Odporna eksploracja danych z wykorzystaniem pakietu DepthProc

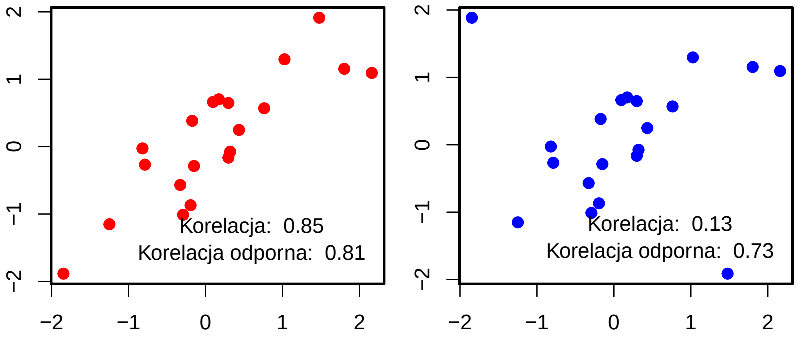
Daniel Kosiorowski, Zygmunt Zawadzki

7 października 2014

Statystyka odporna

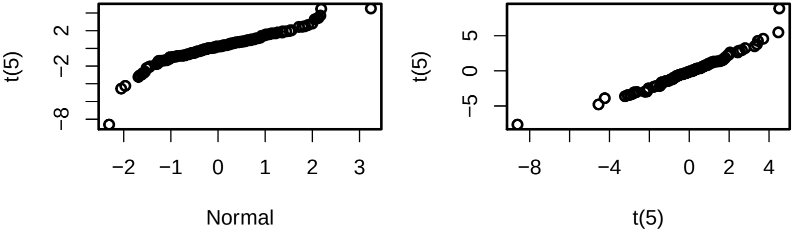
Przykład - korelacja

Dwadzieścia obserwacji z dwuwymiarowego rozkładu noramalnego 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.



Wykres kwantyl-kawntyl

Bardzo użyteczną metodą na etapie eksploracji danych 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 trudno zastosowanie tego typu wykresu do danych wielowymiarowych jest niemożliwe.

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.

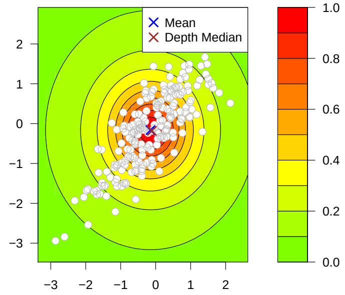
Najprostszy przykład - głebia Euklidesa

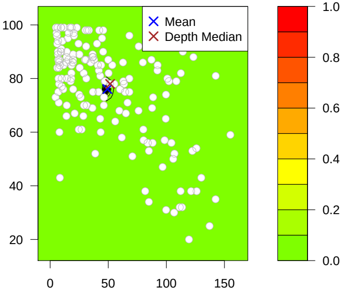
Głębia Euklidesa:

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

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.



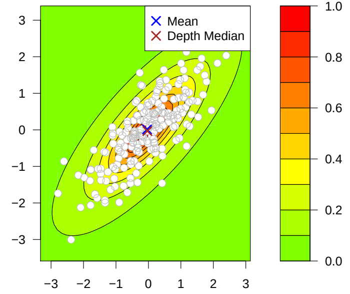


Głębia Mahalanobisa

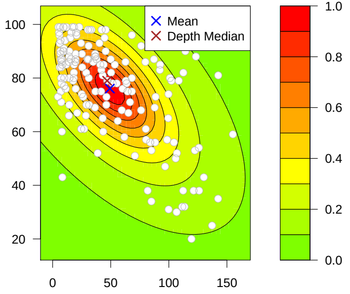
Głębia Mahalanobisa zdefiniowana jest jako:

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

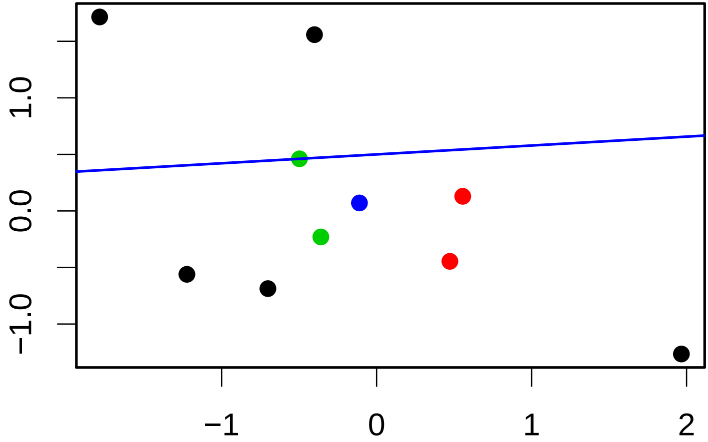
gdzie $\bar{x}$ to wektor średnich z próby X^n.



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

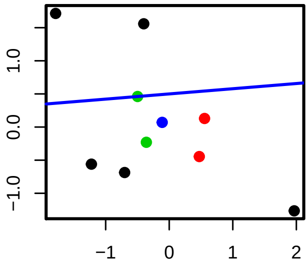


Głębia Tukey'a



Głębia Tukey'a

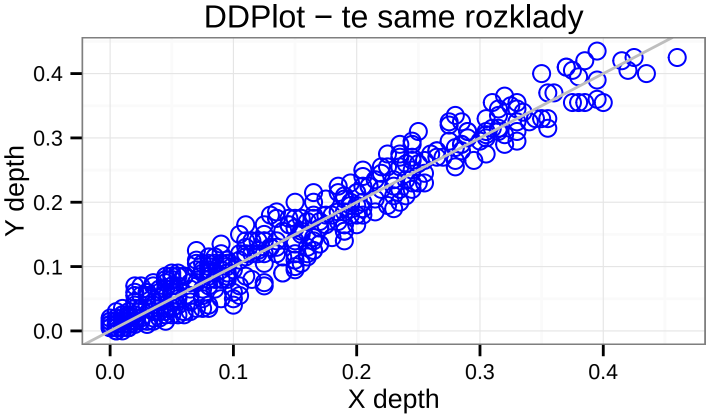
Funkkcją głębi przyporządkowywującej punktowi x najmniejsze prawdopodobieństwo zgromadzonena domkniętej półprzestrzeni, do której brzegu należy ten punkt, nazywamy głębią domkniętej półprzestrzeni Tukeya.



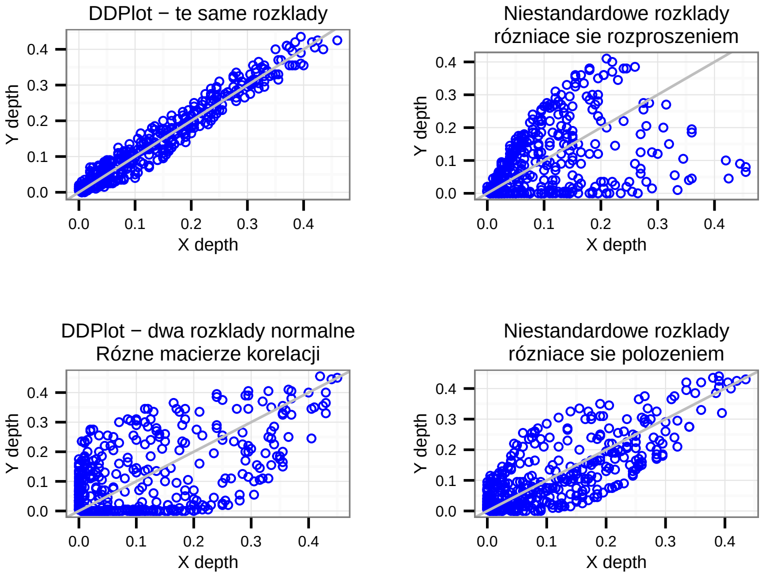
Ta funckja głębi nie opiera się na infrmacji metrycznej, dotyczącej odległości pomiędzy punktami, tylko na ich wzajemnym położeniu. Dlatego też głębia punktu leżącego z dala od chmury punktów, będzie taka sama jak punktu leżącego na jej skraju.

Wykres głębia-versus-głebia

Wykres głębia-versus-głębia (ddPlot) porównuje wartości funckji 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 $(1, 1)$.



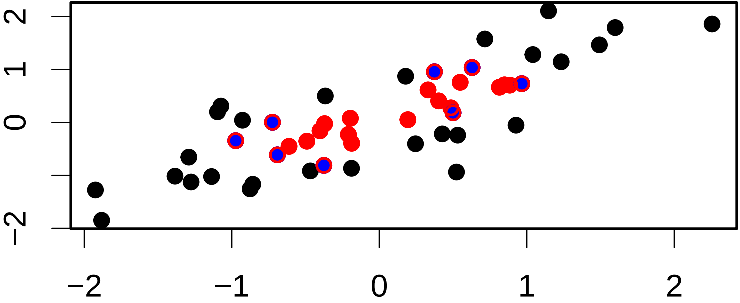
Wykres głębia-versus-głebia CD.



Obszar centralny

Definition

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



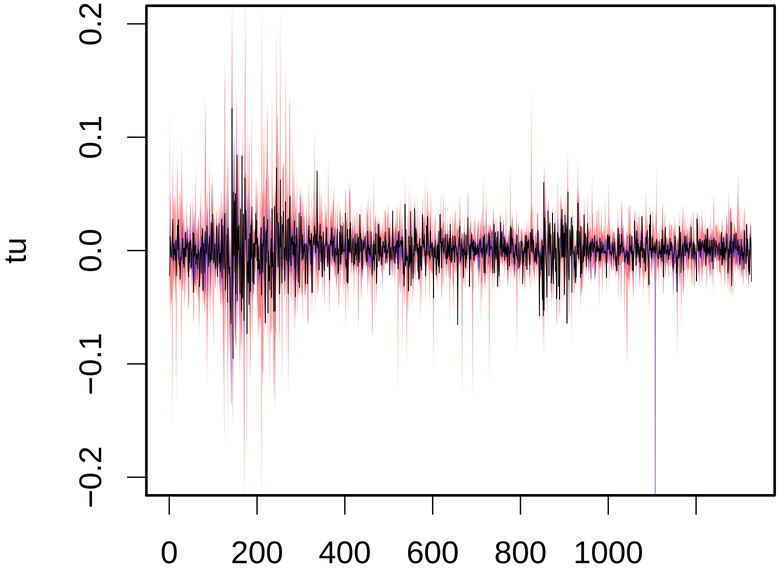
Wielowymiarowa mediana i obszar centralny

Definition

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

Definition

Zbiór punktów


tu
0.2
0.1
0.0
−0.1
−0.2
0
200
400
600
800
1000