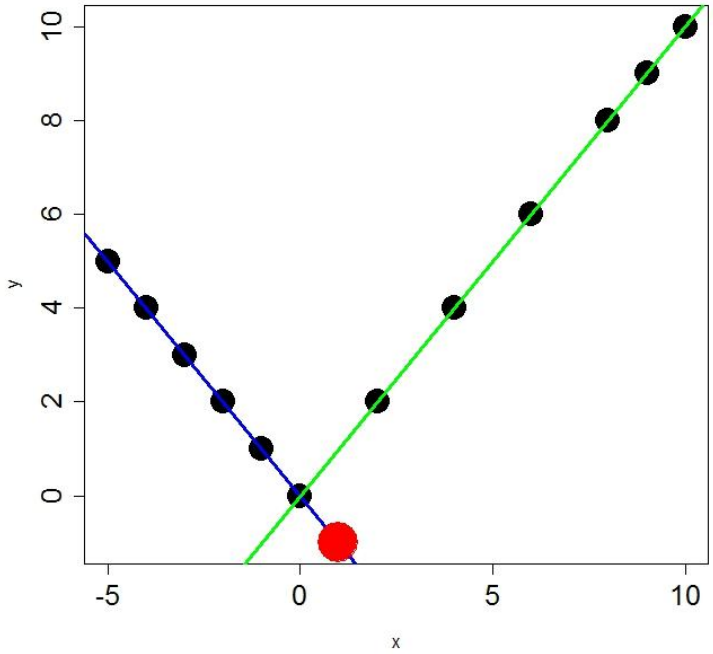
$$DR(\alpha,\mathbf{Z}^n) = \arg\max_{\alpha \neq \mathbf{0}} RD(\alpha,\mathbf{Z}^n)$$



NIE NALEŻY JEDNAKŻE BEZKRYTYCZNIE STOSOWAĆ REGRESJI ODPORNEJ!

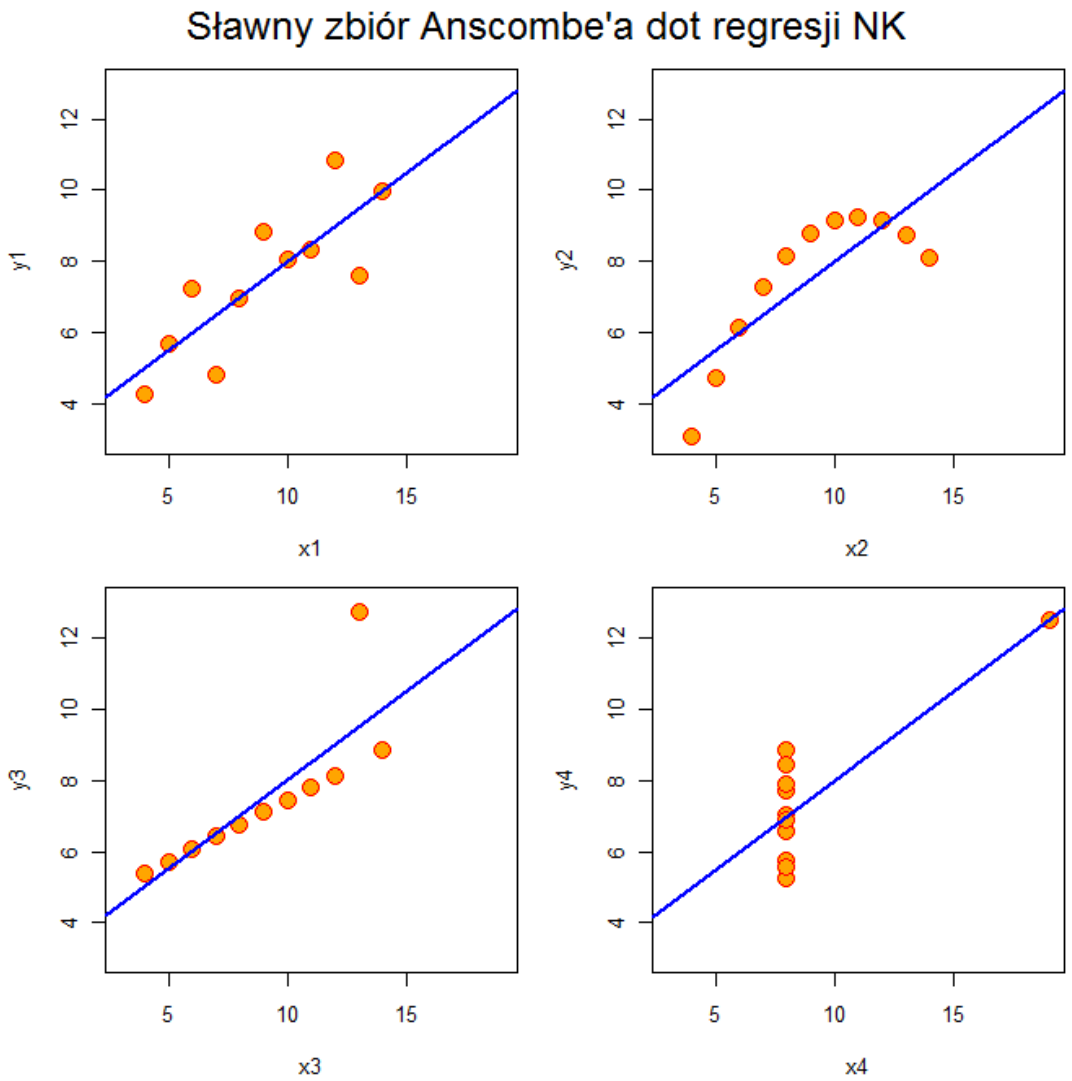


Zarówno metody odporne jak i tradycyjne bywają zawodne w wykrywaniu zależności pomiędzy zmiennymi ekonomicznymi.



Nieodporność wysoce odpornej regresji na zmianę większości obserwacji w próbie.

ISTNIEJĄ TEŻ INNE PROBLEMY DOTYCZĄCE ZASTOSOWAŃ REGRESJI…



PROPOZYCJE ODPORNEGO MONITOROWANIA STRUMIENIA DANYCH

Nasze pierwsze dwie propozycje dotyczące detekcji zmian położenia/rozrzutu w strumieniu wiążą się wielowymiarowym uogólnieniem wykresy kwanty-kwantyl – z wykresem głębia vs. głębia (ang. DD-plot i.e. depth vs. depth plot).

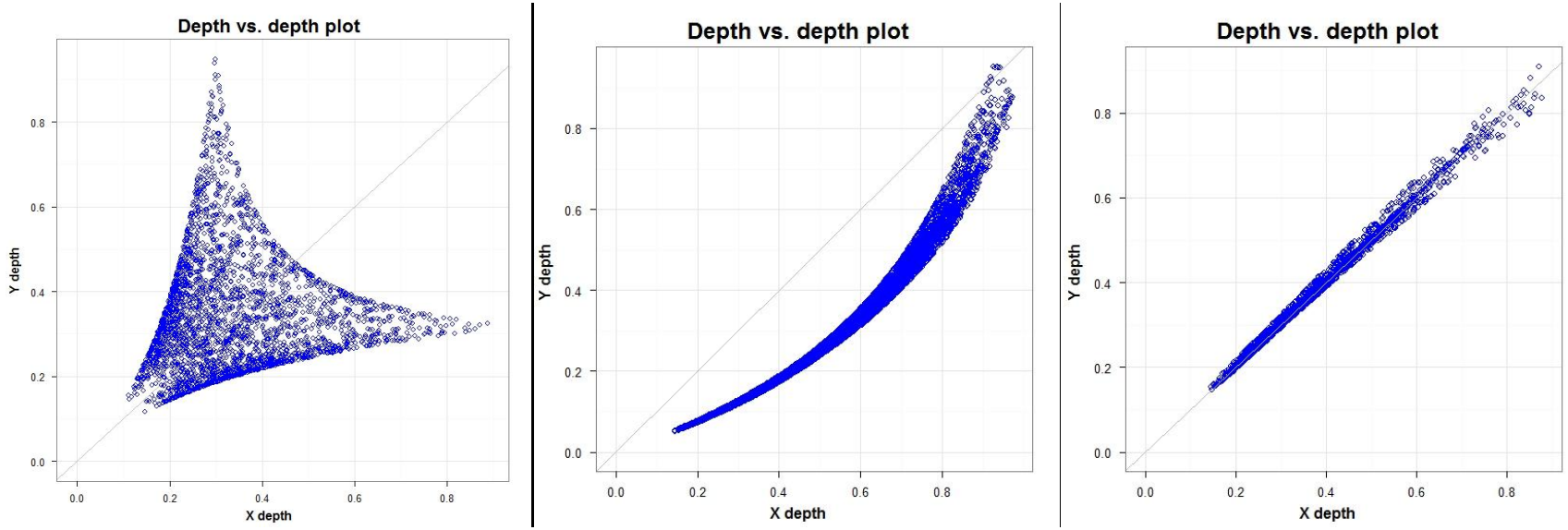
Wykres głębia vs. głębia zaproponowany prze Liu i in. (1999) jest przyjazną dla użytkownika dwuwymiarową metodą graficznego porównywania dwóch prób dowolnego wymiaru. Różnego rodzaju różnice co do rozkładów (położenie, rorzut, skośność, kurioza) związane są z odmiennymi wzorcami, powierzchniami na wykresie głębia vs. głębia.

Dla dwóch rozkładów F i G, obydwu w $\mathbb{R}^d$, **wykres głębia vs. głębia** definiowany jest jako

$$DD(F,G) = \left\{ \left(D(\mathbf{z},F), D(\mathbf{z},G) \right), \mathbf{z} \in \mathbb{R}^d \right\},$$

Natomiast jego wersja empiryczna dla prób $\mathbf{X}^n$ i $\mathbf{Y}^n$ definiowana jest jako

$$DD(F_n, G_m) = \left\{ \left(D(\mathbf{z}, F_n), D(\mathbf{z}, G_m) \right), \mathbf{z} \in \left\{ \mathbf{X}^n \cup \mathbf{Y}^m \right\} \right\}.$$



Wykresy głębia vs. głębia w przypadku różnic co do położenia (po lewej), różnic co do rozrzutu (w środku) i dla prób pobranych z tej samej populacji (po prawej).

Dla dwóch prób $\mathbf{X}^n$ and $\mathbf{Y}^m$ wykorzystując dowolną funkcję głębi możemy policzyć wartości głębi w połączonej próbie $\mathbf{Z}^{n+m} = \mathbf{X}^n \cup \mathbf{Y}^m$, zakładając przy tym rozkład empiryczny wyznaczony w oparciu o wszystkie obserwacje bądź jedynie w oparciu o obserwacje należące do próby $\mathbf{X}^n$ bądź $\mathbf{Y}^m$.

Dla przykładu, jeżeli obserwujemy, że głębie *X-ów* mają większą skłonność do skupiania się wokół centrum połączonej próby podczas gdy głębie *Y-ów* raczej są rozrzucone po peryferiach próby – konstatujemy, że próba $\mathbf{Y}^m$ została pobrana z populacji odznaczającej się większym rozrzutem.

Można na podstawie rang obserwacji w połączonej próbie rozpatrywać szereg wielowymiarowych uogólnień jednowymiarowego testu sumy rang Wilcoxona (zobacz Oja, 2010).

Własności wykresu głębia vs. głębia były przedmiotem intensywnych studiów m. in. prowadzonych przez duet Li i Liu (2004) dla niezależnych obserwacji o tym samym rozkładzie. Autorzy zaproponowali szereg statystyk wykorzystujących wykres głębia vs. głębia, przedstawili argument odwołujące się do metody bootstrap dotyczące zgodności I dobrej efektywności proponowanych testów (w porównaniu do testu Hotellinga T^2 i wielowymiarowych testów Ansari-Bradley'a oraz Tukey-Siegel.

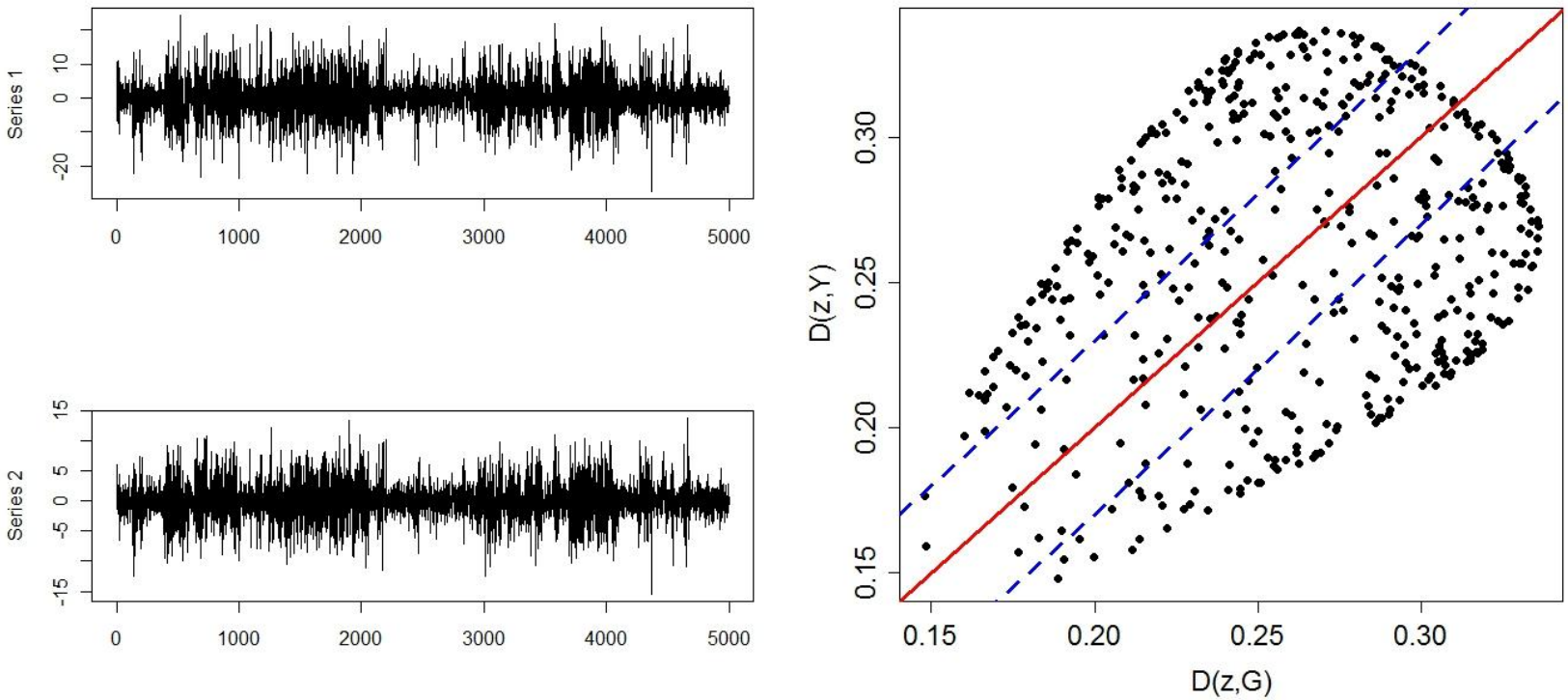
Rozkłady asymptotyczne wielowymiarowych testów Wilcoxona opierających się na wykresie głębia vs. głębia zostały przedstawione przez duet Zuo i He (2006). Wybrane własności testów rangowych opierających się o funkcje głębi zostały krytycznie przedstawione przez duet Jureckova i Kalina (2012).

PROPOZYCJA 1: Nasza pierwsza proponowana statystyka D^1 przeznaczona do wykrywania zmian położenia/rozrzutu opiera się na wprowadzonym powyżej wykresie głębia vs. głębia.

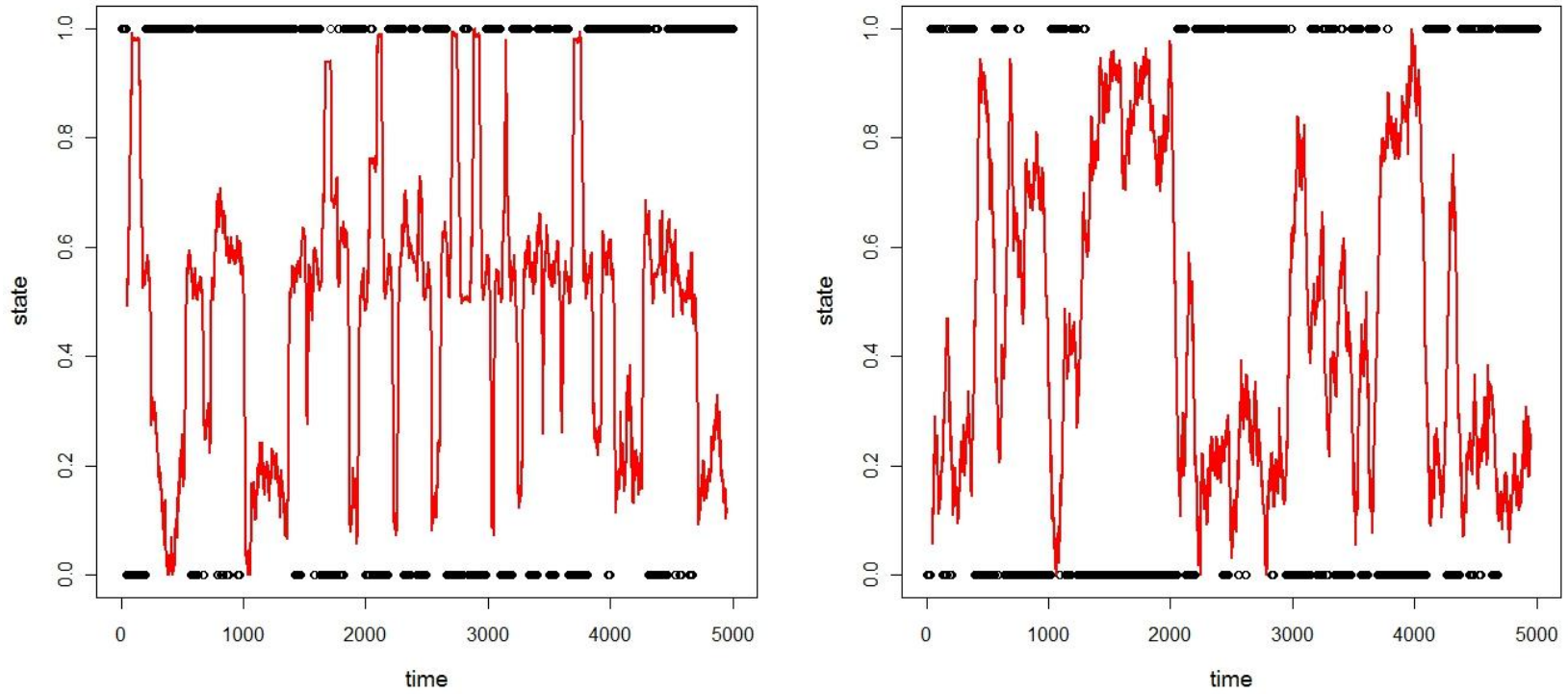
Dla dwóch prób, pierwszej $\mathbf{X}_i^n = \{\mathbf{x}_1,.,\mathbf{x}_n\} \equiv \mathbf{W}_{i,n}$ będącej ruchomym oknem w chwili i o długości n, oraz drugiej będącej oknem referencyjnym $\mathbf{Y}^m = \{\mathbf{y}_1,.,\mathbf{y}_m\}$ pobranym z ustalonego rozkładu G, proponujemy obliczać

$$D_i^1 = \#\left\{\mathbf{z} \in \mathbf{X}_i^n \cup \mathbf{Y}^m : \left| D(\mathbf{z},\mathbf{X}_i^n) - D(\mathbf{z},\mathbf{Y}^m) \right| > const \right\}, i = 1,2,...,$$

gdzie $D(\mathbf{z},\mathbf{X}^n)$ oznacza wartość funkcji głębi z próby, *const* jest ustalonym wcześniej ograniczeniem (np. dla głębi projekcyjnej i okna długości 100-obs proponujemy przyjąć const=0.03).



Ilustracja dla procesu obliczania wartości pierwszej propozycji.



Ilustracja zachowania się pierwszej statystyki w odniesieniu do detekcji zmiany położenia (po lewej) i zmiany rozrzutu (po prawej).

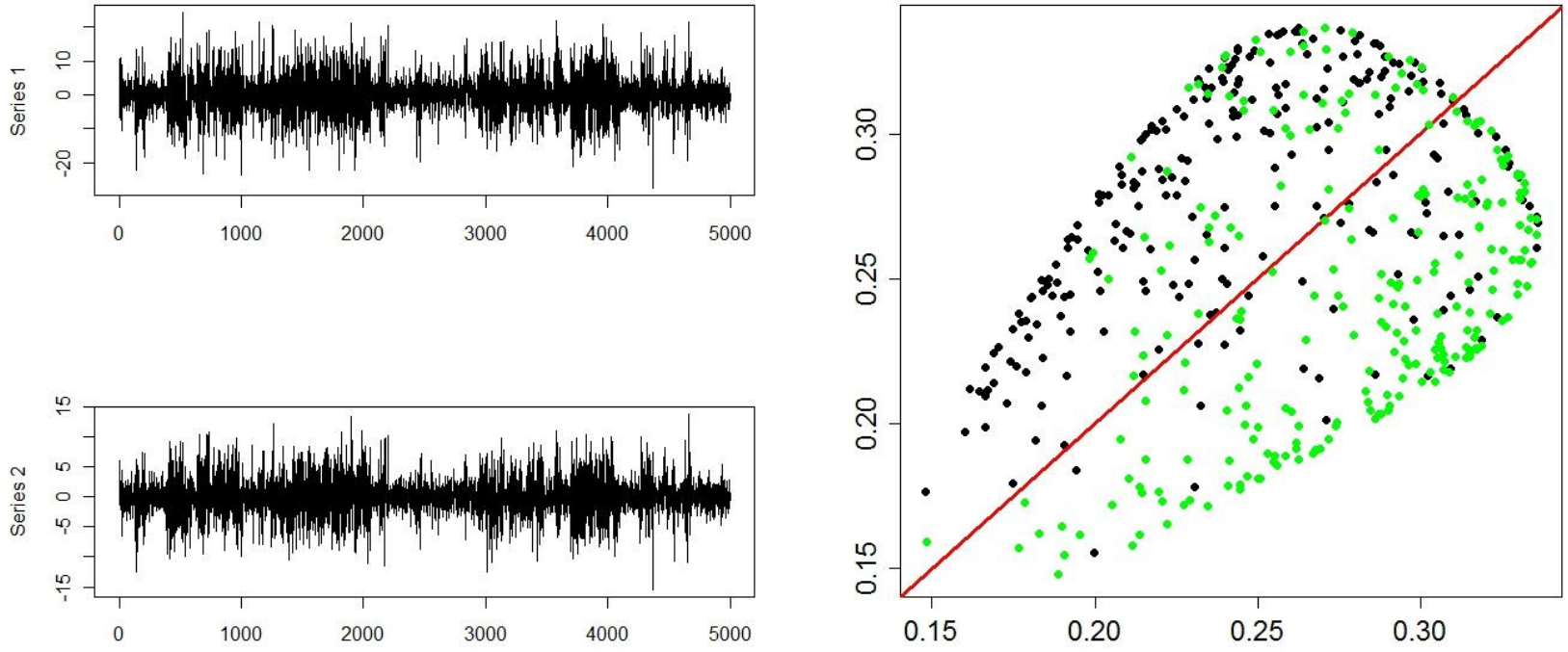
PROPOZYCJA 2: W celu detekcji zmian co do rozrzutu wielowymiarowego strumienia danych proponujemy wykorzystać wielowymiarową statystykę sumy rang Wilcoxona wprowadzoną między innymi przez Liu i Singh (2003) oraz gruntownie badaną przez Jureckowa i Kalina (2012) oraz Zuo i He (2006) w przypadku obserwacji niezależnych o tym samym rozkładzie.

Dla dwóch prób, pierwszej $\mathbf{X}_i^n = \{\mathbf{x}_1,...,\mathbf{x}_n\} \equiv \mathbf{W}_{i,n}$ będącej ruchomym oknem w chwili t o długości n, oraz drugiej będącej oknem referencyjnym $\mathbf{Y}^m = \{\mathbf{y}_1,...,\mathbf{y}_m\}$ pobranym z rozkładu G, $\mathbf{Z}_i^{n+m} = \mathbf{X}_i^n \cup \mathbf{Y}^m$ proponujemy obliczać

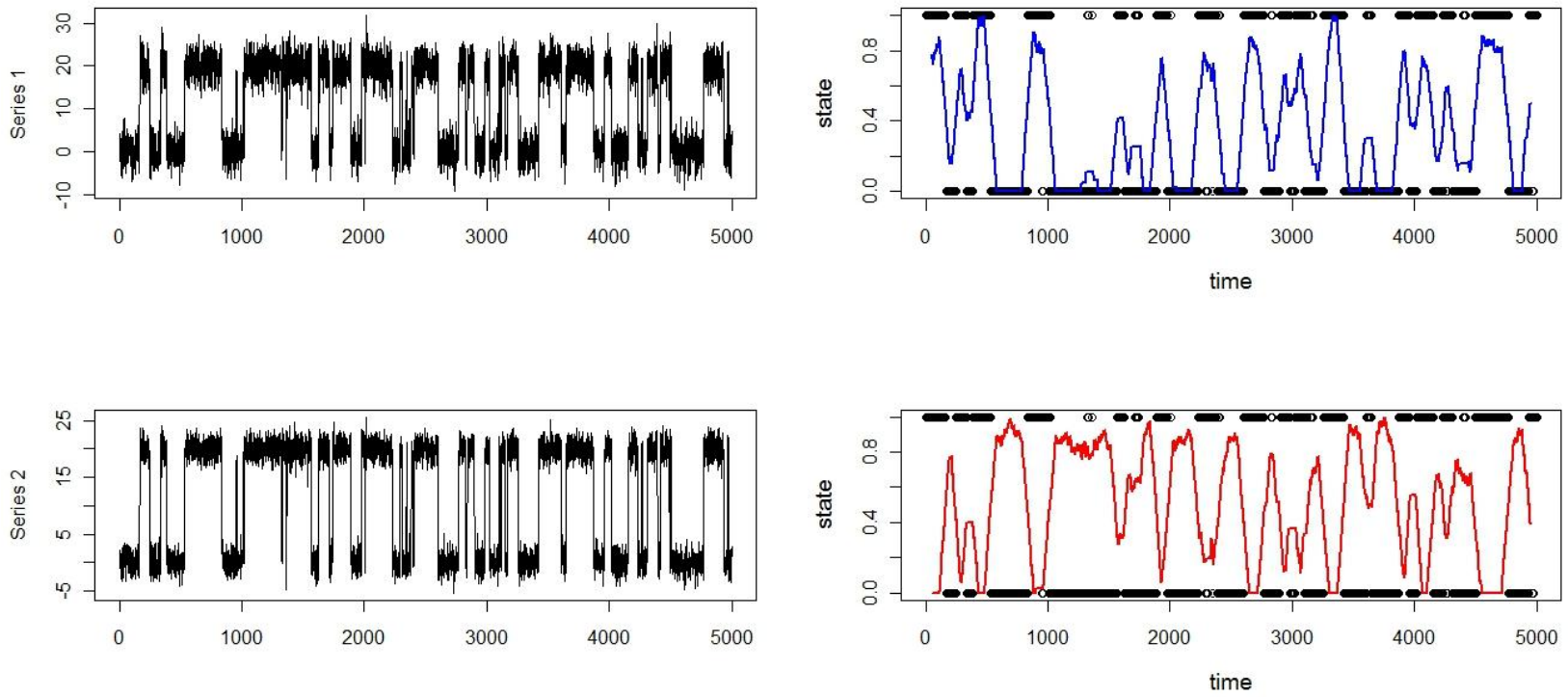
$$R(\mathbf{y}_l) = \#\left\{\mathbf{z}_j \in \mathbf{Z}_i^{n+m} : D(\mathbf{z}_j, \mathbf{Y}^m) \le D(\mathbf{y}_l, \mathbf{Y}^m)\right\}, l = 1,...,m,$$

$$D_i^2 = \sum_{l=1}^{m} R(\mathbf{y}_l), i = 1,2,...$$

W przypadku tej statystyki proponujemy wykorzystać głębię Oja bądź głębię domkniętej półprzestrzeni (zobacz Serfling, 2006).



Ilustracja dla procesu obliczania wartości drugiej propozycji



Ilustracja zachowania się pierwszej statystyki w odniesieniu do detekcji zmiany położenia.

INNE MOŻLIWOŚCI? Objętość obszarów centalnych, odległość pomiędzy wielowymiarowymi medianami? (por. Kosiorowski i Snarska, 2012)

STUDIA MONTE CARLO WŁASNOŚCI NASZYCH PROPOZYCJI W PRZYPADKU PRÓB SKOŃCZONYCH.

W celu zbadania statystycznych własności naszych propozycji – 500 razy generowano próby każda długości 10000 obserwacji z modelu VCHARME składającego się z dwóch dwuwymiarowych modeli VAR(1) oznaczonych dalej za pomocą M1 i M2 (rozważano model strumienia z dwoma reżimami), i macierzą przejścia P dla ukrytego łańcucha Markowa Q(2), o następujących wierszach [0.99; 0.01] i [0.03; 0.97].

Rozważano ruchome okno ustalonej długości 100 obserwacji I ustalone okna referencyjne wygenerowane z modeli M1 i M2, składające się ze 100 obserwacji. Badano strumienie bez jednostek odstających ora strumienie zawierające do 5% obserwacji odstających typu (AO).

Próby generowano za pomocą pakietu środowiska R {Dynamic Systems Estimation}, {DSE}, opublikowanego przez Paula Gilberta. Wykorzystywano modele VAR(1) o następujących specyfikacjach

$$\mathbf{M1}: \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix} . \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix},$$

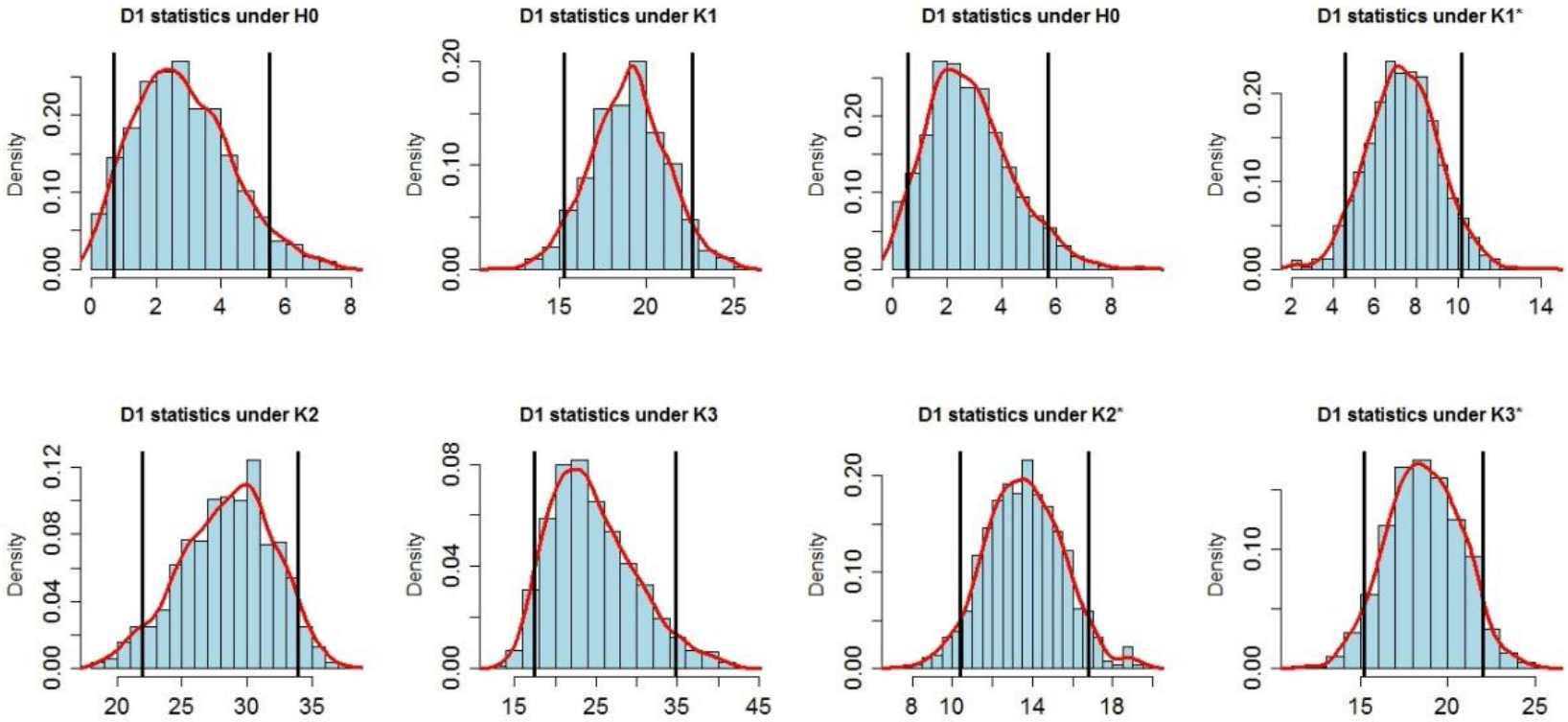
$$\mathbf{M2}: \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix} . \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \times C.$$

W przypadku detekcji zmian położenia rozważano sytuacje gdy dane generowane były przez model o jednym reżimie M1, wektorze trendu $(m_1,m_2)=(0,0)$ i o ustalonym rozrzucie w przypadku hipotezy zerowej oraz $(m_1,m_2)=(0,0)$ dla M1 i $(m_1,m_2)=(0{,}5;0{,}5)$, (1,1), (1,5;1,5) dla hipotez H1, H2, H3, odpowiednio.

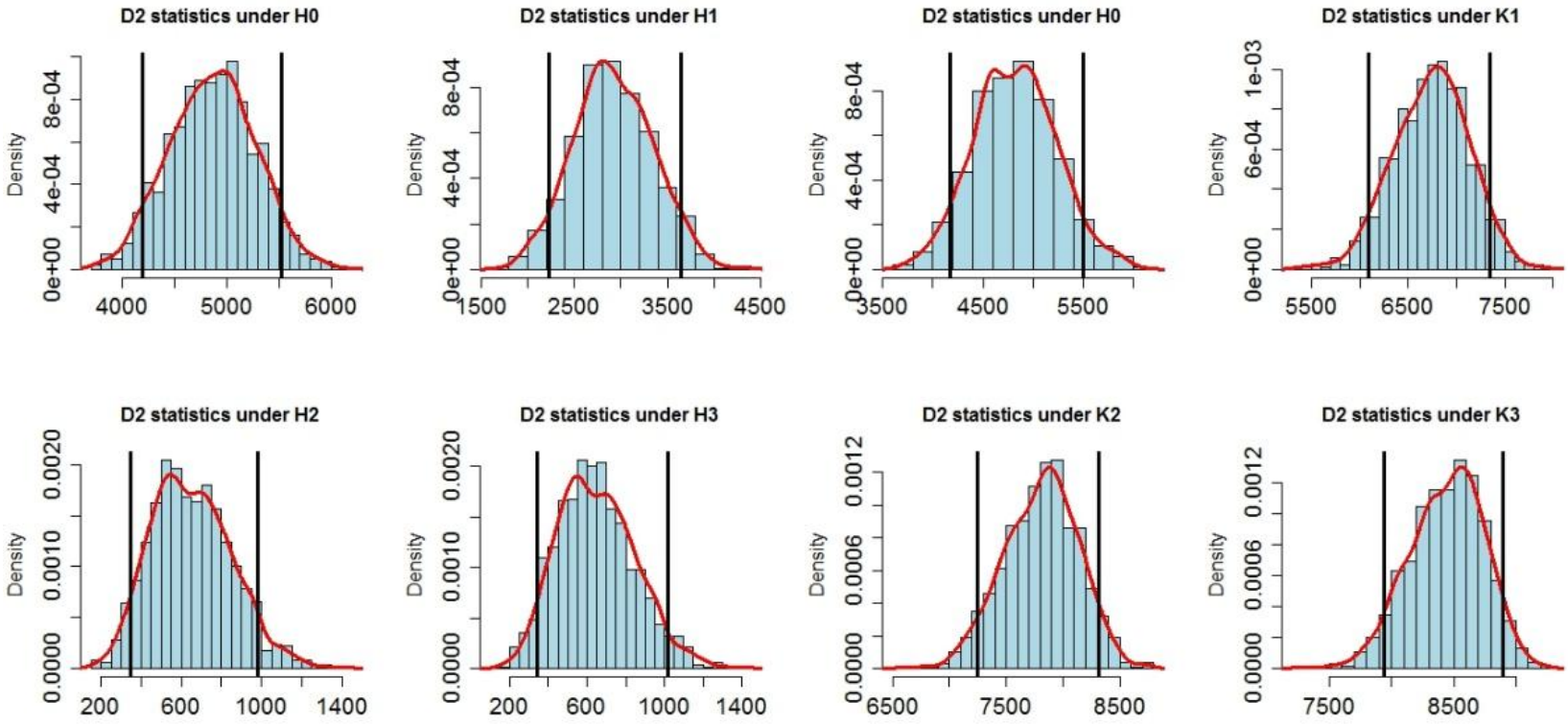
Podobnie w przypadku detekcji zmian rozrzutu, rozważano sytuację gdy dane generował model o jednym reżimie M1 z wektorem trendu (m_1, m_2) = (0,0) i ustalonym rozrzutem C = 1. Taka sytuacja oznaczała dla nas hipotezę zerową H0 głoszącą, że nie nasępuje zmiana rozrzutu strumienia. Rozważano następnie hipotezy alternatywne K1, K2, K3 poprzez generowanie obserwacji wykorzystując dwa reżimy – oba z C = 1 dla M1, i C = 1.5, 2, 2.5, dla K1, K2, K3, odpowiednio.

Rysunki przedstawiają histogramy sporządzone dla statystyk z próby D^1, D^2, wraz z oszacowaniem jądrowym gęstości tych statystyk. Lewe części rysunków przedstawiają sytuacje bez obserwacji odstających, natomiast prawe strony przedstawiają sytuacje gdzie w strumieniu występowało do 5% obserwacji odstających typu. Dodatkowo na rysunkach za pomocą lini pionowych zaznaczono kwantyle rzędu 5% i 95%.

WYNIKI SYMULACJI: Rysunki wskazują na dobre statystyczne własności w kategoriach dyskryminacji pomiędzy rozpatrywanymi “hipotezami”. Rysunki dają podstawy aby sądzić, że nasze procedury testowe odznaczają się “nieobciążonością” – to znaczy w każdej chwili empiryczna moc testu jest nie mniejsza niż założone prawdopodobieństwo błędu pierwszego rodzaju. Statystyki wydają się być odporne na występowanie obserwacji odstających. Kwestie złożoności obliczeniowej naszych propozycji podejmuje praca Alupis (2006). W naszych symulacjach wykorzystywano algorytm do przybliżonego obliczania pewnej klasy funkcji głębi autorstwa Dyckerhoff (2004).



Oszacowanie rozkładu statystyki D1 z próby. Sytuacja, gdy w próbie nie występowały obserwacje odstające (po lewej) oraz sytuacja, gdy w próbie występowało do 5% obserwacji odstających (po prawej).



Oszacowanie rozkładu statystyki D1 z próby. Sytuacja, gdy w próbie nie występowały obserwacje odstające (po lewej) oraz sytuacja, gdy w próbie występowało do 5% obserwacji odstających (po prawej).

PODSUMOWANIE I KONKLUZJE

Analiza, monitorowanie wielowymiarowych strumieni danych zyskuje ogromne zainteresowanie w ostatnich latach ze strony społeczności statystyków i informatyków. W badaniach procedur wykorzystywanych w takich analizach obok klasycznych kryteriów jak nieobciążoność, efektywność, zgodność – na pierwszy plan wysuwają się kryteria stabilności próbkowej i złożoności obliczeniowej procedury.

W opracowaniu przedstawiono pewne propozycje odpornej analizy wielowymiarowego strumienia danych odwołujące się do koncepcji wielowymiarowych statystyk porządkowych I rang indukowanych przez tzw. statystyczne funkcje głębi.

LITERATURA

1. Aggarwal, Ch.C (2003), *A framework for diagnosing changes in evolving data streams*, ACM SIGMOD Conference Proceedings , 575-586

2. Aggerwal, Ch., C., (ed.), (2007), Data streams – models and algorithms, Springer, New York.

3. Alupis, G. (2006), Geometric measures of data depth. In: Liu R.Y., Serfling R., Souvaine D. L. (Eds.): *Series in Discrete Mathematics and Theoretical Computer Science*, AMS, 72,147--158.

4. Bocian, Kosiorowski, Wegrzynkiewicz, Zawadzki (2012), R package {depthproc 1.0} for depth based procedures calculation. https://r-forge.r-project.org/projects/depthproc/

5. Dasu T., Krishnan S., Pomann G.M. (2011), *Robustness of change detection algorithms*, Advances in Intelligent Data Analysis X. Lecture Notes in Computer Science, Springer, 125-137

6. Dasu, T.,Krishnan, S., Venkatasubramanian, S., Yi K. (2006), *An information-theoretic approach to detecting changes in multi-dimensional data streams*. Proceedings of the 38th

Symposium on the Interface of Statistics, Computing Science, and Applications (Interface '06)}, Pasadena, CA.

7. Dyckerhoff, R. (2004), Data depths satisfying the projection property. *Allgemeines Statistisches Archiv*, 88, 163--190.

8. Fan, J. Yao, Q. (2005), *Nonlinear time series : nonparametric and parametric methods*, Springer, New York.

9. Hardle, W. K., Simar, L. (2012), *Applied multivariate statistical analysis*, third edition, Springer, New York

10. Genton M. G., Lucas A. (2003), Comprehensive definitions of breakdown points for independent and dependent observations, *Journal of the Royal Statistical Society Series* B, 65(1), 81-–84.

11. Hall, P., Rodney, C. L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94(445), 154--163.

12. Hastie T., Tibshiriani R., Friedman J., (2009), *The elements of statistical learning: data Mining, inference, and prediction*. Second Edition, Springer.

13. Hahsler, M., Dunhamr, H. M. (2010), EMM: Extensible Markov model for data stream clustering in R, *Journal of Statistical Software*, 35(5), 2-–31.

14. Huber, P. (2011), *Data analysis: what can be learned from the past 50 Years*, John Wiley & Sons. New York.

15. Jacod, J., Shiryaev, A.N., (2003), *Limit theorems for stochastic processes*, second ed., Springer-Verlag, New York.

16. Jureckova, J., Kalina, J., (2012). Nonparametric multivariate rank tests and their unbiasedness. *Bernoulli*, 18(1), 229--251.

17. Li, J., Liu, R. Y. (2004). New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science*, bf 19(4), 686--696.

18. Liu, R. Y. (1995). Control charts for multivariate processes. *Journ. of Amer. Stat. Assoc.*, 90, 1380--1387.

19. Liu, R. Y., Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journ. of Amer. Stat. Assoc.*, 88, 252--260.

20. Kosiorowski, D. (2012), Student depth in robust economic data stream analysis, Colubi A.(Ed.) Proceedings of COMPSTAT'2012, The International Statistical Institute/International Association for Statistical Computing.

21. Maronna, R. A., Martin, R. D., Yohai, V. J. (2006), Robust statistics - theory and methods. Chichester: John Wiley & Sons.

22. Mosler, K. (2002). *Multivariate dispersion, central regions and depth: The lift zonoid approach*, Springer, New York.

23. Muthukrishnan S., van den Berg E., Wu Y. (2007), Sequential change detection on data streams, ICDM Workshops 2007, 551-550.

24. Oja, H. (2010). *Multivariate nonparametric methods with R. An approach based on spatial signs and ranks*. *Lecture Notes in Statistics*, 199, Springer, New York.

25. Oja, H. and Randles R. H. (2004). Multivariate nonparametric tests. *Statistical Science*, 19, 598--605.

26. Rousseeuw, P. J., Hubert, M. (1999), Regression depth, *Journal of the American Statistical Association*, 94, 388 -- 433.

27. Shalizi C. R., Kontorovich, A. (2007), {\it *Almost none of the theory of stochastic processes - A course on random processes, for students of measure-theoretic probability, with a view to applications in dynamics and statistics*, http://www.stat.cmu.edu/~cshalizi/almost-none/

28. Serfling, R. (2006). Depth functions in nonparametric multivariate inference, In: Liu R.Y., Serfling R., Souvaine D. L. (Eds.): *Series in Discrete Mathematics and Theoretical Computer Science*, AMS, 72, 1 -- 15.

29. Song X., Wu M., Jermaine C., Ranka S. (2006), *Statistical change detection for multidimensional data*, ACM SIGKDD, 667-676.

30. Stockis, J-P.,Franke, J., Kamgaing, J. T. (2010). On geometric ergodicity of CHARME models, *Journal of the Time Series Analysis*, 31, 141--152.

31. Szewczyk, W. (2010), Streaming data, *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(1), (on-line journal).

32. Van Aelst, S., Rousseeuw, P. J. (2000), Robustness properties of deepest regression, *J. Multiv. Analysis*, 73, 82-106.

33. Zuo, Y. and He, X. (2006). On the limiting distributions of multivariate depth-based rank sum statistics and related tests. *Ann. Statist*., 34, 2879--2896.