Odporna eksploracja danych z wykorzystaniem pakietu DepthProc

Daniel Kosiorowski[1], **Zygmunt Zawadzki**[2]

[1]Uniwersytet Ekonomiczny w Krakowie, Katedra Statystyki

[2]Uniwersytet Ekonomiczny w Krakowie, student Analityki Gospodarczej

16 października 2014

Statystyka odporna

Procedury statystyczne konstruuje się przy założeniu, że spełniane są określone warunki dotyczące mechanizmu generującego dane (np. dane generowane są przez rozkład normalny i są niezależne od siebie).

W praktyce, możemy mieć do czynienia z odstępstwem od przyjmowanych założeń. Przykładowo w danych mogą występować obserwacje odstające, znacząco odbiegające od reszty danych. W takiej systuacji jakość procedury statystycznej może znacząco się obniżyć (utrata efektywności estymatora, wzrost obciążenia, itp).

Celem statystyki odpornej jest zaproponowanie procedur dających wiarygodne oszacowania również w przypadku gdy rozkład generujący dane odbiega od zakładanego rozkładu.

Statystyka odporna

"*Robustness theories can be viewed as stability theories of statistical inference. Robust statistics deals with stability, relative to model perturbation.* ***Hampel et all (1986)***"

Statystyka odporna - Punkt Załamania

Niech X^n będzie próbą o rozmiarze n. **Punkt załamania próby skończonej** (BP - ang. breakdown point) estymatora T zdefiniowany jest jako:

$$BP(T, X^n) = \left\{ \frac{m}{n} : \| T(X^n_m) - T(X^n) \| > \delta \right\},$$

gdzie X^n_m jest zanieczyszczoną próbą powstałą przez zastąpienie m punktów z X^n dowolnymi wartościami, $\|\cdot\|$ określa normę, δ to określony próg.

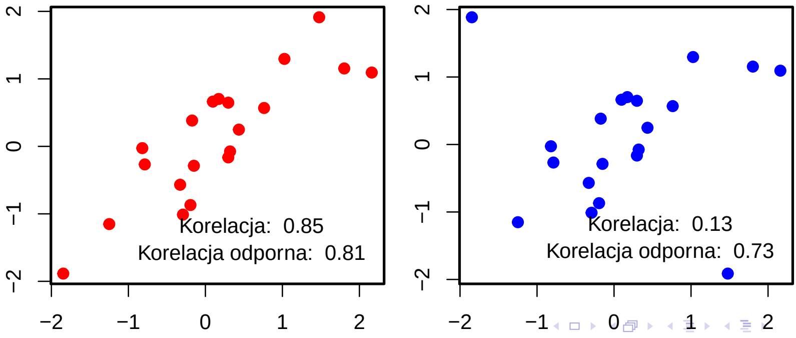
PUNKT ZAŁAMANIA PRÓBY SKONCZONEJ ESTYMATORA – najmniejsza frakcja złych obserwacji w próbie, która sprawia, że estymator staje się bezużyteczny – np. jego obciążenie staje się zbyt wysokie.

Przykładowe BP

BP dla odchylenia standardowego wynosi $\frac{1}{n} \approx 0$
BP dla mediany wynosi $\approx 50\%$

Statystyka odporna - przykład - korelacja

Dane jest dwadzieścia obserwacji z dwuwymiarowego rozkładu normalnego ze współczynnikiem korelacji 0.8. Dwie obserwacje zastąpiono obserwacjami odstającymi. Zmiana jedynie 10% oberwacji spowodowała drastyczną różnicę we wskazaniach klasycznego estymatora, natomiast w przypadku metody odpornej, wpływ obserwacji odstających został mocno ograniczony.



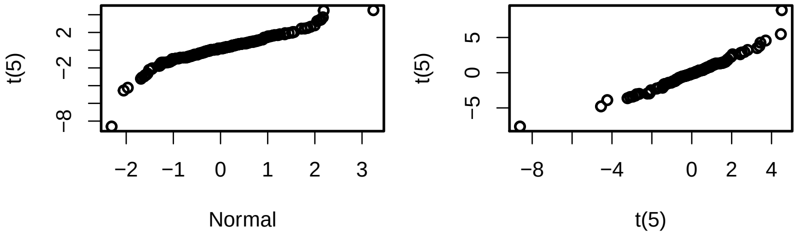
Eksploracyjna analiza danych

Techniki **Eksploracyjnej Analizy Danych** (**EAD**) pozwalają lepiej zrozumieć dane z którymi ma się do czynienia. Stosunkowo proste wykresy jak histogram czy boxplot pozwalają w przybliżony sposób określić rozkład danych, lub sprawdzić czy występują obserwacje odstające[4].

Wiele użytecznych i prostych w interpretacji technik EAD jest ograniczona jedynie do danych jednowymiarowych (histogram, boxplot, wykres kwantyl-kwantyl), bądź dwuwymiarowych (heatmapy). Z praktycznego punktu widzenia wartościowe byłoby uogólnienie ich na większą liczbę wymiarów, zachowując jednocześnie prostotę interpretacji.

Wykres kwantyl-kawntyl

Bardzo użyteczną techniką EAD jest wykres kwantyl-kwantyl, pozwalający w graficzny sposób sprawdzić, czy dane mogą być generowane przez zakaładny rozkład, lub czy dwie próby pochodzą z tego samego rozkładu.



Posiada on jednak tę kluczową wadę, że do jego konstrukcji wymagany jest kwantyl - przez co trudno uogólnić taki wykres na przypadek wielowymiarowy. Definicja kwantyla opiera się na porządku liniowym liczb rzeczywistych w jednym wymierze. W wielu wymiarach nie ma prostej definicji takiego porządku, przez co wykres kwantyl-kwantyl ograniczony jest jedynie do rozkładów jednowymiarowych.

Koncepcja głębi danych

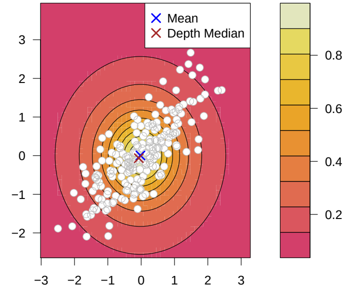
Statystyczna funkcja głębi ma na celu kompensacje braku porządku liniowego w $\mathbb{R}^d, d \leq 2$. Zakładając pewien rozkład prawdopodobieństwa F na $\mathbb{R}^d$, funkcja głębi $D(x, F)$ umożliwia porządkowanie punktów na zasadzie odstawania od centrum rozkładu.
W przypadku próby $X^n = \{x_1, ..., x_n\}$ rozkład F zastępowany jest rozkładem F_n wyznaczonym na podstawie X^n.
Formalną definicję funkcji głębi można znaleźć w [3] i [6].

Najprostszy przykład - głebia Euklidesa

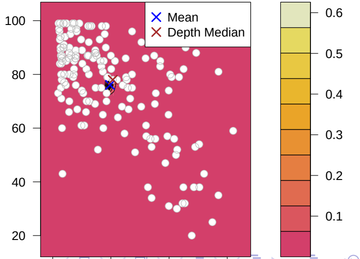
Głębia Euklidesa:

$$D_E(x, X^n) = \frac{1}{1 + ||x - \bar{x}||}, \qquad (1)$$

gdzie $\bar{x}$ to wektor średnich z próby X^n.



Zaletą tej głębi jest jedynie szybkość obliczeń. Jednak w praktycznych przypadkach głębia Euklidesa nie ma zastosowania - nie radzi sobie z eliptycznym kształtem danych, lub skośnymi danymi.



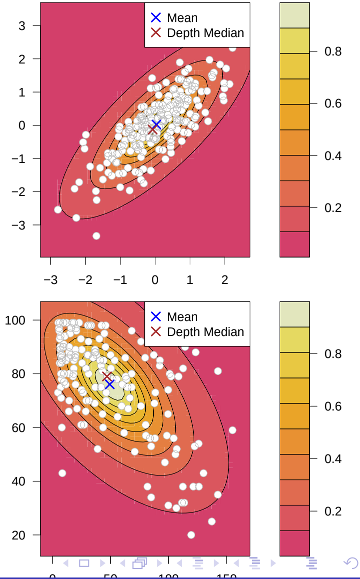
Głębia Mahalanobisa

Głębia Mahalanobisa zdefiniowana jest jako:

$$D_E(x, X^n) = \frac{1}{1 + (x - \bar{x})^T \Sigma (x - \bar{x})}, \quad (2)$$

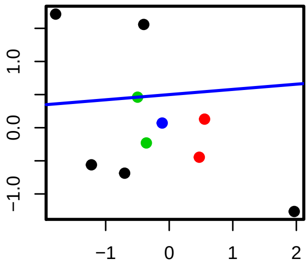
gdzie $\bar{x}$ to wektor średnich z próby X^n.

Głębia Mahalanobisa podobnie jak głębia Euklidesa opiera się na odległości. W tym przypadku by otrzymać wiarygodne wartości, należałoby zastosowac odporny estmator macierzy kowariancji, jak również odporną miarę położenia. W takim przypadku ciężko zdefiniować nową miarę położenia opartą na funkcji głębi.



Głębia Tukey'a

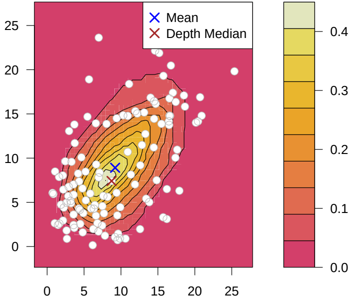
Funkcją głębi przyporządkowująca punktowi x najmniejsze prawdopodobieństwo zgromadzone na domkniętej półprzestrzeni, do której brzegu należy ten punkt, nazywamy głębią domkniętej półprzestrzeni Tukeya.

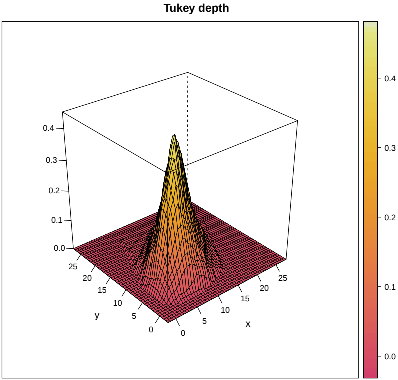


Ta funkcja głębi nie opiera się na infrmacji metrycznej, dotyczącej odległości pomiędzy punktami, tylko na ich wzajemnym położeniu. Dlatego też głębia punktu leżącego z dala od chmury punktów, będzie taka sama jak punktu leżącego na jej skraju.

Głębia Tukey'a. cd - wywołanie funkcji

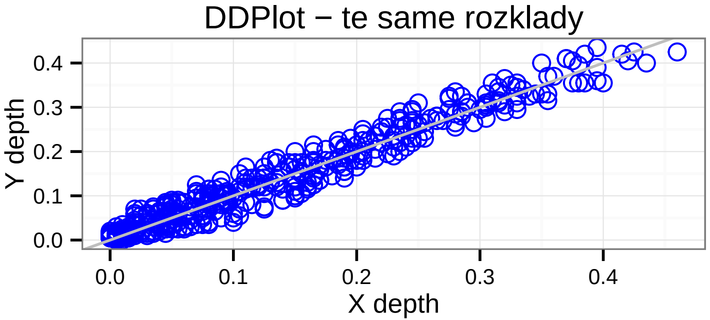
```
depthContour(x, method = "Tukey", points = TRUE)
depthPersp(x, method = "Tukey")
```



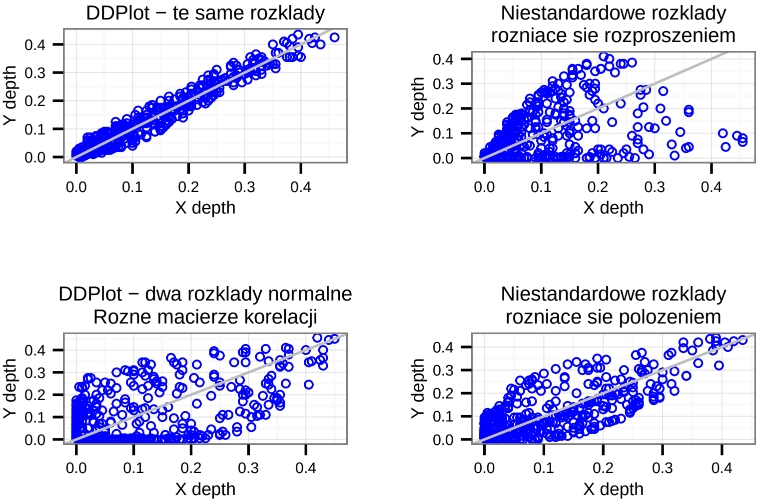


Wykres głębia-versus-głebia

Wykres głębia-versus-głębia (ddPlot) porównuje wartości funkcji głębi punktu x, przy założeniu, że punkt generowany jest przez rozkład F, lub G (w praktycznych przypadkach F i G są zastępowane estymatorami z próby). Jeżeli $F = G$, wtedy ddPlot jest odcinkiem o końcach w $(0, 0)$ i $(1, 1)$.
Interpretacja tego wykresu jest więc zbliżona do interpretacji wykresu kwantyl-kwantyl.



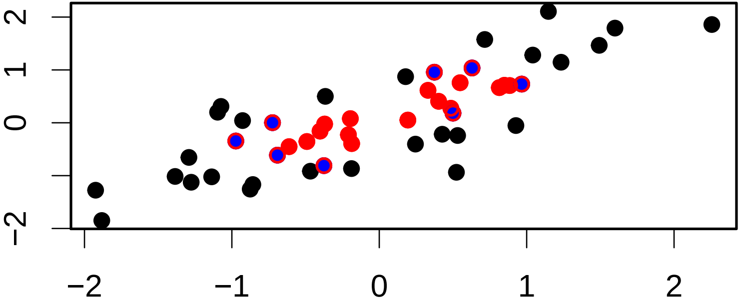
Wykres głębia-versus-głębia CD.



Obszar centralny

Definition

Obszarem centralnym rzędu α, $PC_F(\alpha)$ nazywamy zbiór punktów $x \in \mathbb{R}^d$, takich, że $D(x, F) \geq \alpha$.



Krzywa skali

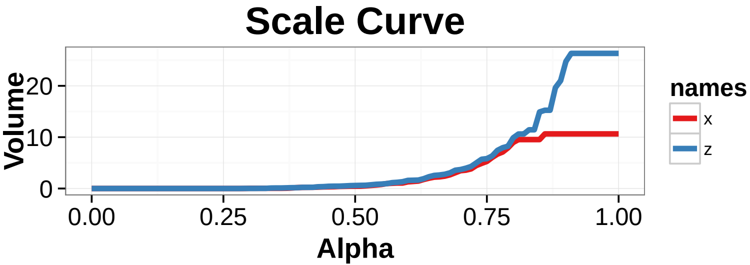
Krzywa skali zdefiniowana jest jako:

$$SC(\alpha) = (\alpha, vol(D_\alpha(Z^n)) \subset \mathbb{R}^2, dla \alpha \in [0, 1], \tag{3}$$

gdzie $vol(D_\alpha(Z^n)$, jest to objętość otoczki wypukłej wznaczonej dla punktów znajdujących się w obszarze centralnym rzedu α. Taka definicja pozwala na bardzo intuicyjną interpretację - im większe wartości na tej krzywej, tym zbiór jest bardziej rozproszony.

Krzywa skali - cd.

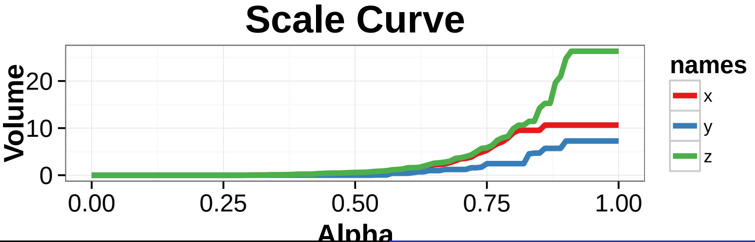
```
set.seed(123)
sigma = cbind(c(1,0.8),c(0.8,1))
x = mvrnorm(180, mu = c(0,0), sigma)
y = mvrnorm(20, mu = c(1,-2), sigma*1.5)
z = rbind(x,y)
sc = scaleCurve(z,x, name = "z", name_y = "x")
# Wykresy wspolpracuja z ggplot2:
getPlot(sc) + scale_color_brewer(palette = "Set1")
```



Krzywa skali - cd. łączenie wykresów

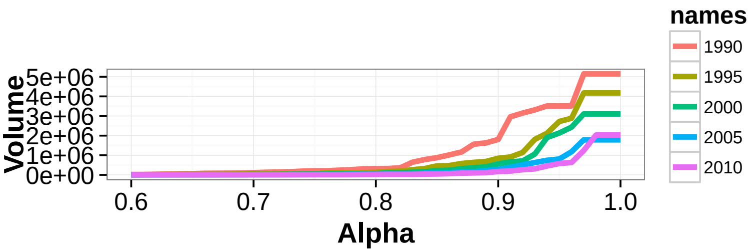
Standardowo funkcja scaleCurve może posłużyć do estymacji krzywej skali dla jednego, lub dwóch zbiorów danych. By zestawić ze sobą więcej krzywych należy skorzystać z operatora $\% + \%$, który pozwala na "dodawanie" do siebie wykresów.

```
sc_z = scaleCurve(z,name = "z")
sc_x = scaleCurve(x,name = "x")
sc_y = scaleCurve(y,name = "y")

sc = sc_x %+% sc_z %+% sc_y
getPlot(sc) + scale_color_brewer(palette = "Set1")
```



Krzywa skali. Łączenie wykresów, cd.

```
scWrap = function(x) scaleCurve(x[,3:5],
              name = as.character(x[1,2] %>% unlist))
sc_curves = all_data %>% group_by(Year) %>%
  filter(as.numeric(Year)%%5 == 1) %>%
  do(scale_curves = scWrap(.))

scale_curves = sc_curves$scale_curves
scurves = Reduce("%+%", scale_curves)
getPlot(scurves) + xlim(c(0.6,1)) +
  ggtitle("") + guides(col = guide_legend(nrow = 7))
```



Wielowymiarowa mediana

Definition

Punkt o najwyższej wartości funkcji głębi będziemy utożsamiać z wielowymiarową medianą.

```
data2010 = all_data %>% filter(Year == "2010") %>%
    select(maesles.imm, under5.mort, inf.mort) %>%
  na.omit

depthMedian(data2010, method = "Tukey")

## maesles.imm under5.mort    inf.mort
##        60.0        20.2        15.7
```

Wielowymiarowy test Wilcoxona

Dla próby $X^m = \{X_1, ..., X_m\}$, $Y^n = \{Y_1, ..., Y_n\}$, i połączonej próby $Z = X^n \cup Y^m$ **statystyka Wilcoxona** zdefiniowana jest jako

$$S = \sum_{i=1}^{m} R_i, \tag{4}$$

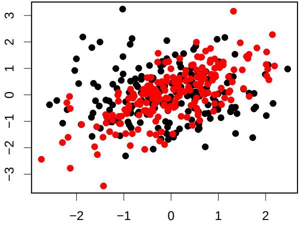
gdzie R_i oznacza rangę i-tej obserwacji, $i = 1, ..., m$ w połączonej próbie Z:

$$R(x_l) = \#\left\{z_j \in Z : D(z_j, Z) \leq D(x_l, Z)\right\}, l = 1, ..., m. \tag{5}$$

Rozkład S jest symetryczny względem $E(S) = 1/2m(m+n+1)$, a jego wariancja wynosi $D^2(S) = 1/12\, mn(m+n+1)$. Wiecej na ten temat można znaleźć w pracach [2] i [5].

Wielowymiarowy test Wilcoxona - przykład

```r
set.seed(123)
x  = mvrnorm(200, c(0,0), diag(2))
sigma = cbind(c(1,0.7), c(0.7,1))
y = mvrnorm(200, mu = c(0,0), sigma)

mWilcoxonTest(x,y,
    alternative = "two.sided")
```



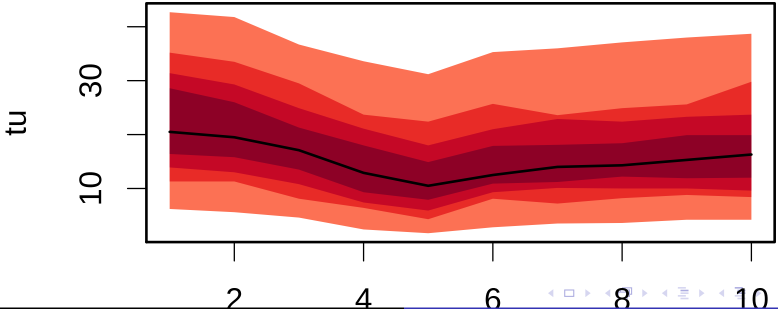

```
##
##  Multivariate Wilcoxon test for equality of dispersion
##
## data:  dep_x and dep_y
## W = 17133, p-value = 0.01316
## alternative hypothesis: true dispersion ratio is not equal to 1
```

Uogólniona głębia pasma

Uogólniona głębia pasma została zaproponowana w kontekście odpornej klasyfikacji obserwacji będących funkcjami.
Intuicyjnie określa średnią frakcję czasu jaką dana trajktoria znajduje się pomiędzy dwiema innymi trajektoriami z dostępnego zbioru.
Formalną definicję można znaleźć w [1].

MBD - przykład

Poniżej zaprezentowano trajektorie bezrobocia rejestrowanego dla powiatów w Polsce od roku 2004 do 2013. Kolorami zaznaczono obszary w których znajduje się odpowiednio 100%, 75%, 50% i 25% najbardziej centralnych krzywych. Szerokość poszczególnych pasm zmienia się w podobny sposób, co sugeruje, że pod względem zatrudnienia różnice pomiędzy powiatami na przestrzeni lat utrzymują się na zbliżonym poziomie.



Szczegóły implementacyjne

Pakiet został napisany w dużej mierze z pomocą **Rcpp** i **RcppArmadillo**, co oznacza, że tak naprawdę spora część kodu jest napisana nie w **R**, tylko w **C++**.
Gdzie to możliwe wykorzystywane są obliczenia równoległe, zrealizowane na poziomie C++ przy pomocy **OpenMP** (jeżeli OpenMP jest niedostępne, DepthProc będzie ograniczony do jednego rdzenia). Domyślnie wykorzystywane są wszystkie dostępne rdzenie procesora, można jednak kontrolować to przy pomocy parametru *threads*:

```
x = matrix(rnorm(800000), ncol = 20)
system.time(d <- depth(x))

##    user  system elapsed
##   9.584   0.156   6.900

# jeden watek:
system.time(d <- depth(x, threads = 1))

##    user  system elapsed
##   8.269   0.020   8.371
```

Co dalej?

- **Odporna regresja wykorzystująca głębię regresyjną** - w DepthProc dostępna wersja 2d.
- **Krzywa asymetrii.**
- **Głębie lokalne.**

D. Kosiorowski.
Statystyczne funkcje głębi w odpornej analizie ekonomicznej.
Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, 2012.

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

R. Y. Liu, J. M. Parelius, and K. Singh.
Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion).
The Annals of Statistics, 27:783–858, 1999.

John W. Tukey.
We Need Both Exploratory and Confirmatory.
The American Statistician, 34(1):23–25, 1980.

Y. Zuo and X. He.
On the limiting distributions of multivariate depth-based rank sum statistics and related tests.
The Annals of Statistics, 34:2879–2896, 2006.

Y. Zuo and R. Serfling.
General notions of statistical depth function.
The Annals of Statistics, 28:461–482, 2000.