Wydział Podstawowych Problemów Techniki

# PRACA DYPLOMOWA

## Teoretyczne i praktyczne własności metod identyfikacji punktów zmiany rozkładu

## (Theoretical and practical issues in change point detection)

Marta Karaś

Promotor: dr hab. inż. Małgorzata Bogdan

słowa kluczowe:
identyfikacja punktów zmiany, dekompozycja wieloskalowa, progowanie falkowe, FDR ...

streszczenie:
W pracy rozważano możliwość zastosowania idei dekompozycji wieloskalowej obiektu jako metody identyfikacji segmentów średnich (identyfikacji punktów zmiany) rozkładu zmiennych losowych, których realizacje obserwujemy. Działanie proponowanej procedury zostało porównane z wynikami otrzymanymi dla dwóch wybranych metod referencyjnych. Zaproponowano nowe podejście w wyznaczaniu FDR i mocy metod identyfikacji punktów zmiany.

Wrocław 2015

# Spis treści

# Wstęp

Identyfikacja zmian w rozkładzie zmiennych losowych, których realizacje obserwujemy, jest problemem, który od dekad znajduje się w obrębie zainteresowania statystyków. Metody identyfikacji tych punktów zmiany (ang. *change point detection*) mogą być podzielone na dwie zasadnicze kategorie: *identyfikację sekwencyjną* (wykrywanie w czasie rzeczywistym) oraz *identyfikację segmentową* (wykrywanie retrospektywne). Identyfikacja segmentowa znajduje zastosowanie w wielorakich dziedzinach – z metodami tymi mamy do czynienia w problemach detekcji zmian klimatycznych, w analizie danych genetycznych, segmentacji sygnału czy wykrywaniu włamań w sieciach komputerowych.

W niniejszej pracy ograniczamy się do metod identyfikacji segmentowej, które służą wykrywaniu zmian wielokrotnych w poziomie średniej rozkładu prawdopodobieństwa. Ten konkretny przypadek identyfikacji punktów zmiany ma zastosowanie między innymi w tzw. hybrydyzacji genomowej, będącej jedną z metod analizy struktury genomu. W niniejszej pracy przedstawiamy dwa algorytmy identyfikacji punktu zmiany, dedykowane do stosowania w procesie hybrydyzacji genomowej. Proponujemy także autorską ideę metody identyfikacji zmian, polegającą na dekompozycji wieloskalowej obiektu, której wynikiem jest reprezentacja tego obiektu (tu: ciągu obserwacji) w postaci tzw. współczynników falkowych. W proponowanej przez nas procedurze współczynniki falkowe poddawane są procedurze tzw. progowania, która ma na celu pozostawienie w reprezentacji współczynników, które oddają najważniejsze wzorce współtworzące dany obiekt, a pozbycie się tych, które mogą korespondować z szumem w danych. W rezultacie można otrzymać obiekt, który oddaje jedynie "zgrubny" charakter wejściowych obserwacji, i na jego podstawie wnioskować o punktach, w których następuje ewentualna istotna zmiana w poziomie wartości obserwacji.

Istotną częścią niniejszej pracy są rozważania dotyczące miar, przy pomocy których dokonujemy oceny poprawności dokonanych estymacji punktu-/punktów zmiany. Miary, które wykorzystujemy, to przede wszystkim frakcja fałszywych odkryć (ang. *false discovery rate (FDR)*) oraz czułość (inaczej: moc) metody; w pracy zwracamy uwagę na istotne naszym zdaniem aspekty stosowania tych miar w problemie identyfikacji punktów zmiany średniej rozkładu – twierdzimy, że standardowa procedura może zostać poprawiona w sposób, który zwiększa interpretowalność otrzymanych wyników.

Niniejsza praca składa się z 5 rozdziałów.

W rozdziale pierwszym zamieszczamy wprowadzenie do tematyki identyfikacji punktów zmiany. Czytelnik odnajdzie tam rozszerzony opis zasygnalizowanego już możliwego podziału metod identyfikacji na metody sekwencyjne i segmentowe, a także schematyczny opis procedury hybrydyzacji genomowej.

W rozdziale drugim przedstawione są dwa z istniejących algorytmów identyfikacji punktów zmiany, które zostały wybrane na potrzeby niniejszej pracy jako metody referencyjne w stosunku do metody przez nas proponowanej. W rozdziale trzecim przedstawiona jest teoria niezbędna do zrozumienia idei proponowanej przez nas metody identyfikacji punktów zmiany, polegającej na wykorzystaniu reprezentacji falkowej w dekompozycji wieloskalowej obiektu.

W rozdziale czwartym zdefiniowane są miary, które zastosowaliśmy do oceny dobroci otrzymywanych estymacji punktów zmiany; w szczególności, rozdział ten zawiera opis procedury wyznaczania miar przez nas proponowanych, funkcjonujących pod nazwami $FDR.smooth$ oraz $POWER.smooth$.

Rozdział piąty zawiera wyniki przeprowadzonej analizy symulacyjnej. Analiza ta składa się z dwóch części. Część pierwsza koncentruje się na badaniu własności metod referencyjnych, opisanych w rozdziale drugim niniejszej pracy. Część druga analizy symulacyjnej skupia się na porównaniu wyników otrzymywanych przy zastosowaniu proponowanej przez nas metody identyfikacji punktów zmiany oraz przy zastosowaniu metod referencyjnych.

Rozdział 1

# Problem identyfikacji punktów zmiany rozkładu

W niniejszym rozdziale zamieszczamy wprowadzenie to tematyki identyfikacji punktów zmiany rozkładu.

Sekcja pierwsza rozdziału rozpoczyna się od zdefiniowania problemu identyfikacji punktów zmiany rozkładu w ogólnej postaci. Następnie określony jest podział zagadnienia na dwa typy problemów – *identyfikację sekwencyjną* oraz *identyfikację segmentową*, które różnią się specyfiką problemu oraz obszarami zastosowań.

W sekcji drugiej skupiamy się na szczególnym przypadku problemu identyfikacji sekwencyjnej, z którym mamy do czynienia w procesie tzw. *porównawczej hybrydyzacji genomowej*, będącej jedną z metod analizy struktury genomu. Sekcja ta zawiera szkic opisu procedry hybrydyzacji oraz wskazanie jej związku z problemem identyfikacji punktów zmiany rozkładu. Konkluzją tej części rozprawy jest określenie specyficznego zagadnienia praktycznego, na którym koncentruje się niniejsza praca.

## 1.1. Ogólne sformułowanie problemu identyfikacji punktów zmiany rozkładu

### 1.1.1. Zdefiniowanie problemu

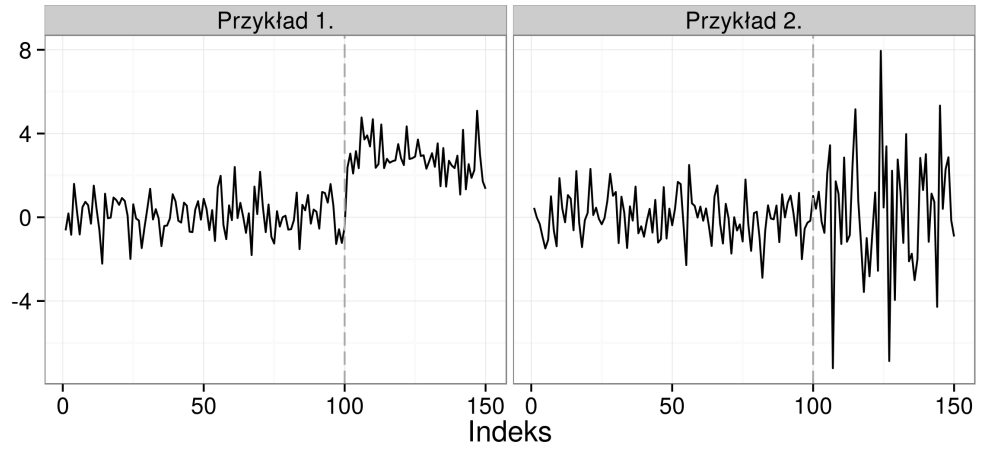
Problem identyfikacji punktów zmiany w danym ciągu obserwacji pojawia się w wielu problemach statystycznych. W ogólnym zapisie problemu (por. [1]) zakładamy, że dysponujemy ciągiem obserwacji $x_1, x_2, ...$, które są realizacjami zmiennych losowych $X_1, X_2, ...$ i które podlegają gwałtownej jednej lub kilku zmianom rozkładu w nieznanych punktach zmiany (ang. *change points*) $\tau_1, \tau_2, \dots$ .

Najczęściej zakładamy, że zmienne losowe są niezależne i pochodzą z tego samego rozkładu między każdą parą punktów zmiany. Możemy zapisać rozkład zmiennych $X_1, X_2, ...$ jako:

$$X_i \sim \begin{cases} F_0 & \text{gdy } i \leqslant \tau_1, \\ F_1 & \text{gdy } \tau_1 < i \leqslant \tau_2, \\ F_2 & \text{gdy } \tau_2 < i \leqslant \tau_3, \\ \dots , \end{cases} \tag{1.1}$$

gdzie $F_k$ oznaczają rozkład, jakiemu podlegają zmienne losowe w w *segmencie* $(X_{\tau_k+1}, ..., X_{\tau_{k+1}})$.

Na zamieszczonym poniżej Rysunku 1.1 widzimy dwa przykłady zmiennych z rozkładu normalnego, które podlegają zmianom w średniej i wariancji rozkładu, odpowiednio.




Rysunek 1.1: Przykłady zmian w rozkładzie zmiennych gaussowskich, z naniesionym punktem zmiany (linia przerywana).

Problemy identyfikacji punktów zmiany rozkładu różnią się od siebie w tym, co zakładamy o rozkładach $F_k$ zmiennych losowych $X_i$. W praktyce, najczęściej mamy do czynienia z sytuacją, gdy parametry tych rozkładów są nieznane; w niektórych przypadkach nie ma nawet informacji o rodzinie rozkładów $F_i$ ([1]).

W zależności od specyfiki problemu, możemy być zainteresowani poszukiwaniem punktów zmiany średniej rozkładu, wariancji rozkładu lub średniej i wariacji rozkładu jednocześnie.

### 1.1.2. Identyfikacja segmentowa i identyfikacja sekwencyjna

Zdefiniowany w Równaniu 1.1 problem identyfikacji punktów zmiany stanowi przedmiot intensywnych badań od okresu lat 50-tych XX wieku. Z uwagi na bardzo ogólną naturę problemu, literatura dot. badań nad tymi zagadnieniami jest zróżnicowana i obejmuje różnorodne obszary zastosowań.

Wiele popularnych metod identyfikacji punktów zmiany rozkładu ma swoje źródło w zastosowaniach związanych z kontrolą jakości, gdzie celem jest monitorowanie produktów wynikowych procesu przemysłowego i wykrywanie ewentualnych usterek możliwie szybko (por. [2]). Inne obszary zastosowań metod identyfikacji punktów zmiany rozkładu obejmują: badanie zmienności

liczby kopii DNA (ang. *copy-number variations (CNVs)*) w analizie struktury genomu ([3]), wykrywanie zakłóceń w sieciach komputerowych ([4]) czy dopasowywanie wielodziedzinowych modeli (ang. *multiple regime models*), popularnych w zastosowaniach ekonomicznych i finansowych ([5]).

Możemy wyodrębnić dwa główne typy problemów identyfikacji zmiany rozkładu, *segmentowy* (ang. *batch*) i *sekwencyjny* (ang. *sequential*) (por. [1]).

**Identyfikacja segmentowa**

W tym przypadku mamy do czynienia ze skończonym ciągiem $n$ obserwacji, będących realizacjami zmiennych losowych $X_1, ..., X_n$. Często brak jest założenia, że w danej sekwencji występuje co najmniej 1 punkt zmiany. Ten typ identyfikacji jest retrospektywny w tym sensie, że decyzja o zidentyfikowaniu bądź nie punktu zmiany jest podejmowana w oparciu o wszystkie dostępne obserwacje, zarówno te występujące przed, jak i po potencjalnym punkcie zmiany w danym ciągu obserwacji.

**Identyfikacja sekwencyjna**

W tym przypadku nie zakładamy, że mamy do czynienia ze skończonym ciągiem obserwacji. Odmiennie do identyfikacji segmentowej, obserwacje są dostarczane i przetwarzane w sposób ciągły, wraz z upływem czasu. Kiedy nowa obserwacja jest dostarczona, decyzja o zidentyfikowaniu bądź nie punktu zmiany jest podejmowana tylko na podstawie obserwacji otrzymanych do tego momentu. Jeśli nie stwierdzamy wystąpienia punktu zmiany, to przechodzimy do przetwarzania następnej obserwacji. Jeśli stwierdzamy wystąpienie punktu zmiany, to "restartujemy" detektor zmian w punkcie następnej obserwacji.

Z reguły narzędzia stosowane w przypadku obu tych typów są różne. Popularne podejścia do problemu identyfikacji segmentowej obejmują testy ilorazu wiarygodności ([6], [7]) czy wnioskowanie bayesowskie ([8], [30]). W przypadku identyfikacji sekwencyjnej, stosowane podejścia wykorzystują tzw. *karty kontrolne* (ang. *control charts*), na przykład CUSUM ([10]), Wykładnicze Ważone Ruchome Średnie (ang. *Exponential Weighted Moving Average*) ([11]) czy sekwencyjne metody bayesowskie ([12] , [13]).

W niniejszej pracy koncentrujemy się na metodach, które mogą być stosowane w szczególnym przypadku identyfikacji segmentowej, jakim jest identyfikacja punktów zmiany w procesie porównawczej hybrydyzacji genomowej. W kolejnej sekcji bieżącego rozdziału przedstawione są podstawowe informacje dotyczące tego procesu. Następne dwa rozdziały niniejszej pracy zawierają opis wybranych istniejących algorytmów, które mogą być stosowane w tym konkretnym zagadnieniu oraz opis proponowanej przez nas nowej metody.

## 1.2. Problem identyfikacji punktów zmiany rozkładu w procesie porównawczej hybrydyzacji genomowej

W niniejszej sekcji przedstawiamy podstawowe informacje dotyczące porównawczej hybrydyzacji genomowej (ang. *comparative genomic hybridization*), będącej jedną z metod analizy struktury genomu. Zawarte w tej sekcji informacje podajemy głównie za dwoma źródłami: [14] i [7].

### 1.2.1. Motywacja

Tworzenie się i wzrost nowotworów są związane z nagromadzeniem się zmian genetycznych i epigenetycznych w danym regionie organizmu. W wyniku tego nagromadzenia zmienia się poziom ekspresji[1] pewnych genów, a w konsekwencji – zmianie ulega normalny proces wzrostu i istnienia komórek organizmu. Wiele z tych zmian pociąga za sobą amplifikację i/lub ubytek partii genomu[2] – w pracy [15] stwierdzony został związek pomiędzy liczbą kopii DNA a aktywnością transkrypcyjną genów w zmienionych chorobowo regionach organizmu.

Zastosowanie licznych i wielorakich tzw. metod cytogenicznych oraz metod analizy mikromacierzowej pozwoliło ([16], [17], [18]) na wskazanie wielu i o różnej naturze aberracji (amplifikacji lub ubytków) w liczbie kopii DNA, występujących w komórkach nowotworowych u ludzi i u gryzoni.

Podsumowując, zaobserwowanie występujących w genomie aberracji może być sygnałem świadczącym o zwiększonym wzroście i zwiększonym czasie przeżycia komórek organizmu i tym samym może wskazywać na zmieniony nowotworowo charakter tych komórek. Ponadto uważa się, że analiza tych aberracji może pomóc w identyfikacji mechanizmów prowadzących do tych szkodliwych zmian.

### 1.2.2. Porównawcza hybrydyzacja genomowa

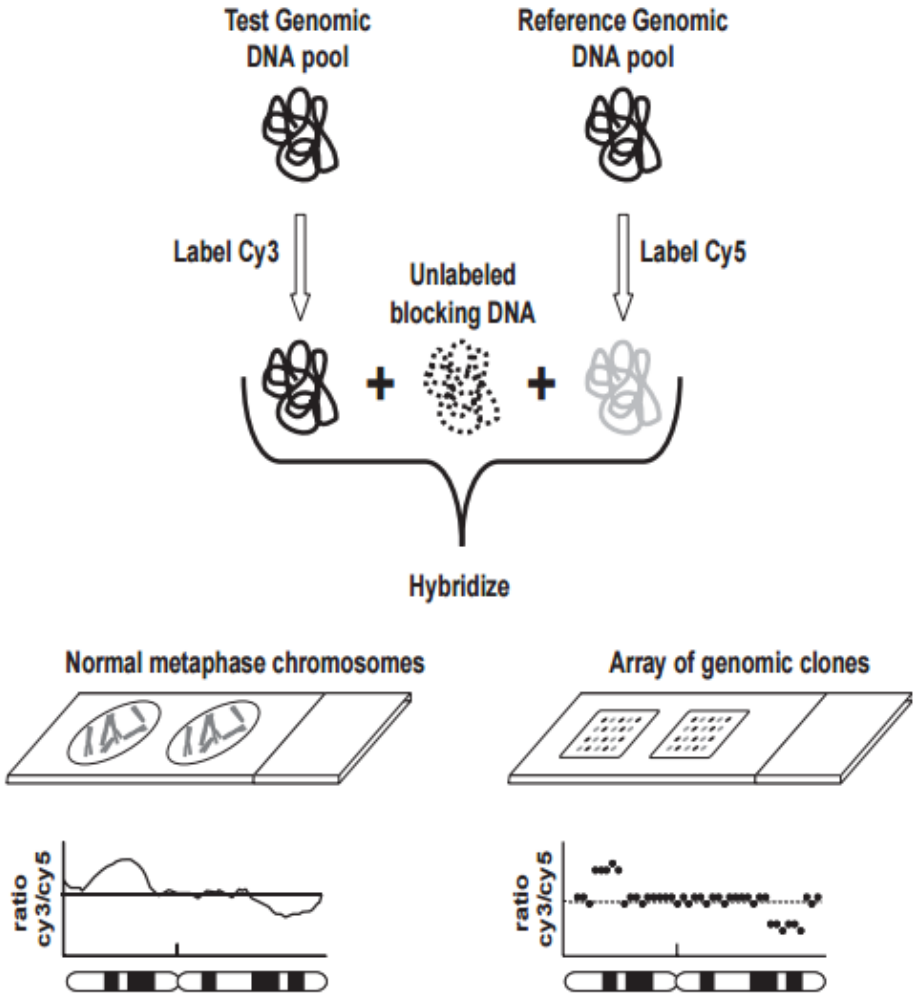
Z wymienionych powyżej powodów jesteśmy zainteresowani detekcją aberracji występujących w genomie. Jedną z metod, które powalają na ilościowy pomiar aberracji w liczbie kopii DNA oraz mapowanie tych aberracji na sekwencje genomowe, jest *porównawcza hybrydyzacja genomowa z wykorzystaniem mikromacierzy* (ang. *microarray-based comparative genomic hybridization (array CGH)*).

Standardowo, w porównawczej hybrydyzacji genomowej z wykorzystaniem mikromacierzy stosuje się *testową pulę DNA* (ang. *test genomic DNA pool*)

[1] *Ekspresja genów* – proces wyrażania informacji genetycznej. Proces ten obejmuje dwa etapy:
— *transkrypcję* – przepisywanie informacji genetycznej z DNA na mRNA,
— *translację* – tłumaczenie informacji genetycznej z mRNA na białko.
(Na podst. [19].)

[2] *Genom* – suma wszystkich kodujących i niekodujących sekwencji DNA zawartych w haploidalnej komórce organizmu. DNA genomu człowieka podzielone jest na 23 jednostki strukturalno-funkcjonalne, zwane chromosomami. (Na podst. [20].)

np. z obszaru zmienionego nowotworowo oraz *referencyjną pulę DNA* (ang. *reference genomic DNA pool*) ze zdrowego obszaru organizmu. Schematyczna reprezentacja procesu CGH przedstawiona jest na Rysunku 1.2.




Rysunek 1.2: Schematyczna reprezentacja procesu porównawczej hybrydyzacji genomowej z wykorzystaniem mikromacierzy. Źródło: [14], str. 134.

W procesie hybrydyzacji (por. Rysunek 1.2), DNA testowe i referencyjne są oznaczane dwoma różnymi barwnikami (cyjanina o sygnaturze $Cy3$ i $Cy5$, odpowiednio), a następnie łączone w procesie hybrydyzacji na mikromacierz. Stosunek intensywności fluorescencji barwników każdej z otrzymanych plamek mikromacierzy jest wskaźnikiem względnej różnicy w liczbie kopii DNA w próbce testowej w stosunku do próbki referencyjnej.

### 1.2.3. Modelowanie intensywności fluorescencji barwników w porównawczej hybrydyzacji genomowej

W niniejszej części pracy formalizujemy powyższy szkic procesu porównawczej hybrydyzacji genomowej oraz formułujemy związek między tym procesem a problemem identyfikacji punktów zmiany rozkładu.

Niech $X_1, ..., X_n$ będzie ciągiem zmiennych losowych. Analogicznie jak w Równaniu 1.1, indeks $\tau$ nazywamy punktem zmiany, jeśli zmienne $X_1, ..., X_\tau$ mają jednakowy rozkład prawdopodobieństwa $F_0$ oraz zmienne $X_{\tau+1}, ...$ mają jednakowy rozkład prawdopodobieństwa $F_1$, do momentu następnego punktu rozkładu (jeśli taki istnieje).

W analizie mikromacierzowej mającej na celu detekcję różnic liczby kopii DNA w próbce testowej i próbce referencyjnej, analizowane dane są uszeregowane zgodnie z lokalizacją tzw. *markerów*, wyróżnionych wzdłuż chromosomu, z którego próbki pochodzą. Jesteśmy zainteresowani badaniem relacji między intensywnościami fluorescencji próbki testowej i referencyjnej dla każdego markera.

Niech $I_{tm}$ i $I_{rm}$ oznaczają wartości intensywności fluorescencji próbki testowej i referencyjnej odpowiednio, dla markera o indeksie $m$. Intensywności te są związane z liczbą kopii DNA w próbce testowej i referencyjnej, którą oznaczamy przez $C_{tm}$ i $C_{rm}$, odpowiednio. Relacja między $(I)$ i $(C)$ może być modelowana następująco:

$$I_{tm} = \beta_{tm} C_{tm}(1 + \epsilon) \tag{1.2}$$

oraz

$$I_{rm} = \beta_{rm} C_{rm}(1 + \epsilon), \tag{1.3}$$

gdzie parametry $(\beta)$ są zależne od czynników związanych z konkretnym eksperymentem (por. [7], str. 559), a $\epsilon$ są błędami losowymi.

W analizie zakładamy, że logarytm ilorazu współczynników $\beta$ w modelach 1.2 i 1.3 jest stały oraz że próbka referencyjna nie zawiera aberracji w liczbie kopii DNA. Ponadto, przeprowadza się dodatkową normalizację danych, aby w rezultacie otrzymać wartości wyrażenia:

$$log(\beta_{tm} C_{tm} / \beta_{rm} C_{rm}) \tag{1.4}$$

takie, że średnia po wszystkich wyrażeniach 1.4 jest równa 0.

Mając dane powyższe założenia możemy wnioskować, że zaobserwowanie odchylenia od poziomu zerowego wartości wyrażenia 1.4 dla danego indeksu obserwacji $m$ oznacza zaobserwowanie zmiany w liczbie kopii DNA w próbce testowej w markerze o lokalizacji przypisanej do indeksu $m$.

Podsumowując, jesteśmy zainteresowani poszukiwaniem punktów zmiany w wartościach logarytmu ilorazu znormalizowanych intensywności fluorescencji, indeksowanych wg lokalizacji markerów w danym chromosomie. Zgodnie ze specyfiką tej analizy, możemy spodziewać się występowania więcej niż jednego punktu zmiany w danym chromosomie. Naszym celem jest identyfikacja wszystkich punktów zmiany i podział chromosomu na segmenty, w których liczba kopii DNA w próbce testowej jest stała. Innymi słowy, problem identyfikacji punktów zmiany w procesie porównawczej hybrydyzacji genomowej jest problemem identyfikacji wielokrotnych punktów zmiany w średniej

rozkładu zmiennych losowych $X_1, ..., X_n$, gdzie ciągiem realizacji $x_1, ..., x_n$ z tego rozkładu są wartości logarytmu ilorazu intensywności, zdefiniowane w Równaniu 1.4.

## Rozdział 2

# Wybrane istniejące metody identyfikacji punktów zmiany rozkładu

W niniejszym rozdziale przedstawiamy wybrane istniejące metody identyfikacji punktów zmiany rozkładu, które wykorzystaliśmy w analizie porównawczej, której wyniki przedstawione są w dalszej części pracy.

Jak zostało zasygnalizowane w poprzednim rozdziale, problem identyfikacji punktów zmiany rozkładu jest szerokim zagadnieniem; nie dziwi zatem fakt, że powstało dotychczas wiele różnych algorytmów dedykowanych temu problemowi, odpowiadających na jego szczególne przypadki. Na moment pisania niniejszej pracy, do dyspozycji czytelnika jest m.in. prawie 30 różnych bibliotek z oprogramowaniem służącym do identyfikacji punktów zmiany, dostępnych w darmowym środowisku do obliczeń statystycznych i wizualizacji wyników `R` ([21]). Lista tych bibliotek oraz referencje do materiałów zawierających opis działania danej metody zgromadzone są na portalu o nazwie: *The Changepoint Repository. Fostering the exchange of knowledge and software related to changepoint analysis.* ([24]).

Na potrzeby niniejszej pracy wybrane zostały dwie istniejące metody identyfikacji punktów zmiany rozkładu, które posłużą jako metody referencyjne do metody przez nas proponowanej (por. Rozdział 3.). Są to:

1. algorytm *Circular Binary Segmentation*, zaimplementowany w bibliotece `DNAcopy` ([22]), dostępnej w środowisku `R`,
2. algorytm *BH*, zaimplementowany w bibliotece `bcp` ([23]), dostępnej w środowisku `R`.

Powyższe dwie metody referencyjne zostały wybrane z kilku powodów. Po pierwsze, są to metody odpowiednie do pracy ze szczególnym przypadkiem problemu identyfikacji punktów zmiany, który wskazaliśmy poprzednim rozdziale (problem występujący w procesie porównawczej hybrydyzacji genomowej). Po drugie, są to metody zbudowane wg różnych podejść matematycznych – testów statystycznych opartych na ilorazie wiarygodności i analizie bayesowskiej, odpowiednio, i jak wykazały przeprowadzone analizy – dostarczają też rezultatów o odmiennym charakterze. Po trzecie, są to metody, które pojawiają się stosunkowo często jako metody referencyjne w literaturze opisującej inne metody identyfikacji punktów zmiany (por. [24], zakładka *Software*).

W dalszej części bieżącego rozdziału znajduje się opis wybranych metod.

## 2.1. Algorytm Circular Binary Segmentation (CBS)

Algorytm *Circular Binary Segmentation (CBS)* to metoda statystyczna dedykowana do identyfikacji istotnie niższych/wyższych wartości intensywności fluorescencji w próbce testowej w badaniu CGH, co jest oznaką detekcji ubytków/amplifikacji liczby kopii DNA. Idea metody polega na dzieleniu chromosomu na regiony o równej liczbie kopii DNA, przy uwzględnieniu występowania szumu w danych.

Przedstawiony poniżej opis działania algorytmu przedstawiamy za [7]. Opis rozpoczyna się od opisania algorytmu *Binary Segmentation*, który leży u podstaw idei algorytmu $CBS$.

### 2.1.1. Binary Segmentation

Niech $X_1, ..., X_n$ będą logarytmami ilorazów intensywności fluorescencji, indeksowanymi według lokalizacji $n$ markerów, którymi zajmujemy się w danej analizie. Niech $S_i = X_1 + ... + X_i$, $1 \leqslant i \leqslant n$ oznaczają sumy częściowe ciągu $(X_1, ..., X_n)$.

Jeśli zmienne $X_1, ..., X_n$ pochodzą z rozkładu normalnego o znanej wariancji (bez utraty ogólności zakładamy, że wartość nieznanej wariancji wynosi 1), to statystyka testowa w teście ilorazu wiarygodności weryfikującym prawdziwość hipotezy zerowej $H_0$ o tym, że nie ma żadnego punktu zmiany rozkładu, przeciwko hipotezie alternatywnej $H_1$ o tym, że istnieje dokładnie jeden punkt zmiany rozkładu w nieznanym położeniu $i$, dana jest ([25]) formułą:

$$Z_B = max_{1 \leqslant i < n} |Z_i|, \tag{2.1}$$

gdzie

$$Z_i = \left\{ \frac{1}{i} + \frac{1}{n-i} \right\}^{-1/2} \left\{ \frac{S_i}{i} - \frac{S_n - S_i}{n-i} \right\}. \tag{2.2}$$

Hipoteza zerowa $H_0$ jest odrzucona, jeśli wartość statystyki danej formułą 2.1 przekracza kwantyl rzędu $\alpha$ rozkładu statystyki 2.1 przy założeniu prawdziwości hipotezy zerowej $H_0$ i jako miejsce punktu zmiany wskazany zostaje ten indeks $i$, dla którego $Z_B = |Z_i|$.

A. Sen i M. S. Srivastava w pracy [25] uzyskują wartość krytyczną rozkładu statystyki testowej za pomocą symulacji Monte Carlo. Wartość ta może być także wyznaczona dla zadanego poziomu istotności $\alpha$ dzięki wykorzystaniu aproksymacji rozkładu ogonów statystyki testowej 2.1, danych przez D. Siegmunda w pracy [26].

Procedura *binary segmentation* wykonuje test w sposób rekursywny do czasu, aż żadna kolejna zmiana nie jest stwierdzana w segmentach otrzymanych w wyniku dotychczasowych podziałów chromosomu.

Wykazano, że przy zachodzeniu odpowiednich warunków regularności ([27]), procedura jest zgodna.

W przypadku, gdy wariancja rozkładu zmiennych $X_1, ..., X_n$ jest nieznana, stosujemy estymator wariancji uzyskany z próby; wtedy statystyki $Z_i$ są zastąpione odpowiadającym im $t$-statystykom i główna statystyka, 2.1, jest zastąpiona przez odpowiadającą jej wartość maksimum z wartości bezwzględnych $t$-statystystyk.

Procedura *binary segmentation* jest oparta na teście służącym do identyfikacji *pojedynczej* zmiany; z faktu tego wynika możliwość pojawienia się trudności metody ([28]) w detekcji niewielkich segmentów "ukrytych" między dużymi segmentami. Aby zaradzić temu problemowi, do początkowego algorytmu *binary segmentation* została wprowadzona modyfikacja, opisana w następnej subsekcji bieżącego rozdziału.

### 2.1.2. Circular Binary Segmentation

Nakreślony powyżej problem w działaniu procedury *binary segmentation* ma swoje źródło w tym, że procedura szuka tylko jednego punktu zmiany w danym kroku algorytmu. B. Levin i J. Kline w pracy [29] zaproponowali postać statystyki w teście weryfikującym prawdziwość hipotezy $H_0$ o braku punktu zmiany przeciwko hipotezie alternatywnej $H_1$ o występowaniu dwóch punktów zmiany. Poniżej znajduje sie opis modyfikacji *binary segmentation*, w której wykorzystana jest postać statystyki podana przez Levin'a i Kline'a.

Rozważmy segment wartości "sklejony" na swoich końcach tak, aby formował okrąg. Statystyka testowa w teście ilorazu wiarygodności do testowania hipotezy zerowej $H_0$ o braku punktu zmiany przeciwko hipotezie alternatywnej $H_1$ twierdzącej, że kąt wyznaczony od punktu $i+1$ do punktu $j$ oraz jego dopełnienie mają różne średnie, jest postaci:

$$Z_{ij} = \left\{ \frac{1}{j-1} + \frac{1}{n-j+i} \right\}^{-1/2} \left\{ \frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i} \right\}. \tag{2.3}$$

Modyfikacja procedury *binary segmentation*, nazwana *circular binary segmentation (CBS)*, jest oparta o statystykę:

$$Z_C = max_{1 \leqslant i < j \leqslant n} |Z_{ij}|. \tag{2.4}$$

Można zauważyć, że statystyka 2.4 może być sprowadzona do postaci pozwalającej na detekcję pojedynczej zmiany, jeśli przyjmiemy $j = n$.

Podobnie jak poprzednio, identyfikację punktu/punktów zmiany stwierdzamy, jeśli wartość statystyki 2.4 przekroczy ustalony próg, wyznaczony w oparciu o rozkład statystyki 2.4 przy założeniu prawdziwości hipotezy zerowej $H_0$ i dla zadanego poziomu istotności $\alpha$. Jeśli zmienne losowe $X_1, ..., X_n$ mają rozkład normalny, to – ponownie – ta progowa wartość może być wyznaczona przy

użyciu symulacji Monte Carlo lub przy zastosowaniu aproksymacji ogonów rozkładu statystyki 2.4, danych przez D. Siegmunda w pracy [26]. Gdy hipoteza zerowa $H_0$ o braku punktu zmiany jest odrzucona, jako punkty zmiany są wskazywane te indeksy $i$ i $j$ obserwacji, dla których spełniona jest równość $Z_C = |Z_{ij}|$. Procedura jest powtarzana w sposób rekursywny do momentu identyfikacji wszystkich punktów zmiany.

### 2.1.3. Modyfikacje algorytmu wprowadzone w implementacji metody dostępnej w środowisku `R`

Jak wspomniano powyżej, w przeprowadzonych analizach porównawczych stosujemy implementację algorytmu $CBS$ pochodzącą z biblioteki `bcp` ([23]), dostępnej w środowisku `R`. Implementacja ta, jak podano w [7], uwzględnia kilka dodatkowych modyfikacji, wymienionych poniżej.

— Korekta "efektu krawędzi".

Jednym z problemów, który może pojawić się przy stosowaniu metody opartej o algorytm $CBS$, jest "efekt krawędzi" (ang. *edge effect*) w estymacji punktów zmiany rozkładu. Problem ten polega na tym, że jeśli w danym kroku algorytmu indeksy $i$ i $j$, dla których spełniona jest równość $Z_C = |Z_{ij}|$, są takie, że albo $i$ jest "blisko" 1, albo $j$ jest "blisko" $n$, to możemy mieć do czynienia z istnieniem jednego prawdziwego punktu zmiany rozkładu, w przeciwieństwie do dwóch punktów, których istnienie sugerują dane (por. [7], str. 560.).

Aby temu zaradzić, wprowadzona jest następująca poprawka: w pierwszej kolejności testujemy, czy dane świadczą o tym, że indeks $i$ jest punktem zmiany dla zmiennych losowych $X_1, ..., X_j$ i cofamy uznanie $i$ za punkt zmiany rozkładu na wyjściowym segmencie $X_1, ..., X_n$, jeśli powyższe przypuszczenie okaże się nieprawdziwe. Podobny test przeprowadzamy dla indeksu $j$. Jako że w praktyce nie jest jasne, jak stwierdzić, że punkt zmiany jest "blisko" krawędzi, powyższa procedura wykonywana jest dla każdej pary indeksów $i$ i $j$, które otrzymujemy w takim kroku algorytmu, którego wynikiem jest podział bieżącego segmentu na trzy części.

— Uogólnienie do danych niepochodzących z rozkładu normalnego.

W dotychczasowych rozważaniach koncentrowaliśmy się na przypadku, gdy rozważane zmienne losowe $X_1, ..., X_n$ pochodzą z rozkładu normalnego. W przypadku, gdy założenie o normalności nie jest prawdziwe, stosuje się tzw. podejście permutacyjne.

Przy założeniu prawdziwości hipotezy zerowej o braku punktu zmiany w danych, zmienne $X_i$ pochodzą z jednakowego rozkładu. Definiujemy $X_1^*, ..., X_n^*$ jako losową permutację zmiennych $X_1, ..., X_n$ i oznaczamy przez $Z_C^* = max|Z_{ij}^*|$ statystykę 2.4 otrzymaną z $X_1^*, ..., X_n^*$. Za wartość progową statystyki $Z_C^*$ można przyjąć kwantyl rzędu $\alpha$ w rozkładzie permutacji

zadanym przez $Z_C^*$, przy założeniu odpowiednio dużej liczby permutacji (por. [7], str. 560.).

— Modyfikacja usprawniająca obliczenia w przypadku danych niepochodzących z rozkładu normalnego, dla dużych zbiorów danych (por. [7], str. 560.).

— Modyfikacje wprowadzone z uwagi na szczególny przypadek identyfikacji punktów zmiany rozkładu – analizy *array-CGH* (ang. *array-based comparative genomic hybridization*).

Modyfikacje te uwzględniają specyfikę danych i polegają na dwóch dodatkowych krokach. W pierwszym kroku przeprowadzane jest wygładzanie obserwacji odstających w danych. Wartości odstające mogą być wynikiem błędów technicznych popełnionych w eksperymencie lub aberracją w liczbie kopii DNA obejmującą lokalizację tylko jednego markera (por. [7], str. 560.).

W drugim kroku modyfikacji wykonuje się procedurę tzw. "przycinania", która ma na celu wyeliminowanie pojawiających się w niektórych przypadkach (z powodów nie do końca jasnych) lokalnych trendów w obserwowanych wartościach, które nie są indykatorami prawdziwych punktów zmiany rozkładu (por. [7], str. 560.).

## 2.2. Algorytm BH

Algorytm *BH* zawdzięcza swoją nazwę od nazwisk autorów – D. Barry'ego i J. A. Hartigana, którzy w pracy [30] zaproponowali metodę identyfikacji punktów zmiany rozkładu polegającą na wyznaczaniu prawdopodobieństwa istnienia punktu zmiany dla każdej lokalizacji w badanym ciągu.

### 2.2.1. Założenia

Podobnie jak w podstawowej wersji algorytmu $CBS$, i w tym przypadku zakładamy, że obserwacje są realizacjami niezależnych zmiennych losowych o rozkładzie $N(\mu_i, \sigma^2)$, $1 \leqslant i \leqslant n$; zakładamy także, że prawdopodobieństwo wystąpienia punktu zmiany wynosi $p$, niezależnie dla każdej lokalizacji $i$, $1 \leqslant i \leqslant n$ elementów ciągu wejściowego.

Rozkład a priori dla $\mu_{ij}$ (średniej zmiennych w segmencie rozpoczynającym się na pozycji $i+1$ i kończącym się na pozycji $i$) jest definiowany jako $N(\mu_0, \sigma_0^2/(j-1))$.

Zgodnie z uwagą zamieszczoną w artykule [31], ścisła implementacja procedury przedstawionej przez D. Barry'ego i J. A. Hartigana w pracy [30] jest możliwa, ale jej złożoność obliczeniowa jest rzędu $O(n^3)$; wersja algorytmu zaimplementowana w bibliotce `bcp` dostępnej w środowisku `R` wykorzystuje aproksymacje uzyskane metodą MCMC (ang. *Markov chain Monte Carlo*) (złożoność obliczeniowa jest rzędu $O(n^2)$).

### 2.2.2. Algorytm

W algorytmie definiujemy tzw. partycję $\rho = (U_1, U_2, ..., U_n)$, gdzie $U_i = 1$ oznacza punkt zmiany na pozycji $i + 1$; inicjalizujemy partycję $\rho$, ustalając $U_i = 0$ dla każdego $i < n$ i $U_n \equiv 1$.

W każdym kroku łańcucha Markowa, dla każdej pozycji $i$ wartość $U_i$ jest generowana z rozkładu warunkowego $U_i$, warunkowanego danymi obserwacjami $x_1, ..., x_n$ i bieżącą partycją $\rho$.

Za [30], oznaczamy $b$ jako liczbę segmentów otrzymanych, gdy $U_i = 0$, pod warunkiem $U_j$ dla $i \neq j$. Prawdopodobieństwo przejścia $p_i$ dla prawdopodobieństwa warunkowego punktu zmiany na pozycji $i + 1$, jest otrzymywane z następującego ilorazu, podanego przez D. Barry'ego i J. A. Hartigana:

$$\frac{p_i}{1-p_i} = \frac{P(U_i = 1|\mathbf{X}, U_j, j \neq i)}{P(U_i = 0|\mathbf{X}, U_j, j \neq i)} = \frac{[\int_0^\gamma p^b(1-p)^{n-b-1}dp]\left[\int_0^\lambda \frac{w^{b/2}}{(W_1+B_1w)^{(n-1)/2}}dw\right]}{[\int_0^\gamma p^{b-1}(1-p)^{n-b}dp]\left[\int_0^\lambda \frac{w^{(b-1)/2}}{(W_0+B_0w)^{(n-1)/2}}dw\right]}, \tag{2.5}$$

gdzie $\mathbf{X}$ oznacza dane, a $W_0$, $B_0$, $W_1$, $B_1$ oznaczają wewnątrz-segmentową i między-segmentową sumę kwadratów uzyskaną, gdy $U_i = 0$ i $U_i = 1$, odpowiednio, przy czym:

$$B = \sum_{i,j \in \rho} (j - i)(\overline{X}_{ij} - \overline{X})^2, \tag{2.6}$$

$$W = \sum_{i,j \in \rho} \sum_{l=i+1}^{j} (X_l - \overline{X}_{ij})^2, \tag{2.7}$$

gdzie $\overline{X} = \sum_{i=1}^n X_i/n$, $\overline{X}_{ij} = \sum_{l=i+1}^j X_l/(j - i)$.

Parametry $\gamma$ i $\lambda$ przyjmują wartości z przedziału $[0, 1]$ i są dobierane tak, by metoda była "efektywna w sytuacjach, gdy nie ma wielu rzeczywistych punktów zmiany ($\gamma$ małe) i gdy wartości zmian nie są relatywnie duże ($\lambda$ małe)", por. [30], str. 312.

W każdym kroku algorytmu, średnie $\mu_{ij}$ a posteriori segmentów są uaktualniane, zależnie od bieżącej partycji $\rho$.

### 2.2.3. Korekta dla danych dużych rozmiarów

Ścisła implementacja algorytmu BH MCMC jest numerycznie niestabilna dla długich ciągów danych, ponieważ funkcje podcałkowe w wyrażeniach:

$$\int_0^\lambda \frac{w^{b/2}}{(W_1 + B_1w)^{(n-1)/2}}dw \tag{2.8}$$

oraz

$$\int_0^\lambda \frac{w^{(b-1)/2}}{(W_0 + B_0w)^{(n-1)/2}}dw \tag{2.9}$$

występujących w Równaniu 2.5 są rozbieżne lub zbiegają do 0 dla długich ciągów danych.

Funkcje te mogą być uproszczone do formy zbliżonej do postaci funkcji beta. Stosując to uproszczenie możemy przepisać 2.5 jako:

$$\begin{aligned}\frac{p_i}{1-p_i} &= \frac{P(U_i=1|\mathbf{X},U_j,j\neq i)}{P(U_i=0|\mathbf{X},U_j,j\neq i)} \\ &= \left(\frac{W_0}{W_1}\right)^{(n-b-2)/2}\cdot\left(\frac{B_0}{B_1}\right)^{(b+1)/2}\cdot\sqrt{\frac{W_1}{B_1}} \\ &\quad\cdot\frac{\int_0^{\frac{B_1\lambda/W_1}{1+B_1\lambda/W_1}} p^{(b+2)/2}(1-p)^{(n-b-3)/2}dp}{\int_0^{\frac{B_0\lambda/W_0}{1+B_0\lambda/W_0}} p^{(b+1)/2}(1-p)^{(n-b-2)/2}dp}\cdot\frac{\int_0^{\gamma} p^{b}(1-p)^{n-b-1}dp}{\int_0^{\gamma} p^{b-1}(1-p)^{n-b}dp}. \quad (2.10)\end{aligned}$$

Powyższa formuła zawiera wyrażenia numerycznie stabilne i umożliwia stosowanie algorytmu $BH$ na ciągach obserwacji dowolnej długości.

**Rozdział 3**

# Wykorzystanie reprezentacji falkowej w dekompozycji wieloskalowej obiektu do identyfikacji punktów zmiany rozkładu

W tej części pracy koncentrujemy się na przedstawieniu podstaw teoretycznych, na których opiera się proponowany przez nas pomysł zastosowania reprezentacji falkowej w dekompozycji wieloskalowej obiektu do identyfikacji punktów zmiany rozkładu.

Niniejszy rozdział zawiera wstęp do teorii zagadnienia dekompozycji wieloskalowej funkcji oraz wektora obserwacji. Opisany został aparat matematyczny, przedstawiony w fundamentalnej pracy S. G. Mallata: "A Theory of Multiresolution Signal Decomposition: The Wavelet Representation" ([32]) z 1989 r. Aparat ten wykorzystuje szczególny przypadek wieloskalowej reprezentacji obiektu, zwanej reprezentacją falkową. Pokazujemy, w jaki sposób reprezentacja falkowa może być wykorzystana w problemie estymacji nieparametrycznej funkcji regresji.

Niniejszy rozdział pracy ma z założenia przybliżyć własności reprezentacji falkowej wektora obserwacji, w szczególności – przypadku z wykorzystaniem tzw. falek Haara, które to własności skłoniły nas do próby wykorzystania transformacji falkowej jako narzędzia do identyfikacji punktów zmiany rozkładu.

## 3.1. Dekompozycja wieloskalowa obiektu

Nie jest wielkim nadużyciem stwierdzenie, że Statystyka oparta jest o różnego rodzaju *transformacje* danych. Podstawowe formuły i narzędzia statystyczne, takie jak średnia próbkowa, wariancja próbkowa, histogram etc., są wynikiem transformacji danych. Również niektóre bardziej zaawansowane metody, takie jak analiza składowych głównych, periodogramy, estymatory jądrowe funkcji gęstości etc., stanowią przykłady przekształceń danych.

Transformacje w Statystyce są stosowane z różnych powodów. Uważa się ([33], Rozdział 7.), że dane poddane odpowiedniej transformacji są wygodniejsze w raportowaniu, przechowywaniu i analizie. Czasami transformowanie danych jest potrzebne, by móc zastosować konkretną metodę statystyczną. W niniejszej pracy korzystamy z kolei z faktu, że transformacja danych często daje nam dodatkowy "wgląd" w zjawisko i pozwala na pojęcie tegoż zjawiska z perspektywy, która nie była możliwa z poziomu danych nieprzetransformowanych.

Niniejsza sekcja zawiera wprowadzenie do zagadnienia jednej z form transformacji obiektu - dekompozycji wieloskalowej.

### 3.1.1. Wstęp

Naszym celem w tej sekcji pracy jest sformułowanie matematyczne reprezentacji wieloskalowej obiektu. W dużej części opisu teoretycznego obiektem tym będą funkcje z przestrzeni $L^2(\mathbb{R})$. W zastosowaniach praktycznych najczęściej mamy do czynienia ze skończoną liczbą obserwacji, które dostarcza nam np. urządzenie pomiarowe. Stąd, w opisie teoretycznym nie zbraknie zdefiniowania problemu reprezentacji wieloskalowej dla przypadku pomiaru dyskretnego.

Wieloskalowa reprezentacja obiektu pozwala na analizę tzw. *detali*, czyli różnic wartości w kolejnych częściach obiektu, na różnych poziomach *rozdzielczości*. Rozdzielczość możemy rozumieć intuicyjnie jako przybliżenie, z którym przyglądamy się obiektowi.

Zgodnie z intuicją, rozpatrywanie różnic pomiędzy wartościami kolejnych części obiektu, gdy wykonywane na bardziej *zgrubnym* poziomie rozdzielczości ("z większej odległości") – odpowiada charakterystyce większych struktur budujących ten obiekt. Z kolei na "drobniejszym" poziomie rozdzielczości – różnica wartości kolejnych części obiektu koduje informację o szczegółach struktury obiektu. Reprezentacja wieloskalowa pozwala więc na *rozpoznawanie wzorca tworzącego dany obiekt* na wybranym poziomie szczegółowości tego wzorca.

W dalszej części pracy przedstawiamy wyniki zawarte w artykule "A Theory of Multiresolution Signal Decomposition: The Wavelet Representation" S. G. Mallata ([32]). W pracy tej zostały opisane własności matematyczne operatora, który transformuje funkcję $f \in L^2(\mathbb{R})$ do jej aproksymacji (przybliżenia)

na poziomie rozdzielczości $2^j$. Pokazano, że różnica informacji zawartej w aproksymacjach funkcji $f$ na dwóch kolejnych poziomach rozdzielczości, $2^{j+1}$ i $2^j$, może być uzyskana w procesie dekompozycji funkcji w tzw. *falkowej bazie ortonormalnej*. Reprezentacja wieloskalowa funkcji $f$ uzyskana w ten sposób nazywana jest *reprezentacją falkową*.

*Falki* są to funkcje wprowadzone przez A. Grossmanna i J. Morleta w pracy "Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape" ([34]) w 1984 r. Falki definiujemy jako funkcje $\psi(x)$, dla których w wyniku operacji *translacji* i *dylatacji* otrzymujemy rodzinę funkcji

$$\{\sqrt{(}s)\psi(sx-t)\}_{(s,t)\in\mathbb{R}^+\times\mathbb{R}} \tag{3.1}$$

będącą bazą ortonormalną w $L^2(\mathbb{R})$. W szczególności, Y. Meyer ([35]) pokazał, że istnieją falki $\psi(x)$ takie, że

$$\{\sqrt{(}2^j)\psi(2^jx-k)\}_{(j,k)\in\mathbb{Z}^2} \tag{3.2}$$

jest bazą ortonormalną w $L^2(\mathbb{R})$.

Ważnym elementem opisu teoretycznego zawartego w tej sekcji pracy jest opis tzw. algorytmu piramidalnego (ang. *pyramidal lagorithm*), pozwalającego wyznaczyć *reprezentację falkową* funkcji $f$.

**Notacja**

Stosować będziemy następujące standardowe oznaczenia, zaczerpnięte z [32].

— Dla $f(x) \in L^2(\mathbb{R})$ i $g(x) \in L^2(\mathbb{R})$, iloczyn skalarny $f(x)$ z $g(x)$ oznaczamy jako

$$\langle g(u), f(u)\rangle = \int_{-\infty}^{+\infty} g(u)f(u)du. \tag{3.3}$$

— Normę $f(x)$ w $L^2(\mathbb{R})$ oznaczamy jako

$$||f(x)|| = \sqrt{\int_{-\infty}^{+\infty} |f(x)|^2 du}. \tag{3.4}$$

— Splot dwóch funkcji $f(x) \in L^2(\mathbb{R})$ i $g(x) \in L^2(\mathbb{R})$ oznaczamy jako

$$\begin{aligned} f * g(x) &= (f(u) * g(u))(x) \\ &= \int_{-\infty}^{+\infty} f(u)g(x-u)du. \end{aligned} \tag{3.5}$$

— Splot dwóch ciągów sumowalnych (numerowanych liczbami całkowitymi) $f$ i $g$ oznaczamy jako

$$\begin{aligned} (f * g)[n] &= \sum_{m=-\infty}^{\infty} f[m]\, g[n-m] \\ &= \sum_{m=-\infty}^{\infty} f[n-m]\, g[m]. \end{aligned} \tag{3.6}$$

— Transformatę Fouriera funkcji $f(x) \in L^2(\mathbb{R})$ oznaczamy przez $\widehat{f}(\omega)$ i definiujemy jako

$$\widehat{f}(\omega) = \int_{-\infty}^{+\infty} f(x) e^{-i\omega x} dx. \tag{3.7}$$

— Przestrzeń ciągów sumowalnych z kwadratem oznaczamy jako $I^2(\mathbb{Z})$ i definiujemy

$$I^2(\mathbb{Z}) = \left\{ (\alpha_i)_{i\in\mathbb{Z}} : \sum_{i=-\infty}^{+\infty} |\alpha_i|^2 < \infty \right\}. \tag{3.8}$$

### 3.1.2. Aproksymacja wieloskalowa $L^2(\mathbb{R})$

Oznaczmy przez $A_{2^j}$ operator, o którym powiemy, że aproksymuje on pewien sygnał $f(x)$ na poziomie rozdzielczości $2^j$. Zakładamy, że sygnał wejściowy $f(x)$ jest funkcją mierzalną i o skończonej energii, tj. $f \in L^2(\mathbb{R})$. Poniżej definiujemy własności, których intuicyjnie oczekiwalibyśmy od takiego operatora aproksymującego.

1) $A_{2^j}$ jest operatorem liniowym. Dalej, jeśli $A_{2^j} f(x)$ jest aproksymacją pewnej funkcji $f(x)$ na poziomie rozdzielczości $2^j$, to $A_{2^j} f(x)$ nie jest zmodyfikowany, jeśli ponownie aproksymujemy go na poziomie rozdzielczości $2^j$. Innymi słowy,

$$A_{2^j} \circ A_{2^j} = A_{2^j}.$$

Oznacza to, że operator $A_{2^j}$ jest *operatorem projekcyjnym* na pewną przestrzeń wektorową $V_{2^j} \subset L^2(\mathbb{R})$. Przestrzeń $V_{2^j}$ może być interpretowana jako *zbiór wszystkich możliwych aproksymacji na poziomie rozdzielczości* $2^j$ funkcji z $L^2(\mathbb{R})$.

2) W zbiorze wszystkich funkcji aproksymowanych na poziomie rozdzielczości $2^j$, $A_{2^j} f(x)$ jest funkcją, która jest najbardziej zbliżona do $f(x)$. Formalnie mamy:

$$\forall\ g(x) \in V_{2^j},\ ||g(x) - f(x)|| \geqslant ||A_{2^j} f(x) - f(x)||. \tag{3.9}$$

Oznacza to, że operator $A_{2^j}$ jest projekcją ortogonalną na przestrzeń wektorową $V_{2^j}$.

3) Aproksymacja sygnału $f(x)$ na poziomie rozdzielczości $2^{j+1}$ zawiera wszelkie potrzebne informacje, aby wyznaczyć aproksymację tego samego sygnału $f(x)$ na poziomie rozdzielczości $2^j$. Własność ta w literaturze nazywana jest *przyczynowością* (ang. *causality property*).

Ponieważ $A_{2^j}$ jest operatorem projekcyjnym na $V_{2^j}$, ta własność jest równoważna z:

$$\forall\ j \in \mathbb{Z},\ V_{2^j} \subset V_{2^{j+1}}. \tag{3.10}$$

4) Przestrzenie aproksymowanych funkcji mogą być otrzymywane jedna z drugiej poprzez skalowanie odpowiedniej funkcji o współczynnik będący stosunkiem poziomu rozdzielczości tych przestrzeni, tj.:

$$\forall\ j,k \in \mathbb{Z},\ f(x) \in V_{2^j} \Leftrightarrow f(2^k x) \in V_{2^{j+k}}. \tag{3.11}$$

W szczególności, dla $k = 1$ mamy:

$$\forall\ j \in \mathbb{Z},\ f(x) \in V_{2^j} \Leftrightarrow f(2x) \in V_{2^{j+1}}.$$

5) Aproksymacja $A_{2^j} f(x)$ sygnału wejściowego $f(x)$ jest określana przy użyciu $2^j$-elementowej próbki wartości tej funkcji ($2^j$-elementowej próbki na jednostkę długości sygnału, dla którego wyznaczamy aproksymację). Jeżeli poddamy $f(x)$ operacji *translacji* o odległość proporcjonalną do $2^{-j}$, to po tej operacji $A_{2^j} f(x)$ jest przesunięta o tę samą wartość i jest opisywana przez tę samą $2^j$-elementową próbkę, co przed translacją $f(x)$.

Z uwagi na własność 4) (por. równoważność 3.11), własność 5) możemy opisać w przypadku rozdzielczości dla $j = 0$. Matematyczna formuła operacji translacji przedstawia się następująco.

— Charakteryzacja w przypadku dyskretnym:

$$\text{Istnieje izomorfizm } I \text{ z } V_1 \text{ do } I^2(\mathbb{Z}), \tag{3.12}$$

gdzie $I^2(\mathbb{Z})$ jest zdefiniowane w 3.8.

— Translacja aproksymacji:

$$\forall\ k \in \mathbb{Z},\ A_1 f_k(x) = A_1 f(x-k),\ \text{ gdzie } f_k(x) = f(x-k). \tag{3.13}$$

— Translacja próbki z wartości funkcji:

$$I(A_1 f(x)) = (\alpha_i)_{i\in\mathbb{Z}} \Leftrightarrow I(A_1 f_k(x)) = (\alpha_{i-k})_{i\in\mathbb{Z}}. \tag{3.14}$$

6) Wyznaczanie aproksymacji wejściowego sygnału $f(x)$ na poziomie rozdzielczości $2^j$ wiąże się z utratą części informacji na temat $f(x)$. Gdy zwiększamy poziom rozdzielczości do $+\infty$, aproksymacja sygnału $A_{2^j} f(x)$ zbiega do postaci wejściowego sygnału $f(x)$. I na odwrót, gdy poziom rozdzielczości maleje do 0, aproksymacja sygnału zawiera coraz mniej informacji o wejściowym sygnale $f(x)$ (zbiega do 0).

Ponieważ aproksymacja sygnału $f(x)$ na poziomie rozdzielczości $2^j$ jest równa projekcji ortogonalnej sygnału $f(x)$ na przestrzeń wektorową $V_{2^j}$, własność 6) może być zapisana jako:

$$\lim_{j\to+\infty} V_{2^j} = \bigcup_{j=-\infty}^{+\infty} V_{2^j} \text{ jest gęsty w } L^2(\mathbb{R}) \tag{3.15}$$

oraz

$$\lim_{j\to-\infty} V_{2^j} = \bigcap_{j=-\infty}^{+\infty} V_{2^j} = \{0\}. \tag{3.16}$$

**Definicja 1** (Aproksymacja wieloskalowa)**:** Każdy zbiór przestrzeni wektorowych

$$(V_{2^j})_{j\in\mathbb{Z}} \tag{3.17}$$

spełniający własności (3.10)-(3.16) nazywamy *aproksymacją wieloskalową* (ang. *multiscale approximation (MSA)*) $L^2(\mathbb{R})$. Równoważnie możemy mówić: *analiza wielorozdzielcza* (ang. *multiresolution analysis (MRA)*) $L^2(\mathbb{R})$.

Stowarzyszony (w sensie własności 2)) z $(V_{2^j})_{j\in\mathbb{Z}}$ zbiór operatorów $A_{2^j}$ spełniający własności 1) - 6), *aproksymuje* dowolną funkcję $f \in L^2(\mathbb{R})$ na poziomie rozdzielczości $2^j$.

Zobaczyliśmy, że operator aproksymujący $A_{2^j}$ jest projekcją ortogonalną na przestrzeń wektorową $V_{2^j}$. Aby móc opisać ten operator formułą matematyczną, potrzebujemy znaleźć bazę ortonormalną w przestrzeni wektorowej $V_{2^j}$. Następujące twierdzenie pokazuje, że taką bazę ortonormalną można otrzymać w wyniku nakładania operacji dylatacji i translacji na pewną jednoznacznie zdefiniowaną funkcję $\phi(x)$.

**Twierdzenie 1:** Niech $(V_{2^j})_{j\in\mathbb{Z}}$ będzie wieloskalową aproksymacją $L^2(\mathbb{R})$. Istnieje jednoznacznie zdefiniowana funkcja

$$\phi(x) \in L^2(\mathbb{R}), \tag{3.18}$$

nazywana *funkcją skalującą*, taka, że dla $\phi_{2^j}(x)$:

$$\phi_{2^j}(x) = 2^j\phi(2^j x) \text{ dla } j \in \mathbb{Z} \text{ (dylatacja funkcji } \phi(x) \text{ o } 2^j) \tag{3.19}$$

mamy, że rodzina funkcji

$$\left(\sqrt{2^{-j}}\phi_{2^j}(x - 2^{-j}n)\right)_{n\in\mathbb{Z}} \tag{3.20}$$

jest bazą ortonormalną w $V_{2^j}$.

Szkic dowodu powyższego twierdzenia znajduje się w *Appendiksie B* w pracy [32].

Twierdzenie 1. pokazuje, że możemy zbudować bazę ortonormalną każdej przestrzeni wektorowej $V_{2^j}$ poprzez dylatację funkcji $\phi(x)$ o współczynnik $2^j$ oraz translację wynikowej funkcji do przedziału, którego długość jest proporcjonalna do $2^{-j}$.

Funkcje $\phi_{2^j}$ są normalizowane tak, aby elementy $\sqrt{2^{-j}}\phi_{2^j}(x - 2^{-j}n)$ rodziny ortonormalnej (3.20) miały normę w $L^2(\mathbb{R})$ równą 1; stąd obecność współczynnika $\sqrt{2^{-j}}$ w (3.20).

Stwierdziliśmy, że dla danej wieloskalowej aproksymacji $(V_{2^j})_{j\in\mathbb{Z}}$ istnieje zdefiniowana jednoznacznie funkcja skalująca $\phi(x)$. Należy mieć na uwadze, że dla różnych wieloskalowych aproksymacji mamy różne funkcje skalujące.

### 3.1.3. Aproksymacja dyskretna sygnału $f(x) \in L^2(\mathbb{R})$

Korzystając z Twierdzenia 1. możemy zapisać projekcję ortogonalną sygnału wejściowego $f(x)$ na przestrzeń wektorową $V_{2^j}$ poprzez rozpisanie $f(x)$ w bazie ortonormalnej (3.20):

$$\forall\ f(x) \in L^2(\mathbb{R}),\quad A_{2^j} f(x) = 2^{-j} \sum_{n=-\infty}^{+\infty} \Big\langle f(u), \phi_{2^j}(u - 2^{-j}n) \Big\rangle\, \phi_{2^j}(x - 2^{-j}n). \tag{3.21}$$

Z (3.21) otrzymujemy, że aproksymacja $A_{2^j} f(x)$ sygnału $f(x)$ na poziomie rozdzielczości $2^j$ jest charakteryzowana przez zbiór iloczynów skalarnych, które oznaczamy w następujący sposób:

$$A^d_{2^j} = \Big( \Big\langle f(u), \phi_{2^j}(u - 2^{-j}n) \Big\rangle \Big)_{n \in \mathbb{Z}}. \tag{3.22}$$

**Definicja 2** (Aproksymacja dyskretna $f(x)$ na poziomie rozdzielczości $2^j$)**:** Reprezentację $A^d_{2^j} f$ sygnału $f(x)$ zadaną równaniem (3.22) nazywamy *aproksymacją dyskretną sygnału $f(x)$ na poziomie rozdzielczości $2^j$*.

Jesteśmy zainteresowani rozważaniem aproksymacji dyskretnych z racji faktu, że komputery mogą przetwarzać tylko sygnały *dyskretne*.

Każdy iloczyn skalarny ze zbioru (3.22) może być interpretowany jako splot funkcji, ewaluowany w punkcie $2^{-j}n$:

$$\begin{aligned} \Big\langle f(u), \phi_{2^j}(u - 2^{-j}n) \Big\rangle &= \int_{-\infty}^{+\infty} f(u) \phi_{2^j}(u - 2^{-j}n) du \\ &= \left(f(u) * \phi_{2^j}(-u)\right)(2^{-j}n). \end{aligned} \tag{3.23}$$

Stąd, możemy przepisać $A^d_{2^j} f$ następująco:

$$A^d_{2^j} = \Big( \left(f(u) * \phi_{2^j}(-u)\right)(2^{-j}n) \Big)_{n \in \mathbb{Z}}. \tag{3.24}$$

Podsumujmy intuicje dot. przedstawionej powyżej teorii. Dla danego obiektu mamy do czynienia z pewną zdolnością rozdzielczą jego reprezentacji, którą możemy operować. Dla ciągu obserwacji zdolność rozdzielczą określa najmniejsza odległość między dwoma elementami ciągu (związana z możliwościami przyrządu pomiarowego, specyfiką przedmiotu badań, sposobem przetwarzania danych etc.). Tzw. *filtracja* pozwala wyodrębnić interesujące nas zjawiska. Filtracja oznacza wydobywanie z sygnału tych danych pomiarowych, które nas interesują; najczęściej chodzi o wyodrębnienie składowych o skali większej niż standardowa odległość między punktami pomiarowymi.

Procedura precyzyjnego określenia skali wymaga przedstawienia funkcji w *rozkładzie spektralnym* (por. (3.21)). Rozkładając reprezentowaną funkcję na składniki, którym przypisujemy skale, na ogół korzystamy z jakiejś bazy funkcji ortogonalnych, jako że takie funkcje mają charakter oscylacyjny. W

rozważanym przez nas powyżej przypadku tą bazą jest rodzina zdefiniowana w (3.20).

Do operacji filtracji służą tzw. *filtry* (filtry częstotliwości). Filtrem częstotliwości nazywamy układ, który przepuszcza bez tłumienia lub z małym tłumieniem sygnały o określonym paśmie częstotliwości, a tłumi sygnały leżące poza tym pasmem. W grupie filtrów wyróżniamy m.in.:

— filtry *dolnoprzepustowe* – przepuszczają bez tłumienia (lub z małym tłumieniem) sygnały o częstotliwości równej od 0 do pewnej częstotliwości granicznej,

— filtry *górnoprzepustowe* – przepuszczają bez tłumienia (lub z małym tłumieniem) sygnały o częstotliwości o wartościach od pewnej granicznej do nieskończoności.

Zauważmy, że operacja splotu sygnału $f(x)$ z dowolnym *jądrem* z wagami dodatnimi (z funkcją nieujemną) prowadzi do otrzymania średnich ważonych wartości wejściowego sygnału $f(x)$. Uśrednianie sygnału jest formą filtrowania dolnoprzepustowego.

Powyższe intuicje możemy teraz powiązać z opisem teoretycznym operacji aproksymacji funkcji $f(x)$, przedstawionym powyżej. Zauważmy, że w wyniku operacji aproksymacji funkcji $f(x)$, (3.21), tj. w wyniku projekcji ortogonalnej $f(x)$ na przestrzeń wektorową $V_{2^j}$, usuwamy detale (szczegóły) funkcji $f(x)$ mniejsze, niż $2^{-j}$; pozbywamy się w ten sposób wysokich częstotliwości występujących w częstotliwościowym opisie wejściowej funkcji $f(x)$. Pokazaliśmy, że aproksymację dyskretną $f(x)$ na poziomie rozdzielczości $2^j$ możemy zapisać, wykorzystując splot sygnału wejściowego z funkcją skalującą $\phi(x)$ (poddaną dylatacji) – por. (3.24).

Ponieważ $\phi(x)$ jest filtrem dolnoprzepustowym ([32], str. 677), to ciąg $A^d_{2^j}$ możemy traktować jako sygnał dyskretny, który jest wynikiem operacji filtrowania dolnoprzepustowego wejściowego sygnału $f(x)$, oraz, w dalszej kolejności, próbkowania równomiernego w odstępach równych $2^j$. W tej interpretacji:

— funkcja $\phi_{2^j}(x)$ jest określana jest jako tzw. *odpowiedź impulsowa* (równoważnie: *funkcja odpowiedzi impulsowej*, *charakterystyka impulsowa*),

— funkcja $f(x)$ jest określany jako tzw. *sygnał pobudzający (wejściowy)*,

— splot $(f(u) * \phi_{2^j}(-u))$ to tzw. *sygnał na wyjściu*.

W następnej subsekcji niniejszej pracy podajemy za Mallatem [32] tzw. *algorytm piramidalny* (ang. *pyramidal algorithm*) na wyznaczenie aproksymacji dyskretnej sygnału $f(x)$ na poziomie rozdzielczości $2^j$.

### 3.1.4. Implementacja transformacji wieloskalowej – algortym piramidalny

Jak wspomniano powyżej, w praktyce, urządzenia miernicze dostarczają pomiarów sygnału $f(x)$ w postaci ciągu dyskretnego, tj. na pewnym skończonym poziomie rozdzielczości. Za [32], oznaczmy ten wejściowy, najdrobniejszy poziom rozdzielczości jako równy 1.

Niech $A_1^d f$ będzie dyskretną aproksymacją sygnału $f(x)$ na poziomie rozdzielczości 1, o którym to pomiarze zakładamy, że nim dysponujemy. Z własności przyczynowości (por. 3) powyżej), mając $A_1^d f$ możemy wyznaczyć wszystkie aproksymacje $A_{2^j}^d f$ dla $j < 0$. Mallat w swojej pracy: "A Theory of Multiresolution Signal Decomposition: The Wavelet Representation" ([32]) przedstawia tzw. *algorytm piramidalny* na wyznaczanie tych aproksymacji.

W algorytmie aproksymacje dyskretne $A_{2^j}^d f$ sygnału $f(x)$ na poziomie rozdzielczości $2^j$ są wyznaczane w sposób iteracyjny, przy wykorzystaniu operacji splotu $A_{2^{j+1}}^d f$ z pewnym filtrem $\widetilde{H}$. Szczegóły implementacji algorytmu umieszczamy poniżej.

Niech $(V_{2^j})_{j\in\mathbb{Z}}$ będzie aproksymacją wieloskalową i niech $\phi(x)$ będzie powiązaną z nią funkcją skalującą (por. Twierdzenie 1.). Rozważmy składnik:

$$\phi_{2^j}(x - 2^{-j}n),$$

występoujący w formule aproksymacji sygnału $f(x)$ na poziomie rozdzielczości $2^j$ (3.21).

Rodzina funkcji:

$$\left(\sqrt{2^{-j-1}}\ \phi_{2^{j+1}}(x - 2^{-j-1}k)\right)_{k\in\mathbb{Z}}$$

jest bazą ortonormalną w $V_{2^{j+1}}$. Wiemy, że dla każdego $n \in \mathbb{Z}$ funkcja $\phi_{2^j}(x - 2^{-j}n)$ jest elementem przestrzeni wektorowej $V_{2^j}$, która zawiera się w $V_{2^{j+1}}$. Możemy więc $\phi_{2^j}(x - 2^{-j}n)$ rozwinąć w bazie ortonormalnej w $V_{2^{j+1}}$:

$$\phi_{2^j}(x-2^{-j}n) = 2^{-j-1} \sum_{k=-\infty}^{+\infty} \left\langle \phi_{2^j}(u - 2^{-j}n), \phi_{2^{j+1}}(u - 2^{-j-1}k) \right\rangle \cdot \phi_{2^{j+1}}(x-2^{-j-1}k). \tag{3.25}$$

Wykorzystując formułę dylatacji (3.19) funkcji $\phi(x)$ i wykonując zamianę zmiennych w całce z definicji iloczynu skalarnego, otrzymujemy:

$$\begin{aligned}
&2^{-j-1}\left\langle \phi_{2^j}(u-2^{-j}n), \phi_{2^{j+1}}(u-2^{-j-1}k)\right\rangle = \\
&= 2^{-j-1}\int_{-\infty}^{+\infty} \phi_{2^j}(u-2^{-j}n)\phi_{2^{j+1}}(u-2^{-j-1}k)du \overset{\phi_{2^j}(x)=2^j\phi(2^jx)}{=} \\
&= 2^{-j-1}\int_{-\infty}^{+\infty} 2^{j+1}\phi_{2^{-1}}(2^{j+1}(u-2^{-j}n))2^{j+1}\phi(2^{j+1}(u-2^{-j-1}k))du \\
&= 2^{j+1}\int_{-\infty}^{+\infty} \phi_{2^{-1}}(2^{j+1}u-2n)\phi(2^{j+1}u-k)du \overset{(2^{j+1}u-2n)=u}{=} \\
&= \int_{-\infty}^{+\infty} \phi_{2^{-1}}(u)\phi(u-(k-2n))du = \\
&= \langle \phi_{2^{-1}}(u), \phi(u-(k-2n))\rangle . \qquad (3.26)
\end{aligned}$$

Korzystając z powyższego, możemy zapisać równość 3.25 jako:

$$\phi_{2^j}(x-2^{-j}n) = \sum_{k=-\infty}^{+\infty} \langle \phi_{2^{-1}}(u), \phi(u-(k-2n))\rangle \cdot \phi_{2^{j+1}}(x-2^{-j-1}k). \quad (3.27)$$

Następnie wyznaczamy iloczyn skalarny $f(x)$ z lewą i prawą stroną równania (3.27), otrzymując:

$$\left\langle f(u), \phi_{2^j}(u-2^{-j}n)\right\rangle = \sum_{k=-\infty}^{+\infty} \langle \phi_{2^{-1}}(u), \phi(u-(k-2n))\rangle \cdot \left\langle f(u), \phi_{2^{j+1}}(u-2^{-j-1}k)\right\rangle . \quad (3.28)$$

Chcemy dodatkowo usystematyzować powyższy zapis. Zdefiniujmy $H$ – filtr dyskretny, którego odpowiedź impulsowa $h$ zadana jest jako

$$\forall n \in \mathbb{Z},\ h(n) = \langle \phi_{2^{-1}}(u), \phi(u-n)\rangle . \quad (3.29)$$

Niech $\widetilde{H}$ będzie filtrem dyskretnym *lustrzanym* do $H$, tj. takim, że:

$$\forall n \in \mathbb{Z},\ \widetilde{h}(n) = h(-n) = \langle \phi_{2^{-1}}(u), \phi(u+n)\rangle . \quad (3.30)$$

Wykorzystując postać odpowiedzi impulsowej $\widetilde{h}$ w zapisie (3.28), otrzymujemy:

$$\left\langle f(u), \phi_{2^j}(u-2^{-j}n)\right\rangle = \sum_{k=-\infty}^{+\infty} \widetilde{h}(2n-k) \cdot \left\langle f(u), \phi_{2^{j+1}}(u-2^{-j-1}k)\right\rangle . \quad (3.31)$$

Powyższe równanie pokazuje, że aproksymację dyskretną $A^d_{2^j}f$ sygnału $f(x)$ na poziomie rozdzielczości $2^j$, tj.:

$$A^d_{2^j} = \left(\left\langle f(u), \phi_{2^j}(u-2^{-j}n)\right\rangle\right)_{n\in\mathbb{Z}},$$

możemy wyznaczyć, wykorzystując operację splotu $A_{2^{j+1}}f$ z $\widetilde{H}$. Widzimy ponadto, że każda dyskretna aproksymacja $A_{2^j}f$ dla $j<0$ może być wyznaczona z użyciem wejściowego dyskretnego pomiaru na poziomie rozdzielczości równym 1, $A_1 f$, poprzez interacyjne powtarzanie tego procesu. Operację tę nazywamy *algorytmem piramidalnym*.

## 3.2. Reprezentacja falkowa sygnału

Jak wspomniane zostało na początku tego rozdziału, dążymy do zbudowania reprezentacji wieloskalowej sygnału, która wyznaczana jest w oparciu o *różnicę informacji zawartą między dwoma następującymi po sobie poziomami rozdzielczości*, $2^j$ oraz $2^{j+1}$. W niniejszej części pracy pokazujemy (za [32]), że reprezentacja ta może być wyznaczona poprzez dekompozycję sygnału w *falkowej bazie ortonormalnej*.

### 3.2.1. Sygnał detalu

**Definicja 3** (Sygnał detalu)**:** *Sygnałem detalu* na poziomie rozdzielczości $2^j$ nazywamy różnicę informacji między aproksymacją funkcji $f(x)$ na poziomach rozdzielczości $2^{j+1}$ i $2^j$.

Aproksymacje sygnału $f(x)$ na poziomach rozdzielczości $2^{j+1}$ i $2^j$ są równe projekcjom ortogonalnym $f(x)$ na przestrzenie wektorowe $V_{2^{j+1}}$ i $V_{2^j}$, odpowiednio. Wynika stąd, że sygnał detalu na poziomie rozdzielczości $2^j$ jest zadany przez projekcję ortogonalną sygnału wejściowego $f(x)$ na dopełnienie ortogonalne $V_{2^j}$ w $V_{2^{j+1}}$.

Oznaczmy przez $O_{2^j}$ dopełnienie ortogonalne $V_{2^j}$ w $V_{2^{j+1}}$:

$$O_{2^j} \text{ jest ortogonalne do } V_{2^j}, \tag{3.32}$$

$$O_{2^j} \oplus V_{2^j} = V_{2^{j+1}}. \tag{3.33}$$

Aby wyznaczyć projekcję ortogonalną sygnału wejściowego $f(x)$ na dopełnienie ortogonalne $O_{2^j}$, potrzebujemy znaleźć bazę ortonormalną w $O_{2^j}$. Receptę na wyznaczenie bazy ortonormalnej w $O_{2^j}$ dostarczy Twierdzenie 3., zamieszczone w dalszej części niniejszej pracy. Przed jego sformułowaniem podamy Twierdzenie 2., które mówi nam o pewnych charakterystykach transformaty Fouriera funkcji skalującej $\phi(x)$, oraz zdefiniujemy tzw. *filtr sprzężony* $H$.

### 3.2.2. Transformata Fouriera funkcji skalującej $\phi(x)$

Niech $(V_{2^j})_{j\in\mathbb{Z}}$ będzie wieloskalową aproksymacją $L^2(\mathbb{R})$, $\phi(x)$ - odpowiadającą jej funkcją skalującą. Nakładamy na funkcję skalującą dodatkowe warunki regularności: niech $\phi(x)$ będzie różniczkowalna w sposób ciagły oraz niech $\phi(x)$ i $\phi'(x)$ zanikają w $+\infty$ i $-\infty$ w sposób spełniający:

$$|\phi(x)| = O(x^{-2}) \tag{3.34}$$

oraz

$$|\phi'(x)| = O(x^{-2}). \tag{3.35}$$

**Twierdzenie 2:** Niech $\phi(x)$ będzie funkcją skalującą, $H$ - filtrem dyskretnym z odpowiedzią impulsową $h(n) = \langle \phi_{2^{-1}}, \phi(u-n) \rangle$. Niech $H(\omega)$ będzie szeregiem Fouriera zdefiniowanym jako:

$$H(\omega) = \sum_{-\infty}^{\infty} h(n) e^{-in\omega}. \tag{3.36}$$

$H(\omega)$ spełnia następujące dwie własności:

$$|H(0)| = 1 \text{ oraz } h(n) = O(n^{-2}) \text{ w nieskończoności,} \tag{3.37}$$

$$|H(\omega)|^2 + |H(\omega + \pi)|^2 = 1. \tag{3.38}$$

I odwrotnie, niech $H(\omega)$ będzie funkcją spełniającą (3.37) oraz (3.38) oraz taką, że

$$|H(\omega)| \neq 0 \text{ dla } \omega \in [0, \pi/2]. \tag{3.39}$$

Wtedy funkcja zdefiniowana jako

$$\widehat{\phi}(\omega) = \prod_{p=1}^{+\infty} H(2^{-p}\omega) \tag{3.40}$$

jest transformatą Fouriera funkcji skalującej $\phi(x)$.

Szkic dowodu powyższego Twierdzenia znajduje się w *Appendiksie C* w pracy [32].

**Definicja 4** (Filtr sprzężony)**:** Filtr, który spełnia warunek (3.38) nazywamy *filtrem sprzężonym* (ang. *conjugate filter*).

### 3.2.3. Baza ortonormalna dopełnienia ortogonalnego $V_{2^j}$ w $V_{2^{j+1}}$

**Twierdzenie 3:** Niech $(V_{2^j})_{j \in \mathbb{Z}}$ będzie wieloskalową aproksymacją $L^2(\mathbb{R})$, $\phi(x)$ - odpowiadającą jej funkcją skalującą. Niech $\psi(x)$ będzie funkcją, której transformata Fouriera zadana jest w następujący sposób:

$$\widehat{\psi}(\omega) = G\left(\frac{\omega}{2}\right) \widehat{\phi}\left(\frac{\omega}{2}\right), \tag{3.41}$$

gdzie

$$G(\omega) = e^{-i\omega} \overline{H(\omega + \pi)}. \tag{3.42}$$

Niech $\psi_{2^j}(x) = 2^j \psi(2^j x)$ oznacza dylatację funkcji $\psi(x)$ o $2^j$. Wtedy

$$\left(\sqrt{2^{-j}}\ \psi_{2^j}(x - 2^{-j}n)\right) n \in \mathbb{Z} \tag{3.43}$$

jest bazą ortonormalną w $O_{2^j}$ oraz

$$\left(\sqrt{2^{-j}}\ \psi_{2^j}(x - 2^{-j}n)\right)_{(n,j) \in \mathbb{Z}^2} \tag{3.44}$$

jest bazą ortonormalną w $L^2(\mathbb{R})$.

Szkic dowodu powyższego Twierdzenia znajduje się w *Appendiksie D* w pracy [32].

Z powyższego Twierdzenia otrzymujemy, że baza ortonormalna w $O_{2^j}$ może być otrzymana poprzez skalowanie falki $\psi(x)$ przez współczynnik $2^j$ oraz jej translację do przedziału, który jest proporcjonalny do $2^{-j}$.

**Definicja 5** (Falka ortogonalna)**:** Funkcję $\psi(x)$ określoną jak Twierdzeniu 3. nazywamy *falką ortogonalną*.

Poniżej przedstawiamy przykład, łączący wprowadzone w tym rozdziale pojęcia – wieloskalową aproksymację $L^2(\mathbb{R})$ oraz związane z nią: funkcję skalującą $\phi(x)$ i falkę ortogonalą $\psi(x)$.

**Przykład 1** Rozważmy $V_1$ – przestrzeń wektorową wszystkich funkcji z $L^2(\mathbb{R})$, które są stałe na każdym odcinku postaci $[k, k+1),\ k \in \mathbb{Z}$.

Równanie (3.11) charakteryzujące wieloskalową aproksymację $L^2(\mathbb{R})$:

$$\forall\ j,k \in \mathbb{Z},\ f(x) \in V_{2^j} \Leftrightarrow f(2^k x) \in V_{2^{j+k}}$$

implikuje, że $V_{2^j}$ jest przestrzenią wektorową funkcji z $L^2(\mathbb{R})$, które są stałe na każdym odcinku postaci $[k2^{-j}, (k+1)2^{-j}),\ k \in \mathbb{Z}$.

Równanie (3.10):

$$\forall\ j \in \mathbb{Z},\ V_{2^j} \subset V_{2^{j+1}}$$

jest łatwo weryfikowalne.

Można zdefiniować izomorfizm $I$, który każdej funkcji $f(x) \in V_1$ przyporządkowuje ciąg $(\alpha_k)_{k\in\mathbb{Z}}$ taki, że $\alpha_k$ równe jest wartości funkcji $f(x)$ na przedziale $[k, k+1)$. Izomorfizm ten spełnia warunki wieloskalowej aproksymacji $L^2(\mathbb{R})$, definiowane równaniami (3.12), (3.13), (3.14).

Wiadome jest, że przestrzeń wektorowa funkcji kawałkami stałych jest gęsta w $L^2(\mathbb{R})$, stąd otrzymujemy, że $\bigcup_{j=-\infty}^{+\infty} V_{2^j}$ jest gęsta w $L^2(\mathbb{R})$. Jednocześnie mamy $\bigcap_{j=-\infty}^{+\infty} V_{2^j} = \{0\}$.

Ostatecznie możemy stwierdzić, że zbiór przestrzeni wektorowych $(V_{2^j})_{j\in\mathbb{Z}}$ jest wieloskalową aproksymacją $L^2(\mathbb{R})$. Łatwo widać, że związana z nią, wyznaczona jednoznacznie (por. Twierdzenie 3.18.) funkcja skalująca $\phi(x)$ jest postaci:

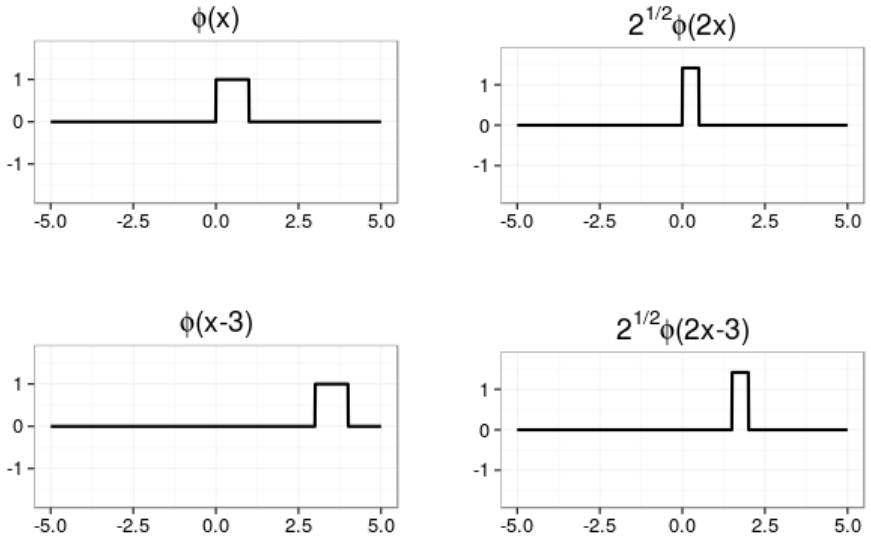
$$\phi(x) = \begin{cases} 1 & 0 \leqslant x < 1, \\ 0 & \text{w pozostałych przypadkach.} \end{cases} \tag{3.45}$$

Odpowiadająca $\phi(x)$ falka ortonormalna $\psi(x)$ jest postaci:

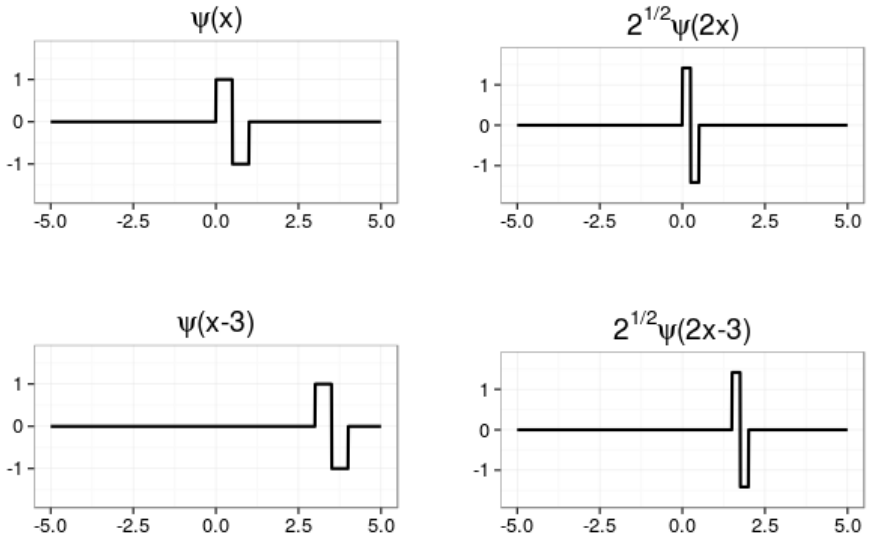
$$\psi(x) = \begin{cases} 1 & 0 \leqslant x < \frac{1}{2}, \\ -1 & \frac{1}{2} \leqslant x < 1, \\ 0 & \text{w pozostałych przypadkach.} \end{cases} \tag{3.46}$$

**Definicja 6** (Falka Haara)**:** Falkę ortogonalną zdefiniowaną w 3.46 nazywamy *falką Haara*.

Wykresy funkcji $\phi(x)$, $\psi(x)$ oraz przykłady tych funkcji po nałożeniu operacji dylatacji lub/i translacji przedstawione są na Rysunkach poniżej.




Rysunek 3.1: Wykres funkcji skalującej $\phi(x)$ oraz przykłady tej funkcji po nałożeniu operacji dylatacji lub/i translacji.




Rysunek 3.2: Wykres falki ortonormalnej $\psi(x)$ oraz przykłady tej funkcji po nałożeniu operacji dylatacji lub/i translacji.

### 3.2.4. Ortogonalna falkowa reprezentacja sygnału

Niech $P_{O_{2^j}}$ będzie projekcją ortogonalną na przestrzeń wektorową $O_{2^j}$. Z Twierdzenia 3. wynika, że operator ten możemy zapisać przy użyciu elementów bazy ortonormalnej przestrzeni $O_{2^j}$ w następujący sposób:

$$P_{O_{2^j}} f(x) = 2^{-j} \sum_{n=-\infty}^{+\infty} \left\langle f(u), \psi_{2^j}(u - 2^{-j}n) \right\rangle \cdot \psi_{2^j}(x - 2^{-j}n). \tag{3.47}$$

$P_{O_{2^j}} f(x)$ wyraża sygnał detalu $f(x)$ na poziomie rozdzielczości $2^j$ (por. Definicja 3.). Sygnał ten jest charakteryzowany przez zbiór iloczynów skalarnych:

$$D_{2^j} f = \left( \left\langle f(u), \psi_{2^j}(u - 2^{-j}n) \right\rangle \right)_{n \in \mathbb{Z}}. \tag{3.48}$$

**Definicja 7** (Dyskretny sygnał detalu)**:** Zbiór iloczynów $D_{2^j} f$ nazywamy *dyskretnym sygnałem detalu* funkcji $f(x)$ na poziomie rozdzielczości $2^j$.

Podobnie jak w 3.24, można pokazać, że każdy z iloczynów skalarnych z 3.48 jest równoważny splotowi funkcji $f(x)$ z $\psi_{2^j}(-x)$, ewaluowanym w punkcie $2^{-j}$:

$$D_{2^j} f = \left( \left( f(u) * \psi_{2^j}(-u) \right) (2^{-j}n) \right)_{n \in \mathbb{Z}}. \tag{3.49}$$

$D_{2^j} f$ koduje różnicę informacji między $A^d_{2^{j-1}} f$ i $A^d_{2^j} f$.

Można pokazać ([32], str. 680.) za pomocą indukcji (wyznaczając kolejno różnice: między $A^d_{2^{j-2}} f$ a $A^d_{2^{j-1}} f$, między $A^d_{2^{j-3}} f$ a $A^d_{2^{j-2}} f$ itd.), że dla dowolnego $J > 0$, aproksymacja dyskretna $A^d_1 f$ sygnału $f(x)$ (np. pomiar dyskretny funkcji $f(x)$, którym dysponujemy) jest charakteryzowana przez następujący zbiór:

$$\left( A^d_{2^{-J}} f, (D_{2^j} f)_{-J \leqslant j \leqslant -1} \right). \tag{3.50}$$

Powyższy zbiór składa się z "bazowego" sygnału dyskretnego $A^d_{2^{-J}} f$, tj. aproksymacji dyskretnej wejściowego sygnału $f(x)$ na najbardziej "zgrubnym" poziomie rozdzielczości $2^{-J}$, oraz dyskretnych sygnałów detalu na kolejnych poziomach rozdzielczości $2^j$, dla $-J \leqslant j \leqslant -1$.

**Definicja 8** (Ortogonalna reprezentacja falkowa)**:** Zbiór 3.50 sygnałów dyskretnych nazywamy *ortogonalną falkową reprezentacją* (ang. *orthogonal wavelet representation*) sygnału $f(x)$.

#### Reprezentacja "teleskopowa" aproksymacji $A_1 f(x)$

Na zbiór 3.50 możemy też patrzyć jako na zbiór iloczynów skalarnych, które występują w następującym – "teleskopowym" – rozwinięciu $A_1 f(x)$:

$$\begin{aligned} A_1 f(x) = {} & 2^{-J} \sum_{n=-\infty}^{+\infty} \left\langle f(u), \phi_{2^J}(u - 2^{-J}n) \right\rangle \phi_{2^J}(x - 2^{-J}n) \\ & + \sum_{j=-J}^{-1} \left( 2^{-j} \sum_{n=-\infty}^{+\infty} \left\langle f(u), \psi_{2^j}(u - 2^{-j}n) \right\rangle \cdot \psi_{2^j}(x - 2^{-j}n) \right). \end{aligned} \tag{3.51}$$

Powyższe rozwinięcie wykorzystuje projekcję ortogonalną sygnału $f(x)$ na przestrzeń wektorową $V_{2^{-J}}$ oraz na przestrzenie wektorowe $O_{2^j}$ dla $-J \leqslant j \leqslant -1$.

**Reprezentacja "teleskopowa" funkcji** $f(x)$

Rozwinięcie "teleskopowe" 3.51 można uogólnić (por. [36], str. 40.) do rozwinięcia "teleskopowego" ciągłej funkcji $f(x)$, przechodząc do nieskończoności z $j$ indeksujący, kolejne przestrzenie wektorowe $0_{2^j}$, na które wykonujemy projekcję ortogonalną sygnału $f(x)$:

$$f(x) = 2^{-J} \sum_{n=-\infty}^{+\infty} \left\langle f(u), \phi_{2^J}(u - 2^{-J}n) \right\rangle \phi_{2^J}(x - 2^{-J}n) + \sum_{j=-J}^{\infty} \left( 2^{-j} \sum_{n=-\infty}^{+\infty} \left\langle f(u), \psi_{2^j}(u - 2^{-j}n) \right\rangle \cdot \psi_{2^j}(x - 2^{-j}n) \right). \quad (3.52)$$

**Uwaga dot. notacji**

Alternatywną do stosowanej dotychczas notacji jest notacja wykorzystującą oznaczenia: $c_{j,k}$, $d_{j,k}$, $\phi_{j,k}$, $\psi_{j,k}$. Oznaczenia te spotykane są w części literatury z bibliografii niniejszej pracy (por. [33], [36]).

Za powyższymi oznaczeniami stoją następujące wyrażenia (por. [36], str. 2.):

$$\phi_{j,k} = 2^{j/2}\phi(2^j x - k), \quad (3.53)$$

$$\psi_{j,k} = 2^{j/2}\psi(2^j x - k), \quad (3.54)$$

$$c_{j,k} = \int_{-\infty}^{\infty} f(x)\phi_{j,k}(x)dx = \langle f, \phi_{j,k} \rangle, \quad (3.55)$$

$$d_{j,k} = \int_{-\infty}^{\infty} f(x)\psi_{j,k}(x)dx = \langle f, \psi_{j,k} \rangle. \quad (3.56)$$

Korzystając z oznaczeń: $c_{j,k}$, $d_{j,k}$, $\phi_{j,k}$, $\psi_{j,k}$, formułę 3.52 reprezentacji "teleskopowej" funkcji f(x) możemy przedstawić następująco:

$$f(x) = \sum_{k\in\mathbb{Z}} c_{J,k}\phi_{J,k}(x) + \sum_{j=J}^{\infty} \sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(x). \quad (3.57)$$

### 3.2.5. Implementacja reprezentacji falkowej sygnału – algorytm piramidalny

W tej subsekcji przedstawiamy za [32], str. 681. algorytm piramidalny do wyznaczania reprezentacji falkowej sygnału $f(x)$. W sposób podobny do tego przedstawionego w części "Implementacja transformacji wieloskalowej – algorytm piramidalny" pokażemy, że $D_{2^j}f$ może być wyznaczony poprzez splot $A^d_{2^{j+1}}f$ z pewnym filtrem dyskretnym $G$.

Dla każdego $n \in \mathbb{Z}$, funkcja $\psi_{2^j}(x - 2^{-j}n)$ jest elementem przestrzeni wektorowej $O_{2^j} \subset V_{2^{j+1}}$. Podobnie jak w 3.25, funkcję tę możemy rozwinąć w bazie ortonormalnej przestrzeni $V_{2^{j+1}}$ w następujący sposób:

$$\psi_{2^j}(x-2^{-j}n) = 2^{-j-1} \sum_{k=-\infty}^{+\infty} \left\langle \psi_{2^j}(u - 2^{-j}n), \phi_{2^{j+1}}(u - 2^{-j-1}k) \right\rangle \cdot \phi_{2^{j+1}}(x-2^{-j-1}k). \tag{3.58}$$

Podobnie jak w 3.26, przeprowadzając zamianę zmiennych w całce w definicji iloczynu skalarnego można pokazać, że:

$$2^{-j-1} \left\langle \psi_{2^j}(u - 2^{-j}n), \phi_{2^{j+1}}(u - 2^{-j-1}k) \right\rangle = \left\langle \psi_{2^{-1}}(u), \phi(u - (k - 2n)) \right\rangle . \tag{3.59}$$

Następnie, wyznaczając iloczyn skalarny funkcji $f(x)$ z obiema stronami równości 3.58, otrzymujemy:

$$\left\langle f(u), \psi_{2^j}(u - 2^{-j}n) \right\rangle = \sum_{k=-\infty}^{+\infty} \left\langle \psi_{2^{-1}}(u), \phi(u - (k - 2n)) \right\rangle \cdot \left\langle f(u), \phi_{2^{j+1}}(u - 2^{-j-1}k) \right\rangle . \tag{3.60}$$

Niech $G$ będzie filtrem dyskretnym o odpowiedzi impulsowej:

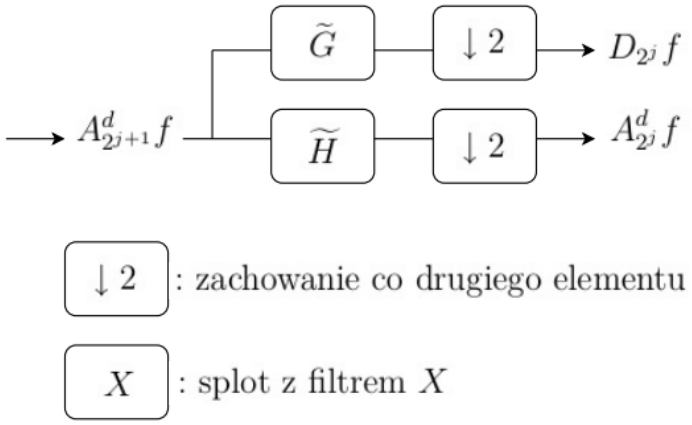
$$g(n) = \left\langle \psi_{2^{-1}}(u), \phi(u - n) \right\rangle \tag{3.61}$$

i niech $\widetilde{G}$ będzie filtrem do niego *symetrycznym*, tj. o odpowiedzi impulsowej postaci $\widetilde{g}(n) = g(-n)$.

Przepisanie 3.60 przy użyciu wyrażenia 3.61 prowadzi do następującej formuły:

$$\left\langle f(u), \psi_{2^j}(u - 2^{-j}n) \right\rangle = \sum_{k=-\infty}^{+\infty} \widetilde{g}(2n - k) \cdot \left\langle f(u), \phi_{2^{j+1}}(u - 2^{-j-1}k) \right\rangle . \tag{3.62}$$

Równanie 3.62 pokazuje, że możemy wyznaczać sygnał detalu $D_{2^j}f$ poprzez splot $A^d_{2^{j+1}}$ z filtrem $\widetilde{G}$ i zachowywanie co drugiego elementu wynikowego ciągu wartości. Ortogonalna reprezentacja falkowa sygnału dyskretnego $A^d_1 f$ (por. Definicja 8.) może być więc wyznaczona poprzez sukcesywną dekompozycję $A^d_{2^{j+1}}f$ do $A^d_{2^j}f$ i $D_{2^j}$ dla $-J \leqslant j \leqslant -1$. Algorytm ten jest zilustrowany na diagramie przedstawionym na Rysunku 3.3.




Rysunek 3.3: Dekompozycja aproksymacji dyskretnej $A^d_{2^{j+1}} f$ do aproksymacji na bardziej "zgrubnym" poziomie rozdzielczości $A^d_{2^j} f$ i sygnału detalu $D_{2^j} f$. Poprzez powtarzanie powyższego kroku dla $-J \leqslant j \leqslant -1$, wyznaczamy reprezentację dyskretnego sygnału wejściowego $A^d_1 f$ na $J$ poziomach rozdzielczości.

Jak wspomniano wcześniej, w praktyce, wejściowy dyskretny sygnał $A^d_1 f$ ma skończoną liczbę elementów. Jeśli dyskretny sygnał wejściowy ma $N$ elementów, to sygnały dyskretne $A^d_{2^j} f$ i $D_{2^j} f$ – na bardziej "zgrubnych" poziomach rozdzielczości – mają po $2^j N$ elementów każdy. Innymi słowy, $2^j N$ elementów bazy z każdej z przestrzeni wektorowych $V_{2^j}$ i $O_{2^j}$ jest potrzebnych, by wyrazić projekcję ortogonalną $f(x)$ na daną przestrzeń. Stąd, reprezentacja falkowa

$$\left(A^d_{2^{-J}} f, (D_{2^j} f)_{-J \leqslant j \leqslant -1}\right)$$

ma taką samą liczbę elementów, jak wejściowy dyskretny sygnał $A^d_1 f$.

W Przykładzie 2. prezentujemy (za [36], str. 18.) schemat dekompozycji dyskretnego sygnału przy użyciu opisanego powyżej algorytmu piramidalnego. W dekompozycji sygnału wykorzystujemy falkę Haara (por. Definicja 6.) jako falkę ortogonalną $\psi(x)$. Przykład wykorzystuje notację opisaną równaniami 3.53 – 3.56.

**Przykład 2**

Załóżmy, że dany jest 8-elementowy sygnał dyskretny $y = (y_1, ..., y_n) = (1, 1, 7, 9, 2, 8, 8, 6)$. Oznaczmy przez $J = 3$ poziom rozdzielczości najbardziej "drobnej" aproksymacji tego sygnału, odpowiadającej w tym przypadku danym wejściowym $y$. $J$ jest takie, że $n = 2^J$, gdzie $n$ jest liczbą elementów $y$.

Zgodnie z opisem algorytmu piramidalnego implementacji reprezentacji falkowej sygnału, aby wyznaczyć aproksymacje sygnału/sygnały detalu na bardziej "zgrubnych" poziomach rozdzielczości ($j = 2, 1, 0$), potrzebujemy znać formułę na splot aproksymacji $c_j$ z filtrem $\widetilde{H}$ (por.

Równanie 3.31) oraz z filtrem $\widetilde{G}$ (por. Równanie 3.62) dla przypadku zastosowania falki Haara.

Oznaczmy za [36], str. 21. $c_{j,k}$ jako $k$-ty element aproksymacji sygnału oraz $d_{j,k}$ jako $k$-ty element sygnału detalu na poziomie rozdzielczości $j$ oraz $c_{j+1,2k}$ jako $2k$-ty element aproksymacji sygnału na poziomie rozdzielczości $j+1$.

Sploty aproksymacji $c_{j+1}$ z filtrem $\widetilde{H}$ oraz z filtrem $\widetilde{G}$ możemy zapisać za [36] w ogólnej postaci jako:

$$c_{j,k} = \sum_{l=-\infty}^{\infty} h_l c_{j+1,2k-l} \tag{3.63}$$

oraz

$$d_{j,k} = \sum_{l=-\infty}^{\infty} g_l c_{j+1,2k-l}, \tag{3.64}$$

odpowiednio.

Za [36], str. 21, podajemy formułę na postać $h_l$ oraz $g_l$ w szczególnym przypadku dekompozycji sygnału przy wykorzystaniu falki Haara:

$$h_l = \begin{cases} 2^{-1/2} & l = 0, \\ 2^{-1/2} & l = 1, \\ 0 & \text{w pozostałych przypadkach,} \end{cases} \tag{3.65}$$

$$g_l = \begin{cases} -2^{-1/2} & l = 0, \\ 2^{-1/2} & l = 1, \\ 0 & \text{w pozostałych przypadkach.} \end{cases} \tag{3.66}$$

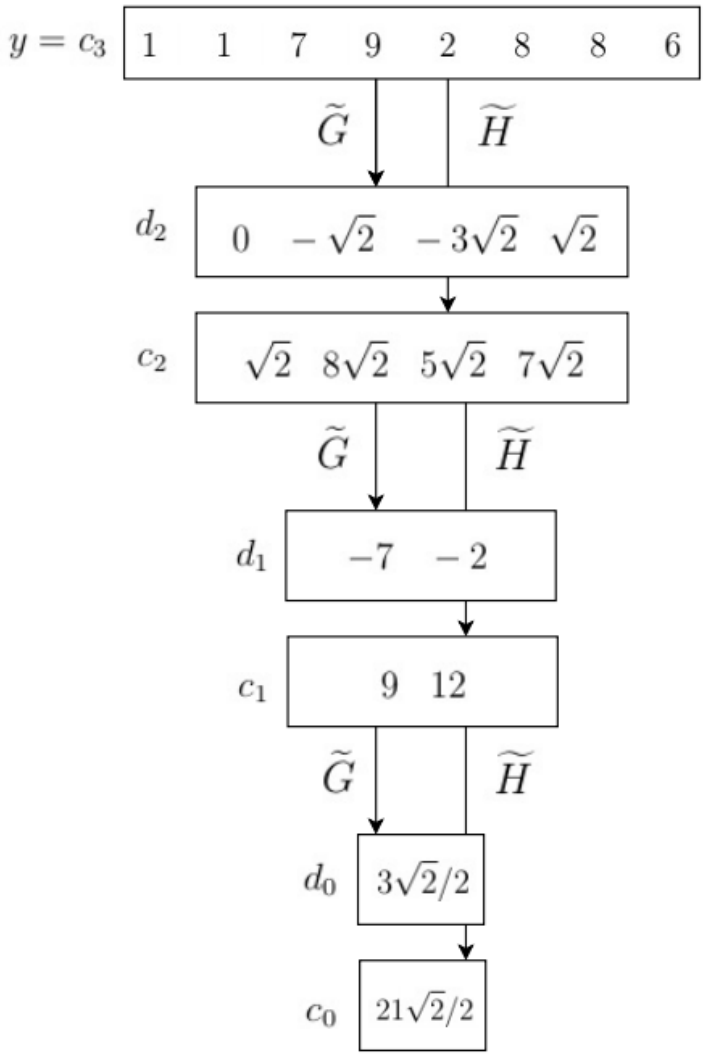
Łatwo widać, że w tym szczególnym przypadku wykorzystaniu falki Haara wyrażenia 3.63 i 3.64 sprowadzają się do postaci:

$$c_{j,k} = (c_{j+1,2k} + c_{j+1,2k-1})/\sqrt{2} \tag{3.67}$$

oraz

$$d_{j,k} = (c_{j+1,2k-1} - c_{j+1,2k})/\sqrt{2}, \tag{3.68}$$

stanowią więc odpowiednio przeskalowane sumy i różnice par elementów z aproksymacji sygnału na poziomie rozdzielczości $j+1$. Wyznaczając $c_{j,k}$ i $d_{j,k}$ dla każdego $k = 1, ..., 2^j$, wykonujemy jeden krok algorytmu piramidalnego, którego schemat przedstawiony jest na Rysunku 3.3. Wyznaczenie wszystkich $c_{j,k}$, $d_{j,k}$ dla $j = 0, ..., J-1$, gdzie $J$ - poziom rozdzielczości sygnału wejściowego $y$, oznacza wykonanie pełnej dekompozycji sygnału wejściowego $y$. Schematyczny obraz pełnej dekompozycji sygnału $y = (1, 1, 7, 9, 2, 8, 8, 6)$ znajduje się na Rysunku 3.4. W poniższej dekompozycji otrzymujemy $(c_0, d_2, d_1, d_0) = (2\sqrt{2}/2,\ 0,\ -\sqrt{2},\ -3\sqrt{2},\ \sqrt{2},\ -7,\ -2,\ 3\sqrt{2}/2)$.




Rysunek 3.4: Schematyczny obraz pełnej dekompozycji sygnału $y = (1, 1, 7, 9, 2, 8, 8, 6)$.

**Definicja 9** (Współczynniki falkowe)**:** Otrzymane w procesie dekompozycji sygnału wejściowego współczynniki:

$$d = (c_0, d_{J-1}, ..., d_0) \tag{3.69}$$

nazywane są *współczynnikami falkowymi*.

## 3.3. Przycinanie falkowe

W niniejszej sekcji przedstawiamy wprowadzenie do konceptu przycinania falkowego. Koncept ten został wporwadzony w artykułach D. L. Donoho ([37] w 1993, [38] w 1995), D. L. Donoho i I. M. Johnstone'a ([39] w 1994, [40] w 1995) oraz D. L. Donoho, I. M. Johnstone'a, G. Kerkyachariana i D. Picarda ([41] w 1995) jako element pomysłu na wykorzystanie reprezentacji falkowej sygnału $f(x)$ w regresji nieparametrycznej.

Pomysł metody identyfikacji punktów zmiany rozkładu proponowany w niniejszej pracy polega na przeprowadzeniu regresji nieparametrycznej z wykorzystaniem reprezentacji falkowej sygnału $f(x)$ w celu wydobycia "szkieletu" struktury sygnału $f(x)$. Indeks/indeksy wartości, w których "szkielet", tj. otrzymana w ten sposób estymacja sygnału zmienia poziom wartości, uznajemy za lokalizacje punktu/punktów zmiany.

### 3.3.1. Reprezentacja macierzowa transformaty falkowej

Do omówienia metody regresji nieparametrycznej z wykorzystaniem reprezentacji falkowej sygnału $f(x)$ potrzebna jest znajomość reprezentacji macierzowej transformaty falkowej. Pojęcie to wprowadzimy, wykorzystując (za [36], str. 25.) kontynuację Przykładu 2.

W Przykładzie 2 przedstawiona jest dekompozycja wektora danych wejściowych $y = (1, 1, 7, 9, 2, 8, 8, 6)$, w wyniku której otrzymujemy współczynniki falkowe (por. Definicja 9.), które możemy zapisać w postaci następującego wektora:

$$d = (c_0, d_2, d_1, d_0) = (2\sqrt{2}/2,\ 0,\ -\sqrt{2},\ -3\sqrt{2},\ \sqrt{2},\ -7,\ -2,\ 3\sqrt{2}/2). \tag{3.70}$$

Z uwagi na fakt, iż współczynniki te zostały wyznaczone z wejściowego wektora danych $y$ w wyniku operacji dodawania, różnicy i/lub skalowania, to jasne jest, że wynikowy wektor współczynników falkowych $d$ może być wyznaczony z $y$ poprzez mnożenie macierzowe. Łatwo sprawdzić, że dla danych z Przykładu 2 następująca macierz $W$:

$$W = \begin{bmatrix} \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 1/2 & 1/2 & -1/2 & -1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 & -1/2 & -1/2 \\ \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & \sqrt{2}/4 & -\sqrt{2}/4 & -\sqrt{2}/4 & -\sqrt{2}/4 & -\sqrt{2}/4 \end{bmatrix} \tag{3.71}$$

jest macierzą, dla której spełnione jest:

$$d = Wy. \tag{3.72}$$

Można poczynić spostrzeżenie (por. [36], str. 26.), że macierz $W$ jest macierzą ortogonalną, tj.:

$$W^T W = W W^T = I. \tag{3.73}$$

W uwagi na ortogonalność macierzy $W$ otrzymujemy:

$$||d||^2 = d^T d = (Wy)T(Wy) = y^T(W^T W)y = y^T y = ||y||^2. \tag{3.74}$$

Innymi słowy, długość wynikowego wektora $d$ jest taka sama, jak długość wejściowego wektora danych $y$ oraz spełniona jest tożsamość Parsewala dla transformacji zadanej macierzą $W$ (por. 3.75).

### 3.3.2. Model regresji

W niniejszej subsekcji konkcetrujemy się na idei *przycinania falkowego* (ang. *wavelet shrinkage*). W dużej ogólności, idea ta przedstawia się następująco: obserwujemy pewną funkcję $y$, o której wiemy, że składa się z funkcji $f$ i addytywego element szumu $\epsilon$. Jesteśmy zainteresowani odzyskaniem $f$. Wykonujemy dekompozycję $y$ do postaci reprezentacji falkowej, a następnie *przycinamy* te współczynniki falkowe, które związane z szumem $\epsilon$ występującym w $y$. Otrzymujemy w ten sposób estymację funkcji $f$.

Poniżej przedstawiamy podstawowy model statystyczny opisanej powyżej procedury.

Zakładamy, że dysponujemy wektorem obserwacji $y = (y_1, ..., y_n)$, które podlegają następującemu modelowi:

$$y_i = f(x_i) + \epsilon_i, \quad \text{dla } i = 1, ..., n, \tag{3.75}$$

gdzie $x_i = i/n$.

Naszym celem jest estymacja nieznanej funkcji $f(x)$ dla $x \in [0, 1]$ na podstawie "zaszumionych" obserwacji $y_i$. W praktyce, często jesteśmy zainteresowani estymacją $f(x)$ dla $x$ należącego do pewnego zbioru punktów $x_i$, $i = 1, ..., m$. W naszym problemie zakładamy dodatkowo, że $\epsilon \sim N(0, \sigma^2)$.

### 3.3.3. Przycinanie falkowe w regresji

Niech $W$ oznacza macierz definiującą transformację dekompozycji $y$ do postaci reprezentacji falkowej $d$ (por. Równanie 3.72.). Niech $y$, $f$ i $\epsilon$ oznaczają wektor danych wejściowych, nieznaną prawdziwą funkcję oraz szum, odpowiednio.

Z uwagi na fakt, że transformacja opisana przez $W$ jest liniowa (por. [36], str. 84.), możemy zapisać model w reprezentacji falkowej jako:

$$d^* = d + \varepsilon, \tag{3.76}$$

gdzie $d^* = Wy$, $d = Wf$ oraz $\varepsilon = W\epsilon$.

Za [36], str. 84. przytaczamy trzy własności 3.76, które są kluczowe w kwestii powodzenia omawianej metody estymacji nieznanej funkcji $f(x)$.

1. W przypadku wielu funkcji $y$ – funkcji gładkich, funkcji gładkich z pewną liczbą skoków wartości etc. – wektor współczynników falkowych $d$ jest wektorem *rzadkim* (ang. *sparse vector*).

2. Z uwagi na fakt zachodzenia równoważności Parsewala w transformacji dekompozycji (por. 3.75), tzw. *energia* zdefiniowana w domenie funkcyjnej jako $\sum_i f(x_i)^2$ jest równa sumie kwadratów współczynników falkowych $\sum_{j,k} d_{j,k}^2$. Jednocześnie, w przypadku zachodzenia rzadkości wektora $d$ (por. własność 1)), energia wejściowego sygnału $f$ jest po transformacji skoncentrowana w "kilku" współczynnikach z $d$.

3. Z uwagi na fakt, że macierz transformacji $W$ jest macierzą ortogonalną, to również $\varepsilon$ – transformacja białego szumu $\epsilon$ – jest białym szumem (wyjaśnienie tego faktu znajduje się w [42]). Stąd szum nie jest skumulowany w wybranych kilku współczynnikach falkowych (jak $f$), ale jest rozproszony "równomiernie" we wszystkich współczynnikach falkowych.

W oparciu o powyższe własności, D. L. Donoho i I. M. Johnstone zaproponowali ([39], 1994) technikę przycinania falkowego jako metodę estymacji $f(x)$. Idea metody polega na tym, że spodziewamy się, że współczynniki falkowe $d^*$ o dużych wartościach są związane z rzeczywistym sygnałem oraz szumem, podczas gdy współczynniki $d^*$ o małych wartościach związane są tylko z szumem. Stąd, aby skutecznie estymować $d$, stosujemy *progowanie* (ang. *thresholidng*) współczynników $d^*$, polegające na usuwaniu tych z nich, które są mniejsze, niż zadany *próg* (ang. *threshold*), i jednoczesnym zachowywaniu pozostałych współczynników. Wynikowy wektor współczynników falkowych oznaczamy jako $\widehat{d}$.

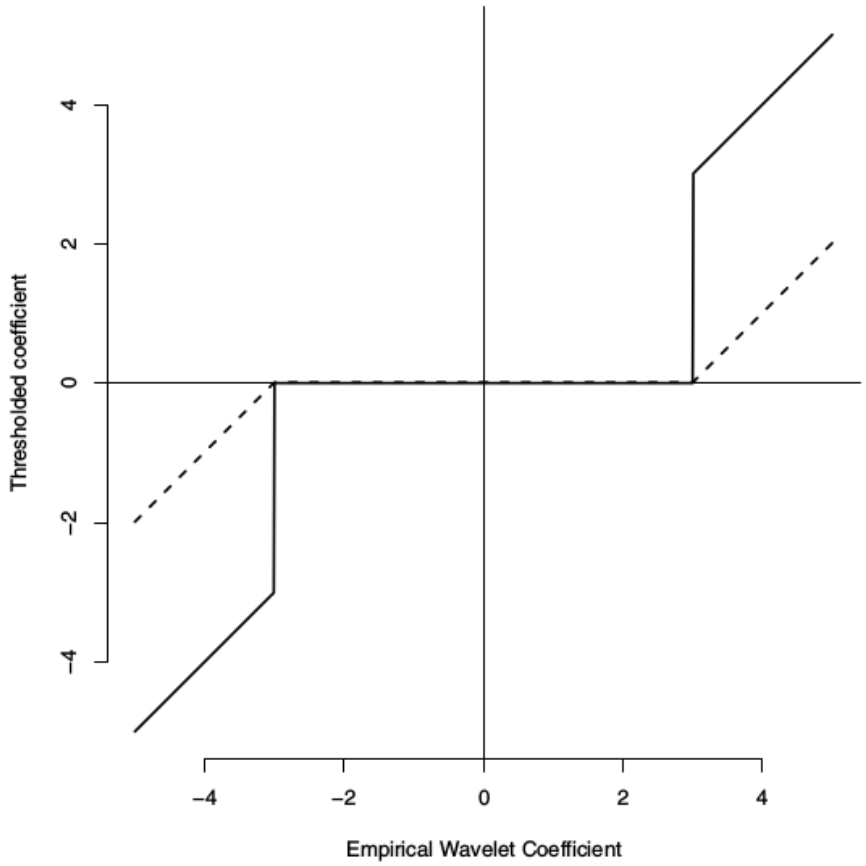
Donoho i Johnstone definiują w [39] funkcje tzw. *łagodnego* i *ostrego* progowania (ang. *soft thresholding*, *hard thresholding*) jako:

$$\widehat{d} = \eta_H(d^*, \lambda) = d^* \mathbb{I}\{|d^*| > \lambda\}, \tag{3.77}$$

$$\widehat{d} = \eta_S(d^*, \lambda) = sgn(d^*)(|d^*| - \lambda)\mathbb{I}\{|d^*| > \lambda\}, \tag{3.78}$$

gdzie $\mathbb{I}$ jest funkcją charakterystyczną zbioru, $d^*$ jest empirycznym wektorem współczynników falkowych, które chcemy progować, a $\lambda$ jest wartością progową. Przykład wykresów zdefiniowanych powyżej funkcji progowania znajduje się na Rysunku 3.5.

Inne metody przycinania współczynników falkowych obejmują *mocne przycinanie* (ang. *firm shrinkage*) H.-Y. Gao i A. G. Bruce'a ([43]) czy metody bayesowskie (por. [36], Rozdział 3.10).




Rysunek 3.5: Wykres funkcji progowania dla progu $\lambda = 3$. Na osi poziomej umieszczone są wartości empirycznych współczynników falkowych (ang. *empirical wavelet coefficients*), na osi pionowej – wartości współczynników po procedurze progowania (ang. *thresholded wavelet coefficients*). Linią ciągłą naniesiona jest funkcja ostrego progowania $\eta_H$ (ang. *hard thresholding*), linią przerywaną – funkcja łagodnego progowania $\eta_S$ (ang. *soft thresholding*).

W dalszej części niniejszej pracy przedstawiamy wyniki stosowania przycinania falkowego z wykorzystaniem funkcji ostrego oraz łagodnego progowania. Eksperymentujemy także podejściem progowania polegającym na zachowaniu $r$ największych (co do wartości bezwzględnej) współczynników falkowych.

Rozdział 4

# Miary poprawności estymacji punktów zmiany rozkładu

W niniejszym rozdziale definiujemy miary, które zastosowaliśmy do oceny dobroci otrzymywanych estymacji punktów zmiany średniej rozkładu w przeprowadzonej analizie symulacyjnej (por. Rozdział 5.). Miary te wykorzystaliśmy do scharakteryzowania i porównania analizowanych metod.

## 4.1. Wstęp

Miary, na których bazujemy, to:

— $FDR$ (ang. *False Discovery Rate*) – frakcja identyfikacji (odkryć) fałszywie dodatnich, wskazanych przez metodę,

— *moc* metody (inaczej: *czułość* metody) – frakcja prawdziwych identyfikacji (odkryć), które zostały wskazane przez metodę,

— $MSE$ (ang. *Mean Squared Error*) – wartość oczekiwana kwadratu błędu estymacji; stosujemy ją do oceny estymacji segmentów średnich rozkładu, na które metoda dzieli dane.

Wymienione powyżej miary należą do podstawowych, powszechnie stosowanych mierników jakości otrzymywanych estymacji.

W niniejszym rozdziale zwracamy uwagę na istotne naszym zdaniem aspekty stosowania $FDR$ oraz czułości w problemie identyfikacji punktów zmiany średniej – twierdzimy, że standardowa procedura użycia tych miar może zostać poprawiona w sposób, który zwiększa interpretowalność otrzymywanych wyników.

W niniejszej pracy proponujemy nowe miary poprawności identyfikacji punktu zmiany – są to miary o roboczych nazwach: $FDR.smooth$ oraz $POWER.smooth$, odpowiednio. Proponowane miary pozwalają na "wygładzoną" ("uciągloną") ocenę poprawności identyfikacji punktu/punktów zmiany. Jest to podejście różne od stosowanego przy wyznaczaniu $FDR$ czy mocy w sposób standardowy, gdzie dana identyfikacja podlega ocenie binarnej (stwierdza się, że jest lub nie jest prawdziwym odkryciem).

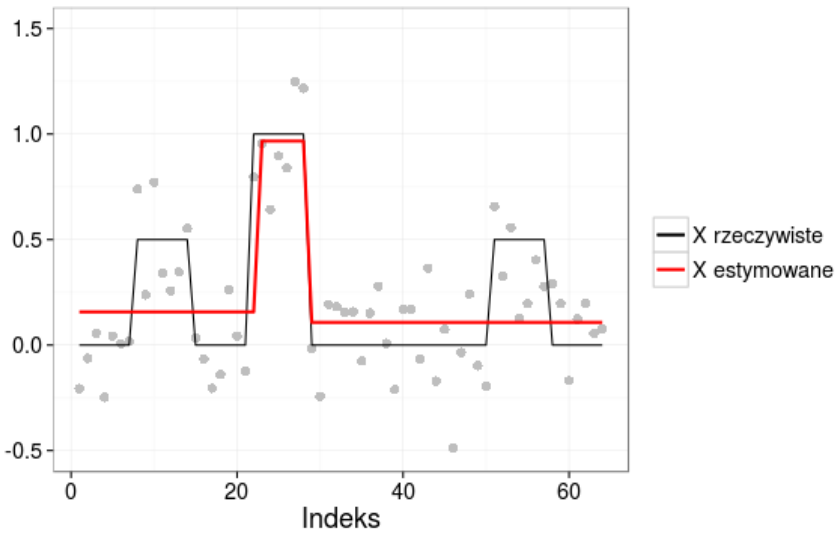
Miary $FDR.smooth$ oraz $POWER.smooth$ zostały zaimplementowane w środowisku R na potrzeby niniejszej pracy. Z uwagi na dostrzeżone interesujące własności tych miar oraz niestwierdzenie istnienia metod implementujących takie rozwiązanie w środowisku R, autorka niniejszej pracy planuje w niedalekiej przyszłości obudowanie ich w formę pakietu celem udostępnienia społeczności użytkowników R.

## 4.2. FDR i moc metody – podejście "klasyczne"

W niniejszej sekcji prezentujemy standardową procedurę wyznaczania wartości FDR i mocy (czułości) metody w problemie identyfikacji punktów zmiany średniej rozkładu oraz zwracamy uwagę na aspekty interpretacji otrzymywanych wartości, które były motywacją do zaproponowania przez nas ulepszeń podejścia klasyczego.

### 4.2.1. Procedura

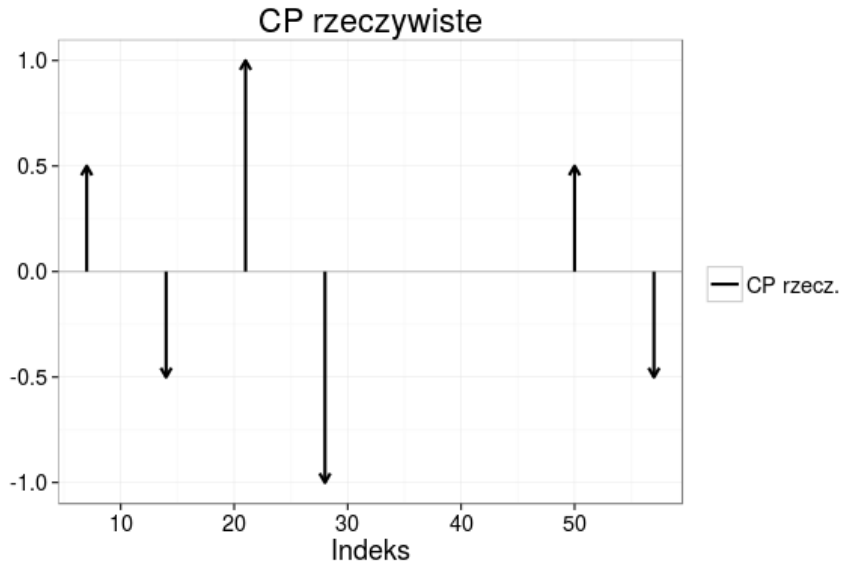
Kroki standardowej procedury wyznaczania wartości FDR i mocy metody w problemie identyfikacji punktów zmiany opisujemy na przykładowych danych, przedstawionych na Rysunku 4.1. Zakładamy, że dane są: wartości rzeczywiste średnich rozkładu (czarna linia ciągła), dane zaszumione (szare punkty) oraz pewna estymacja segmentów średnich, otrzymana na podstawie zaszumionych danych (czewona ciągła linia). Zarówno teraz jak i w dalszej części niniejszej pracy, zakładamy, że zmienne losowe, których realizacje obserwujemy, pochodzą z rozkładu normalnego o stałej wariancji.




Rysunek 4.1: Przykładowe dane: wartości rzeczywiste średnich rozkładu (czarna linia ciągła), dane zaszumione (szare punkty) i pewna estymacja segmentów średnich rozkładu, otrzymana na podstawie zaszumionych danych (czerwona ciągła linia).

**Krok 1.**

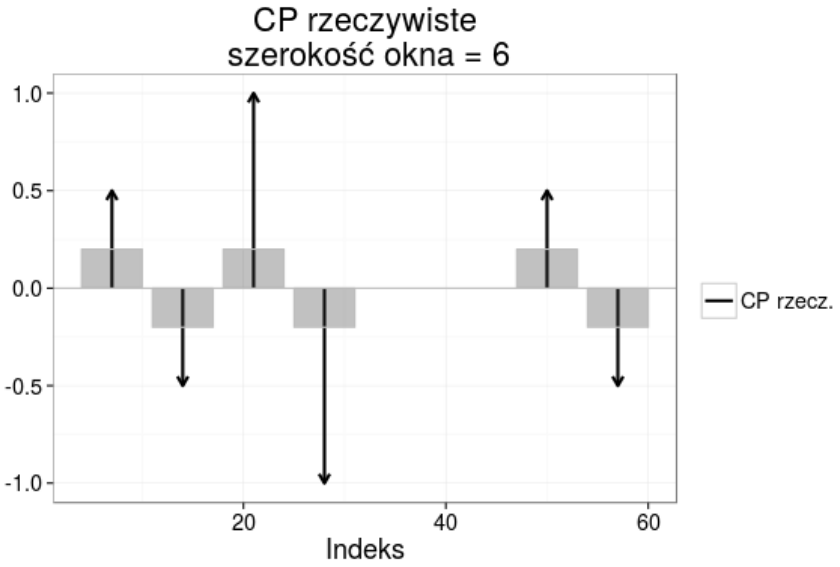
W kroku 1. klasycznej procedury wyznaczania wartości $FDR$ i mocy identyfikujemy rzeczywiste punkty zmiany średniej. W analizowanym przykładzie punkty te zlokalizowane są w miejscach o indeksach: 7, 14, 21, 28, 50 i 57. Punkty zmiany 7, 21, i 50 związane są ze wzrostem średniej rozkładu zmiennych, punkty 14, 28 i 57 – z jej spadkiem. Rzeczywiste punkty zmiany są naniesione w sposób symboliczny na wykresie zamieszczonym na Rysunku 4.2.; w schemacie tym uwzględniono wartości zmiany (por. długość strzałek) oraz kierunek zmiany średniej (por. zwrot strzałek). Akronim "CP" zastosowany w tytule wykresu jest skrótem od wyrażenia *change point(s).*




Rysunek 4.2: Identyfikacja rzeczywistych punktów zmiany średniej rozkładu. Wartości zmiany są wyrażone przez długość strzałek, kierunek zmiany – przez zwrot strzałek.

**Krok 2.**

W kroku 2. klasycznej procedury wyznaczania $FDR$ i mocy ustalamy szerokość okna, które definiuje nam obszar akceptacji estymowanego punktu zmiany jako prawdziwego odkrycia. Okna te są rozmieszczone symetrycznie na lewo i na prawo od każdego z rzeczywistych punktów zmiany, co w sposób symboliczny przedstawione jest na Rysunku 4.3. Szerokość okna jest dobierana arbitralnie w zależności od specyfiki danego problemu; heurystyczny wybór szerokości okna może być motywowany wiedzą dotyczącą rozkładu estymacji punktów zmiany w danym problemie.




Rysunek 4.3: Okna określające obszar akceptacji estymowanego punktu zmiany jako prawdziwego odkrycia; zastosowano okna o szerokości równej 6.

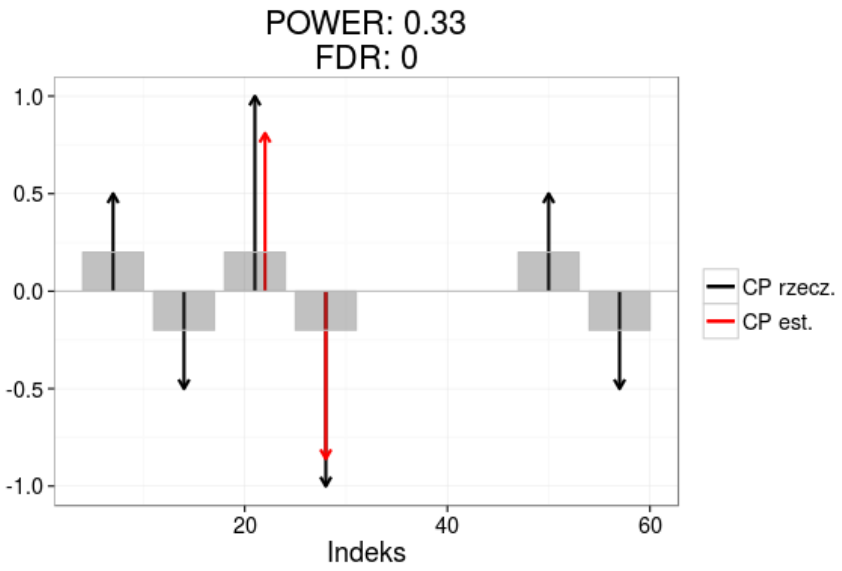
**Krok 3.**

W kroku 3. klasycznej procedury wyznaczania $FDR$ i mocy metody sprawdzamy, w jaki sposób zwrócone przez daną metodę estymacje punktów zmiany korespondują z rzeczywistymi punktami zmiany. Zwracamy uwagę na fakt, czy otrzymane estymacje znalazły się w obszarze okien akceptacji wyznaczonych w kroku 2. i czy ich kierunek jest zgodny z kierunkiem rzeczywistego punktu zmiany – jeśli tak, to stwierdzamy, że dokonana przez metodę identyfikacja jest prawdziwym odkryciem oraz że korespondujący z danym odkryciem rzeczywisty punkt zmiany został zidentyfikowany. Następnie wyznaczamy wartości FDR i mocy metody zgodnie ze standardową formułą:

$$FDR = \left( \frac{\#\text{ fałszywych odkryć CP}}{\#\text{ wszystkich odkryć CP}} \right), \tag{4.1}$$

$$POWER = \left( \frac{\#\text{ prawdziwych odkryć CP}}{\#\text{ rzeczywistych CP}} \right). \tag{4.2}$$

Otrzymane w analizowanym przykładzie estymacje punktów zmiany są naniesione w sposób symboliczny (czerwona strzałka) na wykresie zamieszczonym na Rysunku 4.3. Dla tej estymacji wartości FDR oraz mocy metody wynoszą 0 oraz 0.33, odpowiednio.




Rysunek 4.4: Schematyczne przedstawienie estymowanych punktów zmiany średniej (czerwone strzałki) oraz wyznaczenie wartości FDR oraz mocy metody.

### 4.2.2. Wybrane aspekty interpretacji

Analizując przedstawiony przykład, można zwrócić uwagę na pewne aspekty interpretacji wyników otrzymywanych w opisanej powyżej standardowej procedurze wyznaczania FDR i mocy metody.

— Standardowa procedura nie uwzględnia *stopnia odległości* między rzeczywistym a estymowanym punktem zmiany, poza faktem znalezienia się estymowanego punktu w odległości nie większej, niż określona przez okno akceptacji (por. krok 2. procedury). Przykładowo, za tak samo "dobrą" estymację uznany jest punkt zmiany, którego estymowana lokalizacja jest *zgodna* z lokalizacją rzeczywistego punktu zmiany, jak ten, którego estymowana lokalizacja znajduje się np. na krawędzi okna akceptacji.

— Przedstawiona powyżej procedura nie uwzględnia stosunku wielkości skoku/spadku wartości rzeczywistych średnich do wielkości skoku/spadku estymowanych średnich rozkładu (por. różnice w długościach korespondujących czerwonych i czarnych strzałek na Rysunku 4.4.).

Powyższe aspekty są przez nas postrzegane jako obszary, które można poprawić, aby zwiększyć interpretowalność otrzymywanych wyników. Procedura, którą w tym celu proponujemy, przedstawiona jest w następnej sekcji niniejszego rozdziału.

## 4.3. *FDR.smooth* i *POWER.smooth*

Proponowana przez nas procedura wyznaczania FDR oraz mocy metod służących do identyfikacji punktów zmiany średniej rozkładu wykorzystuje ideę splotu funkcji gęstości rozkładu normalnego $N(0, \sigma^2)$ z funkcją delta Diraca zlokalizowaną w punkcie zmiany. Koncept zastosowania takiego splotu przedstawiony jest w prezentacji "Towards a Mathematical Theory of Super-resolution" z 2013 r. ([44]) jako metoda używana w problemie estymacji "szpiców" (ang. *spikes*) w przetwarzaniu sygnałów wielowymiarowych.

### 4.3.1. Procedura

Kroki proponowanej procedury wyznaczania wartości FDR i mocy opisujemy na tych samych przykładowych danych, tóre wykorzystywaliśmy w poprzedniej sekcji (por. wykres na Rysunku 4.1.).

**Krok 1.**

W kroku 1. proponowanej procedury postępujemy jak poprzednio – identyfikujemy rzeczywiste punkty zmiany (por. wykres na Rysunku 4.2.).

**Krok 2.**

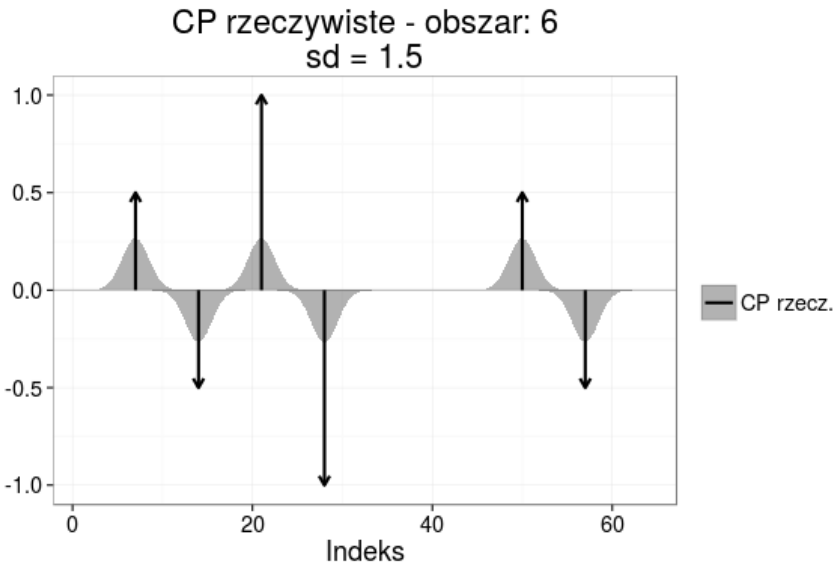
W kroku 2. proponowanej procedury wyznaczamy splot funkcji gęstości rozkładu normalnego $N(0, \sigma^2)$ z funkcją delta Diraca zlokalizowaną w punkcie zmiany, dla każdego rzeczywistego punktu zmiany. Odchylenie standardowe $\sigma$ rozkładu normalnego jest – podobnie jak szerokość okna w procedurze klasycznej – dobrane arbitralnie w zależności od specyfiki analizy (tu: zakładamy wartość parametru $\sigma = 1.5$). Na wykresie na Rysunku 4.5 przedstawione są czarne strzałki reprezentujące rzeczywiste punkty zmiany rozkładu oraz zaznaczony jest (kolorem szarym) obszar pod/nad wykresem funkcji splotu (w zależności od kierunku zmiany średniej w danym punkcie).

Suma pól oznaczonych kolorem szarym obszarów pod/nad wykresem funkcji splotu (nad/pod osią OX, odpowiednio) wynosi 6.

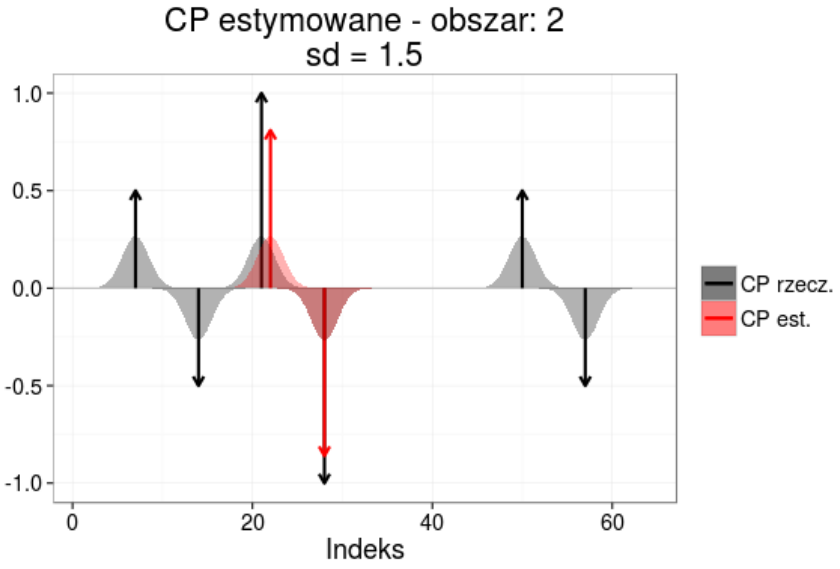
**Krok 3.**

W kroku 3. proponowanej procedury wyznaczamy splot funkcji gęstości rozkładu normalnego $N(0, \sigma^2)$ z funkcją delta Diraca zlokalizowaną w punkcie zmiany, dla każdego estymowanego punktu zmiany. Stosujemy tę samą wartość odchylenia standardowego $\sigma$ rozkładu normalnego co w poprzednim kroku (tu: $\sigma = 1.5$). Na wykresie na Rysunku 4.6 naniesione są czerwone strzałki reprezentujące estymowane punkty zmiany rozkładu oraz odpowiadający im (oznaczony kolorem czerwonym) obszar pod/nad wykresem funkcji splotu (w zależności od kierunku zmiany średniej w danym punkcie).

Pole oznaczonego kolorem czerwonym obszaru pod/nad wykresem funkcji splotu (nad/pod osią OX, odpowiednio) wynosi 2.




Rysunek 4.5: Strzałki reprezentujące rzeczywiste punkty zmiany rozkładu oraz (oznaczony kolorem szarym) obszar reprezentujący pole pod/nad wykresem funkcji splotu rozkładu normalnego $N(0, (1.5)^2)$ z funkcją delta Diraca zlokalizowaną w rzeczywistym punkcie zmiany.




Rysunek 4.6: Strzałki reprezentujące estymowane punkty zmiany rozkładu oraz (oznaczony kolorem czerwonym) obszar reprezentujący pole pod/nad wykresem funkcji splotu rozkładu normalnego $N(0, (1.5)^2)$ z funkcją delta Diraca zlokalizowaną w każdym z estymowanych punktów zmiany.

**Krok 4.**

W 4. kroku proponowanej procedury wyznaczamy pole części wspólnej obszaru oznaczonego kolorem szarym i obszaru oznaczonego kolorem czerwonym. Następnie wykorzystujemy otrzymaną wartość do wyznaczenia wartości miar *FDR.smooth* oraz *POWER.smooth*, zgodnie z następującą formułą:

$$FDR.smooth = \left(1 - \frac{\text{pole części wspólnej}}{\text{pole części związanej z estymowanymi CP (obszar szary)}}\right), \tag{4.3}$$

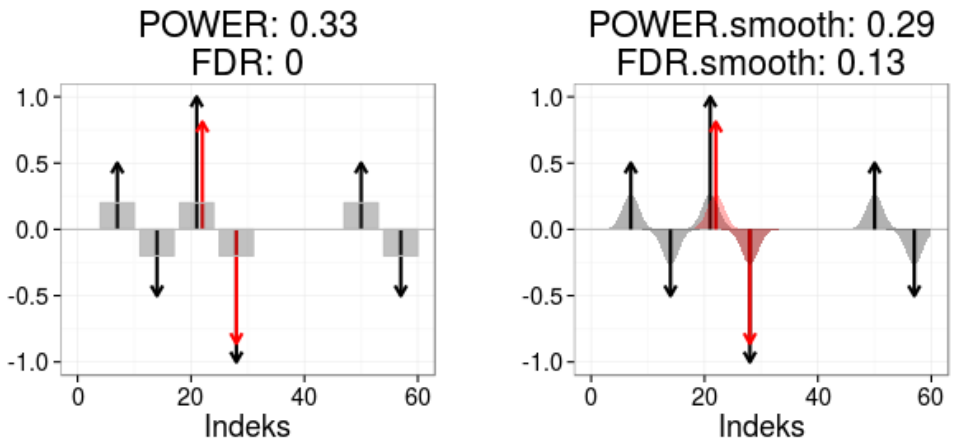
$$POWER.smooth = \left(\frac{\text{pole części wspólnej}}{\text{pole części związanej z rzeczywistymi CP (obszar czerwony)}}\right). \tag{4.4}$$

W analizowanym przez nas przykładzie, pole części wspólnej obszaru oznaczonego kolorem szarym i obszaru oznaczonego kolorem czerwonym wynosi 1.74. Otrzymujemy więc następujące wartości *FDR.smooth* i *POWER.smooth*:

$$FDR.smooth = 1 - \frac{1.74}{2} = 0.13,$$
$$POWER.smooth = \frac{1.74}{6} = 0.29.$$

### 4.3.2. *FDR.smooth* i *POWER.smooth* – Wybrane aspekty interpretacji

Wby uwypuklić różnicę interpretacyjną między podejściem standardowym a podejściem przez nas proponowanym, przyjrzyjmy się zestawieniu wykresów przedstawiających wyniki końcowe obu procedur, które umieszczone zostało na Rysunku 4.7.




Rysunek 4.7: Wartości *POWER* i *FDR* oraz *POWER.smooth* i *FDR.smooth*.

Analiza wykresów zamieszczonych na Rysunku 4.7 pozwala na następujące spostrzeżenia:

— Wartości $POWER$ i $POWER.smooth$ oraz wartości $FDR$ i $FDR.smooth$ różnią się. Różnica ta wynika z faktu, iż pierwszy z estymowanych punktów zmiany ma swoją lokalizację przesuniętą w stosunku do korespondującego rzeczywistego punktu zmiany.

— Metoda przez nas proponowana uwzględnia niedokładność w *estymacji lokalizacji* pierwszego z estymowanych punktów zmiany – niedokładność ta ma przełożenie w mniejszej wartości oceny mocy metody ($POWER.smooth$) oraz większej wartości oceny frakcji odkryć fałszywie dodatnich ($FDR.smooth$) w porównaniu z metodą klasyczną.

— Spodziewamy się, że zmiana przyjętego odchylenia standardowego $\sigma$ rozkładu normalnego, którego gęstość była użyta w wyznaczaniu splotu w metodzie przez nas proponowanej, będzie miała wpływ na wartości $POWER.smooth$ oraz $FDR.smooth$ – zwiększenie wartości $\sigma$ oznacza "rozluźnienie" kary za niedokładność w *estymacji lokalizacji* estymowanych punktów zmiany i będzie powodować wzrost wartości $POWER.smooth$ oraz spadek wartości $FDR.smooth$; i na odwrót – zmniejszenie wartości $\sigma$ oznacza "zaostrzenie" kary za niedokładność w *estymacji lokalizacji* estymowanych punktów zmiany i będzie powodować spadek wartości $POWER.smooth$ oraz wzrost wartości $FDR.smooth$.

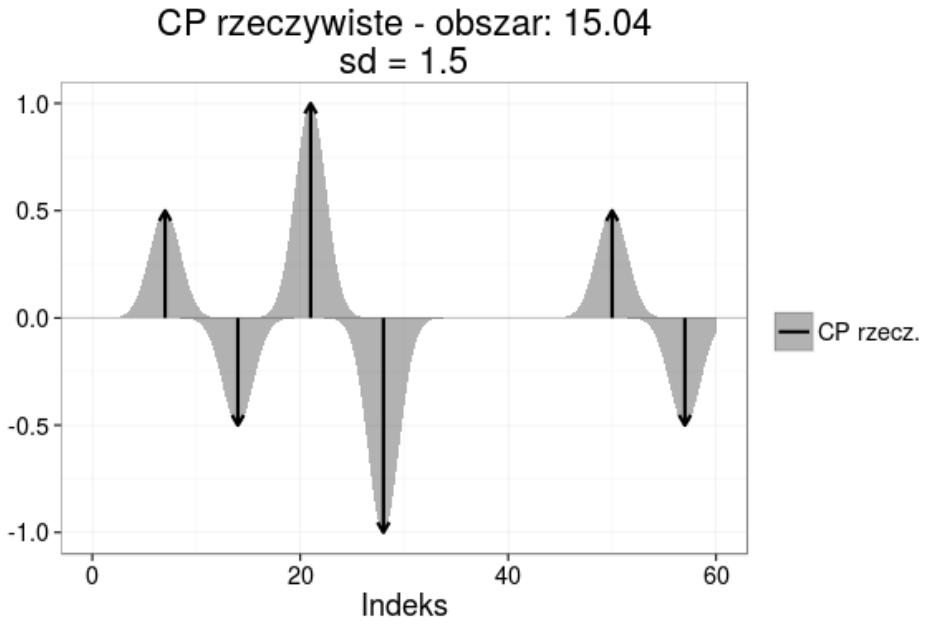
  Wykresy przedstawiające przykłady wpływu doboru wartości $\sigma$ na wartości $POWER.smooth$ oraz $FDR.smooth$ przedstawione są w dalszej części niniejszego rozdziału.

### 4.3.3. $FDR.smooth$ i $POWER.smooth$ – wersja skalowana

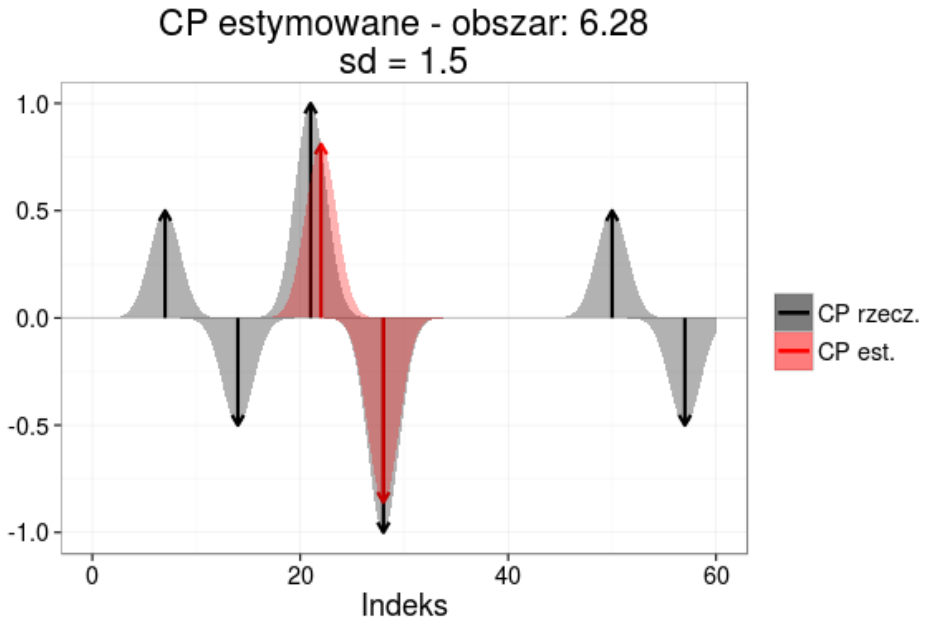
Przedstawiona powyżej wersja proponowanej przez nas procedury adresuje kwestię braku uwzględniania *stopnia odległości* między rzeczywistym a estymowanym punktem zmiany średniej rozkładu w procedurze klasycznej. Jednocześnie, nie uwzględnia ona kwestii stosunku wielkości skoku/spadku wartości rzeczywistych średnich do estymowanych średnich. Odpowiedzią na tę drugą kwestię jest modyfikacja miar $POWER.smooth$ i $FDR.smooth$, polegająca na skalowaniu wartości wynikowego splotu funkcji gęstości rozkładu normalnego i funkcji delta Diraca tak, aby wysokość splotu w jego najwyższym punkcie równa była wysokości strzałki reprezentującej dany punkt zmiany.

Modyfikacja ta dotyczy zarówno splotów związanych z rzeczywistymi, jak i estymowanymi punktami zmiany, tj. kroku 2. i kroku 3. proponowanej przez nas procedury. Pozostałe kroki algorytmu wyznaczania $POWER.smooth$ i $FDR.smooth$, w szczególności – idea wyznaczania wartości miary oparta na wykorzystaniu pól powierzchni odpowiednich obszarów, wyznaczonych przez funkcję splotu – pozostają bez zmian.

Na Rysunku 4.8. i na Rysunku 4.9 przedstawione są wykresy obrazujące zmiany wprowadzane w kroku 2. i kroku 3. proponowanej przez nas procedury. Niezmieniona pozostaje wartość odchylenia standardowego $\sigma$ rozkładu normalnego, który splatany jest z funkcją delta Diraca (wynosi ona $\sigma = 1.5$).




Rysunek 4.8: Krok 2. w procedurze wyznaczania *POWER.smooth* i *FDR.smooth* w wersji skalowanej.



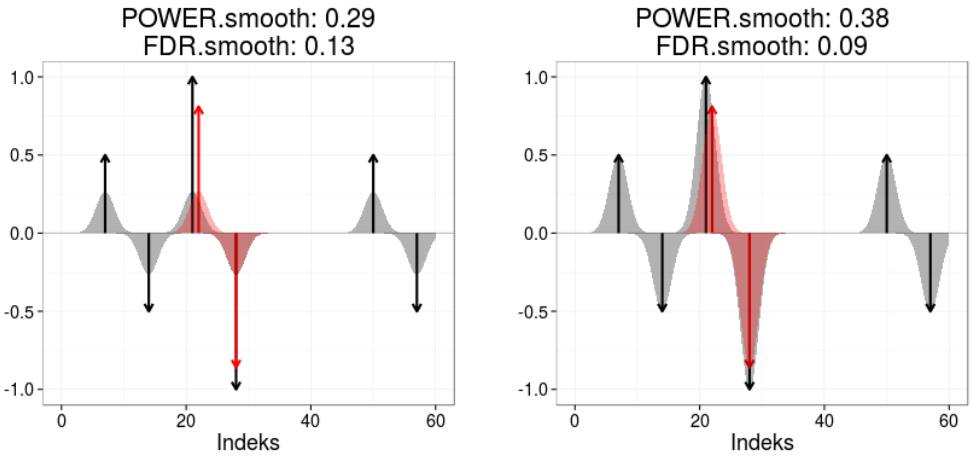

Rysunek 4.9: Krok 3. w procedurze wyznaczania *POWER.smooth* i *FDR.smooth* w wersji skalowanej.

W analizowanym przez nas przykładzie (wersja skalowana procedury wyznaczania $POWER.smooth$ i $FDR.smooth$), pole części wspólnej obszaru oznaczonego kolorem szarym i obszaru oznaczonego kolorem czerwonym wynosi 5.7. Otrzymujemy więc następujące wartości miar $FDR.smooth$ i $POWER.smooth$:

$$FDR.smooth = 1 - \frac{5.7}{6.28} = 0.09,$$
$$POWER.smooth = \frac{5.7}{15.04} = 0.38.$$

### 4.3.4. $FDR.smooth$ i $POWER.smooth$ – wersja skalowana – wybrane aspekty interpretacji

Wby uwypuklić różnicę interpretacyjną między proponowanym podejściem w wersji nieskalowanej i w wersji skalowanej, przyjrzyjmy się zestawieniu wykresów przedstawiających wyniki końcowe obu procedur, które umieszczone zostało na Rysunku 4.10.




Rysunek 4.10: Wartości $POWER.smooth$ i $FDR.smooth$, otrzymane w proponowanej procedurze w wersji nieskalowanej (wykres po lewej stronie) i w wersji skalowanej (wykres po prawej stronie).

Analiza wykresów zamieszczonych na Rysunku 4.10 pozwala na następujące spostrzeżenia:

— Wartości $POWER.smooth$ i $FDR.smooth$ w wersji nieskalowanej oraz w wersji skalowanej różnią się. Różnica w wartości $POWER.smooth$ wynika z faktu, iż w wersji skalowanej bardziej "doceniamy" identyfikację punktów zmiany, z którymi związane są relatywnie duże zmiany wartości średnich rozkładu. W analizowanym przykładzie metoda estymująca wskazała dwa punkty zmiany korespondujące z rzeczywistymi punktami, z którymi związana jest relatywnie duża zmiana wartości średnich, stąd wartość $POWER.smooth$ jest istotnie większa w przypadku procedury

wykorzystującej wersję skalowaną (wynosi 0.38, podczas gdy w wersji nieskalowanej wynosi 0.29).

— Innym istotnym spostrzeżeniem jest mniejsza wartość $FDR.smooth$ w wersji skalowanej w porównaniu z wersją nieskalowaną. Oceniamy, że w tym przypadku dzieje się tak dlatego, że skalowanie wynikowych splotów dla punktów rzeczywistych i estymowanych tak, że oddalamy "czubki" obydwu splotów od osi OX, powoduje, że zwiększa się powierzchnia części wspólnej obszarów ograniczonych tymi splotami, skutkiem czego wartość $FDR.smooth$ wzrasta.

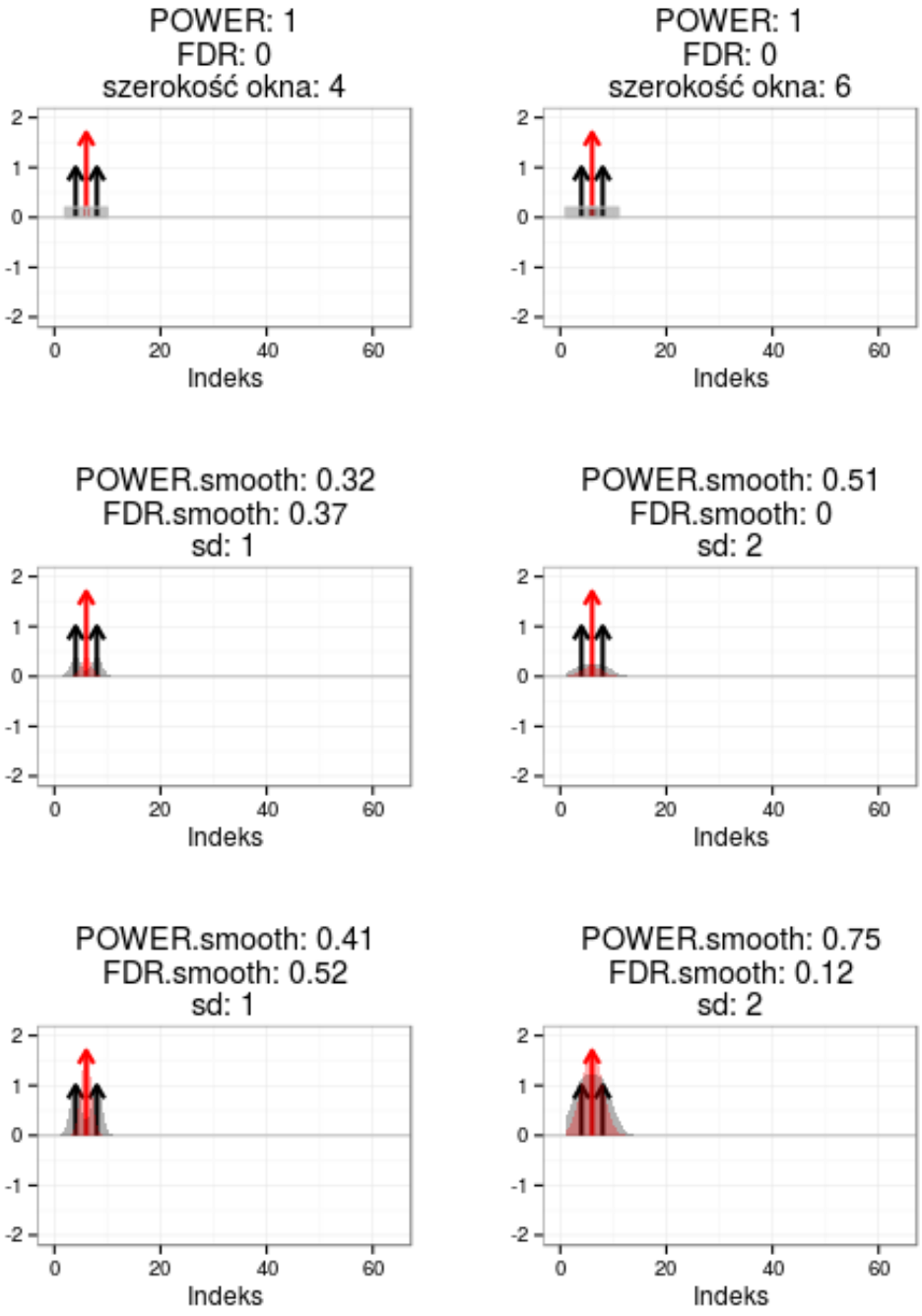
## 4.4. Przykłady

Opisując procedurę wyznaczania miar $FDR$ i $POWER$ oraz $FDR.smooth$ i $POWER.smooth$ (wersja nieskalowana i skalowana), bazowaliśmy na jednym zestawie przykładowych danych – wartościach rzeczywistych średnich rozkładu i przykładowej estymacji, przedstawionych na Rysunku 4.1. Dane te zostały dobrane tak, aby móc pokazać na nich charakterystyczne własności każdego z prezentowanych podejść.

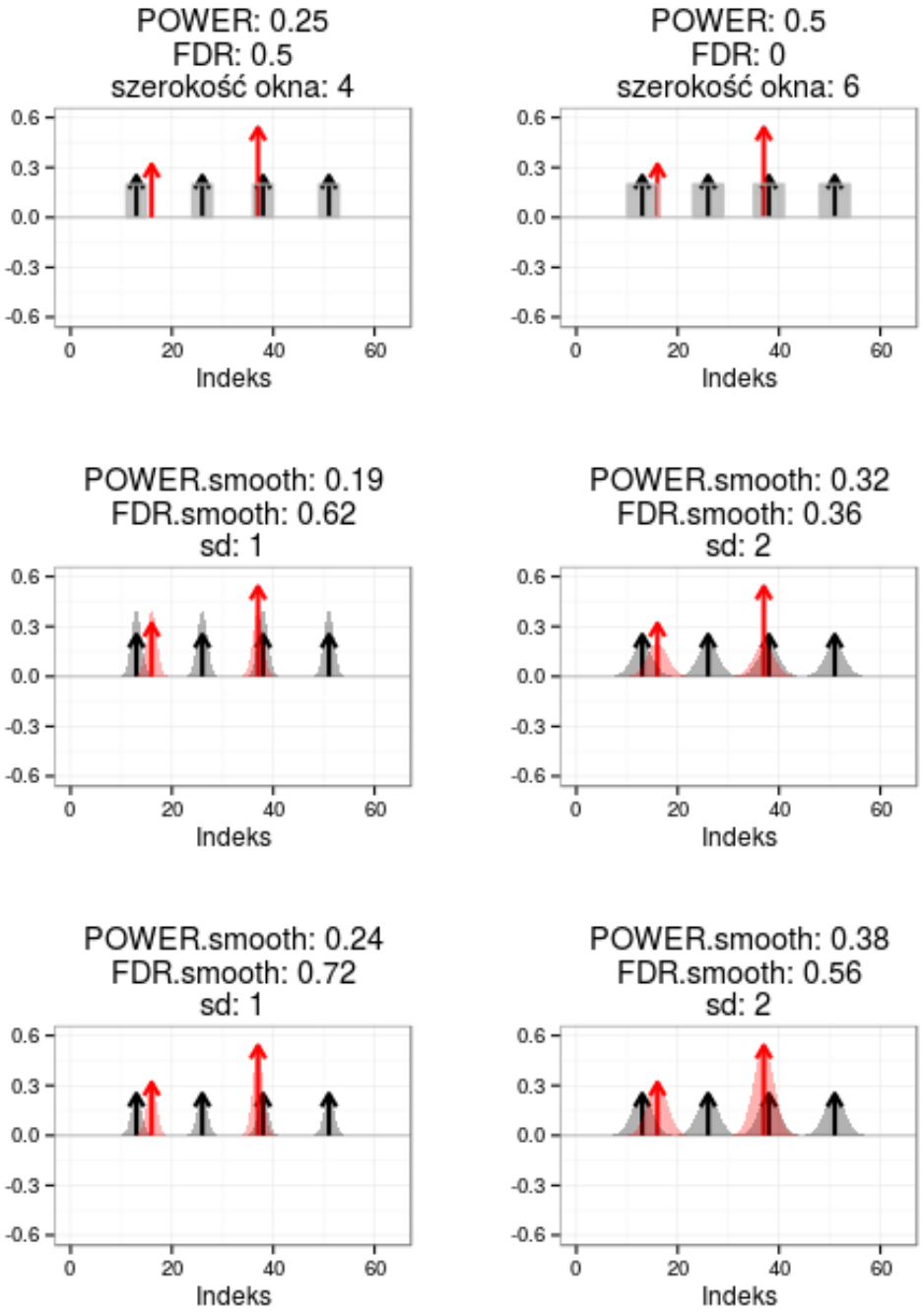
Autorka niniejszej pracy jest świadoma, iż zainteresowany Czytelnik mógłby chcieć przeanalizować więcej przykładów, aby móc zbudować sobie pewne intuicje dot. omawianych miar oceny poprawności estymacji punktów zmiany. W związku z tym, w niniejszej sekcji zamieszczone zostały wykresy z innymi przykładami, które uwzględniają:

— 3 różne zestawy przykładowych danych,
— 2 różne szerokości okna (procedura klasyczna) / wartości odchylenia standardowego $\sigma$ (procedura przez nas proponowana).

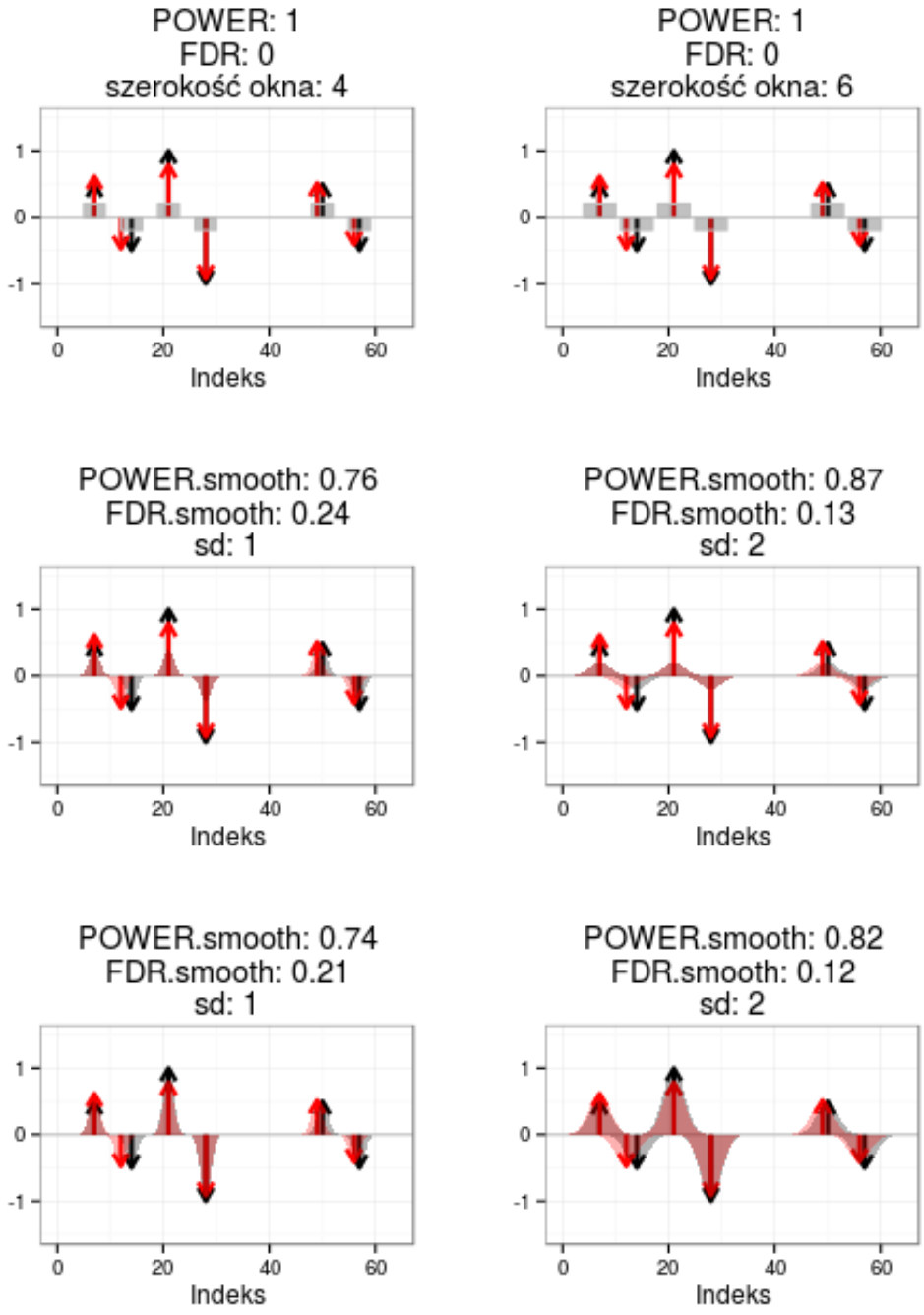
Wynikom dla każdego z 3 zestawów przykładowych danych poświęcamy po jednej stronie niniejszej pracy. Na każdej stronie umieszczamy 3 rzędy wykresów, odpowiadające procedurze klasycznej oraz procedurze przez nas proponowanej w wersji nieskalowanej i skalowanej, odpowiednio. Finalnie, każdy wiersz składa się z 2 wykresów, które różnią się tym, jaką szerokość okna / wartość odchylenia standardowego $\sigma$ zastosowano.




Rysunek 4.11: Przykładowy scenariusz nr 1.




Rysunek 4.12: Przykładowy scenariusz nr 2.




Rysunek 4.13: Przykładowy scenariusz nr 3.

## Rozdział 5

# Analiza symulacyjna

W niniejszym rozdziale przedstawiamy wyniki przeprowadzonej analizy symulacyjnej. Analiza ta podzielona jest na dwie główne części.

Pierwsza część koncentruje się na badaniu własności metod referencyjnych identyfikacji punktów zmiany rozkładu. Analizę rozpoczęliśmy od porównania kształtów trajektorii estymacji otrzymywanych dla obu metod referencyjnych, dla trzech wybranych scenariuszy symulacyjnych oraz przy założeniu różnego poziomu zaszumienia danych, na podstawie których wykonywana jest estymacja. W kroku drugim tej części analizy przeprowadziliśmy symulacje Monte Carlo celem sprawdzenia, jakie średnie wartości miar poprawności identyfikacji (por. Rozdział 4.) otrzymujemy. Krok drugi ma z założenia dwa główne cele – pierwszym z nich jest scharakteryzowanie metod referencyjnych identyfikacji punktów zmiany, drugim – porównanie otrzymywanych wyników w zależności od tego, jaką miarę poprawności identyfikacji stosujemy.

W drugiej części analizy symulacyjnej przenosimy naszą uwagę na metodę identyfikacji wykorzystującą reprezentację falkową w dekompozycji wieloskalowej obiektu (por. Rozdział 3.). Wykorzystujemy popularny w literaturze scenariusz symulacyjny – tzw. funkcję blokową – do porównania wyników otrzymywanych dla różnych metod progowania współczynników falkowych. Rezultaty dla tej metody porównujemy z rezultatami otrzymanymi dla metod referencyjnych.

## 5.1. Część pierwsza – charakteryzacja metod referencyjnych identyfikacji punktów zmiany

W niniejszej sekcji prezentujemy wyniki analiz porównawczych, przeprowadzonych celem charakteryzacji metod referencyjnych identyfikacji punktów zmiany (por. Rozdział 2.).

**Stosunek sygnału do szumu**

W badaniach symulacyjnych korzystamy z trzech scenariuszy symulacyjnych; dodatkowo, zakładamy różne poziomy zaszumienia danych, na podstawie których dokonywana jest estymacja segmentów średnich rozkładu. Poziom zaszumenia danych określamy przez współczynnik stosunku sygnału do szumu (ang. *signal to noise ratio (SNR)*). Współczynnik ten definiujemy jako

$$SNR = \frac{ssig}{sigma}, \tag{5.1}$$

gdzie *ssig* oznacza próbkowe odchylenie standardowe w wektorze $x$ rzeczywistych wartości segmentów średnich w danym problemie symulacyjnym, a *sigma* jest wartością odchylenia standardowego w rozkładzie $N(0, \sigma^2)$, z którego symulujemy wektor szumu $\varepsilon$. Estymacja segmentów średnich rozkładu (estymacja punktów zmiany średniej rozkładu) odbywa się na podstawie wektora danych $y$:

$$y = x + \varepsilon. \tag{5.2}$$

### 5.1.1. Porównanie kształtów trajektorii estymacji segmentów średnich rozkładu

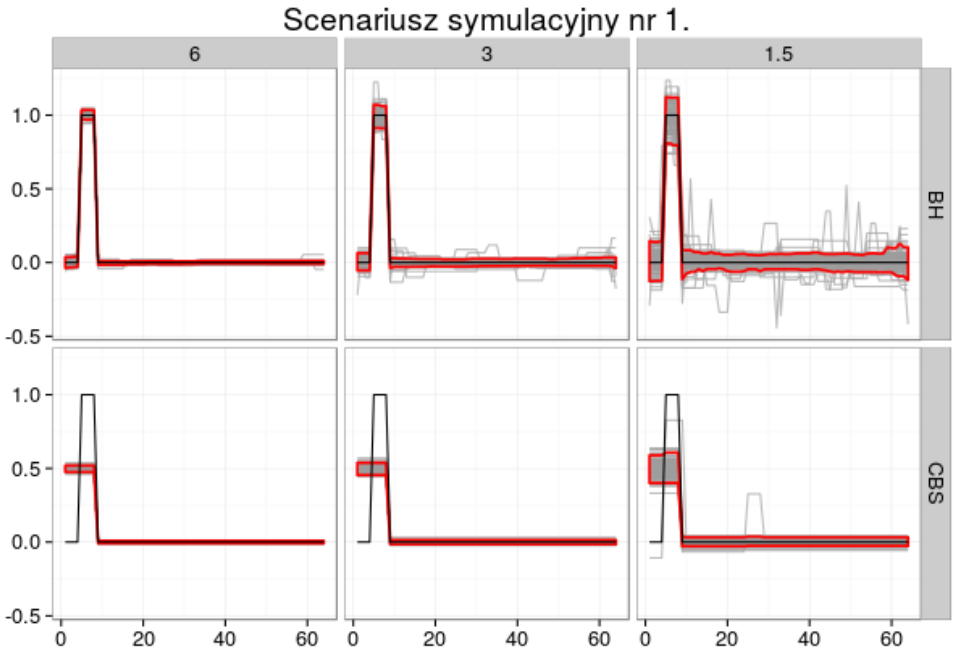
Na Rysunkach 5.1., 5.2. i 5.3. przedstawione są wyniki symulacji po 100 trajektorii estymacji segmentów średnich rozkładu dla:

— 3 scenariuszy symulacyjnych, gdzie każdemu scenariuszowi odpowiada osobny Rysunek; wektor $x$ rzeczywistych wartości segmentów średnich w danym scenariuszu symulacyjnym jest naniesiony czarną pogrubioną linią ciągłą,

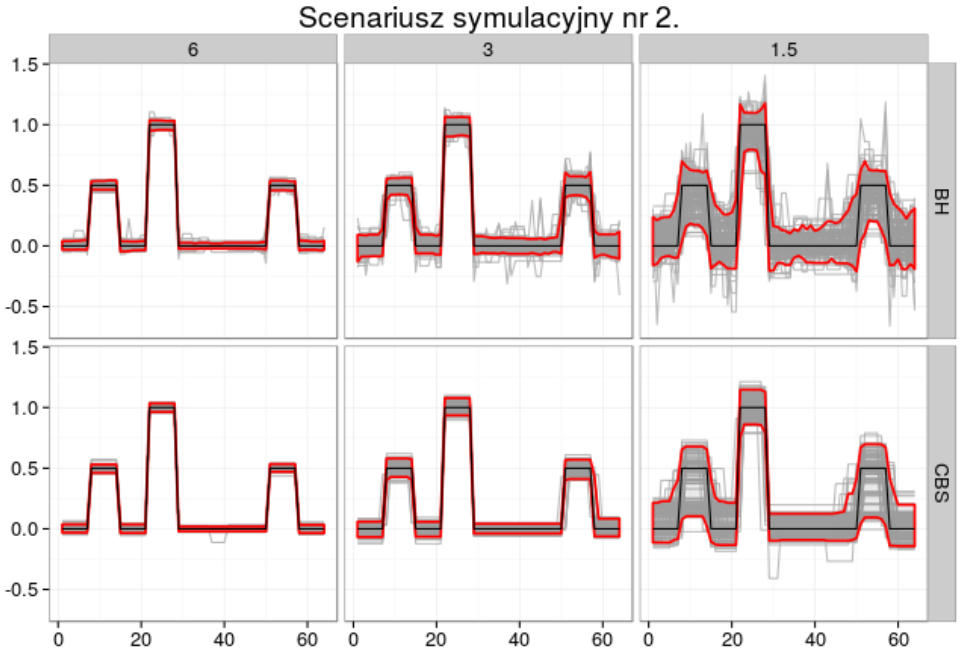
— 2 metod referencyjnych estymacji segmentów średnich rozkładu (estymacji punktów zmiany średniej rozkładu); na każdym z przedstawionych Rysunków danej metodzie przyporządkowany jest jeden z dwóch wierszy wykresów (por. sygnatury metod umieszczone na bocznym panelu po prawej stronie na każdym z Rysunków),

— 3 poziomów zaszumienia danych; poziom zaszumienia danych określony jest przez wartość współczynnika SNR (por. wartości umieszczone na górnym panelu na każdym z Rysunków).

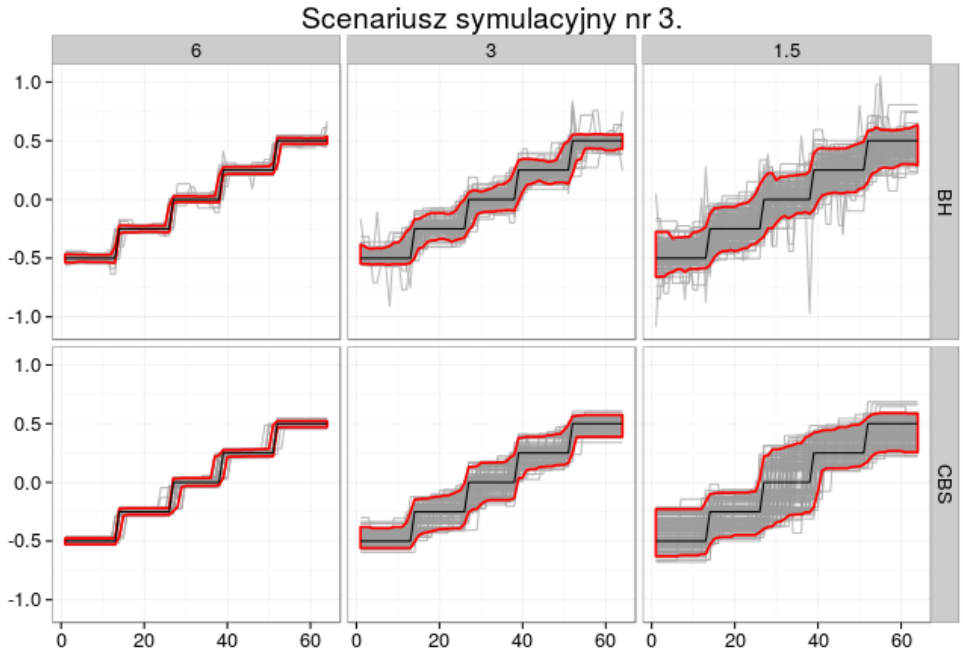
Otrzymane trajektorie naniesione są szarymi ciągłymi liniami. Dwiema czerwonymi liniami ciągłymi ograniczony jest pas, który każdorazowo pokrywa 90% ze 100 otrzymanych trajektorii.




Rysunek 5.1: Trajektorie w scenariuszu symulacyjnym nr 1.




Rysunek 5.2: Trajektorie w scenariuszu symulacyjnym nr 2.




Rysunek 5.3: Trajektorie w scenariuszu symulacyjnym nr 3.

Analiza wykresów zamieszczonych na Rysunkach 5.1., 5.2. i 5.3. pozwala poczynić następujące spostrzeżenia:

— Widoczna jest duża różnica w estymacjach segmentów średnich zwracanych przez obie metody w scenariuszu symulacyjnym nr 1. Metoda wykorzystująca podejście bayesowkie (ozn. $BH$) poprawnie identyfikuje "schodek" wartości nawet dla stosunkowo mocno zaszumionych danych ($SNR = 1.5$); jednocześnie można zauważyć, że trajektorie dla tej metody są nieregularne (widzimy wiele skoków). Metoda wykorzystująca algorytm Circular Binary Segmentation (ozn. $CBS$) w większości przypadków identyfikuje jedynie jeden punkt zmiany, nawet dla relatywnie mało zaszumionych danych ($SNR = 6.0$). Pierwsze intuicje są więc takie, że spodziewamy się, że dla metody $BH$ będziemy otrzymywać większe niż dla $CBS$ wartości $FDR$ i zarazem większe niż dla $CBS$ wartości miary mocy (czułości) metody.

— W przypadku scenariusza symulacyjnego nr 2., porównywane metody podobnie "dobrze" radzą sobie z identyfikacją rzeczywistych punktów zmiany wartości średnich rozkładu. Można poczynić uwagę, iż – podobnie jak w scenariuszu symulacyjnym nr 1. – metoda $BH$ zwraca o wiele bardziej nieregularne trajektorie, niż metoda $CBS$.

— W przypadku scenariusza symulacyjnego nr 3., wraz ze wzrostem poziomu zaszumienia danych obserwujemy szybki wzrost szerokości pasm pokrywających 90% trajektorii. Na podstawie inspekcji wizualnej wykresów możemy stwierdzić, że w przypadku tego scenariusza pasma te są węż-

sze dla metody $BH$ niż dla metody $CBS$ (odmiennie w porównaniu ze scenariuszami symulacyjnymi nr 1. i 2.).

### 5.1.2. Porównanie średnich wartości miar poprawności identyfikacji

W tej części niniejszej pracy prezentujemy szereg wykresów przedstawiających wyniki średnich wartości miar poprawności identyfikacji, otrzymanych dla każdego z badanych scenariuszy symulacyjnych i dla różnego poziomu zaszumienia danych.

Ze względu na mnogość otrzymanych wyników, wykresy przedstawiamy i komentujemy osobno dla każdego scenariusza symulacyjnego (za wyjątkiem komentarza dot. miary MSE, który jest wspólny dla każdego ze scenariuszy, por. końcowa część niniejszej subsekcji).

Zamieszczone poniżej Rysunki z wykresami można przedstawić w sposób następujący:

— na każdym z Rysunków przedstawione są dwa rzędy wykresów, korespondujące z wynikami dla jednej z dwóch metod referencyjnych identyfikacji punktów zmiany,

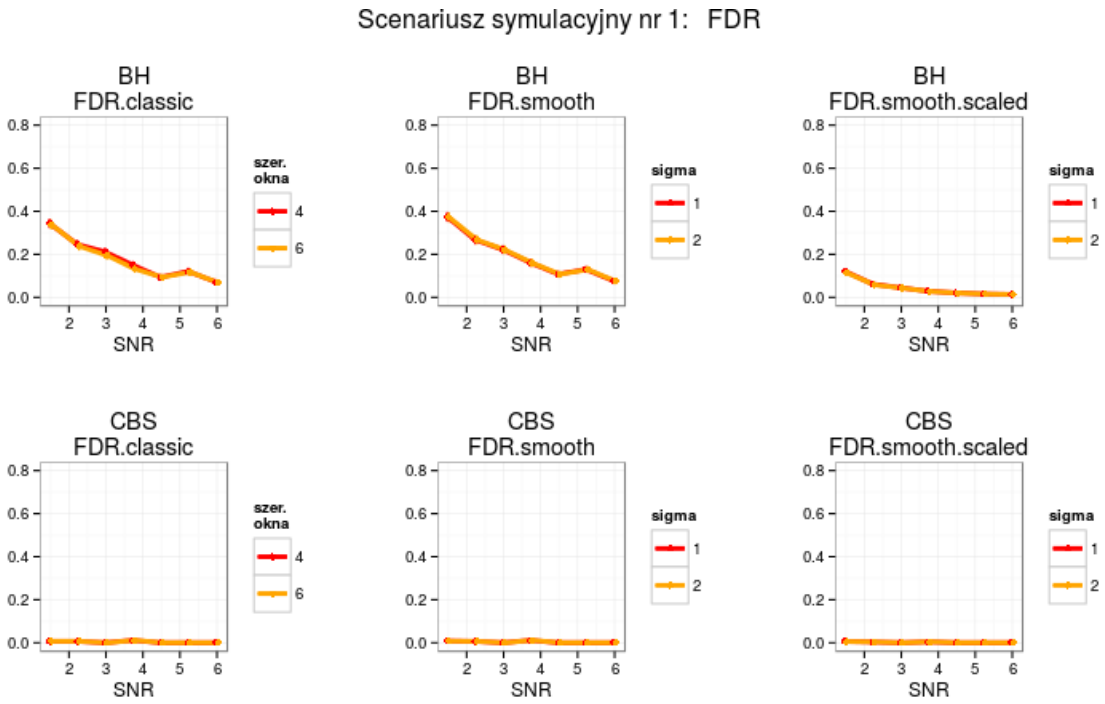
— na osi OX oznaczone są wartości współczynnika SNR zaszumienia danych,

— na osi OY oznaczone są wartości analizowanej miary; linie ciągłe na wykresach odpowiadają wartościom uśrednionym ze 100 symulacji Monte Carlo; dla każdego przypadku symulacyjnego widzimy także naniesione "odcinki" odchodzące od ciągłej linii wykresu, symbolizujące wartość średnią +/- próbkowe odchylenie standardowe dla wartości otrzymanych ze 100 symulacji,

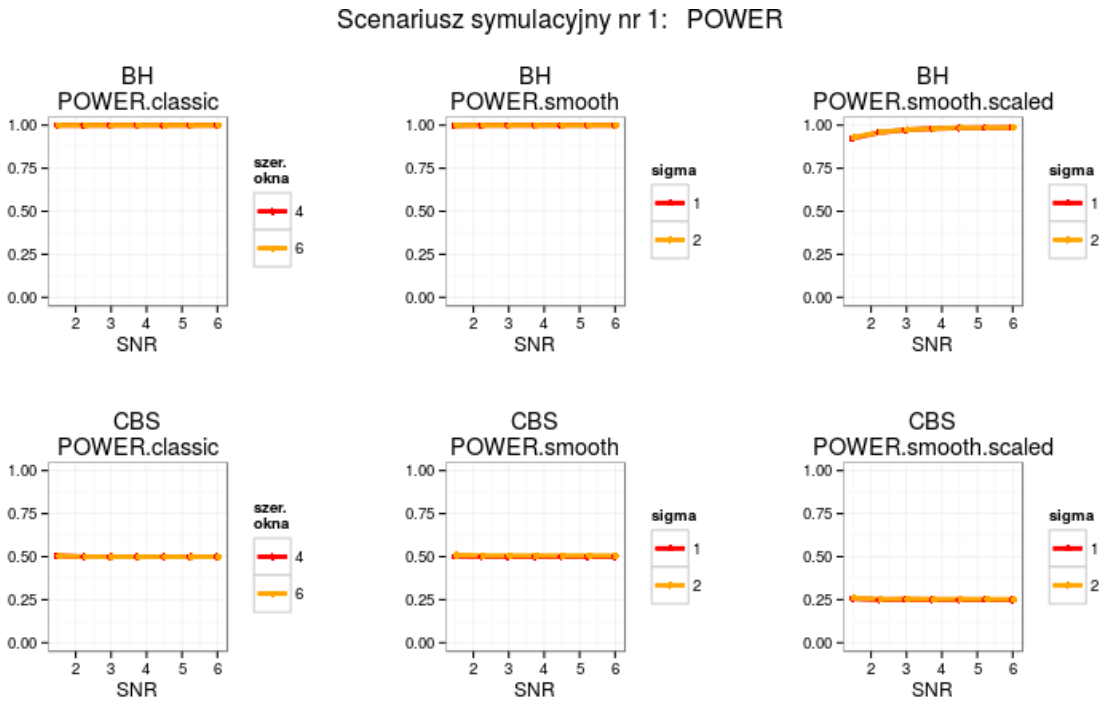
— każdy z Rysunków podzielony jest na 3 kolumny, odpowiadające 3 różnym podejściom w wyznaczaniu danej miary (podeście klasyczne, podejście proponowane przez nas – w dwóch wersjach, nieskalowanej i skalowanej),

— na każdym z wykresów przedstawione są wartości miary wyznaczone w procedurach, różniących się wartością parametru szerokości okna (podejście klasyczne) / $\sigma$ (podejście przez nas proponowane).

**Scenariusz symulacyjny nr 1.**




Rysunek 5.4: Uśrednione wartości miary FDR w scenariuszu symulacyjnym nr 1.



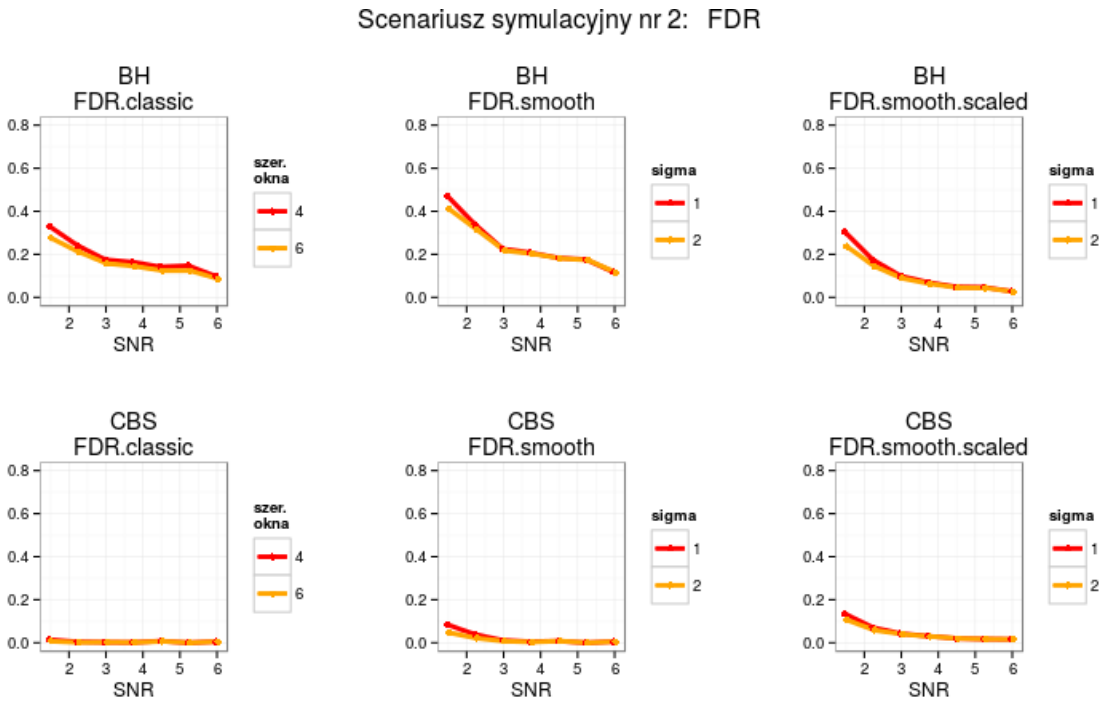

Rysunek 5.5: Uśrednione wartości miary mocy w scenariuszu symulacyjnym nr 1.

Analiza wykresów zamieszczonych na Rysunku 5.4 i Rysunku 5.5 pozwala poczynić następujące spostrzeżenia:

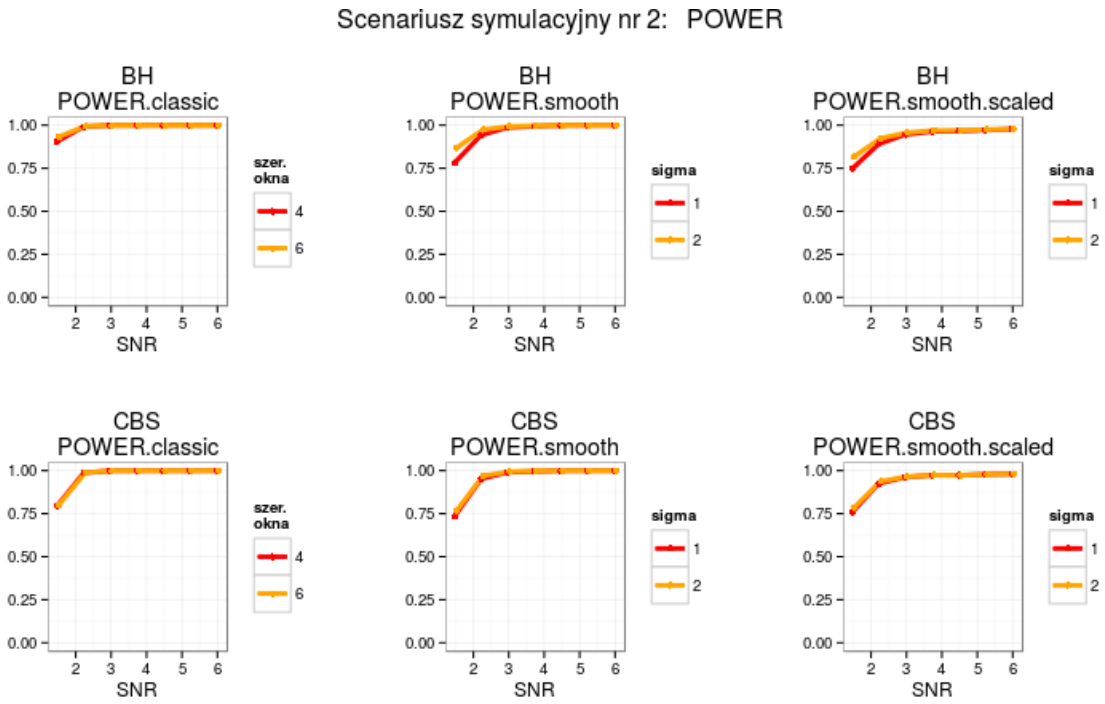
— Zgodnie z sygnalizowanymi wcześniej oczekiwaniami, w scenariuszu symulacyjnym nr 1. widoczne są znaczne różnice między $FDR$ otrzymywanym dla metody $BH$ a $FDR$ otrzymywanym dla metody $CBS$ – niezależnie od typu podejścia stosowanego w wyznaczaniu FDR, jego wartości są bliskie 0 dla metody $CBS$ i relatywnie wysokie dla metody $BH$. Obserwujemy szczególnie wysokie wartości $FDR$ dla metody $BH$ w przypadku estymacji w oparciu o mocno zaszumione dane (por. niskie wartości współczynnika SNR). Można przypuszczać, że w przypadku mocno zaszumionych danych liczba fałszywych odkryć jest na tyle duża, że zatracają się subtelne różnice między wartościami otrzymywanymi w podejściach klasycznym i przez nas proponowanym (wersja nieskalowana); sądzimy, że z tego samego powodu (duża liczba identyfikacji wskazanych przez metodę $BH$) nie dostrzegamy istotnej różnicy w wynikach w zależności od wyboru parametrów szerokości okna / $\sigma$.

— Na wykresach przedstawionych na Rysunku 5.5 widać wyraźnie, że metoda $BH$ w przypadku tego scenariusza symulacyjnego nie ma problemów z identyfikacją rzeczywistych punktów zmiany; ewentualne niewielkie rozminięcia w estymacji lokalizacji tych punktów są wyrażone przez niewielkie spadki wartości na wykresie $FDR.smooth$ w wersji skalowanej. Metoda $CBS$ nie identyfikuje poprawnie dwóch punktów zmiany w tym scenariuszu, czego wyrazem są relatywnie niskie wartości mocy metody.

**Scenariusz symulacyjny nr 2.**




Rysunek 5.6: Uśrednione wartości miary FDR w scenariuszu symulacyjnym nr 2.




Rysunek 5.7: Uśrednione wartości miary mocy w scenariuszu symulacyjnym nr 2.

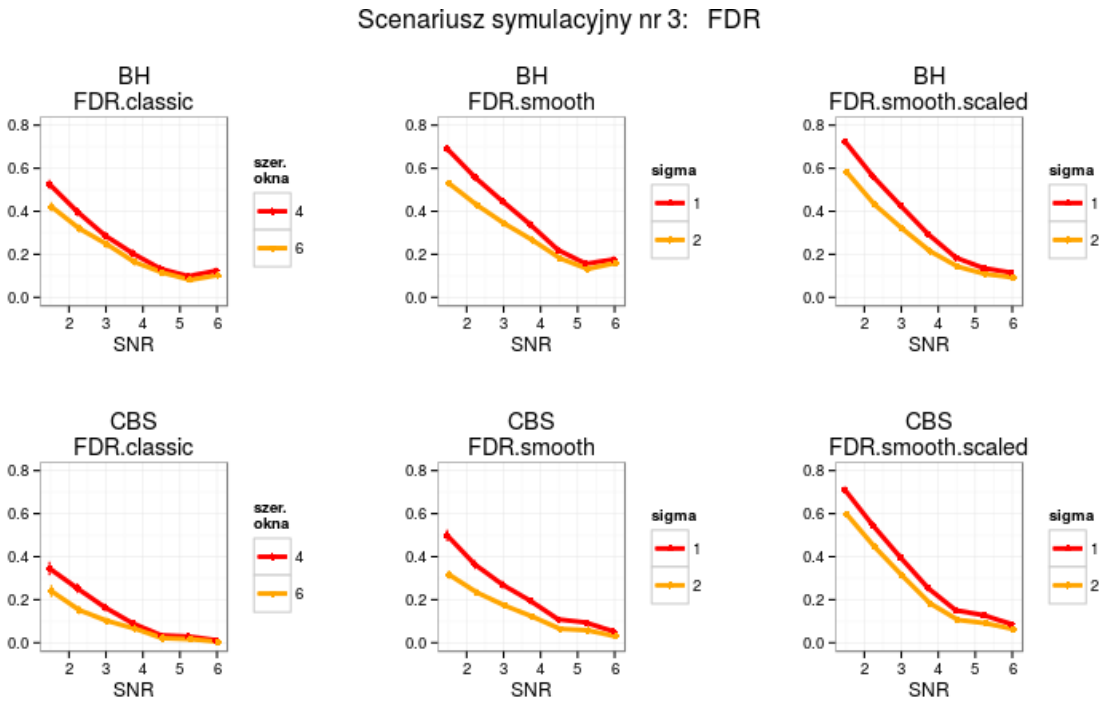
Analiza wykresów zamieszczonych na Rysunku 5.6 i Rysunku 5.7 pozwala poczynić następujące spostrzeżenia:

— Zarówno metoda $CBS$ jak i metoda $BH$ nie mają problemów z identyfikacją wszystkich rzeczywistych punktów zmiany występujących w tym scenariuszu symulacyjnym, czego wyrazem są wysokie (bliskie 1) wartości miar mocy metod (za wyjątkiem przypadków mocno zaszumionych danych). Co więcej, obie metody zwracają stosunkowo dokładne estymacje lokalizacji wszystkich rzeczywistych punktów zmiany, dzięki czemu wykresy wartości dla podejść $POWER$, $POWER.smooth$ oraz $POWER.smooth.scaled$ są do siebie bardzo zbliżone.

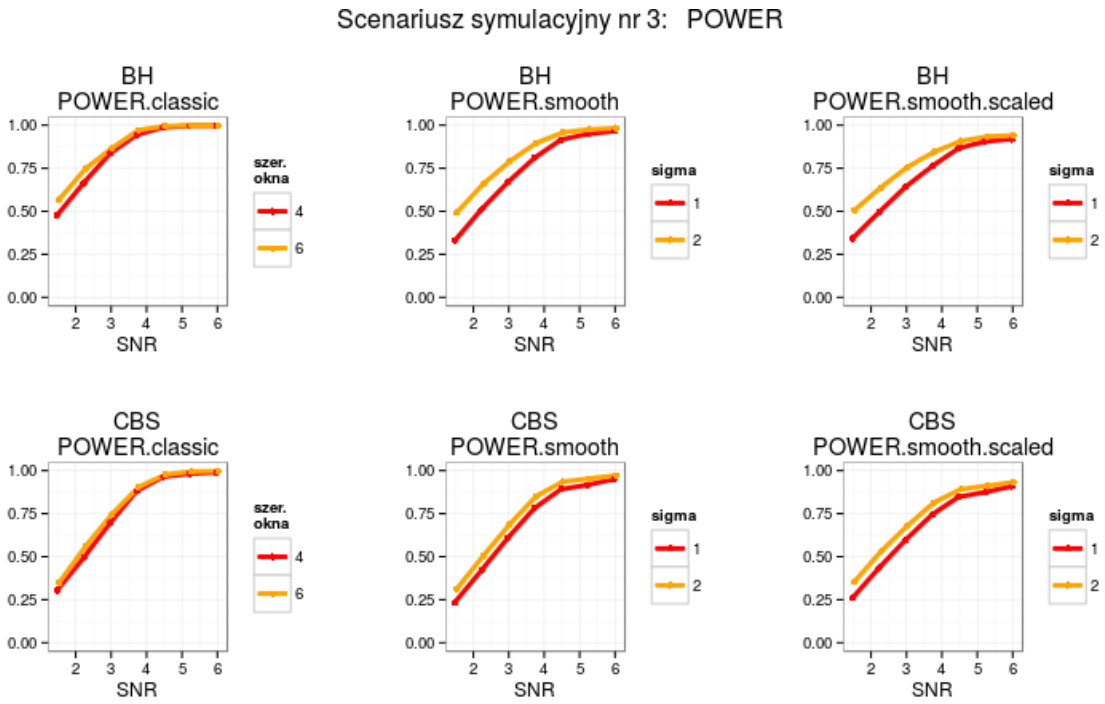
— Ponownie jak w poprzednim scenariuszu symulacyjnym, i w tym przypadku obserwujemy wysokie wartości $FDR$ dla metody $BH$, w szczególności dla mocno zaszumionych danych.

— Można poczynić interesującą obserwację, że wykres wartości $FDR.smooth$ dla metody $BH$ w wersji skalowanej jest zauważalnie bardziej gładki, niż wykresy otrzymane w procedurze klasycznej czy w procedurze przez nas proponowanej w wersji nieskalowanej.

**Scenariusz symulacyjny nr 3.**




Rysunek 5.8: Uśrednione wartości miary FDR w scenariuszu symulacyjnym nr 3.




Rysunek 5.9: Uśrednione wartości miary mocy w scenariuszu symulacyjnym nr 3.

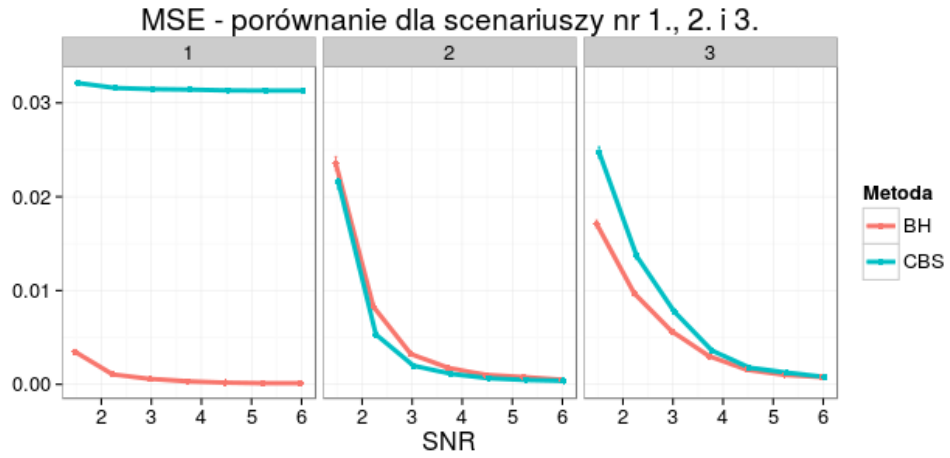
Analiza wykresów zamieszczonych na Rysunku 5.8 i Rysunku 5.9 pozwala poczynić następujące spostrzeżenia:

— Scenariusz symulacyjny nr 3. okazał się być największym "wyzwaniem" w problemie estymacji punktów zmiany średniej spośród trzech rozważanych scenariuszy. Widzimy, że zarówno metoda $BH$ jak i metoda $CBS$ mają problemy z identyfikacją niedużych punktów zmiany (por. schemat scenariusza symulacyjnego nr 3., widoczny na Rysunku 5.3.) w przypadku estymacji na danych, które mają inny niż wysoki współczynnik SNR – dla niskich i średnich z porównywanych wartości SNR obserwujemy relatywnie wysokie wartości $FDR$ i niskie wartości mocy obu metod.

— W przypadku tego scenariusza symulacyjnego widzimy różnice w wynikach wartości miar w zależności od wyboru parametru szerokości okna (podejście klasyczne) / $\sigma$ (podejście przez nas proponowane). Dla bardziej "surowego" wyboru wartości parametru (mniejsza szerokość okna / mniejsza $\sigma$), który na wykresie oznaczony jest linią w kolorze czerwonym, obserwujemy istotnie wyższe wartości $FDR$ i niższe wartości $POWER$, niż dla "łagodniejszego" wyboru wartości parametru (linia w kolorze pomarańczowym).

### MSE

MSE (ang. *Mean Squared Error*) jest ostatnią z miar poprawności identyfikacji punktów zmiany, którą porównywaliśmy. Wykresy na Rysunku 5.10. przedstawiają uśrednione po 100 symulacjach Monte Carlo wartości MSE dla każdej z obu metod referencyjnych, w podziale na 3 analizowane scenariusze symulacyjne i przy założeniu różnego poziomu zaszumienia danych.




Rysunek 5.10: Uśrednione wartości MSE dla 3 porównywanych scenariuszy symulacyjnych.

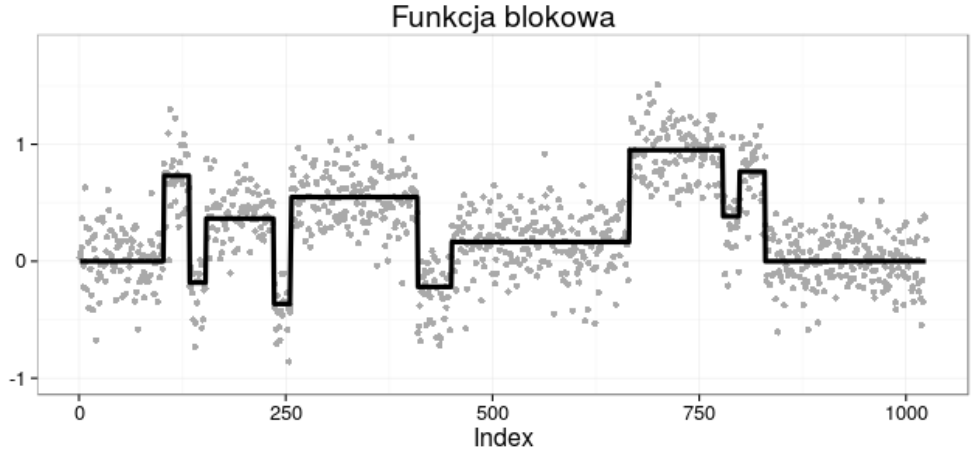
Jak widzimy z wykresów zamieszczonych na Rysunku powyżej, metoda $BH$ charakteryzuje się mniejszą wartością MSE niż metoda $CBS$ dla większości z analizowanych przypadków. Różnica ta jest szczególnie wyraźna w przypadku pierwszego z 3 porównywanych scenariuszy symulacyjnych.

## 5.2. Część druga – przykład identyfikacji punktów zmiany z wykorzystaniem reprezentacji falkowej w dekompozycji wieloskalowej obiektu

W niniejszej sekcji prezentujemy przykład zastosowania reprezentacji falkowej w dekompozycji wieloskalowej obiektu (por. Rozdział 3.) w problemie identyfikacji punktów zmiany rozkładu.

### Funkcja blokowa

Zdecydowaliśmy się na przedstawienie działania metody na przykładzie tzw. *funkcji blokowej* (ang. *block function*); funkcja ta została została zdefiniowana przez Donoho i Johnstone w pracy [39]. Funkcja zdefiniowana jest dla punktów $x_i$, $i = 1, ..., 1064$. Na Rysunku 5.11. przedstawione są wartości funkcji blokowej (czarna ciągła linia) oraz wartości zaszumione (szare kropki), w oparciu o które wyznaczaliśmy estymacje rzeczywistych segmentów w przeprowadzonej analizie porównawczej. Zaszumione dane zostały wygenerowane przy założeniu wartości $SNR = 1.5$.




Rysunek 5.11: Rzeczywiste wartości funkcji blokowej (czarna linia ciągła) oraz wartości zaszumione (szare kropki).

Wybór funkcji blokowej do przeprowadzenia analizy porównawczej zastosowania reprezentacji falkowej w dekompozycji wieloskalowej obiektu był motywowany obserwacją poczynioną we wczesnych analizach (niezamieszczonych w niniejszej pracy), że metoda wykorzystująca reprezentację falkową nie spisuje się "dobrze" na danych małych rozmiarów – specyfika metody nie pozwala jej na poprawną identyfikację pojedynczych punktów zmiany, które występują w lokalizacjach o indeksach innych, niż potęgi liczby 2.

Rozważanym przez nas pomysłem na zaradzenie temu problemowi jest sztuczne "dopisywanie" wartości przed i/lub po ciągu posiadanych obserwacji, by zwiększyć w ten sposób liczebność próbki i podnieść zdolność do poprawnej

identyfikacji. Implementacja i analiza symulacyjna tego pomysłu nie została jednak objęta w ramy niniejszej pracy.

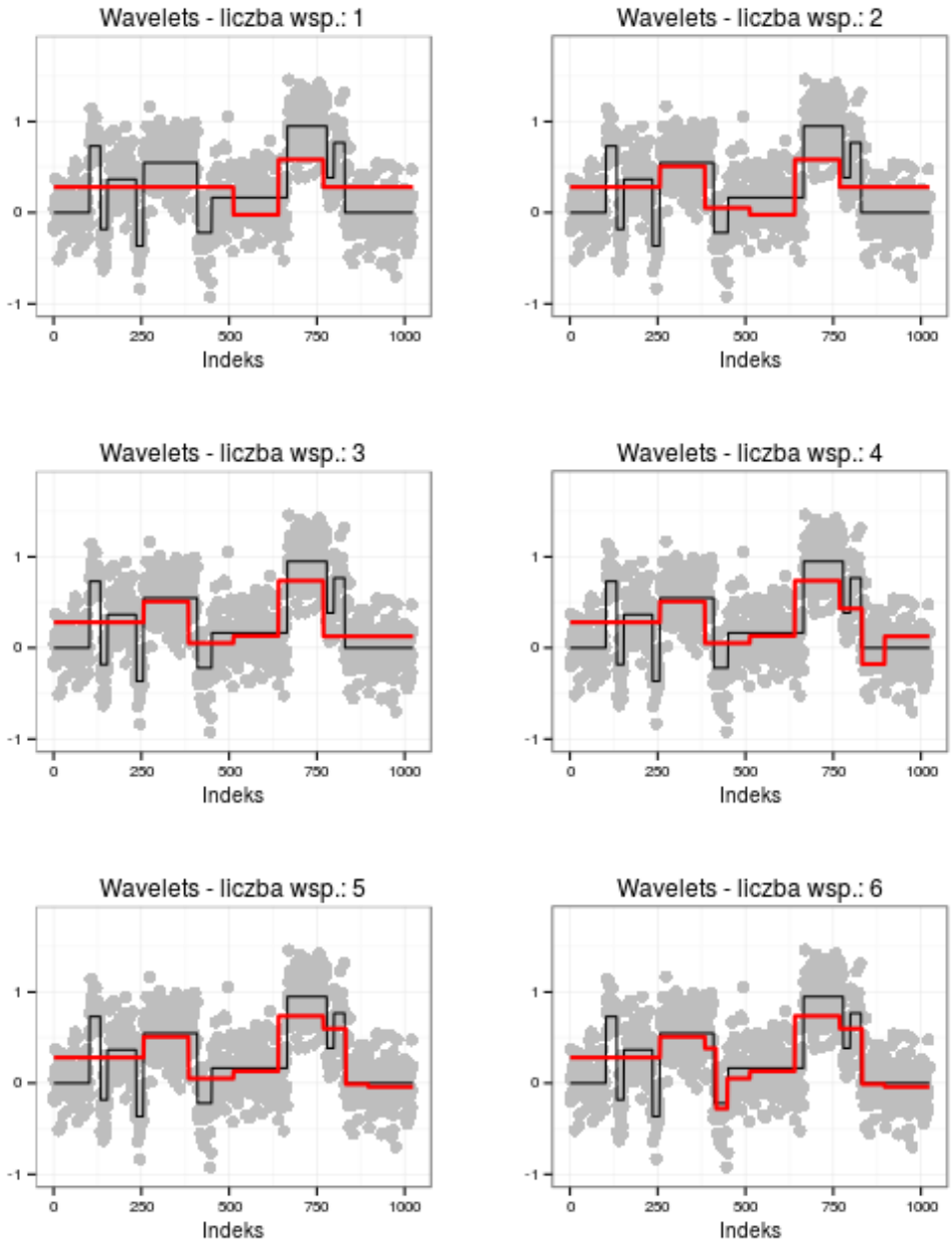
Jak wspomniano w opisie teoretycznym metody (por. Rozdział 3.), stosowane są różne podejścia w procedurze tzw. progowania współczynników falkowych. W niniejszej pracy wykorzystujemy dwa z nich:

1. progowanie polegające na zachowaniu $g$ największych co do wartości bezwzględnej współczynników falkowych,

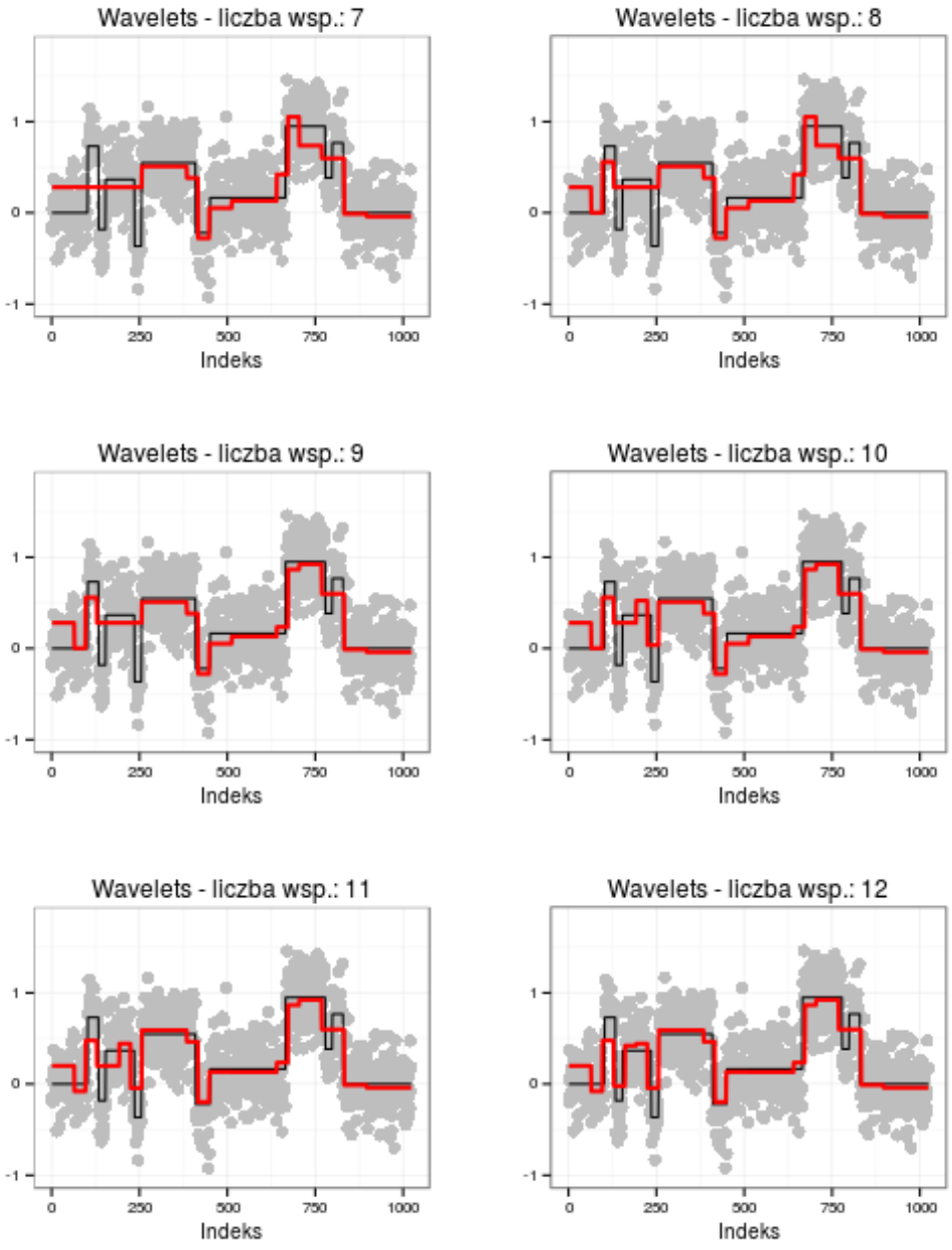
2. progowanie polegające na zachowaniu współczynników, które są większe co do wartości bezwzględnej niż zadany próg; podejście to występuje w dwóch wersjach – wersji ostrej (ang. *hard thresholding*) i wersji łagodnej (ang. *soft thresholding*); w przeprowadzonej analizie porównawczej korzystaliśmy z wersji ostrej.

### 5.2.1. Progowanie z wykorzystaniem $g$ największych współczynników falkowych

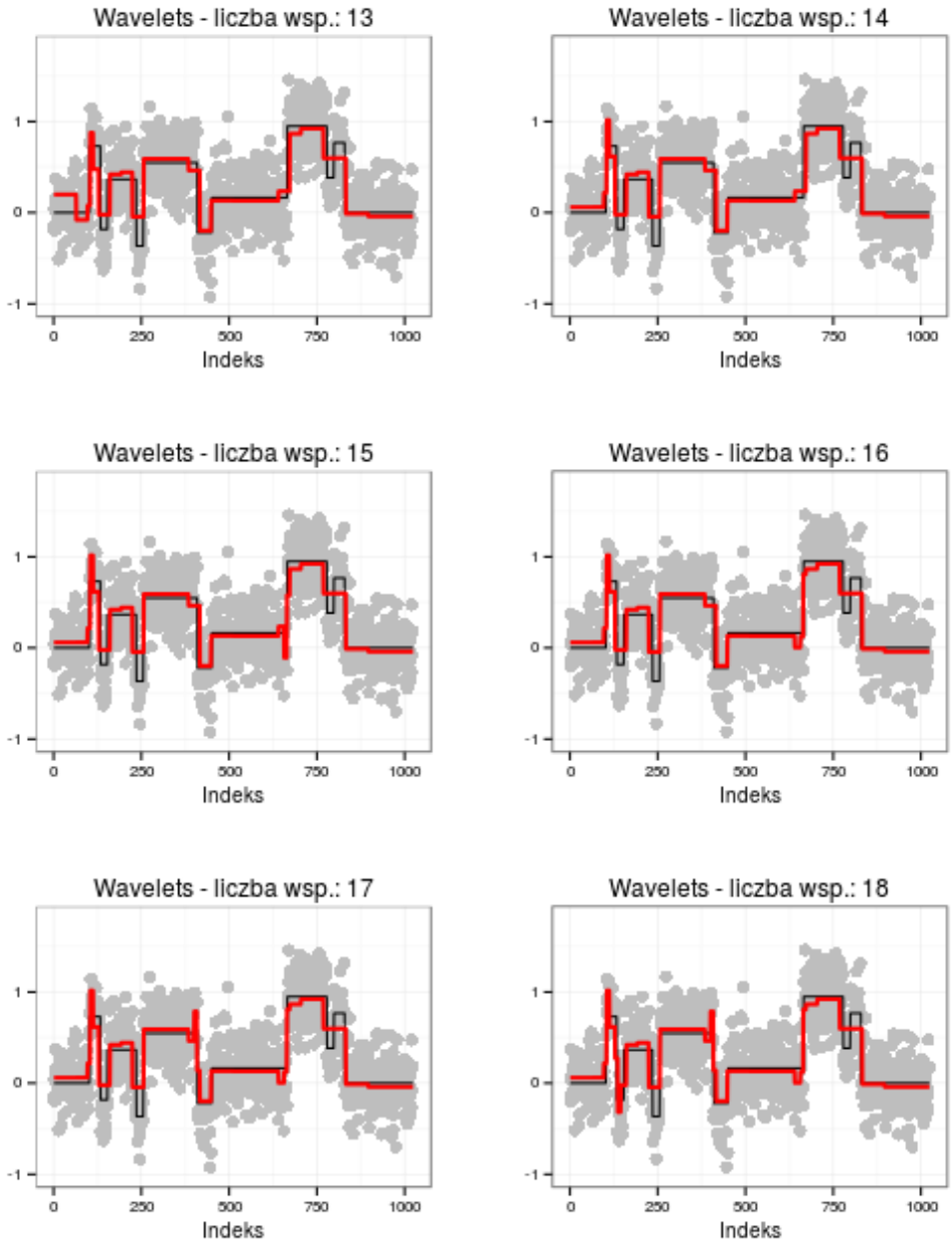
Na kolejnych stronach niniejszej pracy zamieszczone zostały wykresy, przedstawiające wynik estymacji segmentów wartości przy wykorzystaniu reprezentacji falkowej w dekompozycji wieloskalowej obiektu i progowania współczynników falkowych polegającego na zachowaniu $g$ największych co do wartości bezwzględnej współczynników (estymacje oznaczono ciągłą linią czerwoną).




Rysunek 5.12: Estymacja segmentów wartości przy wykorzystaniu reprezentacji falkowej w dekompozycji wieloskalowej obiektu i progowania współczynników falkowych polegającego na zachowaniu $g$ największych co do wartości bezwzględnej współczynników – część pierwsza wyników.




Rysunek 5.13: Estymacja segmentów wartości przy wykorzystaniu reprezentacji falkowej w dekompozycji wieloskalowej obiektu i progowania współczynników falkowych polegającego na zachowaniu $g$ największych co do wartości bezwzględnej współczynników – część druga wyników.




Rysunek 5.14: Estymacja segmentów wartości przy wykorzystaniu reprezentacji falkowej w dekompozycji wieloskalowej obiektu i progowania współczynników falkowych polegającego na zachowaniu $g$ największych co do wartości bezwzględnej współczynników – część trzecia wyników.

Wizualna inspekcja wykresów zmieszczonych na Rysunkach 5.12., 5.13. i 5.14. pozwala stwierdzić, że liczba $g$ największych co do wartości bezwzględnej współczynników, które są zachowywane w procedurze progowania, w istotny sposób wpływa na kształt estymowanych przez metodę segmentów wartości.

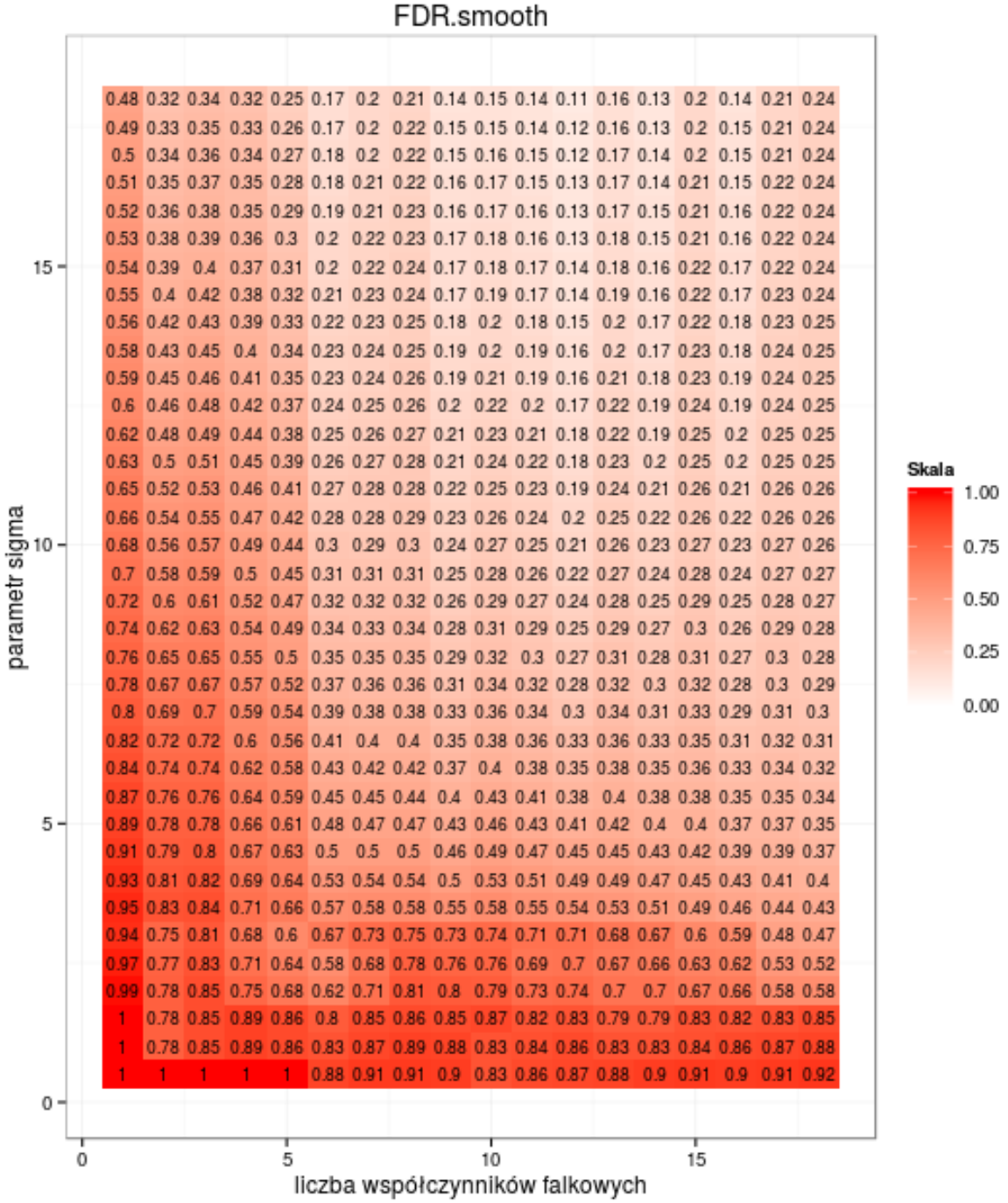
Widzimy, że dla $g = 1, ..., 6$ (por. Rysunek 5.12.) otrzymane estymacje są dość "zgrubne" – nie oddają wielu skoków występujących w rzeczywistych wartościach funkcji blokowej; dla dużych $g$ ($g = 13, ..., 18$, por. Rysunek 5.14.) otrzymane estymacje są już z kolei mocno nieregularne – w wartościach estymowanych widzimy wiele skoków, które nie pozostają w relacji z rzeczywistymi wartościami funkcji blokowej.

**Porównanie wartości miar** $FDR.smooth$ **i** $POWER.smooth$

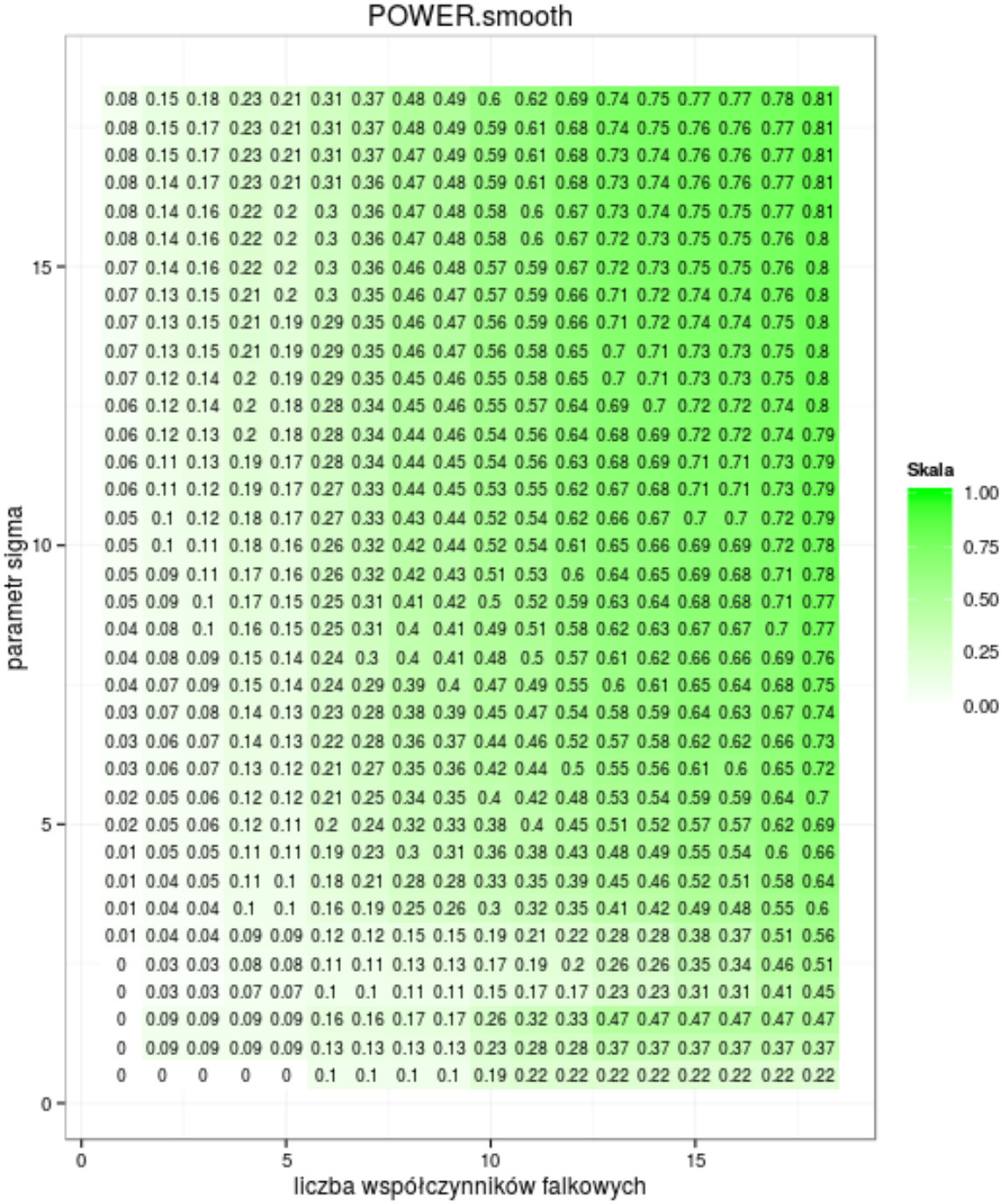
Wizualna ocena dobroci wykresów pokazujących dopasowanie estymacji do danych sugeruje, że relatywnie dobrą estymację otrzymujemy przy zachowaniu np. 12 największych co do wartości bezwzględnej współczynników falkowych w procedurze progowania. W przeprowadzonej analizie interesowało nas, jakie wartości miar oceny dobroci estymacji (zdefiniowanych w Rozdziale 4.) otrzymujemy w zależności od liczby zachowanych w procedurze progowania współczynników falkowych oraz w zależności od wartości parametru $\sigma$, zastosowanego w funkcji gęstości rozkładu normalnego, używanej w funkcji splotu w procedurze wyznaczania $FDR.smooth$ oraz $PWOER.smooth$.

Na Rysunku 5.15 znajduje się wizualizacja otrzymanych wartości $FDR.smooth$ (wersja skalowana) w zależności od liczby $g$ największych co do wartości bezwzględnej współczynników falkowych zachowanych w procedurze progowania oraz w zależności od wartości parametru $\sigma$.

Podobnie, na Rysunku 5.16 znajduje się wizualizacja otrzymanych wartości $POWER.smooth$ (wersja skalowana) w zależności od liczby $g$ największych co do wartości bezwzględnej współczynników falkowych zachowanych w procedurze progowania oraz w zależności od wartości parametru $\sigma$.




Rysunek 5.15: Wartości $FDR.smooth$ (wersja skalowana) dla estymacji funkcji blokowej metodą z wykorzystaniem reprezentacji falkowej, w zależności od liczby $g$ największych co do wartości bezwzględnej współczynników falkowych zachowanych w procedurze progowania oraz w zależności od wartości parametru $\sigma$.




Rysunek 5.16: Wartości $POWER.smooth$ (wersja skalowana) dla estymacji funkcji blokowej metodą z wykorzystaniem reprezentacji falkowej, w zależności od liczby $g$ największych co do wartości bezwzględnej współczynników falkowych zachowanych w procedurze progowania oraz w zależności od wartości parametru $\sigma$.

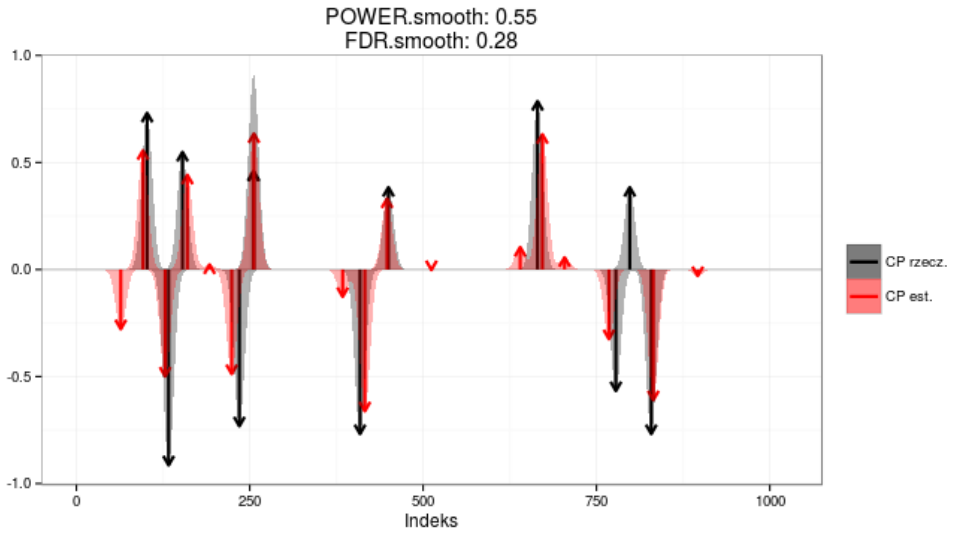
Analiza wizualizacji przedstawionych na Rysunku 5.15 i Rysunku 5.16 pozwala poczynić następujące spostrzeżenia:

— Wartości $POWER.smooth$ rosną wraz ze wzrostem liczby współczynników falkowych $g$, za wyjątkiem pojedynczych i nieznacznych odstępstw od tej reguły, co jest zgodne z naszymi oczekiwaniami wynikającymi z obserwacji wykresów porównujących estymacje dla różnych $g$ (por. Rysunki 5.12, 5.12 i 5.12).

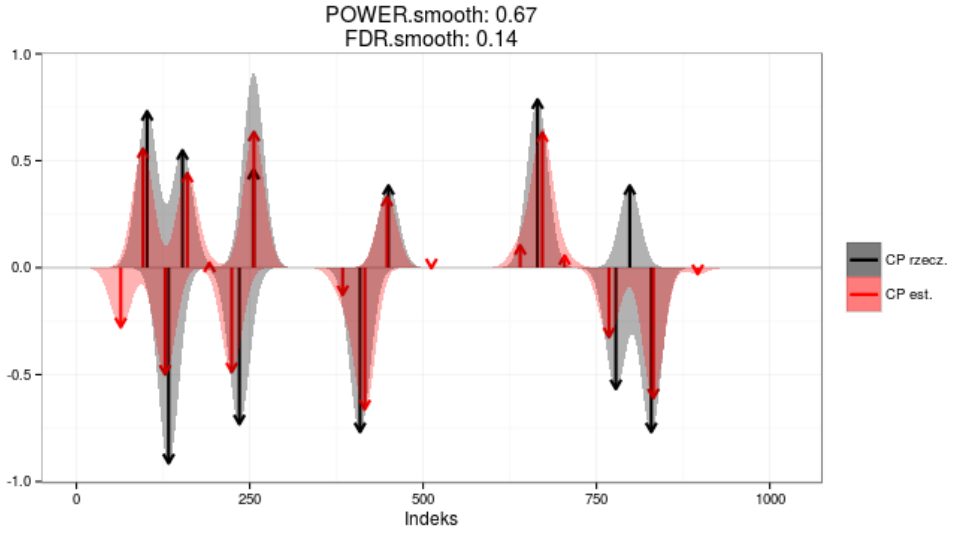
— Podobnie, wartości miary $POWER.smooth$ rosną wraz ze wzrostem wartości parametru $\sigma$; jest to odzwierciedleniem "rozluźniania" surowości w ocenie poprawności estymacji lokalizacji punktu zmiany w tym przypadku.

— Opierając się jedynie na wartościach $POWER.smooth$ należałoby stwierdzić, że im większa liczba $g$ największych co do wartości bezwzględnej współczynników falkowych pozostawionych w procedurze progowania, tym lepsza estymacja funkcji rzeczywistej. Obserwacja wartości widocznych na wizualizacji na Rysunku 5.15 sugeruje jednak, że od pewnej wartości $g$ wartość $FDR.smooth$ zwiększa sie, tj. mamy do czynienia z rosnącą frakcją fałszywych odkryć punktów zmiany. Jest to związane z rosnącą nieregularnością (rosnącą liczbą skoków wartości) występującą w otrzymywanych estymacjach wraz ze wzrostem $g$.

— Analizując wartości na wizualizacji na Rysunku 5.15 możemy zaobserwować, że dla $g$ z przedziału od 9 do 12 mamy "zagęszczenie" jaśniejszych pasm odcieni koloru czerwonego, które odpowiadają mniejszym wartościom $FDR.smooth$, czyli relatywnie dobrym estymacjom otrzymanym metodą z wykorzystanem reprezentacji falkowej. Przykładowe wykresy odpowiadające ostatniemu krokowi w procedurze wyznaczania miar $FDR.smooth$ i $POWER.smooth$ dla $g = 12$ oraz dla $\sigma = 7.5$ oraz $\sigma = 15$ zostały zamieszczone na Rysunku 5.17 i Rysunku 5.18, odpowiednio.




Rysunek 5.17: Wykres wizualizujący ostatni krok w procedurze wyznaczania miar $FDR.smooth$ i $POWER.smooth$, dla $g = 12$ oraz dla $\sigma = 7.5$.

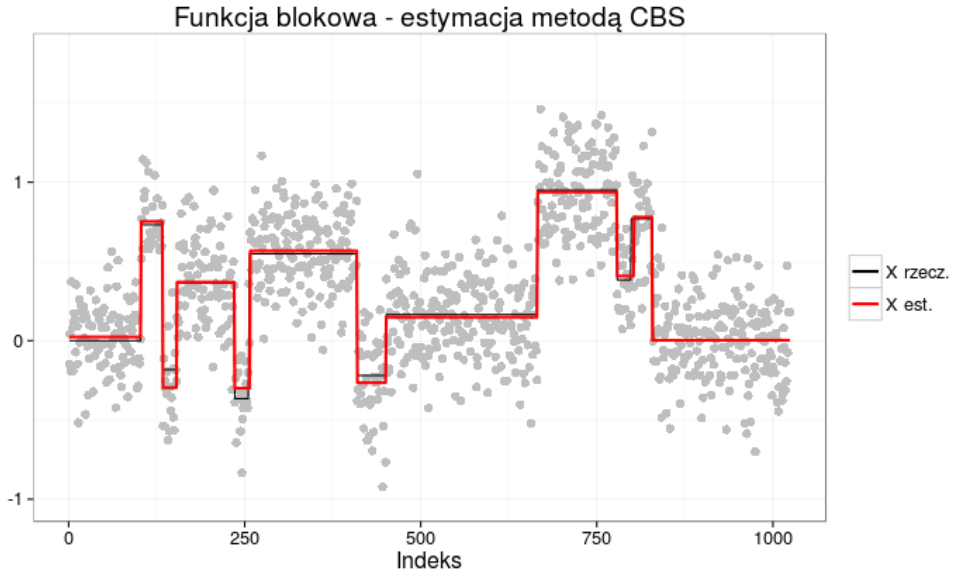



Rysunek 5.18: Wykres wizualizujący ostatni krok w procedurze wyznaczania miar $FDR.smooth$ i $POWER.smooth$, dla $g = 12$ oraz dla $\sigma = 15$.

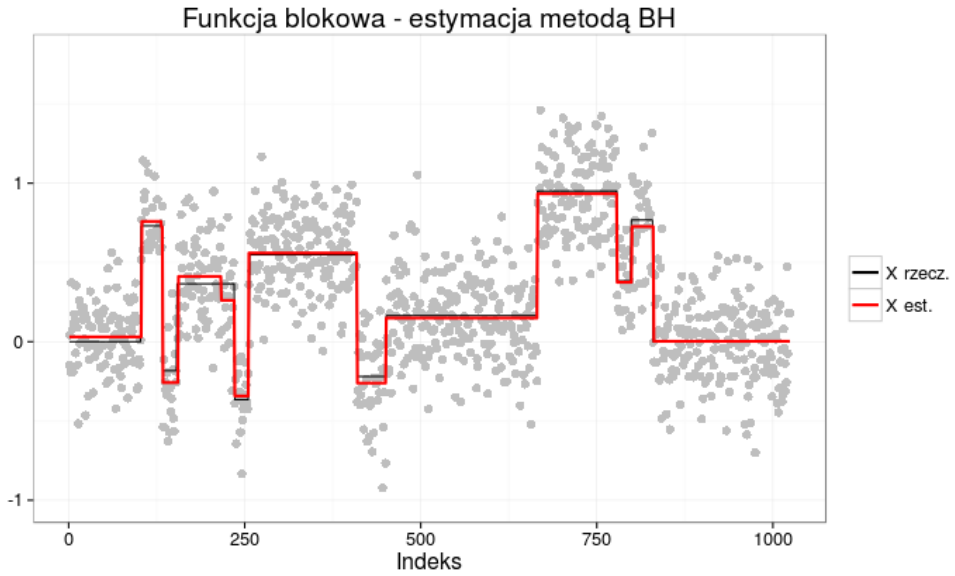
### Prównanie z wynikami otrzymanymi dla metod referencyjnych

W niniejszej subsekcji porównujemy rezultaty otrzymane dla proponowanej przez nas metody wykorzystującej reprezentację falkową w dekompozycji wieloskalowej obiektu z rezultatami, które otrzymaliśmy dla estymacji rozważanej funkcji blokowej przy wykorzystaniu metod referencyjnych – $CBS$ oraz $BH$.

Na Rysunkach 5.19 i 5.20 znajdują się wykresy przedstawiające rzeczywiste wartości funkcji blokowej oraz estymacje otrzymane na podstawie zaszumionych danych przy wykorzystaniu metod referencyjnych $CBS$ i $BH$, odpowiednio.

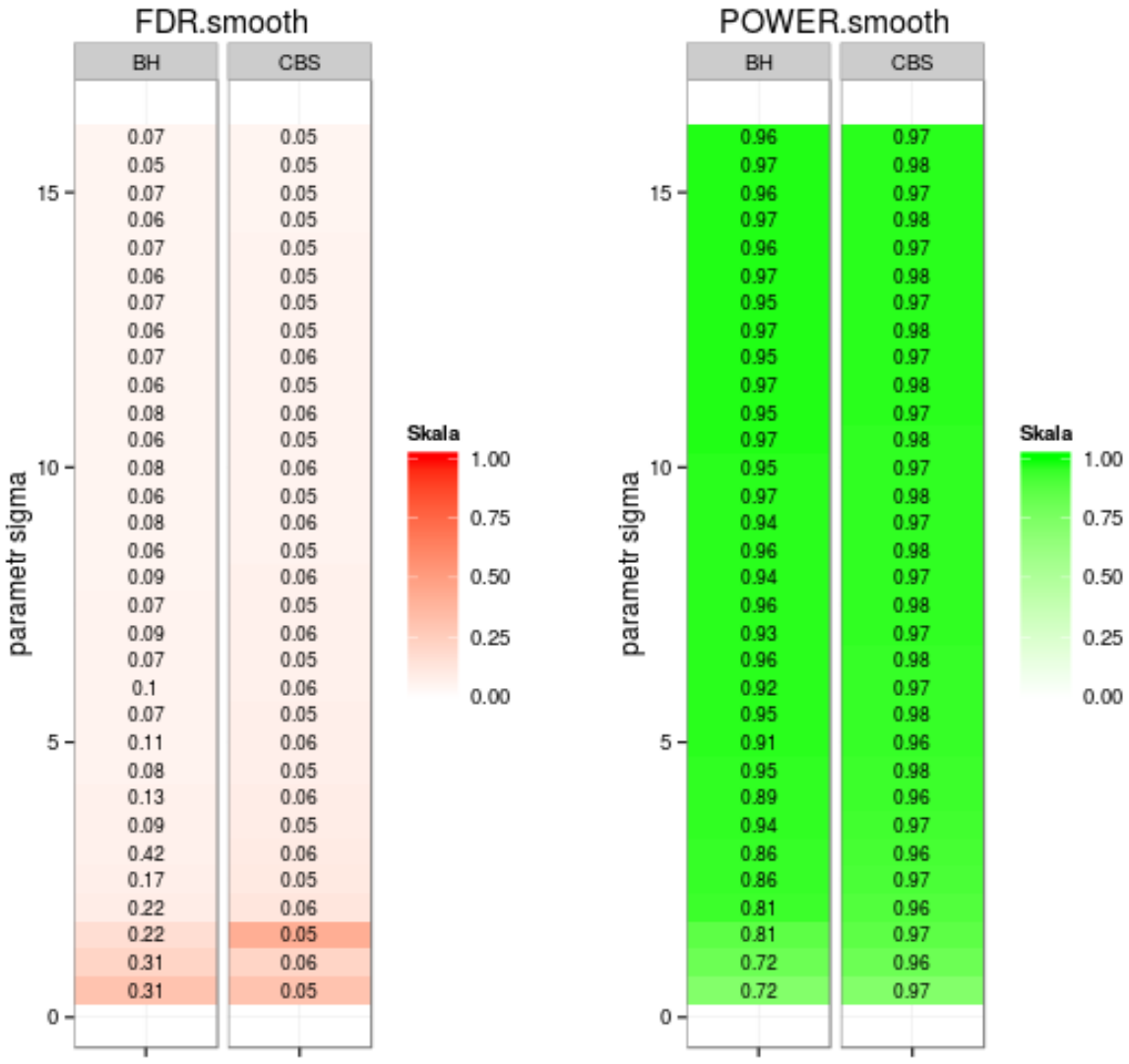



Rysunek 5.19: Rzeczywiste wartości funkcji blokowej (czarna ciągła linia) oraz estymacja (czerwona ciągła linia) otrzymana na podstawie zaszumionych danych (szare kropki) przy wykorzystaniu metody $CBS$.




Rysunek 5.20: Rzeczywiste wartości funkcji blokowej (czarna ciągła linia) oraz estymacja (czerwona ciągła linia) otrzymana na podstawie zaszumionych danych (szare kropki) przy wykorzystaniu metody $BH$.

Inspekcja wizualna wykresów zamieszczonych na Rysunkach 5.19 i 5.20 pozwala stwierdzić, że obie metody referencyjne dają porównywalnie dobre estymacje rzeczywistych wartości funkcji blokowych. Obserwacja ta znajduje potwierdzenie w wartościach miar $FDR.smooth$ oraz $POWER.smooth$, wyznaczonych dla różnych wartości parametru $\sigma$. Wizualizacja tych wartości przedstawiona jest na Rysunku 5.21.




Rysunek 5.21: Wartości $FDR.smooth$ i $POWER.smooth$ (wersja skalowana) dla estymacji funkcji blokowej przy użyciu metod referencyjnych – $CBS$ i $BH$, w zależności od wartości parametru $\sigma$.

Analiza wizualizacji przedstawionej na Rysunku 5.21 pozwala poczynić następujące spostrzeżenia:

— Podobnie jak w przypadku porównania wykonanego dla metody proponowanej w niniejszej pracy, tak i w przypadku obu metod referencyjnych obserwujemy wzrost wartości miary $POWER.smooth$ oraz spadek wartości miary $FDR.smooth$ wraz ze wzrostem wartości parametru $\sigma$.

— Można stwierdzić, że wartości miar $POWER.smooth$ i $FDR.smooth$ otrzymywane dla obu metod referencyjnych w rozważanym przykładzie są podobne; obserwujemy niewielką przewagę metody $CBS$ nad metodą $BH$.

— Zastosowanie obu metod skutkuje otrzymaniem istotnie lepszych rezultatów (rozumianych w sensie wartości miar $POWER.smooth$ i $FDR.smooth$), niż zastosowanie proponowanej przez nas metody wykorzystującej reprezentację falkową w dekompozycji wieloskalowej obiektu.

### 5.2.2. Progowanie metodą *hard thresholding*

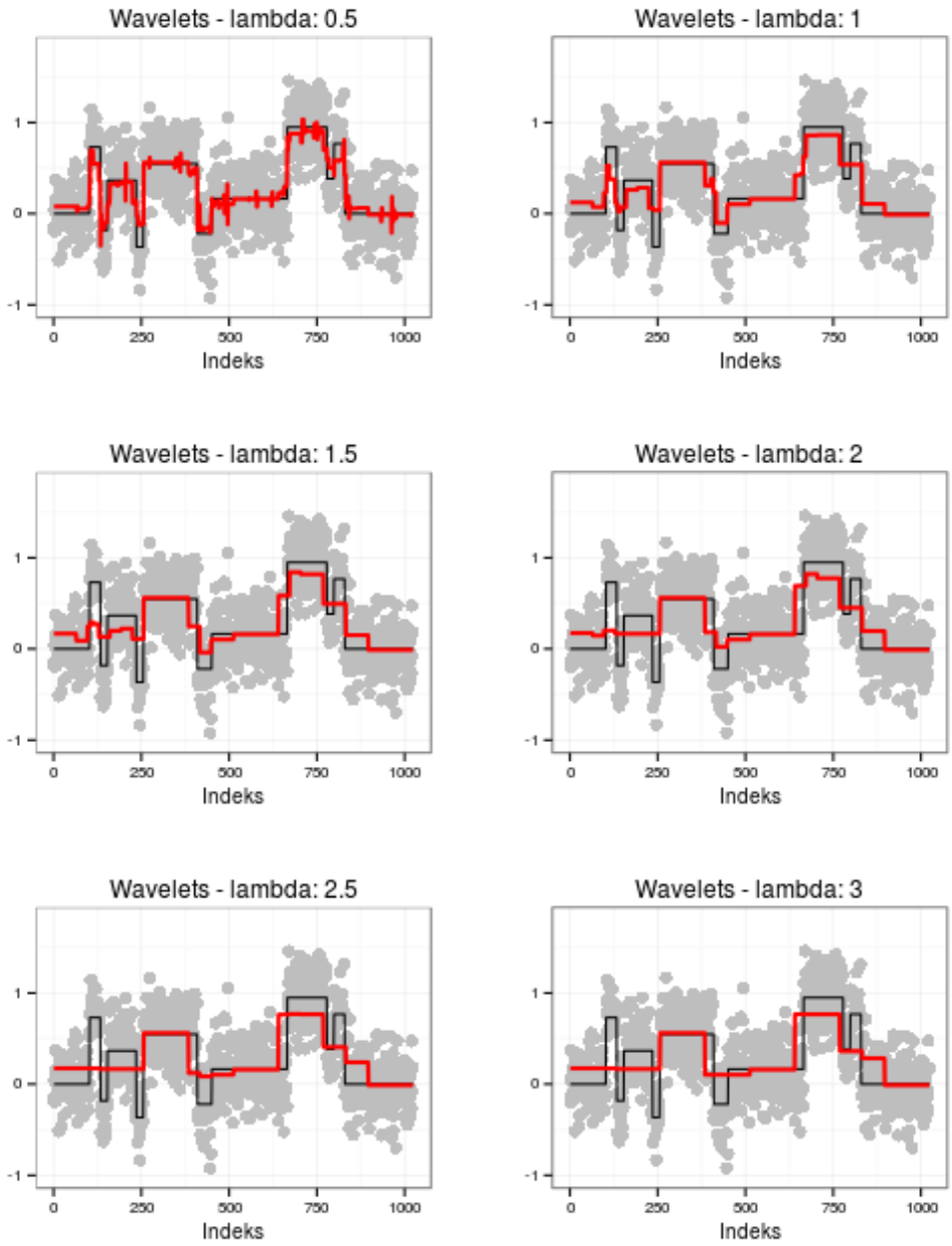
W niniejszej sekcji przyglądamy się rezultatom estymacji rzeczywistych wartości funkcji blokowej, otrzymanych w procedurze identyfikacji punktów zmiany proponowanej w niniejszej pracy. Na Rysunku 5.22. zamieszczone zostały wykresy, przedstawiające wynik estymacji segmentów wartości przy wykorzystaniu reprezentacji falkowej w dekompozycji wieloskalowej obiektu i progowania współczynników falkowych metodą *hard thresholding*, dla różnych wartości progu $\lambda$.

Wizualna inspekcja wykresów zmieszczonych na Rysunku 5.22. pozwala zauważyć, że estymacje otrzymywane przy zastosowaniu tej metody progowania są różne od tych przedstawionych w poprzedniej subsekcji. Podobnie jak w metodzie progowania analizowanej poprzednio, wraz ze zmniejszaniem liczby zachowywanych w procesie progowania współczynników falkowych (tj. wraz ze wzrostem wartości progu $\lambda$), otrzymywane estymacje stają się coraz bardziej "zgrubne" i coraz słabiej oddają nieregularności występujące w wartościach rzeczywistych funkcji blokowej. I odwrotnie, wraz ze zwiekszaniem liczby zachowywanych w procesie progowania współczynników falkowych (tj. wraz ze spadkiem wartości progu $\lambda$), estymacje stają się coraz bardziej nieregularne – obserwujemy wiele skoków ich wartości.

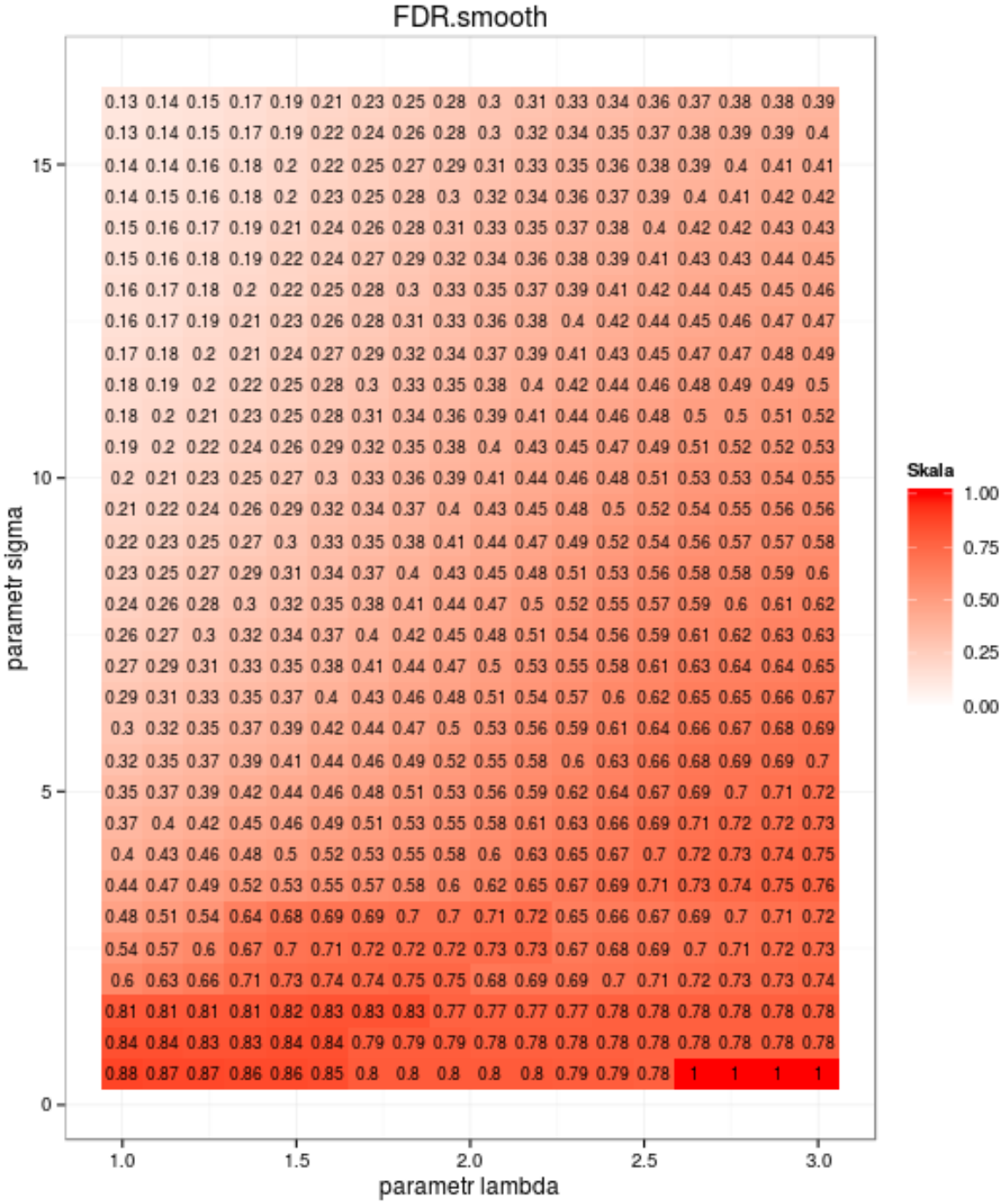
Zauważmy, że gdy zmniejszamy wartość parametru $\lambda$, otrzymujemy estymacje, które są coraz lepiej dopasowane do danych rzeczywistych, ale które wskazują też coraz większą liczbę punktów zmiany (skutkiem czego otrzymujemy coraz więcej fałszywych odkryć punktów zmiany). Innymi słowy, zmniejszanie wartości parametru $\lambda$ będzie skutkowało coraz bardziej dokładnymi estymacjami dużych co do wartości bezwzględnej skoków i coraz większą liczbą fałszywych odkryć powiązanych z małymi skokami estymowanych wartości. Ostatecznie, wartość $FDR.smooth$ maleje wraz ze spadkiem wartości parametru $\lambda$, co widoczne jest na wizualizacji na Rysunku 5.23. W szczególności, nie obserwujemy "doliny" w wartościach $FDR.smooth$, jak w poprzednim przypadku, gdy

porównywaliśmy otrzymywane wartości miary $FDR.smooth$ w zależności od liczby $g$ największych co do wartości bezwzględnej współczynników falkowych zachowywanych w procedurze progowania.

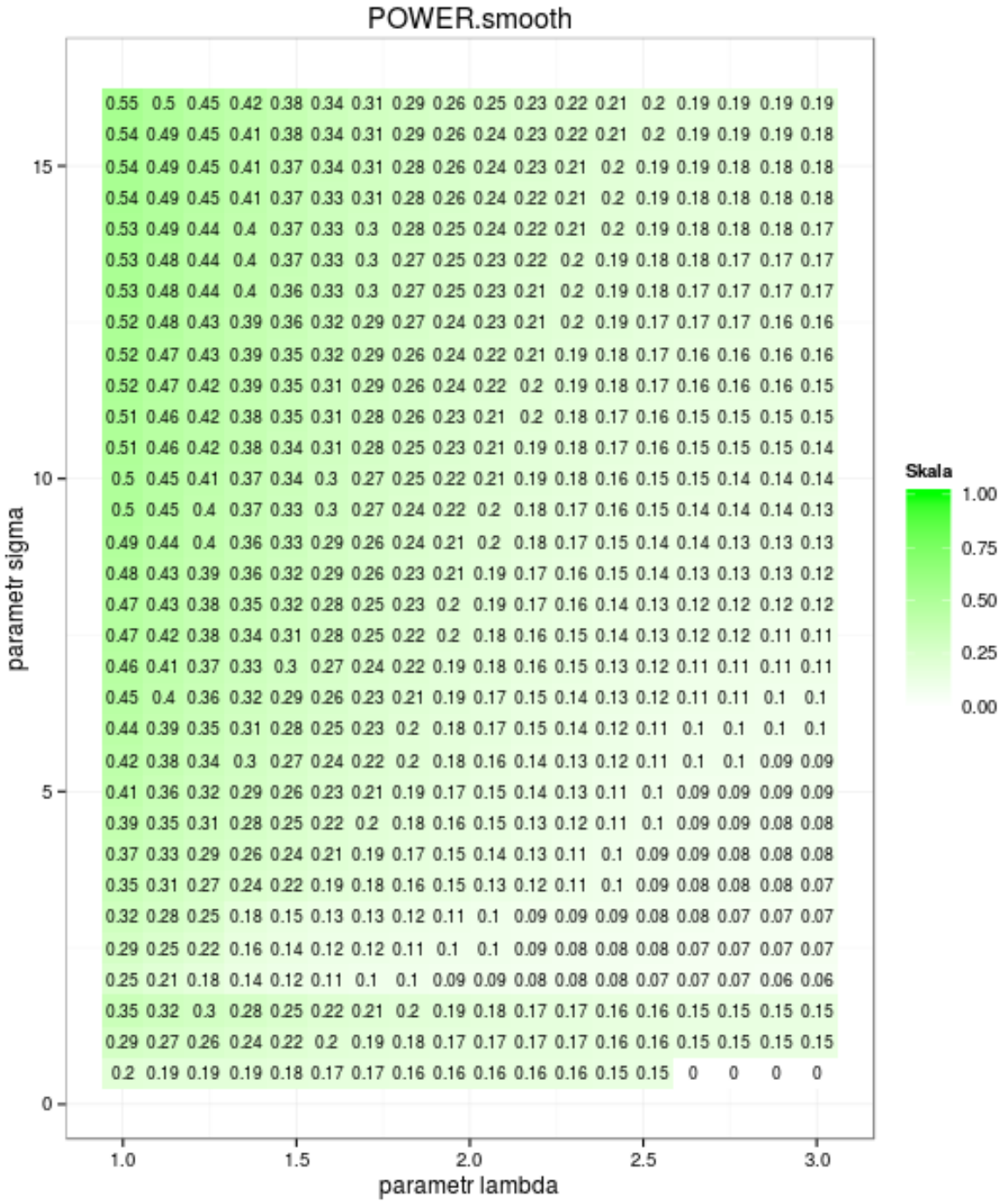
Na wykresie na Rysunku 5.24 przedstawione są wartości miary $POWER.smooth$, otrzymane dla różnych wartości parametru $\lambda$ oraz różnych wartości parametru $\sigma$. Zgodnie z oczekiwaniami wraz ze spadkiem wartości parametru $\lambda$ otrzymywane wartości są coraz większe.




Rysunek 5.22: Estymacja segmentów wartości przy wykorzystaniu reprezentacji falkowej w dekompozycji wieloskalowej obiektu i progowania współczynników metodą *hard thresholding*.




Rysunek 5.23: Wartości *FDR.smooth* (wersja skalowana) dla estymacji funkcji blokowej metodą z wykorzystaniem reprezentacji falkowej, w zależności od wartości parametru $\lambda$ oraz w zależności od wartości parametru $\sigma$.




Rysunek 5.24: Wartości *POWER.smooth* (wersja skalowana) dla estymacji funkcji blokowej metodą z wykorzystaniem reprezentacji falkowej, w zależności od wartości parametru $\lambda$ oraz w zależności od wartości parametru $\sigma$.

### 5.2.3. Podsumowanie części drugiej analizy symulacyjnej

Analiza zamieszczonych powyżej wyników estymacji otrzymanych dla rozważanego przykładu funkcji blokowej pozwala na następujące spostrzeżenia.

— Wybór metody progowania współczynników falkowych ma wpływ na postać otrzymywanych estymacji.

— Wartości parametrów $g$ i $\lambda$ stosowanych w procedurze progowania mają istotny wpływ na postać otrzymanych estymacji, w szczególności – na wartości otrzymywanych miar $FDR.smooth$ i $POWER.smooth$.

— Zarówno w przypadku metody proponowanej, jak i w przypadku metod referencyjnych obserwujemy poprawę w wartościach ww. miar (spadek i wzrost, odpowiednio) wraz ze wzrostem wartości parametru $\sigma$, stosowanego w procedurze wyznaczania $FDR.smooth$ i $POWER.smooth$.

— Metody referencyjne okazały się dawać istotnie większe wartości $POWER.smooth$ i istotnie mniejsze wartości $FDR.smooth$ w porównaniu z metodą przez nas proponowaną, dla każdej z przyjętych wartości parametru $\sigma$.

## Rozdział 6

# Podsumowanie

Przedstawione w pracy wyniki pokazują, że wybrane istniejące algorytmy identyfikacji punktów zmiany rozkładu – choć dedykowane do pracy w tym samym, zawężonym obszarze zastosowań praktycznych – różnią się charakterem zwracanych estymacji. Próba zastosowania narzędzi dekompozycji wieloskalowej obiektu wydawała się pomysłem, który, intuicyjnie, ma szanse dać dobre wyniki, jednak w części symulacyjnej daliśmy przykład na istotną przewagę metod referencyjnych; wydaje się, że zastosowanie proponowanego konceptu wymagałoby większego nakładu pracy, obejmującego m.in. bardziej dogłębną analizę wyboru typu procedury progowania i parametrów tej procedury. Innym z poruszonych zagadnień, którego dalszą analizę można rozważać, jest zastosowanie proponowanych przez nas miar $FDR.smooth$ i $POWER.smooth$.

# Bibliografia


[1] G. J. Ross. Parametric and nonparametric sequential change detection in R: The cpm package. *Journal of Statistical Software*, oczekujący na publikację.

[2] T. L. Lai. Sequential Changepoint Detection in Quality Control and Dynamical Systems. *Journal of the Royal Statistical Society B*, 57(4): 613–658, 1995.

[3] B. Efron, N. R. Zhang. False Discovery Rates and Copy Number Variation. *Biometrika*, 98(2): 251–271, 2011.

[4] A. Tartakovsky, B. Rozovskii, R. Blazek, H. Kim. A Novel Approach to Detection of Intrusions in Computer Networks via Adaptive Sequential and Batch-Sequential ChangePoint Detection Methods. *IEEE Transactions on Signal Processing*, 54(9): 3372–3382, 2006.

[5] G. J. Ross. Modelling financial volatility in the presence of abrupt changes. *Physica A: Statistical Mechanics and its Applications*, 392(2): 350-360, 2012.

[6] D. Hinkley, E. Hinkley. Inference About Change-Point in a Sequence of Binomial Variables. *Biometrika*, 57(3): 477–488, 1970.

[7] A. B. Olshen, E. S. Venkatraman. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4): 557-72, 2004.

[8] D. A. Stephens. Bayesian Retrospective Multiple-Changepoint Identification. *Journal of the Royal Statistical Society C*, 43: 159–178, 1994.

[9] D. Barry, J. A. Hartigan. A Bayesian Analysis for Change Point Problems. *Journal of the American Statistical Association*, 35(3): 309–319, 1993.

[10] E. S. Page. Continuous Inspection Schemes. *Biometrika*, 41(1/2): 100–115, 1954.

[11] S. W. Roberts. Control Chart Tests Based on Geometric Moving Averages. *Technometrics*, 42(1): 97–101, 1959.

[12] S. Chib. Estimation and Comparison of Multiple Change-Point Models. *Journal of Econometrics*, 86(2): 221–241, 1998.

[13] P. Fearnhead, Z. Liu. On-line Inference for Multiple Changepoint Problems. *Journal of the Royal Statistical Society B*, 69(4): 589–605 2007.

[14] J. Fridlyand, A. M. Snijders, D. Pinkel, D. G. Albertson, A. N. Jain. Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 92(2): 132-153, 2005.

[15] J. R. Pollack, T. Sorlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A. L. Borresen-Dale, P. O. Brown. Microarray analysis reveals a major direct role of DNA copy number alteration in the transriptional program of human breast cancers. *Proceedings of the National Academy of Sciences*, 99: 12963-12968, 2002.

[16] G. Hodgson, J. H. Hager, S. Volik, S. Hariono, M. Wernick, D. Moore, D. G. Albertson, D. Pinkel, C. Collins, D. Hanahan, J. W. Gray. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics*, 929: 459-464, 2001.

[17] D. Pinkel, R. D. Segraves, S. C. Sudar, I. P. D. Kowbel, C. Collins, W. L. Kuo,

C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarray. *Nature Genetics*, 20: 207-211, 1998.

[18] A. M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, N. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A. N. Jain, D. Pinkel, D. G. Albertson. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, 29: 4281-4286, 2001.

[19] S. Olszewska. Nauka o genetyce. Ekspresja genów. [Online.] [Dostęp w Internecie: 19 czerwca 2015.] Dostępny w Internecie: <http://www.elis-gen.com/pl/newsy/nauka-o-genetyce/ekspresja-genow.html>.

[20] A. Woźniak. Nauka o genetyce. Genom człowieka. [Online.] [Dostęp w Internecie: 19 czerwca 2015.] Dostępny w Internecie: <http://www.elis-gen.com/pl/newsy/nauka-o-genetyce/genom-czlowieka.html>.

[21] The R Project for Statistical Computing. [Online.] [Dostęp w Internecie: 19 czerwca 2015.] Dostępny w Internecie: <http://www.r-project.org/>.

[22] E. S. Venkatraman, A. Olshen. Bioconductor 3.1. Software Packages. DNAcopy. [Online.] [Dostęp w Internecie: 19 czerwca 2015.] Dostępny w Internecie: <http://www.bioconductor.org/packages/release/bioc/html/DNAcopy.html>.

[23] C. Erdman, J. W. Emerson. bcp: A Package for Performing a Bayesian Analysis of Change Point Problems. [Online.] [Dostęp w Internecie: 19 czerwca 2015.] Dostępny w Internecie: <http://cran.r-project.org/web/packages/bcp/index.html>.

[24] R. Killick, C. F. H. Nam, J. A. D.Aston, I. A. Eckley. The Changepoint Repository. Fostering the exchange of knowledge and software related to changepoint analysis. [Online.] [Dostęp w Internecie: 19 czerwca 2015.] Dostępny w Internecie: <http://www.changepoint.info/>.

[25] A. Sen, M. S. Srivastava. On tests for detecting a change in mean. *Annals of Statistics*, 3: 98-108, 1975.

[26] D. Siegmund. Boundary crossing probabilities and statistical applications. *Annals of Statistics*, 14: 361-404, 1986.

[27] L. J. Vostrikova. Detecting "disorder" in multidimensional random processes. *Soviet Mathematics – Doklady*, 24: 55-59, 1981.

[28] E. S. Venkatraman. Consistency results in multiple change-point situations. Technical report, Department of Statistics, Stanford University, 1992.

[29] B. Levin, J. Kline. The CUSUM test of homogeneity with an application in spontaneous abortion epidemiology. *Statistics in Medicine*, 4: 469-488, 1985.

[30] D. Barry, J. A. Hartigan. A Bayesian Analysis for Change Point Problems. *Journal of the American Statistical Association*, 35(3): 309-319, 1993.

[31] C. Erdman, J. W. Emerson. bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems. *Journal of Statistical Software*, 23(3), 2007.

[32] S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7): 674-693, 1989.

[33] J. E. Gentle, W. K. Härdle, Y. Mori. Handbook of Computational Statistics. Springer, ISBN: 978-3-642-21550-6 (Print) 978-3-642-21551-3 (Online), 2012.

[34] A. Grossmann, J. Morlet. Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape. *SIAM Journal on Mathematical Analysis*, 15(4): 723-736, 1984.

[35] Y. Meyer. Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs. *Séminaire Bourbaki*, 145-146: 209-223, 1985-1986.

[36] G. P. Nason, Wavelet Methods in Statistics with R. Springer, ISBN: 978-0-387-75960-9 e-ISBN: 978-0-387-75961-6, 2008.

[37] D. L. Donoho. Unconditional Bases are Optimal Bases for Data Compression and for Statistical Estimation. *Wavelet Theory and Harmonic Analysis in Applied Sciences*, 1: 100-115, 1993.

[38] D. L. Donoho. De-noising by soft-thresholing. *IEEE Transactions on Information Theory*, 41: 613-627, 1995.

[39] D. L. Donoho, I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81: 425-455, 1994.

[40] D. L. Donoho, I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90: 1200-1224, 1995.

[41] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, D. Picard. Wavelet Shrinkage: Asymptopia? *Journal of the Royal Statistical Society: B*, 57: 301-369, 1995.

[42] K. V. Mardia, J. T. Kent, J. M. Bibby. Multivariate Analysis (Probability and Mathematical Statistics). Academic Press, ISBN 10: 0124712525 ISBN 13: 9780124712522, 1980.

[43] H.-Y. Gao, A. G. Bruce. WaveShrink with firm shrinkage. *Statistica Sinica*, 4: 855-874, 1997.

[44] C. Fernandez-Granda, E. J. Candès (współpraca). Towards a Mathematical Theory of Super-resolution. Information Theory Forum, Information Systems Laboratory, Stanford University, 2013. [Online.] [Dostęp w Internecie: 19 czerwca 2015.] Dostępny w Internecie: <http://www.cims.nyu.edu/ cfgranda-/stuff/superres_IT_forum.pdf>.