Sprawozdanie 2
Modelowanie i identyfikacja

Maciej Kajdak
nr indeksu: 226256

2 maja 2019

1 Wstęp

Empriyczne wyznaczanie własności pewnych sygnałów, które przypuszczalnie rządzą się prawami pewnego nieznanego rozkładu jest zadaniem, które stawiane jest przy problemach identyfikacji systemów. Dzięki poprawnemu zidentyfikowaniu obiektu, który się obserwuje można przewidywać jego zachowania, a następnie odpowiednio taką wiedzę wykorzystywać.

W tym celu potrzebne jest narzędzie – estymator, pozwalający na wyznaczenie a następnie weryfikowanie poprawności przypuszczeń. W niniejszym sprawozdaniu opisano przebieg badań estymatorów pewnych rozkładów prawdopodobieństwa i ich własności. Większość badań została przeprowadzona w środowisku Python w wersji 3.7.1 [3] wraz z pakietami numpy w wersji 1.15.4 [2] oraz matplotlib w wersji 3.0.2 [1]. Wszystkie stworzone w ramach laboratoriów skrypty można znaleźć w repozytorium twórcy w serwisie Github

2 Podstawy estymacji

W celu wyznaczenia cech jakimi charakteryzuje się ciąg zmiennych losowych można posłużyć się empirycznymi metodami wyznaczania estymatorów. Do badania cech estymatorów wygenerowano ciąg zmiennych losowych o rozkładzie normalnym $\mathcal{N}(0, 1)$.

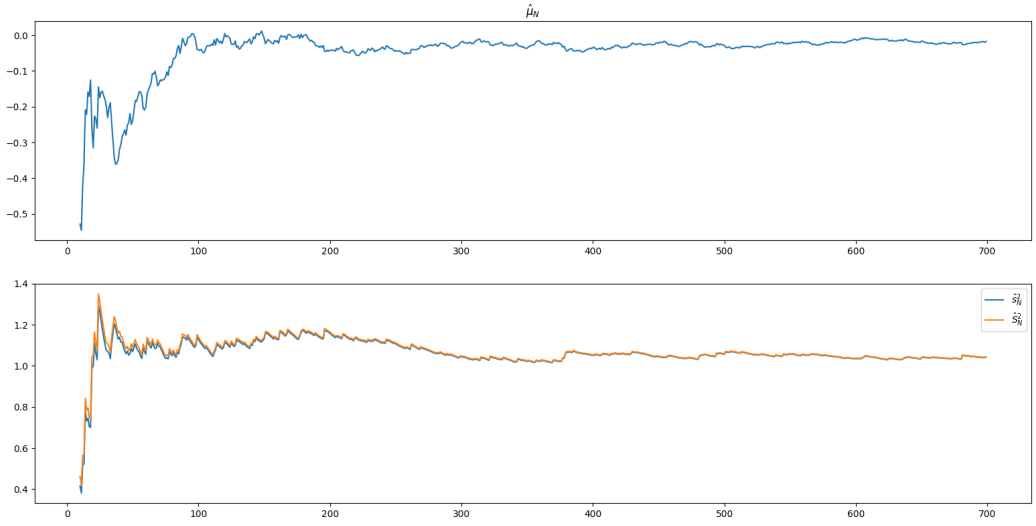
2.1 Estymatory empiryczne

Na podstawie wygenerowanych próbek wyznaczono wartości $\hat{\mu}_N$, obliczanego przy pomocy wyrażenia 1, w funkcji N. Następnym krokiem było wyznaczenie estymatorów wariancji obciążonej (wyrażenie 2 oraz nieobciążonej 3. Sprawdzano w ten sposób jaki wpływ na jakość estymacji wartości oczekiwanej oraz wariancji ma ilość wygenerowanych próbek. Wyniki badań jako funkcje poszczególnych estymatorów od x przedstawiono na rysunku 1. W ten sam sposób wyznaczono estymatory dla rozkładu Cauchyego, zostały one przedstawione na rysunku 2. Dla rozkładu normalnego estymatory w miarę zwiększania ilości badanych próbek stabilizują się na właściwych wartościach, tzn na $\mu = 0$ oraz $\sigma^2 = 1$. Z kolei dla rozkładu Cauchy'ego wyznaczanie estymatorów kończy się ogromnymi błędami. Jest to spowodowane faktem, że rozkład Cauchy'ego nie ma ani wartości średniej ani wariancji, więc wynik eksperymentu był oczekiwany.

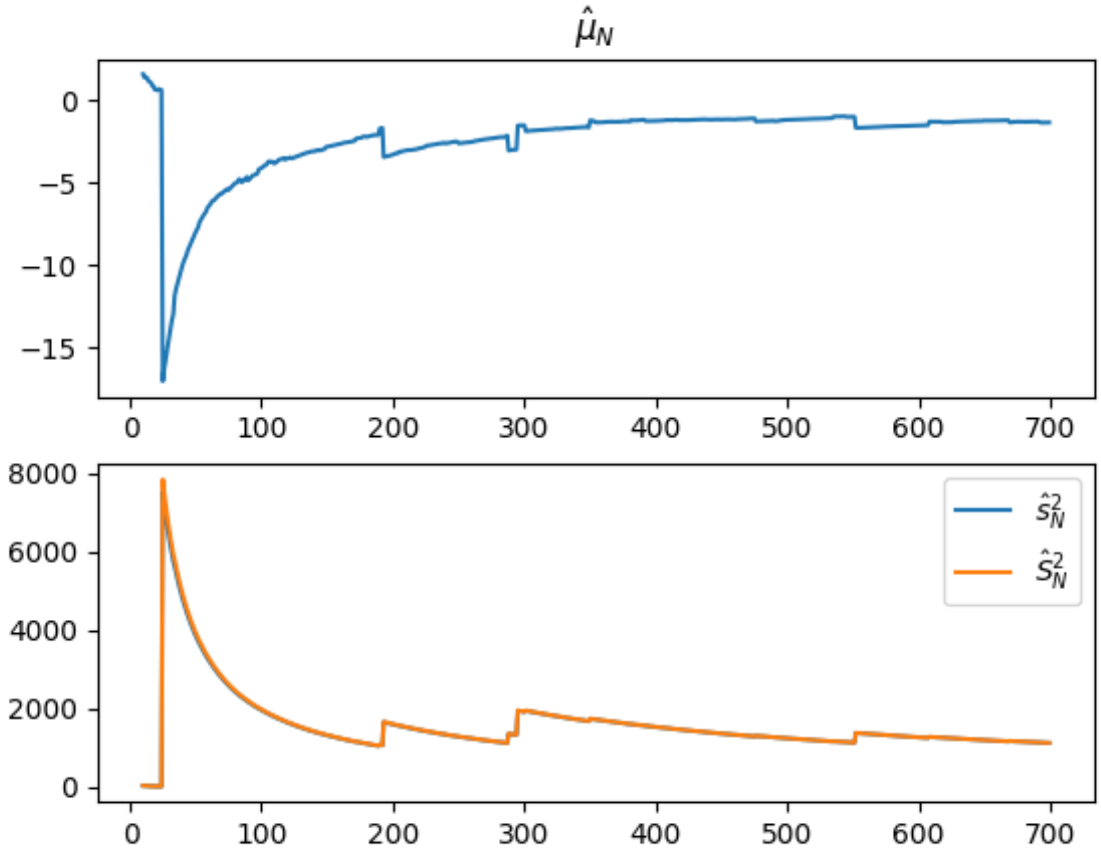
$$\hat{\mu}_N = \frac{1}{N} \sum_{n=1}^{N} X_N \tag{1}$$

$$\hat{s}_N^2 = \frac{1}{N}\sum_{n=1}^{N}(X_N - \hat{\mu}_N) \tag{2}$$

$$\hat{S}_N^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_N - \hat{\mu}_N) \tag{3}$$



Rysunek 1: Wykresy estymatorów empirycznych w funkcji N dla rozkładu normalnego



Rysunek 2: Wykresy estymatorów empirycznych w funkcji N dla rozkładu Cauchy'ego

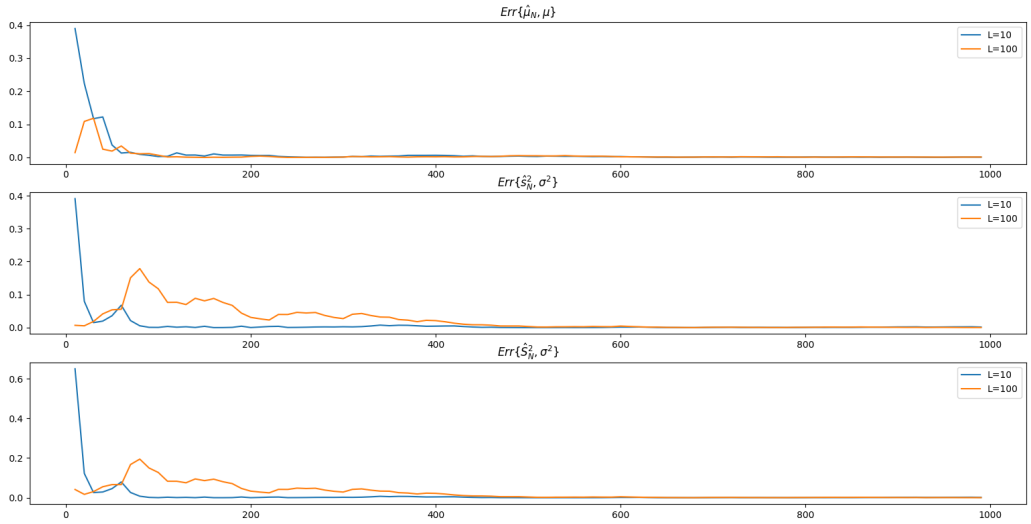
2.2 Błąd empiryczny

Aby udowodnić poprawność i skuteczność działania estymatorów empirycznych można posłużyć się wskaźnikiem wyznaczającym błąd empiryczny. Jest on definiowany wzorem 4 oraz analogicznie dla wariancji obciążonej (5) i nieobciążonej (6).

$$Err\{\hat{\mu}_N;\mu\} = \frac{1}{L}\sum_{l=1}^{L}\left[\hat{\mu}_N^{[l]} - \mu\right]^2 \tag{4}$$

$$Err\{\hat{s}_N;\mu\} = \frac{1}{L}\sum_{l=1}^{L}\left[(\hat{s}_N^{[l]})^2 - \sigma^2\right]^2 \tag{5}$$

$$Err\{\hat{S}_N;\mu\} = \frac{1}{L}\sum_{l=1}^{L}\left[(\hat{S}_N^{[l]})^2 - \sigma^2\right]^2 \tag{6}$$



Rysunek 3: Wykresy błędów estymatorów dla rozkładu normalnego.

3 Dystrybuanta empiryczna i jej własności

W celu przebadania własności estymatora dystrybuanty empirycznej wygenerowano ciąg liczb losowych przy pomocy wcześniej zaimplementowanego generatora. Gęstość rozkładu prawdopodobieństwa, z którego generowane były liczby przedstawiono w postaci równania 7. W wyniku generacji otrzymano ciąg liczb losowych $\{X_1, X_2, \ldots, X_N\}$.

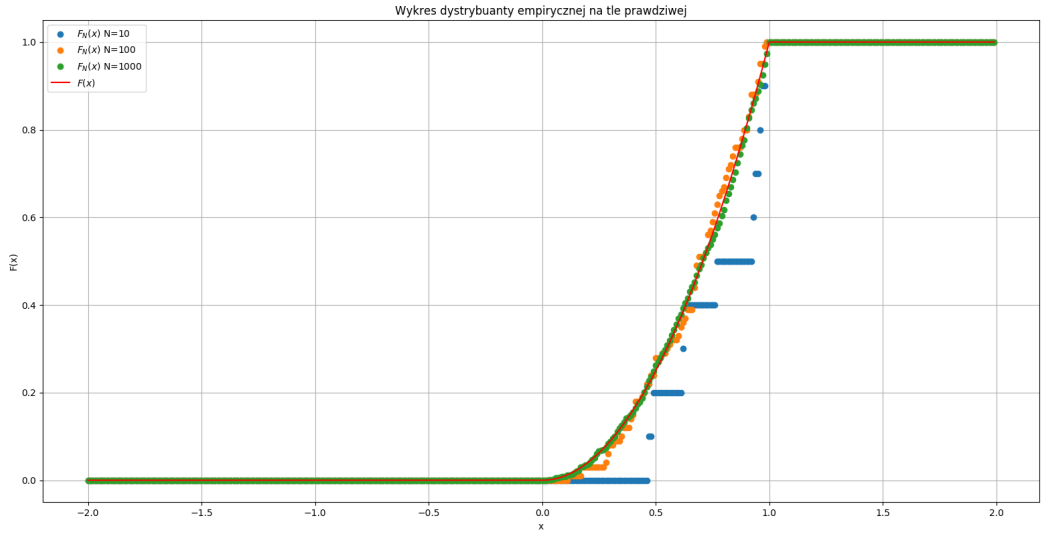
$$f(x) = \begin{cases} 2x, & \text{dla } x \in [0,1] \\ 0, & \text{dla } x \in (-\infty, 0) \cup (1, \infty) \end{cases} \tag{7}$$

Dla tak wygenerowanego przebiegu wyznaczono dystrybuantę empiryczną zaimplementowaną w postaci 4. Wykresy dystrybanty empirycznej dla różnych wartości N przedstawiono na rysunku 4. Jak można się było spodziewać, dla małej ilości próbek dystrybuanta nie działa dobrze, jednak wraz ze wzrostem ilości próbek dystrybuanta empiryczna coraz bardziej przypomina rzeczywisty przebieg.

$$\hat{F}_N(x) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{I}(X_n \leq x) \tag{8}$$

, gdzie

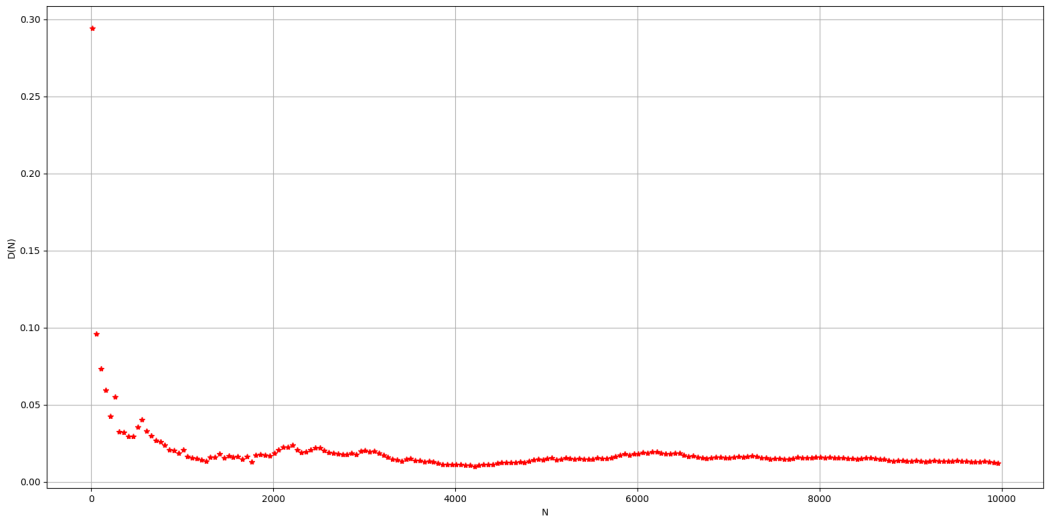
$$\mathbf{I}(a \leq b) = \begin{cases} 1 & \iff a \leq b \\ 0 & \iff a > b \end{cases} \tag{9}$$



Rysunek 4: Wykresy dystrybuanty empirycznej na tle prawdziwego przebiegu dystrybuanty dla różnej wartości N

Tak tworzona dystrybuanta empiryczna może również zostać poddana testom sprawdzającym jej wiarygodność, tzn. czy dystrybuanta empiryczna faktycznie przebiegiem przypomina tę poszukiwaną. Można do tego celu wykorzystać wskaźnik przedstawiany wyrażeniem 10. Mówiąc ściślej, badanie jakości dystrybuanty empirycznej takim wskaźnikiem nazywane jest *testem Kołmogorowa*. Bada on różnice między hipotetyczną dystrybuantą F a dystrybuantą empiryczną $\hat{F}_N$. Dla omawianego problemu wyznaczono taki wskaźnik i przedstawiono go na rysunku 5 w zależności od liczby próbek. Na pierwszy rzut oka widać jak wielki wpływ ma ilość próbek na jakość estymacji, ponieważ wraz ze wzrostem poziomu wiedzy o wyjściu systemu generującego przebieg o takiej dystrybuancie zmniejsza się błąd, a tym samym można przypuszczać, że dystrybuanta empiryczna wyznaczana jest poprawnie, a generowane liczby faktycznie pochodzą z pożądanego rozkładu.

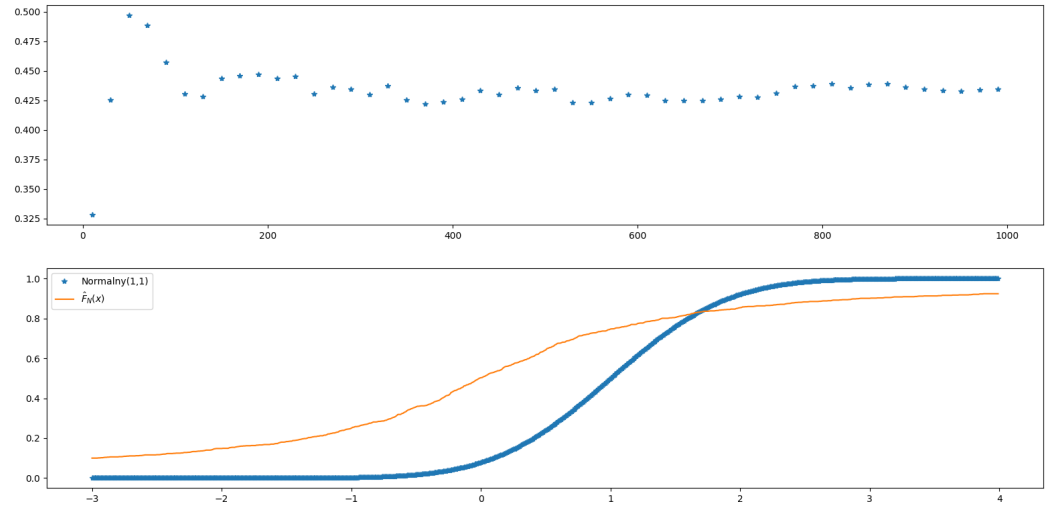
$$D_N = \sup_x |\hat{F}_N(x) - F(x)| \tag{10}$$



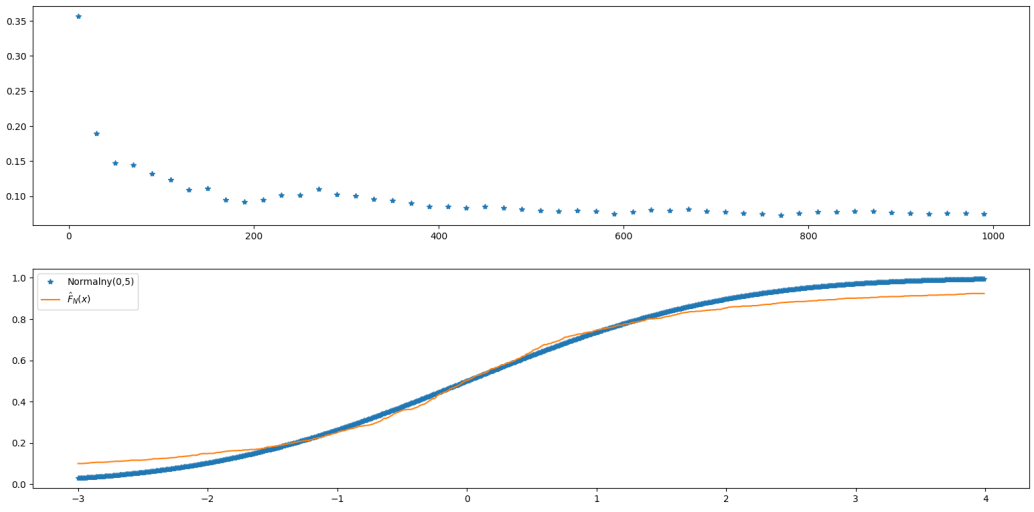
Rysunek 5: Błąd dystrybuanty empirycznej

3.1 Badanie nieznanej próby

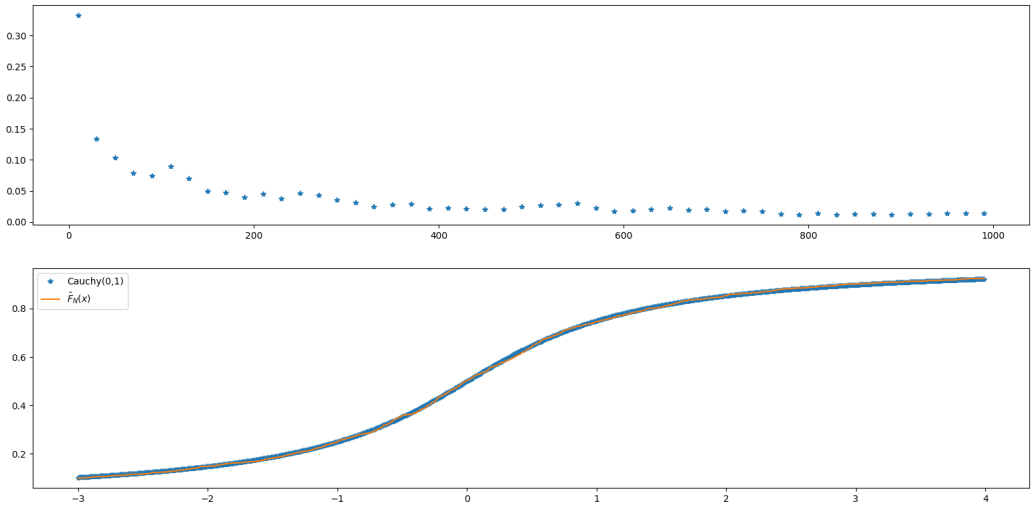
W celu sprawdzenia sposobu działania dystrybuanty empirycznej wykorzystano ciąg liczb losowych pochodzący z nieznanego rozkładu prawdopodobieństwa. Aby określić z z jakiego rozkładu mogą pochodzić próbki wyznaczono dystrybuantę empiryczną oraz przetestowano estymator pod kątem trzech hipotetycznych rozkładów: $\mathcal{N}(1,1)$, $\mathcal{N}(0,5)$ oraz Cauchy'ego z parametrami $x_0 = 0$ oraz $\gamma = 1$. Na rysunkach 6, 7, 8 przedstawiono przebiegi estymatora na tle dystrybuant ww. rozkładów oraz błąd w zależności od ilości wziętych próbek. Na tej podstawie można wnioskować, że najprawdopodobniej próbki pochodzą z rozkładu Cauchy'ego.



Rysunek 6: Porównanie dystrybuanty empirycznej z rozkładem normalnym (1,1)



Rysunek 7: Porównanie dystrybuanty empirycznej z rozkładem normalnym (0,5)

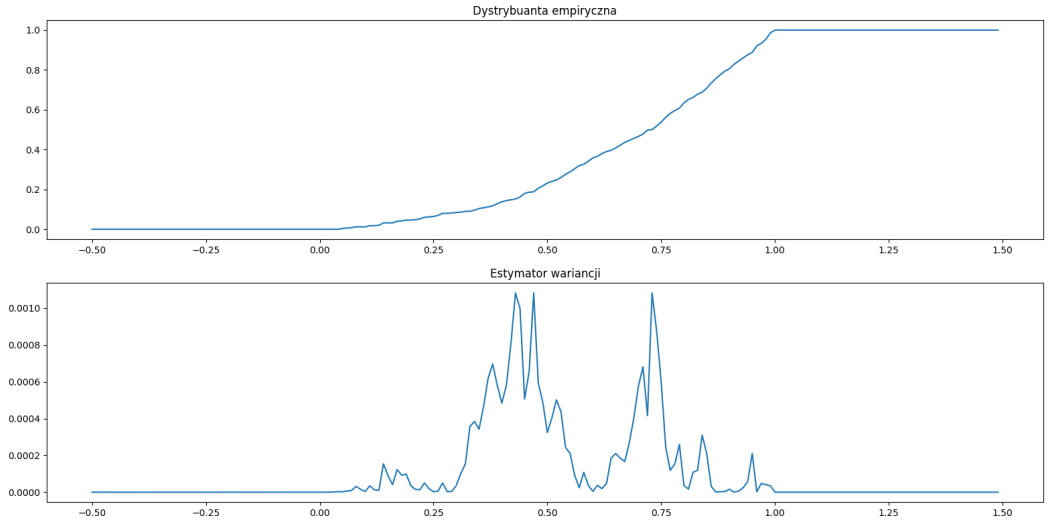


Rysunek 8: Porównanie dystrybuanty empirycznej z rozkładem Cauchy'ego (0,1)

3.2 Estymator wariancji dystrybuanty empirycznej

Do wyznaczenia wariancji dystrybuanty empirycznej wykorzystano fakt, że wariancja jest kwadrat różnicy między próbką z dystrybuanty empirycznej a wartością oczekiwaną rzeczywistej dystrybuanty. Dla rozkładu 7 wykres wariancji dystrybuanty empirycznej przedstawiono na rysunku. Z tego co przedstawia wykres można wnioskować, że najtrudniejsza estymacja dystrybuanty jest tam, gdzie zmiany przebiegu są coraz większe stąd większy rozrzut próbek wokół wartości oczekiwanej – na wykresie widoczne zwiększenie wariancji.

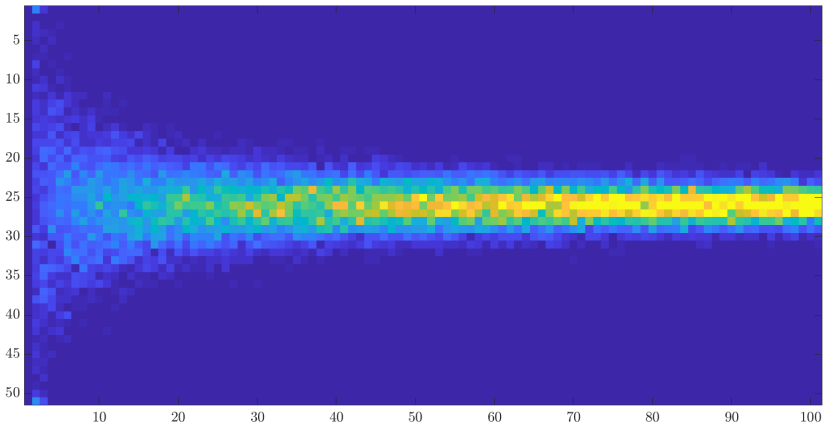
$$\sigma_F^2(x) = (\hat{F}_N(x) - F(x))^2 \tag{11}$$



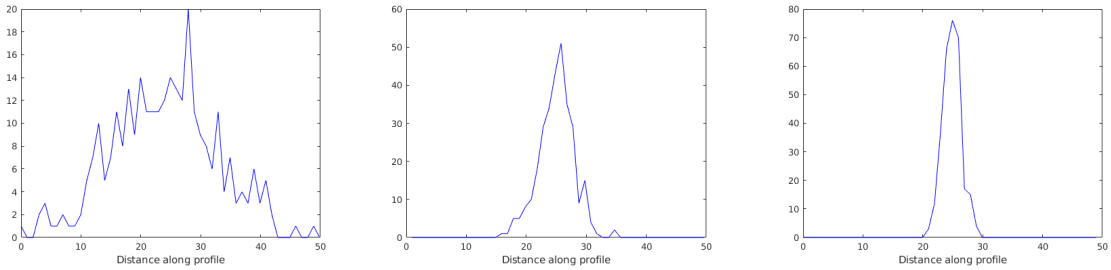
Rysunek 9: Wykres wariancji dystrybuanty empirycznej

4 Centralne twierdzenie graniczne

Jednym z ważniejszych twierdzeń w teorii probabilistyki jest właśnie centralne twierdzenie graniczne, ponieważ uzasadnia ono, dlaczego w przyrodzie tak wiele losowych elementów rządzi się rozkładem normalnym. Generując ciągle próbki o tej samej wartości oczekiwanej oraz tej samej wariancji stale zwiększając liczbę wygenerowanych próbek można zaobserwować, że histogram próbek coraz bardziej przypominać będzie rozkład normalny. Wykres przedstawiony na rysunku 10 przedstawia kolejne generowane ciągi próbek o tej samej wariancji i wartości oczekiwanej. Jak można zauważyć coraz mocniejsza żółć środkująca się w poziomie oznacza, że histogram generowanych próbek zbiega do rozkładu normalnego. Lepiej można ten proces zauważyć na rysunkach , które są przekrojami wykresu z rysunku 10 z początku, połowy i końca generowania kolejnych prób. Wraz ze wzrostem wygenerowanych próbek rozkład budowany na ich podstawie coraz zbiega do rozkładu normalnego.



Rysunek 10: Wykres kształtu rozkładu normalnego w zależności od liczby wygenerowanych próbek



Rysunek 11: A figure with two subfigures

5 Jądrowy estymator gęstości prawdopodobieństwa

Podczas identyfikowania systemów pojawia się zadanie zbadania tego jaki rozkład prawdopodobieństwa może przedstawiać badany ciąg liczb losowych. Aby to sprawdzić potrzebne jest narzędzie, które mogłoby odtwarzać ten rozkład. Do tego celu może posłużyć jądrowy estymator gęstości prawdopodobieństwa. Wyraża się on jako równanie 12, gdzie $K(\cdot)$ jest jądrem estymatora.

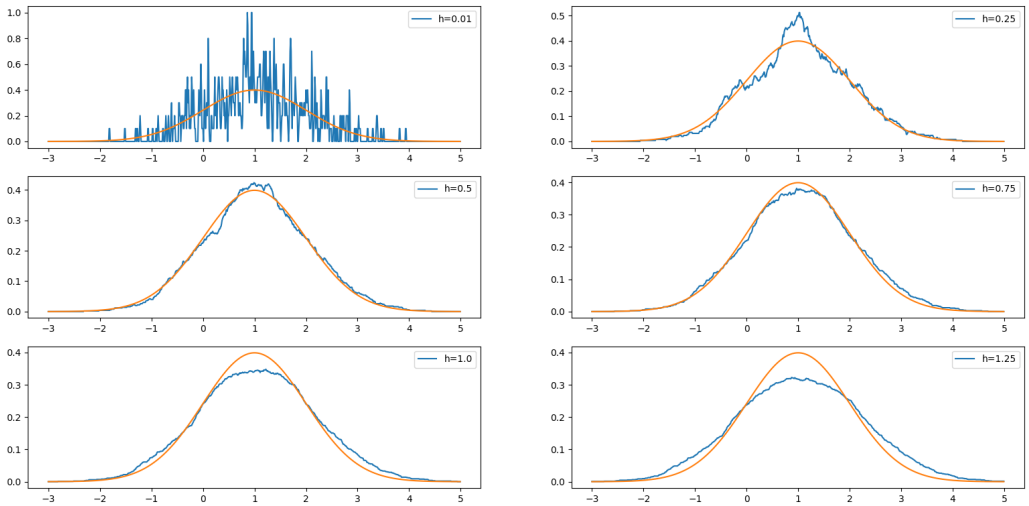
Do badań ww. estymatora wykorzystano próbę o liczności $N = 500$ wygenerowaną z rozkładu normalnego $\mathcal{N}(1, 1)$

$$\hat{f}_N(x) = \frac{1}{Nh_N} \sum_{n=1}^{N} K\left(\frac{X_n - x}{h_N}\right) \tag{12}$$

5.1 Jądro prostokątne

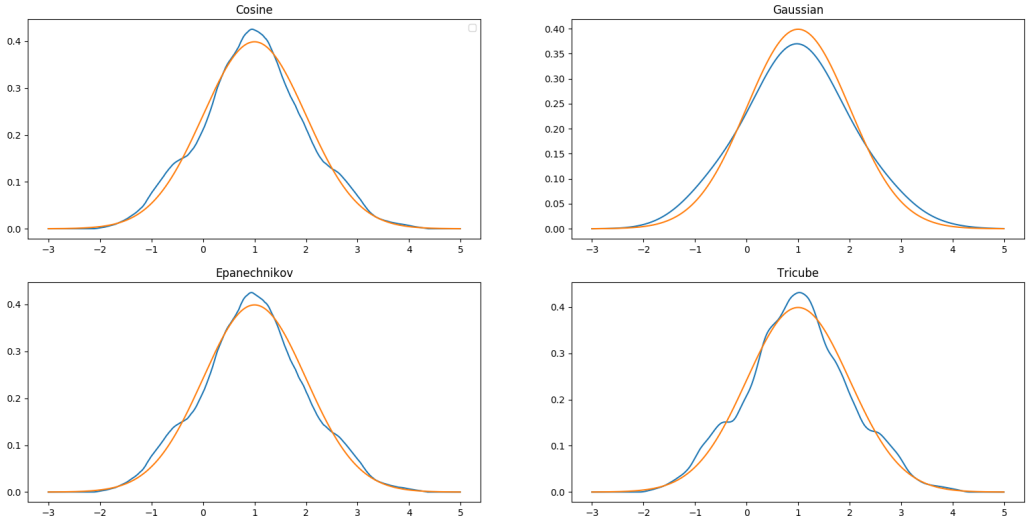
Jedną z najprostszych funkcji jądrowych jest jądro prostokątne. Wyraża się ono jako równanie 13. Wykorzystane zostało do wyznaczenia estymatora jądrowego dla wygenerowanych próbek. Jak widać do wyznaczania jądrowego estymatora gęstości prawdopodobieństwa potrzebny jest parametr h nazywany parametrem wygładzania. Na rysunku 12 przedstawiono wykresy z wyznaczonym estymatorem dla różnych wartości parametru h co obrazuje jaki wpływ na estymator ma ww. parametr. Co więcej, na podstawie przebiegów można wnioskować, że dla takiego problemu optymalna wartość parametru wygładzania znajduje się gdzieś w przedziale $(0.25, 1)$

$$\mathbf{I}(x) = \begin{cases} \frac{1}{2} & , dla\ |x| \leq 1 \\ 0 & , dla\ |x| > 1 \end{cases} \tag{13}$$



Rysunek 12: Jądrowy estymator gęstości prawdopodobieństwa na tle prawdziwego przebiegu dla kilku różnych wartości parametru wygładzania.

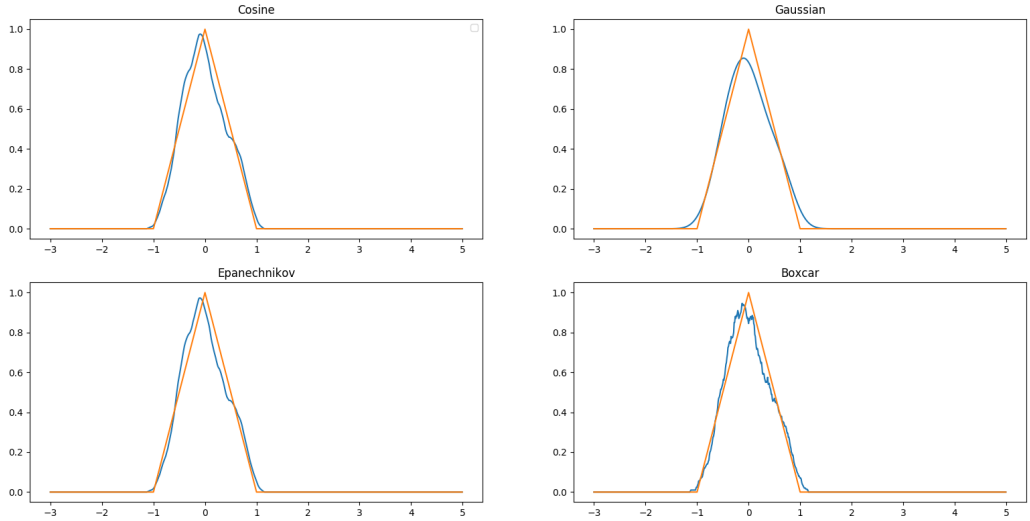
Oczywiście wybór odpowiedniego jądra estymatora również jest kwestią podyktowaną doborem najlepszego z możliwych narzędzi. Oprócz jądra prostokątnego przetestowano także jądro gaussowskie, Epanecznikowa, cosinusowe oraz tricube. Przebiegi wyznaczonych esteymatorów w funkcji x przedstawiono na rysunku 13. Jak można się było spodziewać dość dobrym estymatorem rozkładu normalnego jest estymator oparty na jądrze gaussowskim. Jednak jądro Epanecznikowa również bardzo dobrze dopasowuje się do przebiegu funkcji gęstości rozkładu normalnego, podobnie jądro cosinusowe. Co więcej jądro Epanecznikowa i jądro Cosinusowe są mniej zależne od parametru h, co sprawdzono badając ich przebiegi dla kilku różnych wartości parametru wygładzania co może potwierdzać ich dość dobre własności.



Rysunek 13: Estymatory gęstości prawdopodobieństwa dla różnych funkcji jądra.

Takim samym badaniom poddano próbki wygenerowane z rozkładu trójkątnego. Wyniki eksperymentów przedstawiono na rysunku 14. W tym przypadku widać, że jądro Epanechni-

kowa oraz jądro cosinusowe radzą sobie najlepiej. Jądro Gaussa prawdopodobnie poradziłoby sobie lepiej, gdyby odszukać odpowiedni współczynnik wygładzania.



Rysunek 14: Estymatory jądrowe dla próbek pochodzących z rozkładu trójkątnego.

5.2 Błąd empiryczny

Dla wyznaczanego metodą empiryczną rozkładu można zdefiniować błąd empiryczny, tj. błąd jaki generowany jest podczas estymowania prawdziwego rozkładu, definiowany jako różnica między estymatorem a wartością rozkładu dla danej wartości wejściowej. Błąd empiryczny definiowany jest w postaci wyrażenia 14. Dla wyżej analizowanego rozkładu wyznaczony błąd empiryczny przedstawiono na rysunku 15. W ten sposób można zauważyć, że parametr h przyjmuje optymalną wartość w okolicach 0.2–0.4 jednak, aby stwierdzić to dokładnie, należałoby zbadać pod kątem podobieństwa rozkładów dla parametru wygładzania z tego przedziału.

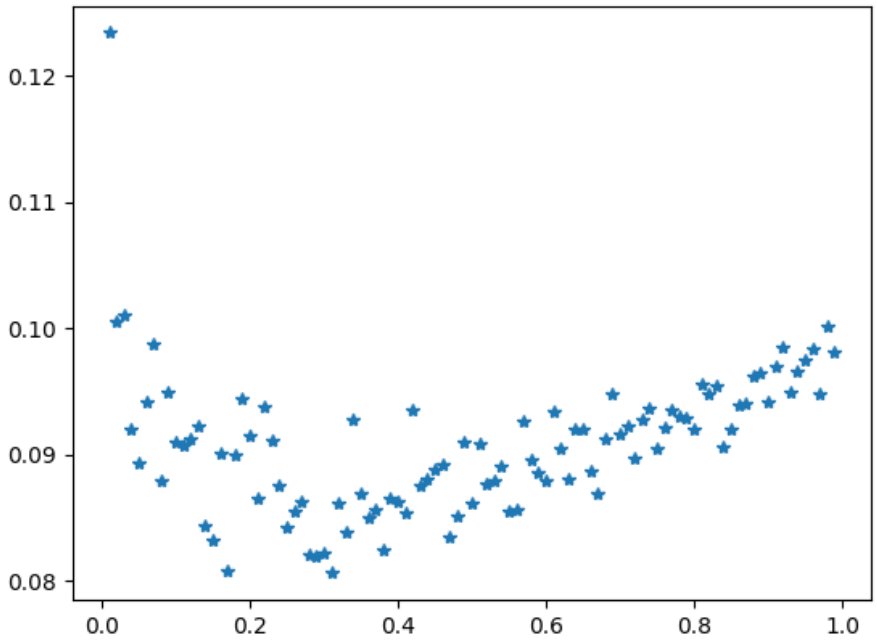
$$Err\{\hat{f}_N\} = \frac{1}{LM} \sum_{l=1}^{L} \sum_{m=1}^{M} \left[\hat{f}_N^{[l]}(x_m) - f(x_m) \right]^2 \tag{14}$$

6 Jądrowy estymator funkcji regresji

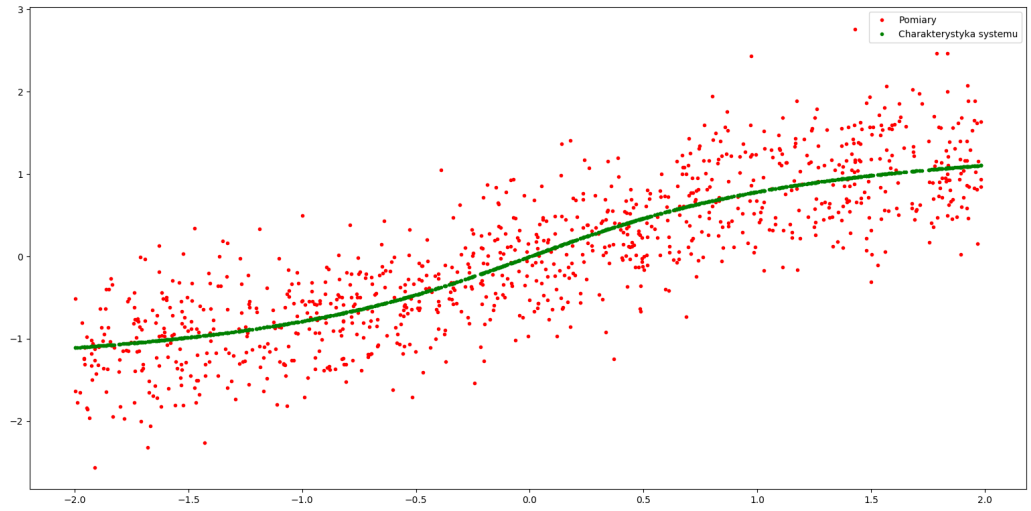
Innym sposobem modelowania nieznanych rozkładów prawdopodobieństwa jest jądrowy estymator funkcji regresji. Również on wykorzystuje w swoim działanie funkcję będącą jądrem estymatora.

6.1 Estymator dla systemu nieliniowego

Tym razem do przeprowadzenia badań przyjęto, że system nieliniowy ma charakterystykę $m(x) = tan(x)$. Wejściem systemu są próbki sygnału losowego posiadającego równomierny rozkład prawdopodobieństwa $\mathcal{U}(-2, 2)$, a wyjście obiektu jest poddane zakłóceniom działającym na wyjście z gęstością rozkładu normalnego $\mathcal{N}(0, 0.5)$. Wykres przedstawiający zależność wyjścia od wejścia obiektu (chmura punktów – Y(x)) została przedstawiona na rysunku 16. Został on przedstawiony na tle rzeczywistego przebiegu wyjścia bez zakłóceń.



Rysunek 15: Błąd empiryczny dla próby pochodzącej z rozkładu normalnego w zależności od parametru wygładzania h

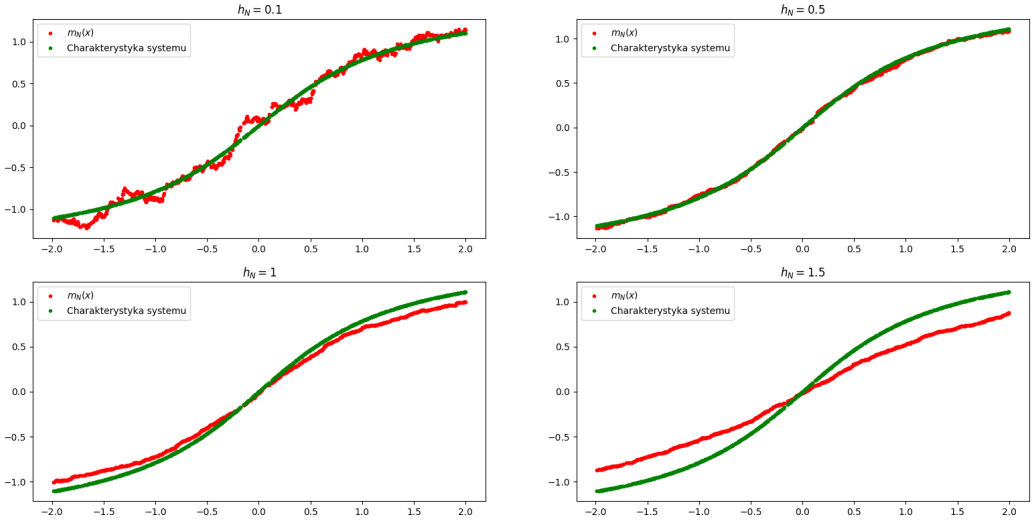


Rysunek 16: Wyjście systemu w zależności od jego wejścia

Jak widać na wykresie chmura punktów układa się w pożądanym kształcie, jednak zakłócenia, którym poddany jest sygnał wyjściowy 'rozmywa' w znaczący sposób dane wyjściowe. W takim wypadku można skorzystać z estymatora, który na podstawie takiej chmury mógłby wyznaczyć przypuszczalnie prawdziwy przebieg odpowiedzi systemu. Jądrowy estymator funkcji regresji wyraża się równaniem 15. Do badań przyjęto, że jądrem estymatora będzie jądro prostokątne. Sprawdzano różnice w działaniu estymatora dla kilku wartości parametru wygładzania, wyniki zostały przedstawione na rysunku 17. Ponownie widać, że pewna optymalna

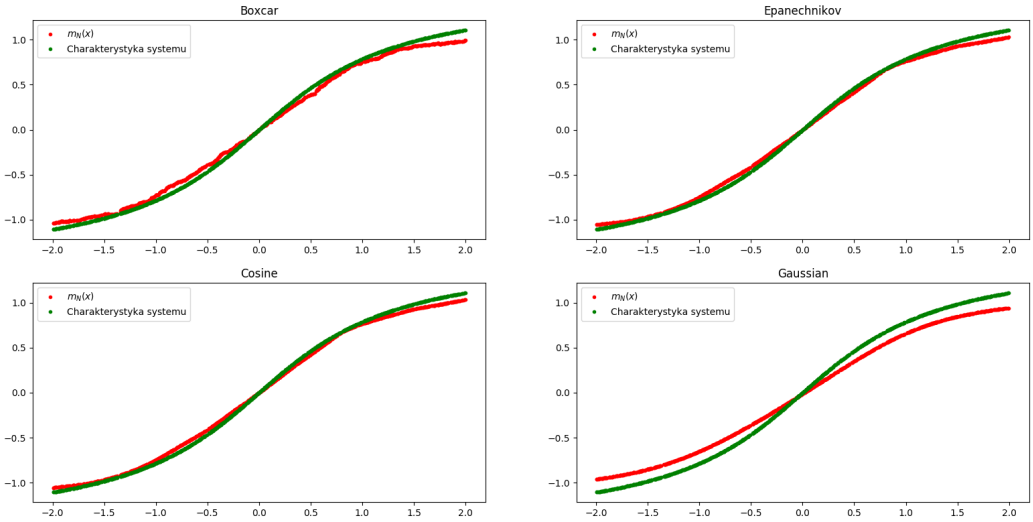
wartość parametru wygładzania znajduje się w okolicach $h_N = 0.5$.

$$\hat{m}_N(x) = \frac{\sum_{n=1}^{N} Y_n K\left(\frac{X_n - x}{h_N}\right)}{\sum_{n=1}^{N} K\left(\frac{X_n - x}{h_N}\right)} \tag{15}$$

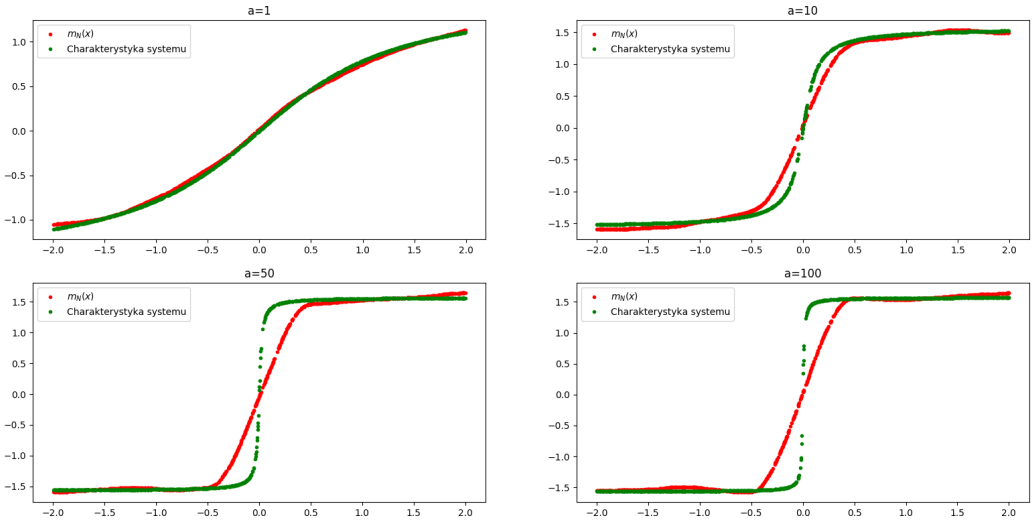


Rysunek 17: Estymator funkcji regresji dla kilku wybranych wartości parametru wygładzania

W celu sprawdzenia wpływu doboru jądra estymatora na jakość działania estymatora wykorzystano poprzednio używane jądra. Ustalając parametr wygładzania na poziomie $h_N = 0.6$ wyznaczono estymator w funkcji x dla 4 różnych jąder. Wyniki estymacji przedstawiono na rysunku 18. Ponownie jądro Epanecznikowa oraz cosinusowe niezwykle dobrze dopasowuje estymator do rzeczywistego przebiegu. W podobny sposób sprawdzono wpływ parametru a(będącego parametrem obiektu), czyli kształt nieliniowości, na jakość estymatora. Na rysunku 19 przedstawiono 4 wykresy pokazujące jak estymator jądrowy funkcji regresji, oparty o jądro Epanecznikowa z parametrem wygładzania $h_N = 0.5$, zmienia swoją jakość w zależności od stopnia nieliniowości obiektu. Jak widać, estymator traci wydajność im bardziej charakterystyka 'przyspiesza' swoje zmiany.



Rysunek 18: Estymatory funkcji regresji używające różnych jąder: funkcja jądra Gaussa, jądra Cosinusowego, Epanecznikowa oraz prostokątnego

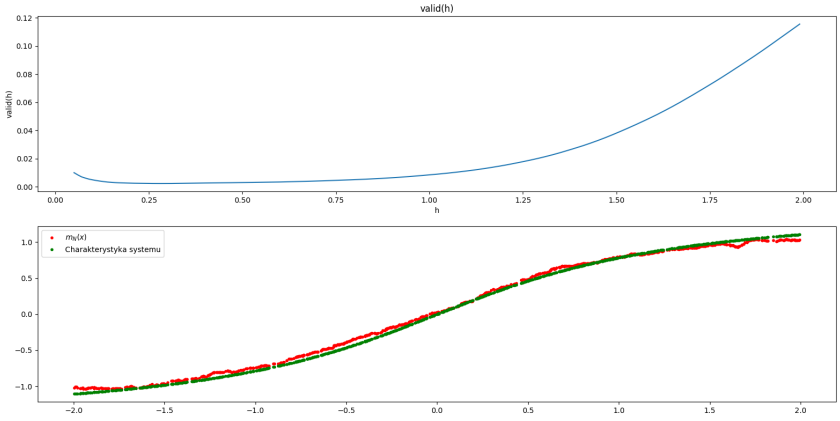


Rysunek 19: Estymatory funkcji regresji dla różnego stopnia nieliniowości obiektu.

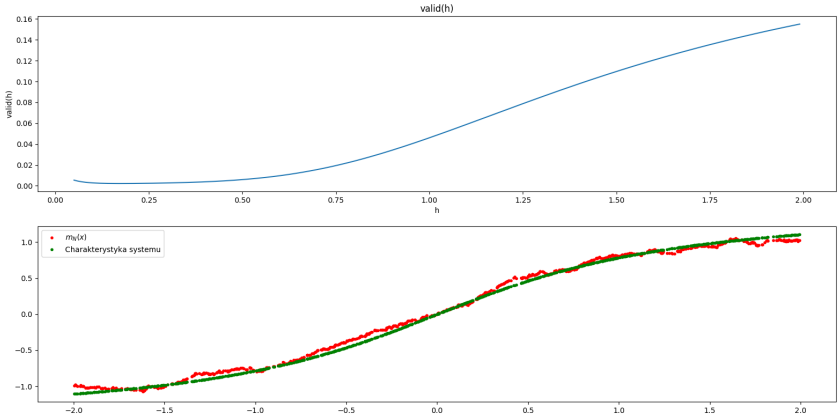
6.2 Optymalna wartość parametru wygładzania

Jako, że parametr wygładzania nie ma pewnej stałej, można próbować optymalizować jego wartość pod kątem jakości estymacji, czyli różnicy pomiędzy estymatorem a estymowanym rozkładem. Do tego celu może posłużyć funkcja wyrażona równaniem 16. Wartości błędu w zależności od parametru wygładzania został przedstawiony na rysunku 16. Na tej podstawie wyznaczono optymalną wartość parametru h dla 4 różnych jąder estymatora. Wyniki badań przedstawiono na rysunkach 20, 21, 22, 23.

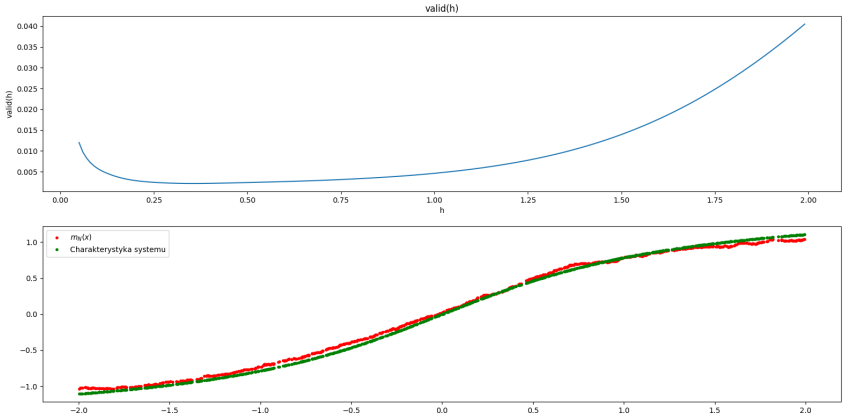
$$valid(h) = \frac{1}{2Q} \sum_{q=-Q}^{Q} \left[\hat{m}_N\left(\frac{q}{Q}\right) - m\left(\frac{q}{Q}\right) \right]^2 \tag{16}$$



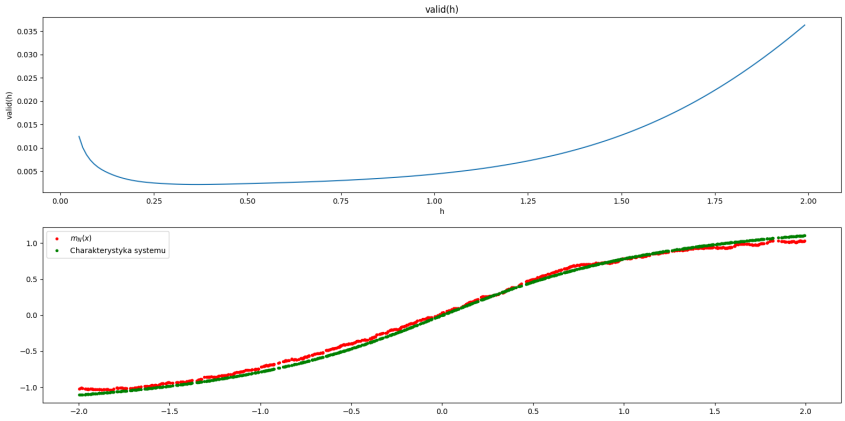
Rysunek 20: Wykres $valid(h)$ oraz estymator oparty o jądro prostokątne dla tak wybranego parametru wygładzania.



Rysunek 21: Wykres $valid(h)$ oraz estymator oparty o jądro Gaussa dla tak wybranego parametru wygładzania.



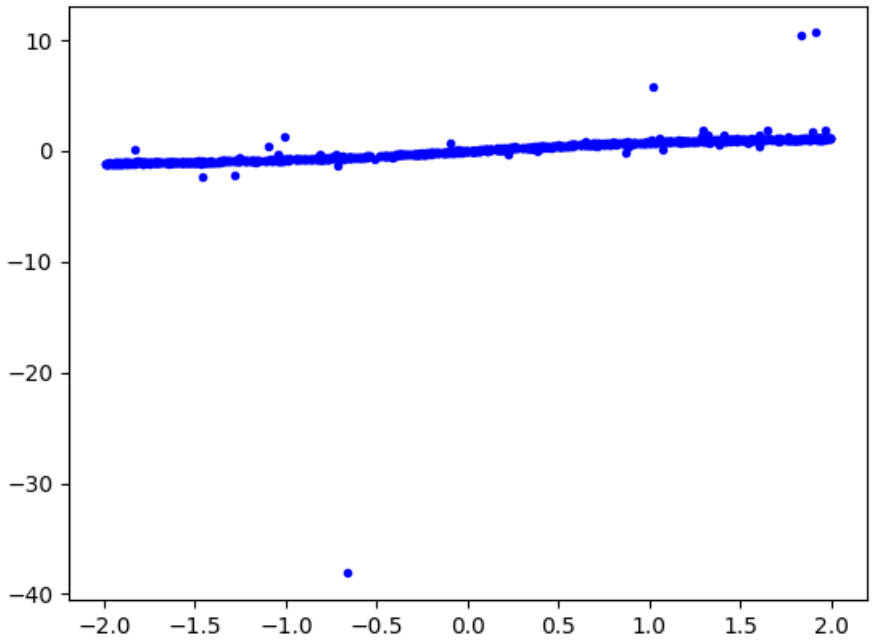
Rysunek 22: Wykres $valid(h)$ oraz estymator oparty o jądro Epanecznikowa dla tak wybranego parametru wygładzania.



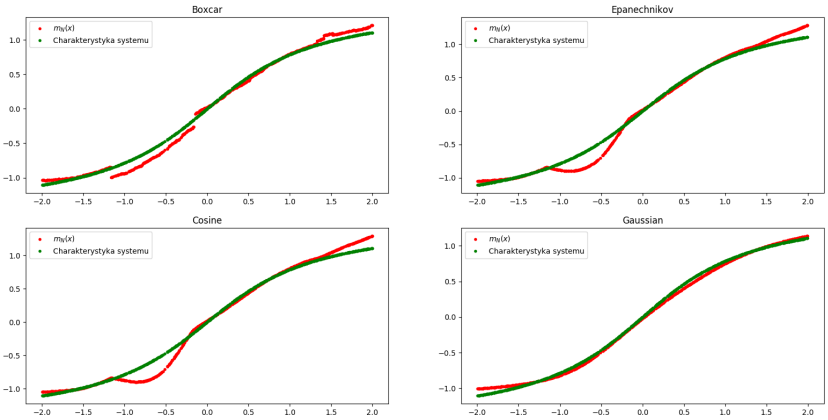
Rysunek 23: Wykres $valid(h)$ oraz estymator oparty o jądro cosinusowe dla tak wybranego parametru wygładzania.

6.3 Zakłócenia z rozkładem Cauchy'ego

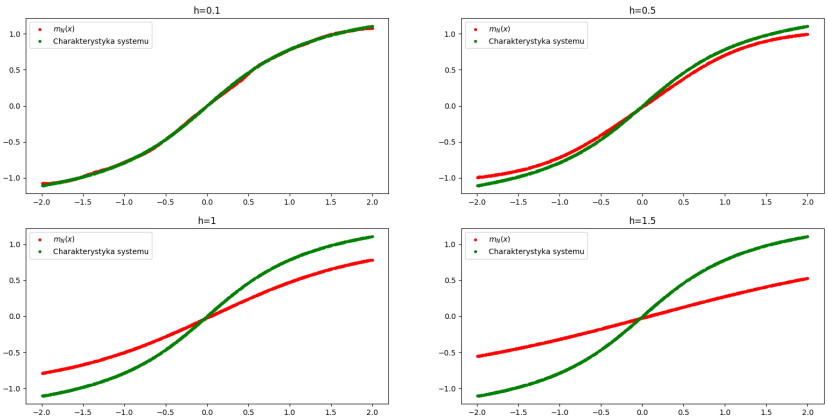
Badania zostały również przeprowadzone dla zakłóceń generowanych z rozkładem Cauchy'ego co oznacza, że nie miały one pewnej stałej wartości oczekiwanej oraz wariancji. Wykres wartości na wyjściu obiektu przedstawiony został na rysunku 24. Jak można przypuszczać, dalekie odchyłki niektórych próbek mogą powodować nieprawidłowe działanie estymatora. Wygenerowane estymatory dla różnych jąder estymatora z parametrem wygładzania $h_N = 0.5$ zostały przedstawione na rysunku 25. Wykresy potwierdzają przypuszczenia – rzadkie, lecz wielkie odchyłki od wartości oczekiwanej powodują, że estymator nie działa wystarczająco dobrze. Jedynie jądro gaussowskie dość dobrze estymuje prawdziwy przebieg, dlatego też w kolejnym badaniu zmianie ulegał parametr wygładzania. Wyniki badań przedstawiono na rysunku 26. Jak widać odpowiednie dobranie parametru wygładzania kompensuje problemy z szumem zakłócającym sygnał. Ostatnim eksperymentem, było badanie zachowania estymatora w zależności od zmieniającego się parametru a. Wynik został przedstawiony na rysunku 27. Ponownie widać, że wraz ze zmieniającą się nieliniowością estymator traci swoje właściwości.



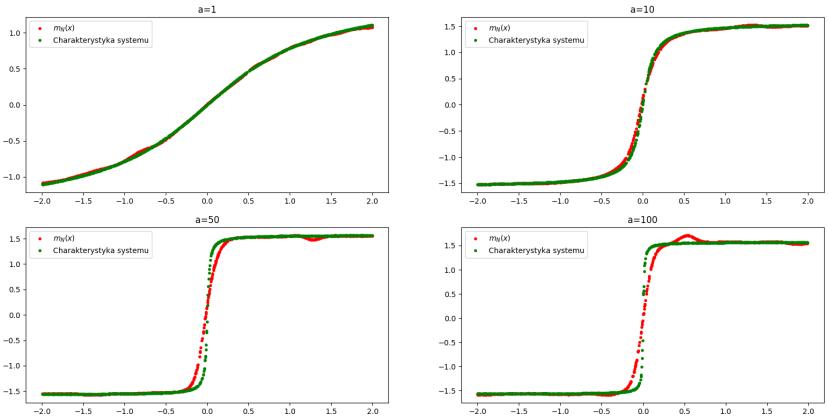
Rysunek 24: Wykres wyjściowy obiektu zaszumiony zakłóceniami o rozkładzie Cauchy'ego.



Rysunek 25: Wykresy dystrybuant empirycznych rozkładu zakłóconego szumem o rozkładzie Cauchy'ego dla różnych jąder estymatora.



Rysunek 26: Wykresy dystrybuant empirycznych rozkładu zakłóconego szumem o rozkładzie Cauchy'ego dla różnych wartości paramteru wygładzania.



Rysunek 27: Wykresy dystrybuant empirycznych rozkładu zakłóconego szumem o rozkładzie Cauchy'ego dla różnych wartości paramteru a, czyli różnego stopnia nieliniowości systemu.

Literatura

[1] Dokumentacja biblioteki matplotlib. `https://matplotlib.org/contents.html`, 03 2019.

[2] Dokumentacja biblioteki numpy. `https://docs.scipy.org/doc/`, 03 2019.

[3] Dokumentacja python 3.7.1. `https://docs.python.org/release/3.7.1/`, 03 2019.

[4] R. S. Jacek Jakubowski. *Wstęp do teorii prawdopodobieństwa.* Jacek Jakubowski, Rafał Sztencel, 2001.

[5] R. S. Jacek Jakubowski. *Rachunek prawdopodobieństwa dla (prawie) wszystkich.* Jacek Jakubowski, Rafał Sztencel, 2006.

[6] R. Z. Robert Wieczorkowski. *Komputerowe generatory liczb losowych.* Wydawnictwa Naukowo-Techniczne, 1997.