Raport Techniczny
Katedry Statystyki Uniwersytetu Ekonomicznego
w Krakowie 2/2016

Prognozowanie warunkowej kowariancji z wykorzystaniem odpornych estymatorów rozrzutu MCD i PCS w analizie portfelowej *

Technical Report
Department of Statistics at Cracow University of Economics
2/2016

Conditional covariance prediction in portfolio analysis using MCD and PCS robust multivariate scatter estimators

Daniel Kosiorowski† Przemysław Jaśko‡

Streszczenie

W przypadku wnioskowania na temat wybranych wielowymiarowych charakterystyk, będących podstawą podejmowania decyzji, bardzo ważna jest możliwość sprawnego uzyskiwania trafnych oszacowań parametrów rozkładów wielowymiarowych, w tym parametru wielowymiarowego rozrzutu, w szczególności, gdy próba będąca ich podstawą zawiera obserwacje odstające, będące wynikiem działania mechanizmów zakłócających właściwy dla zjawiska rozkład. W pracy dokonuje się porównania dwóch odpornych estymatorów wielowymiarowego rozrzutu – MCD (*Minimum Covariance Determinant*) i PCS (*Projection Congruent Subset*), charakteryzujących się afiniczną ekwiwariantnością oraz wysokim punktem załamania. W ramach analizy empirycznej podjęto próbę zastosowania odpornych estymatorów MCD i PCS w procedurze określania struktury wag portfela o minimalnej zmienności oraz portfela ERC (*Equal Risk Contribution*). W badaniach symulacyjnych dotyczących wykrywania obserwacji odstających estymator PCS uzyskiwał przewagę nad MCD.

Słowa kluczowe: *odporne estymatory wielowymiarowego rozrzutu, estymator MCD, estymator PCS, odporna analiza portfelowa, kowariancja zrealizowana ROWCov, portfel o minimalnym ryzyku, portfel ERC*

Abstract

Very important aspect of inference on multivariate characteristics, which are basis for decision making, is ability to make accurate estimates on parameters of multivariate distribution (importantly multivariate scatter), particularly when the sample includes outliers, produced by

* *Praca została dofinansowana ze środków przyznanych Wydziałowi Zarządzania Uniwersytetu Ekonomicznego w Krakowie, w ramach dotacji na utrzymanie potencjału badawczego w roku 2016.*

† Uniwersytet Ekonomiczny w Krakowie, Katedra Statystyki, e-mail: daniel.kosiorowski@uek.krakow.pl

‡ Uniwersytet Ekonomiczny w Krakowie, Katedra Systemów Obliczeniowych, e-mail: jaskop@uek.krakow.pl

interfering mechanism. We make a comparison of two robust multivariate scatter estimators – MCD (Minimum Covariance Determinant) and PCS (Projection Congruent Subset), which are affine equivariant and have high breakdown points. In empirical analysis we make use of mentioned robust scatter estimators in the procedure of weights optimization for minimum risk and equal risk contribution (ERC) portfolios. In simulation studies regarding outlier detection PCS estimator outperforms MCD.

Keywords: *robust multivariate scatter estimators, MCD estimator, PCS estimator, robust portfolio analysis, Realized Outlyingness Weighted Covariation (ROWCov), minimum risk portfolio, Equal Risk Contribution (ERC) portfolio*

Spis treści

1 Wstęp

Ważnym zagadnieniem w wielu praktycznych zastosowaniach analizy danych jest możliwość trafnego oszacowania nieznanych parametrów rozkładów wielowymiarowych, w sytuacji, gdy próba będąca podstawą oszacowania może zawierać obserwacje odstające (ang. *outliers*), często będące wynikiem działania mechanizmu zakłócającego rozpatrywany „główny" rozkład danych.
Wśród popularnych estymatorów wielowymiarowego parametru rozrzutu można wymienić m.in. estymatory: MCD, MVE, SDE, OGK. Ciekawą nową propozycją jest także estymator PCS (ang. *Projection Congruent Subset Estimator*), w którym przyjęto oryginalne podejście do określenia podzbioru danych, dla których zakłada się, że zostały wygenerowane przez „główny" rozkład.
W niniejszej pracy dokonuje się krótkiego omówienia i próby porównania dwóch odpornych estymatorów wielowymiarowego rozrzutu – MCD i PCS, charakteryzujących się wysokim punktem załamania oraz własnością afinicznej ekwiwariantności.
Odporna estymacja parametrów rozkładów wielowymiarowych, w szczególności parametru wielowymiarowego rozrzutu odgrywa bardzo istotną rolę m.in. w odpornej analizie portfelowej (dosyć szerokie ujęcie tego problemu wraz z przykładami empirycznymi można znaleźć w pracy [Pfaff, 2012]). W niniejszej pracy przedstawiony zostanie przykład odpornego podejścia do budowy portfela o minimalnej wariancji oraz portfela ERC (ang. *Equal Risk Contribution*).
Innym obszarem, w którym istotnym wymogiem jest wykorzystanie odpornych estymatorów rozrzutu jest procedura konstrukcji wielowymiarowych kart kontrolnych, np. karty T^2 Hotellinga. W pierwszej fazie procedury budowy wielowymiarowej karty kontrolnej określane są granice kontrolne, odpowiadające wartościom krytycznym testu statystycznego, który stosuje się sekwencyjnie w drugiej fazie, polegającej na monitorowaniu parametrów wielowymiarowych zmiennych procesu.

W pierwszej fazie konstrukcji granice kontrolne (będące kwantylami odpowiednich statystyk testowych) powinny być wyznaczone w oparciu o wielowymiarowe realizacje dotyczące monitorowanych charakterystyk, pochodzące z uregulowanego (ang. *in-control*) procesu $iid(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Jako, że granice kontrolne są kwantylami statystyk testowych, będących funkcjami wielowymiarowych zmiennych tworzących proces, niezbędne jest oszacowanie parametrów wielowymiarowego rozkładu w oparciu o realizacje procesu pozbawione obserwacji odstających, przemawiających za rozregulowaniem procesu. W klasycznym podejściu przyjmuje się za oszacowania parametrów wartości klasycznych estymatorów z próby obserwacji wykorzystywanej w ramach pierwszej fazy i z ich wykorzystaniem wyznacza się granice kontrolne, po czym odrzuca się obserwacje, dla których wartości statystyki testowej są poza granicami kontrolnymi i w oparciu o ograniczoną próbę ponownie szacuje się wartości parametrów – procedurę powtarza się do momentu, gdy powstała podpróba nie zawiera obserwacji z wartościami statystyki testowej poza granicami kontrolnymi. Zastosowanie odpornych estymatorów parametrów wielowymiarowego rozkładu, umożliwia pominięcie opisanej przed chwilą procedury iteracyjnej, gdyż ze względu na swoją konstrukcję, umożliwiają „automatyczne" pominięcie obserwacji odstających przy wyznaczaniu ich wartości. Procedurę budowy karty kontrolnej T^2 Hotellinga z wykorzystaniem odpornego estymatora parametrów $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ można znaleźć m.in. w pracach [Jensen i in., 2007] oraz [Steiner i in., 2009].
Struktura niniejszej pracy przedstawia się następująco: w kolejnym rozdziale wprowadzone zostaną pojęcia podstawowe (m.in. takie jak punkt załamania estymatora, czy afiniczna ekwiwariantność estymatora). W rozdziale 3. zdefiniowane zostaną estymatory MCD i PCS, wraz z omówieniem ich najważniejszych własności. W rozdziale 4. zostaną skrótowo przedstawione algorytmy wyznaczania wartości estymatorów MCD i PCS. Rozdział 5. obejmuje porównanie rozważanych estymatorów w formie tabelarycznej. Pracę zamyka część empiryczna, w ramach której podjęto się próby konstrukcji portfeli o minimalnej zmienności oraz portfeli ERC, z wykorzystaniem odpornych estymatorów MCD i PCS.

2 Pojęcia podstawowe

Niech $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, będzie zbiorem n wektorowych obserwacji $\mathbf{x}_i \in \mathbb{R}^p$, przy czym $n > p+1 > 2$.
Dodatkowo niech $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(\mathbf{X})$ będzie wielowymiarowym estymatorem parametrów $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, wyznaczonym w oparciu o n-elementowy zbiór obserwacji $\mathbf{X}$, gdzie $\boldsymbol{\Theta}$ jest przestrzenią parametrów.
$\boldsymbol{\theta}$ może obejmować wielowymiarowy parametr położenia, odpowiadający p-wymiarowemu wektorowi $\boldsymbol{\mu}$ o wartościach rzeczywistych oraz wielowymiarowy parametr rozrzutu odpowiadający symetrycznej, dodatnio określonej macierzy rzeczywistej $\boldsymbol{\Sigma}$:

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in [\boldsymbol{\Theta}_{\boldsymbol{\mu}} \times \boldsymbol{\Theta}_{\boldsymbol{\Sigma}} = \mathbb{R}^p \times PDS(p)] \tag{1}$$

gdzie $PDS(p)$ jest klasą symetrycznych, dodatnio określonych macierzy o wymiarach $p \times p$
Ponadto niech $\mathcal{Y}_m$ będzie rodziną wszystkich zbiorów $\mathbf{Y}_m = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ złożonych z n wektorów, które to zbiory mają $n - m$ elementów $(0 \leqslant m \leqslant n)$ wspólnych ze zbiorem $\mathbf{X}$:
$\mathcal{Y}_m = \{\mathbf{Y}_m : \#(\mathbf{Y}_m) = n, \#(\mathbf{X} \cap \mathbf{Y}_m) = n - m\}$, gdzie $0 \leqslant m \leqslant n$
Zastąpieniowy punkt załamania próby skończonej (ang. *replacement finite-sample breakdown point*, FBP) estymatora $\widehat{\boldsymbol{\theta}}_n$ dla zbioru $\mathbf{X}$ rozumie się jako najwyższą możliwą frakcję $\varepsilon_n^*(\widehat{\boldsymbol{\theta}}_n, \mathbf{X})$ obserwacji z $\mathbf{X}$, którą można zastąpić arbitralnie przyjętymi obserwacjami odstającymi, tak aby wartość estymatora $\widehat{\boldsymbol{\theta}}_n$ pozostała ograniczona oraz nie należała do brzegu przestrzeni parametrów $\boldsymbol{\Theta}$:

$$\varepsilon_n^*(\widehat{\boldsymbol{\theta}}_n, \mathbf{X}) = \frac{m^*}{n} \tag{2}$$

gdzie $m^* = \max\{m \geqslant 0 : \widehat{\boldsymbol{\theta}}_n(Y_m)$ jest ograniczony i $\widehat{\boldsymbol{\theta}}_n(\mathbf{Y}_m) \notin \boldsymbol{\partial\Theta}\}$
przy czym $\boldsymbol{\partial\Theta}$ jest oznaczeniem brzegu zbioru Θ, będącego przestrzenią parametrów
W celu oceny czy zachodzą przywołane warunki określające wartość punktu załamania, w zależności od charakteru parametrów oraz przestrzeni ich dopuszczalnych wartości, przyjmuje się odpowiednie mierniki.
Rozpatrzmy przykład parametrycznego rozkładu p-wymiarowej zmiennej losowej, określonego za pomocą wielowymiarowych parametrów położenia i skali (rozrzutu) $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, przyjmując dla ich estymatorów wyznaczonych w oparciu o zanieczyszczoną próbę $\mathbf{Y}_m$ następujące oznaczenia: $\widehat{\boldsymbol{\theta}}_n(\mathbf{Y}_m) = \left(\widehat{\boldsymbol{\mu}}_n(\mathbf{Y}_m), \widehat{\boldsymbol{\Sigma}}_n(\mathbf{Y}_m)\right)$.

Jako, że rolę przestrzeni $\mathbf{\Theta_\mu}$ dla wielowymiarowego parametru położenia $\boldsymbol{\mu}$ pełni skończona p-wymiarowa przestrzeń euklidesowa, stąd też aby $\widehat{\boldsymbol{\mu}}_n(\mathbf{Y}_m)$ był ograniczony, wystarczy, że $||\widehat{\boldsymbol{\mu}}_n(\mathbf{Y}_m)|| < \infty$.
Natomiast w przypadku wielowymiarowego parametru rozrzutu $\mathbf{\Sigma}$ przestrzenią $\mathbf{\Theta_\Sigma}$ jest zbiór symetrycznych dodatnio określonych macierzy (PDS), z każdą taką macierzą związany jest zbiór wektorów własnych z odpowiadającymi im rzeczywistymi wartościami własnymi.
Warunek PDS sprowadza się do wymogu, aby wartości własne macierzy $\widehat{\mathbf{\Sigma}}_n(\mathbf{Y}_m)$ były dodatnie ($\widehat{\mathbf{\Sigma}}_n(\mathbf{Y}_m) \notin \boldsymbol{\partial\Theta_\Sigma}$) i skończone ($\widehat{\mathbf{\Sigma}}_n(\mathbf{Y}_m)$ – ograniczony).
Jest to związane z faktem, iż dla dowolnych wektorów $\mathbf{v} \neq \mathbf{0}$ forma kwadratowa $\mathbf{v}'\widehat{\mathbf{\Sigma}}_n(\mathbf{Y}_m)\mathbf{v}$, jest większa od 0 oraz skończona, wtedy i tylko wtedy, gdy (bez utraty ogólności warunek w swej konstrukcji odnosi się do wektorów jednostkowych $||\mathbf{v}|| = 1$) :

$$\max_{||\mathbf{v}||=1} \mathbf{v}'\widehat{\mathbf{\Sigma}}_n(\mathbf{Y}_m)\mathbf{v} = \lambda_1 < \infty \tag{3}$$

$$\min_{||\mathbf{v}||=1} \mathbf{v}'\widehat{\mathbf{\Sigma}}_n(\mathbf{Y}_m)\mathbf{v} = \lambda_p > 0 \tag{4}$$

W przypadku, gdyby $\lambda_1 \to \infty$ mielibyśmy do czynienia z eksplozją, wynikającą z występowania w zbiorze $\mathbf{Y}_m$ nadmiernej (dla przyjętej konstrukcji estymatora wielowymiarowego rozrzutu) liczby, punktów nieskończenie oddalonych od głównego obszaru zmienności danych, czyli tzw. punktów odstających (ang. *outliers*). W sytuacji, gdyby $\lambda_p = 0$, mielibyśmy do czynienia z implozją, czyli z występowaniem w zbiorze $\mathbf{Y}_m$ zbyt dużej (dla przyjętej konstrukcji estymatora) liczby punktów określanych mianem *inliers*, generowanych przez mechanizm zakłócający, będący rozkładem skoncentrowanym w punkcie.
Tak więc, dla przypadku estymatorów wielowymiarowych parametrów położenia i skali $(\boldsymbol{\mu}, \mathbf{\Sigma})$, m^* określające wartość punktu załamania ε_n^*, odpowiada największemu m dla którego istnieją dodatnie, skończone wartości a, b, c takie, że:

$$||\widehat{\boldsymbol{\mu}}_n(\mathbf{Y}_m)|| \leqslant a \text{ oraz } b \leqslant \lambda_p(\widehat{\mathbf{\Sigma}}_n(\mathbf{Y}_m)) \leqslant \lambda_1(\widehat{\mathbf{\Sigma}}_n(\mathbf{Y}_m)) \leqslant c \tag{5}$$

W większości wypadków ε_n^* nie zależy od zbioru X oraz zdąża do asymptotycznego punktu załamania (ang. *asymptotic breakdown point*, BP) przy $n \to \infty$.
Afiniczna ekwiwariantność wielowymiarowych estymatorów położenia i rozrzutu
Rozpatrzmy przekształcenie afiniczne zbioru $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, gdzie poszczególne $\mathbf{x}_i \in \mathbb{R}^p$:

$$\mathbf{AX} + \mathbf{v} = \{\mathbf{A}\mathbf{x}_1 + \mathbf{v}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{v}\} \tag{6}$$

gdzie $\mathbf{A}$ jest ustaloną nieosobliwą macierzą o wymiarach $p \times p$, a $\mathbf{v}$ ustalonym p-elementowym wektorem kolumnowym
Estymator wielowymiarowego położenia $\widehat{\boldsymbol{\mu}}_n$ nazywamy afinicznie ekwiwariantnym, wtedy i tylko wtedy, gdy:

$$\widehat{\boldsymbol{\mu}}_n(\mathbf{AX} + \mathbf{v}) = \mathbf{A}\widehat{\boldsymbol{\mu}}_n(\mathbf{X}) + \mathbf{v} \tag{7}$$

Estymator wielowymiarowego rozrzutu $\widehat{\mathbf{\Sigma}}_n$ nazywamy afinicznie ekwiwariantnym, wtedy i tylko wtedy, gdy:

$$\widehat{\mathbf{\Sigma}}_n(\mathbf{AX} + \mathbf{v}) = \mathbf{A}\widehat{\mathbf{\Sigma}}_n(\mathbf{X})\mathbf{A}' \tag{8}$$

Przy założeniu, że $\mathbf{X}$ znajduje się w położeniu ogólnym (ang. *general position*) w $\mathbb{R}^p$ (tzn. przy $\#\mathbf{X} = n \geqslant p + 1$, żadne z $p + 1$ punktów należących do $\mathbf{X}$ nie zawiera się w hiperpłaszczyźnie o wymiarze niższym niż p), można pokazać [Davies, 1987], iż maksymalny zastąpieniowy punkt załamania próby skończonej (FBP) dla afinicznie ekwiwariantnych estymatorów $\widehat{\mathbf{\Sigma}}_n$ wielowymiarowego rozrzutu wynosi $\varepsilon_n^*(\widehat{\boldsymbol{\theta}}_n, \mathbf{X}) = \frac{1}{n}\left\lfloor \frac{n-p+1}{2} \right\rfloor$.
Obciążoność afinicznie ekwiwariantnych estymatorów wielowymiarowego rozrzutu $\widehat{\mathbf{\Sigma}}_n$, w przypadku próby zanieczyszczonej $\mathbf{Y}_m$, można mierzyć za pomocą wskaźnika uwarunkowania[1]:

$$\text{bias}(\widehat{\mathbf{\Sigma}}_n, \mathbf{Y}_m) = \log\left(\frac{\lambda_1(\widehat{\mathbf{\Sigma}}_n, \mathbf{Y}_m)}{\lambda_p(\widehat{\mathbf{\Sigma}}_n, \mathbf{Y}_m)}\right) \tag{9}$$

[1] Jego postać wynika z uproszczenia bardziej ogólnego miernika obciążenia, w którym mając na uwadze afiniczną ekwiwariantność estymatora zastępuję się nieznaną wartość parametru $\mathbf{\Sigma}$ macierzą jednostkową $\mathbf{I}_p$

Tym samym punkt załamania próby skończonej (FBP) estymatora $\widehat{\boldsymbol{\Sigma}}_n$ można równoważnie zdefiniować z wykorzystaniem miary jego obciążenia:

$$\varepsilon_n^*(\widehat{\boldsymbol{\Sigma}}_n, \mathbf{X}) = \min_{\varepsilon \in (0,1)} \{\varepsilon : \sup_{\mathbf{Y}_m \in \mathcal{Y}_m} \text{bias}(\widehat{\boldsymbol{\Sigma}}_n, \mathbf{Y}_m) = \infty\} \tag{10}$$

Teoretyczne aspekty ekwiwariantnych estymatorów wielowymiarowego rozrzutu wygodniej jest badać w sposób analityczny dla przypadku asymptotycznego $n \to \infty$.
Przez asymptotyczną zgodność estymatora w niniejszej pracy rozumie się klasyczne pojęcie zgodności estymatora. Niech $\underline{\mathbf{X}_1^n} = (\underline{X_1}, \underline{X_2}, \ldots, \underline{X_n})$ będzie procesem *iid* n zmiennych losowych, estymator $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}(\underline{\mathbf{X}_1^n})$ rozważany jako proces losowy $\left(\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_n\right)$ jest asymptotycznie zgodnym estymatorem parametru $\boldsymbol{\theta}$, wtedy i tylko wtedy, gdy $\text{plim}_{n\to\infty} \widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}$ (proces $\widehat{\boldsymbol{\theta}}_n$ jest zbieżny według prawdopodobieństwa do wartości parametru $\boldsymbol{\theta}$).
W przypadku estymatorów odpornych w celu określenia własności asymptotycznych, często wykorzystuje się także podejście z gruntu teorii procesów empirycznych (ang. *Empirical Process Theory*). Dla procesu *iid* zmiennych losowych (możliwie wektorowych) $\left(\underline{X_1}, \underline{X_2}, \ldots, \underline{X_n}\right)$, rozpatruje się proces losowy dystrybuant empirycznych $(F_1(x), F_2(x), \ldots, F_n(x))$, gdzie

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(\underline{X_i}) \tag{11}$$

przy czym I jest funkcją wskaźnikową.
Estymatory rozważa się tutaj jako odpowiednie funkcjonały $\widehat{\boldsymbol{\theta}}(F_n)$.
Zgodnie z twierdzeniem Gliwenki-Cantelliego (ang. *Glivenko-Cantelli theorem*), przy $n \to \infty$ zachodzi jednostajna zbieżność dystrybuant empirycznych $F_n(x)$ do dystrybuanty teoretycznej $F(x)$, rozkładu (o parametrze $\boldsymbol{\theta}$) zmiennych $\underline{X_i}, i = 1, \ldots, n$:

$$\|F_n - F\|_\infty = \sup_x |F_n(x) - F(x)| \overset{p.n.}{\to} 0 \text{ przy } n \to \infty \tag{12}$$

Wykorzystując to podejście można określić własności asymptotyczne estymatora poprzez badanie własności funkcjonału $\widehat{\boldsymbol{\theta}}(F_\infty) = \widehat{\boldsymbol{\theta}}(F)$.
Estymator $\widehat{\boldsymbol{\theta}}(F_n)$ uznaje się za zgodny w sensie Fishera estymator parametru $\boldsymbol{\theta}$, gdy $\widehat{\boldsymbol{\theta}}(F) = \boldsymbol{\theta}$, przy czym F jest dystrybuantą teoretyczną rozkładu o parametrze $\boldsymbol{\theta}$.
W podobny sposób definiuje się asymptotyczny punkt załamania (ang. *asymptotic breakdown point*, BP) estymatora. Wykorzystuje się przy tym pojęcie zanieczyszczonego w stopniu ε sąsiedztwa (ang. ε *contamination neighborhood*) rozkładu parametrycznego z dystrybuantą F i parametrem $\boldsymbol{\theta}$, rozumiane jako zbiór dystrybuant:

$$\mathcal{F}(F, \varepsilon) = \{(1-\varepsilon)F + \varepsilon G : G \in \mathcal{G}\}, \tag{13}$$

$\mathcal{G}$ jest pewnym zbiorem dystrybuant rozkładów (zazwyczaj $\mathcal{G}$ jest zbiorem dystrybuant dowolnych rozkładów, albo zbiorem rozkładów skoncentrowanych w punkcie <ang. *point mass concentration*>).
Asymptotyczny punkt załamania estymatora $\widehat{\boldsymbol{\theta}}$ dla F, oznaczany $\varepsilon^*(\widehat{\boldsymbol{\theta}}, F)$ jest równy największemu $\varepsilon \in (0, 1)$, takiemu, że $\widehat{\boldsymbol{\theta}}_\infty ((1-\varepsilon)F + \varepsilon G)$ jako funkcjonał G pozostaje ograniczony oraz nie należy do brzegu zbioru $\boldsymbol{\Theta}$, będącego przestrzenią parametrów.

3 Afinicznie ekwiwariantne estymatory wielowymiarowego rozrzutu z wysokim punktem załamania

W niniejszym oraz kolejnych punktach pracy przybliżone zostaną wybrane aspekty dotyczące dwóch afinicznie ekwiwariantnych estymatorów wielowymiarowego rozrzutu:

- estymatora najmniejszego wyznacznika macierzy kowariancji, MCD,
- estymatora *Projection Congruent Subset*, PCS.

Wśród pozostałych afinicznie ekwiwariantnych estymatorów wielowymarowego położenia i skali można wymienić m.in. estymator elipsoidy minimalnej objętości MVE (ang. *Minimum Volume Ellipsoid estimate*, [Rousseeuw, 1984]), estymator SDE (ang. *Stahel-Donoho estimate*, [Stahel, 1981], [Donoho, 1982]), estymator M, estymator S [Davies, 1987], estymator P (ang. *P-estimate*, *Projection estimate*, [Maronna i in., 1992]). Przegląd dotyczący własności przywołanych estymatorów można znaleźć m.in. w pracy [Maronna, Martin, 2006].

3.1 Estymator minimalnego wyznacznika macierzy kowariancji (*Minimum Covariance Determinant*, MCD)

Estymator minimalnego wyznacznika macierzy kowariancji (ang. *Minimum Covariance Determinant*, MCD) można wyrazić jako rozwiązanie poniższego problemu optymalizacyjnego z ograniczeniami:

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \text{argmin}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in(\mathbb{R}^p, PDS(p))}|\boldsymbol{\Sigma}| \tag{14}$$

pod warunkiem

$$\frac{1}{hp}\sum_{i=1}^{h} d^2_{MD,(i)} = \frac{1}{hp}\sum_{i=1}^{h}(\mathbf{x}_{(i)} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{(i)} - \boldsymbol{\mu}) = 1 \tag{15}$$

gdzie $d^2_{MD,(i)}$ oznacza i-tą statystykę porządkową dla kwadratu odległości Mahalonobisa, a $\mathbf{x}_{(i)}$ to wektor obserwacji wykorzystany w konstrukcji statystyki porządkowej $d^2_{MD,(i)}$
Zgodnie z przyjętym warunkiem ograniczającym wartość estymatora wielowymiarowego rozrzutu musi odpowiadać próbkowej macierzy kowariancji wyznaczonej dla określonego h-elementowego podzbioru $\mathbf{X}$.
Z geometrycznego punktu widzenia w metodzie MCD poszukuje się elipsoidy $E = \{\mathbf{x} \in \mathbb{R}^p : d^2_{MD}(\mathbf{x}, \mathbf{t}, \mathbf{V}) \leqslant 1\}$, obejmującej co najmniej h obserwacji z $\mathbf{X}$: $\#\{i : 1 \leqslant i \leqslant n \wedge \mathbf{x}_i \in \mathbf{X} \cap E\} \geqslant h$, dla których macierz kowariancji ma najmniejszy wyznacznik.
Dla rzeczonej elipsoidy $\mathbf{t} = \widehat{\boldsymbol{\mu}}$ oraz $\mathbf{V} = d^2_{MD,(h)} \cdot \hat{\boldsymbol{\Sigma}}$, przy czym $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ są wartościami estymatora MCD, a $d^2_{MD,(h)}$ jest h-tą statystyką porządkową dla kwadratów odległości Mahalanobisa wyznaczonych przy parametrach $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, dla obserwacji ze zbioru $\mathbf{X}$.
Jako, że problemu optymalizacyjnego dla MCD nie da się rozwiązać w sposób analityczny, zaproponowano procedurę iteracyjną, której podstawą są tzw. kroki koncentracji (ang. *concentration steps*, *C-steps*).
I tak w danej l-tej iteracji podstawą wyznaczania wartości $\left(\widehat{\boldsymbol{\mu}}^{(l)}, \widehat{\boldsymbol{\Sigma}}^{(l)}\right)$ jest h obserwacji z $\mathbf{X}$ o najmniejszych odległościach Mahalonobisa dla wartości $\left(\widehat{\boldsymbol{\mu}}^{(l-1)}, \widehat{\boldsymbol{\Sigma}}^{(l-1)}\right)$ wyznaczonych w poprzednim kroku.
Niech $H^{(l)}$ będzie zbiorem indeksów obserwacji wykorzystywanych przy wyznaczaniu wartości $\left(\widehat{\boldsymbol{\mu}}^{(l)}, \widehat{\boldsymbol{\Sigma}}^{(l)}\right)$ w l-tej iteracji:

$$H^{(l)} = \left\{i \in \{1, \ldots, n\} : d^2_{MD,i}\left(\widehat{\boldsymbol{\mu}}^{(l-1)}, \widehat{\boldsymbol{\Sigma}}^{(l-1)}\right) \leqslant d^2_{MD,(h)}\left(\widehat{\boldsymbol{\mu}}^{(l-1)}, \widehat{\boldsymbol{\Sigma}}^{(l-1)}\right)\right\} \tag{16}$$

gdzie $d^2_{MD,i}\left(\widehat{\boldsymbol{\mu}}^{(l-1)}, \widehat{\boldsymbol{\Sigma}}^{(l-1)}\right)$ są kwadratami odległości Mahalanobisa dla punktów $\mathbf{x}_i, i = 1, \ldots, n$, wyznaczonymi przy parametrach $\left(\widehat{\boldsymbol{\mu}}^{(l-1)}, \widehat{\boldsymbol{\Sigma}}^{(l-1)}\right)$, a $d^2_{MD,(h)}\left(\widehat{\boldsymbol{\mu}}^{(l-1)}, \widehat{\boldsymbol{\Sigma}}^{(l-1)}\right)$ jest h-tą statystyką porządkową dla tych kwadratów odległości.
Wartości $\left(\widehat{\boldsymbol{\mu}}^{(l)}, \widehat{\boldsymbol{\Sigma}}^{(l)}\right)$ w l-tej iteracji wyznaczane są jako próbkowa wektorowa średnia oraz macierz kowariancji dla h-elementów zbioru $\mathbf{X}$ o indeksach w zbiorze $H^{(l)}$:

$$\left(\widehat{\boldsymbol{\mu}}^{(l)}, \widehat{\boldsymbol{\Sigma}}^{(l)}\right) = (ave_{i\in H^{(l)}}\mathbf{x}_i, cov_{i\in H^{(l)}}\mathbf{x}_i) \tag{17}$$

W pracy [Rousseeuw, Driessen, 1999] wykazano, że takie postępowanie prowadzi do uzyskiwania w kolejnych iteracjach macierzy kowariancji o niewiększych wyznacznikach, niż te dla poprzedniej iteracji, tzn. $\left|\widehat{\boldsymbol{\Sigma}}^{(l)}\right| \leqslant \left|\widehat{\boldsymbol{\Sigma}}^{(l-1)}\right|$.
Powyższe iteracje nazywane krokami koncentracji wykonuje się, aż do momentu uzyskania zbieżności. Osiągnięcie zbieżności algorytmu następuje w sytuacji, gdy $\left|\widehat{\boldsymbol{\Sigma}}^{(L)}\right| = 0$ albo $\left|\widehat{\boldsymbol{\Sigma}}^{(L)}\right| = \left|\widehat{\boldsymbol{\Sigma}}^{(L-1)}\right|$,

wtedy wykonywanie kolejnych kroków koncentracji nie skutkuje zmniejszaniem się wartości funkcji celu $|\mathbf{\Sigma}|$. Jako, że $\left|\widehat{\mathbf{\Sigma}}^{(1)}\right| \geqslant \left|\widehat{\mathbf{\Sigma}}^{(2)}\right| \geqslant \left|\widehat{\mathbf{\Sigma}}^{(3)}\right| \geqslant \ldots$ jest ciągiem nieujemnym, musi być on ciągiem zbieżnym.

Warunkiem koniecznym istnienia minimum globalnego $\widehat{\mathbf{\Sigma}}$ funkcji celu estymatora MCD dla próby $\mathbf{X}$, w punkcie $\widehat{\mathbf{\Sigma}}^{(L)}$ jest $\left|\widehat{\mathbf{\Sigma}}^{(L)}\right| = \left|\widehat{\mathbf{\Sigma}}^{(L-1)}\right|$, nie jest to jednak warunek wystarczający, gdyż procedura charakteryzuje się lokalną zbieżnością do minimum globalnego.

W związku z powyższym procedurę rozpoczyna się z różnych punktów startowych i prowadzi się w każdym przypadku, aż do osiągnięcia zbieżności. Jako wartość estymatora MCD przyjmuje się tą spośród wartości końcowych procedury, dla której wyznacznik macierzy kowariancji (wyznaczonej w dla h-elementowej podpróby) jest najmniejszy.

Wybrane własności estymatora MCD

Dla estymatora wielowymiarowego rozrzutu MCD, w przypadku $\mathbf{X}$ znajdującego się w położeniu ogólnym, punkt załamania próby skończonej wynosi $\varepsilon_n^*(\widehat{\mathbf{\Sigma}}, \mathbf{X}) = \frac{n-h+1}{n}$ i przyjmuje maksymalną możliwą wartość dla estymatorów afinicznie ekwiwariantnych przy $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$.

Afiniczną ekwiwariantność estymatora wielowymiarowego rozrzutu MCD, jako empirycznej macierzy kowariancji z optymalnego h-elementowego podzbioru $\mathbf{X}_{H^*}$ zbioru $\mathbf{X}$, można łatwo pokazać, gdyż dla próbkowej macierzy kowariancji $\mathrm{cov}(\mathbf{X}_{H^*})$ oraz nieosobliwej macierzy kwadratowej $\mathbf{A}$ zachodzi $\mathrm{cov}(\mathbf{A}\mathbf{X}_{H^*}) = \mathbf{A}\mathrm{cov}(\mathbf{X}_{H^*})\mathbf{A}'$. Ponadto w związku z tym, że $|\mathbf{A}\mathrm{cov}(\mathbf{X}_{H^*})\mathbf{A}'| = |\mathbf{A}|^2 |\mathrm{cov}(\mathbf{X}_{H^*})|$, ponieważ $|\mathbf{A}|^2 > 0$, funkcja celu MCD przyjmuje minimum dla macierzy kowariancji wyznaczonej w oparciu o przekształcony afinicznie h-elementowy podzbiór $\mathbf{X}_{H^*}$, dla którego próbkowa macierz kowariancji dla danych nieprzekształconych charakteryzowała się minimalnym wyznacznikiem .

Zakładając wielowymiarowy model normalny generujący dane $\mathbf{x} \sim \mathbf{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, w przypadku braku mechanizmu zakłócającego, można pokazać, że MCD nie jest estymatorem zgodnym (zarówno w sensie asymptotycznym, jak i w sensie Fishera) wielowymiarowego rozrzutu.

Przy powyższym założeniu w celu osiągnięcia asymptotycznej zgodności estymatora MCD wielowymiarowego rozrzutu, jako że $d_{MD}^2(\mathbf{x}, \boldsymbol{\mu}, \mathbf{\Sigma}) \sim \chi^2(p)$, stosuje się korektę na zgodność (ang. *consistency correction*) w postaci mnożnika c dla $\widehat{\mathbf{\Sigma}}$, który można odpornie estymować jako: $\hat{c} = \frac{\mathrm{med}\{d_{MD}^2(\mathbf{x}_1, \widehat{\boldsymbol{\mu}}, \widehat{\mathbf{\Sigma}}), \ldots, d_{MD}^2(\mathbf{x}_n, \widehat{\boldsymbol{\mu}}, \widehat{\mathbf{\Sigma}})\}}{\chi_{0,5,p}^2}$, przy czym med jest empiryczną medianą kwadratów odległości Mahalanobisa przy wartościach $\left(\widehat{\boldsymbol{\mu}}, \widehat{\mathbf{\Sigma}}\right)$ estymatora MCD, a $\chi_{0,5,p}^2$ jest medianą teoretycznego rozkładu chi-kwadrat z p stopniami swobody.

Wyprowadzenie mnożnika – korekty na asymptotyczną zgodność przy wielowymiarowym rozkładzie normalnym można znaleźć w pracy [Maronna, Martin, 2006, s. 186].

Jako, że estymator MCD wielowymiarowego rozrzutu, pomimo zastosowania przytoczonej poprawki na asymptotyczną zgodność, jest w skończonej próbie estymatorem obciążonym stosuje się dodatkowo (wyznaczaną w sposób symulacyjny, przy założeniu wielowymiarowego rozkładu normalnego bez zakłóceń) korektę dla skończonej próby (ang. *finite sample correction*), usuwającą jego obciążenie. Procedurę symulacyjnego wyznaczania wartości rozważanej poprawki przedstawiono w pracy [Pison i in., 2002].

Można również rozpatrywać zgodność estymatora MCD w sensie Fishera. Na gruncie teorii procesów empirycznych, pokazano, iż dla procesu *iid* zmiennych losowych, ciąg ich dystrybuant empirycznych (rozpatrywanych jako zmienne losowe) jest zbieżny do dystrybuanty teoretycznego rozkładu poszczególnych zmiennych tworzących proces *iid*.

Tak więc w tym podejściu estymator MCD, asymptotycznie rozpatruje się jako funkcjonał dystrybuanty określonego rozkładu teoretycznego.

Rozpatrzmy więc estymator MCD parametrów $(\boldsymbol{\mu}, \mathbf{\Sigma})$ dla eliptycznie symetrycznego, jednomodalnego rozkładu wektora losowego $\underline{\mathbf{X}}$ z dystrybuantą teoretyczną $F_{\boldsymbol{\mu},\mathbf{\Sigma}}$.

Jako, że estymator MCD jest afinicznie ekwiwariantny, bez utraty ogólności można dokonać przekształcenia wektora losowego $\underline{\mathbf{X}}$:

$$\mathbf{Z} = \mathbf{\Sigma}^{-1/2}(\underline{\mathbf{X}} - \boldsymbol{\mu}) \tag{18}$$

Powstały wektor losowy $\mathbf{Z}$ (przez $\mathbf{z}$ oznacza się jego realizacje) ma rozkład tego samego typu co wektor losowy $\underline{\mathbf{X}}$, z parametrami $(\mathbf{0}, \mathbf{I})$ oraz dystrybuantą $F_{\mathbf{0},\mathbf{I}}$. Wektor losowy $\mathbf{Z}$ składa się nieskorelowanych zmiennych o jednakowym jednowymiarowym rozkładzie z zerową wartością oczekiwaną i jednostkową wariancją. Dla oznaczenia pojedynczej zmiennej w ramach wektora losowego przyjmuje się symbol Z_1 (przez z_1 oznacza się jej realizację).

W pracy [Butler i in., 1993] pokazano, że problem optymalizacyjny definiujący estymator MCD, dla rozkładu teoretycznego $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ ma jedyne rozwiązanie zadane przez elipsoidę $E(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})$:

$$E(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = \{\mathbf{x} \in \mathbb{R}^p : (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leqslant q_\gamma\} \tag{19}$$

gdzie przy $G(t) = P_{F_{\mathbf{0},\mathbf{I}}}(\mathbf{Z}'\mathbf{Z} \leqslant t)$, $q_\gamma = G^{-1}(1-\gamma)$ jest kwantylem $1-\gamma$ dla rozkładu G, natomiast γ jest asymptotycznym odpowiednikiem odsetka $\frac{n-h}{n}$ obserwacji nieuwzględnianych przy wyznaczaniu wartości estymatora dla próby skończonej
Z elipsoidą związane są estymatory MCD $\widehat{\boldsymbol{\mu}}(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})$ oraz $\widehat{\boldsymbol{\Sigma}}(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})$, które można zadać z wykorzystaniem funkcjonałów standaryzowanego rozkładu teoretycznego $F_{\mathbf{0},\mathbf{I}}$ oraz uwzględnieniem własności afinicznej ekwiwariantności MCD:

$$\widehat{\boldsymbol{\mu}}(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = \frac{1}{1-\gamma}\int_{\mathbf{z}'\mathbf{z}\leqslant q_\gamma} \mathbf{z}\,\mathrm{d}F_{\mathbf{0},\mathbf{I}}(\mathbf{z}) + \boldsymbol{\mu} = \mathbf{0} + \boldsymbol{\mu} = \boldsymbol{\mu} \tag{20}$$

$$\widehat{\boldsymbol{\Sigma}}(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = \left(\frac{c_\gamma}{1-\alpha}\int_{\mathbf{z}'\mathbf{z}\leqslant q_\gamma} \mathbf{z}_1^2\,\mathrm{d}F_{\mathbf{0},\mathbf{I}}(\mathbf{z})\right)\boldsymbol{\Sigma} \tag{21}$$

przy czym $\widehat{\boldsymbol{\Sigma}}(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})$ będzie zgodnym w sensie Fishera estymatorem wielowymiarowego parametru rozrzutu $\boldsymbol{\Sigma}$, gdy przy założonym modelu generującym dane, poprawka c_γ będzie wynosić:

$$c_\gamma = \frac{1-\gamma}{\int_{\mathbf{z}'\mathbf{z}\leqslant q_\gamma} \mathbf{z}_1^2\,\mathrm{d}F_{\mathbf{0},\mathbf{I}}(\mathbf{z})} \tag{22}$$

Rozważania teoretyczne dotyczące własności asymptotycznych estymatora MCD rozumianych w sensie Fishera oraz wartości liczbowe poprawki c_γ dla przypadku wielowymiarowego rozkładu normalnego (bez mechanizmu zakłócającego) przedstawiono w pracy [Croux, Haesbroeck, 1999].

3.2 Estymator Projection Congruent Subset, PCS

Inaczej niż w przypadku estymatora MCD, który wyraża się jako rozwiązanie problemu optymalizacyjnego bezpośrednio względem $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, w celu wyznaczenia wartości estymatora PCS uprzednio poszukuje się pozbawionych obserwacji odstających podprób, rozumianych jako podzbiory $\mathbf{X}$ charakteryzujące się spójnością, której brak mierzony jest poprzez kryterium inkongruencji podlegające minimalizacji. Podzbiór $\mathbf{X}$ o h elementach, minimalizujący kryterium inkongruencji jest podstawą wyznaczania wartości estymatorów PCS, jako średniej arytmetycznej oraz macierzy kowariancji ze wspomnianej podpróby.
W celu zdefiniowania kryterium inkongruencji wykorzystuje się pojęcie odległości punktu od hiperpłaszczyzny oraz wektora do niej normalnego.
Dla zbioru $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{R}^p$ znajdującego się w położeniu ogólnym rozważa się $p-1$-wymiarowe hiperpłaszczyzny rozpinane przez p punktów, tworzących określone podzbiory $\mathbf{X}$.
Wektor normalny do hiperpłaszczyzny zadanej wzorem $\mathbf{a}'_{mk}\mathbf{x} - 1 = 0$, zawierającej w sobie pewien p-elementowy podzbiór punktów zbioru $\mathbf{X}$, to wektor $\mathbf{a}_{mk}$ do niej ortogonalny, taki że:

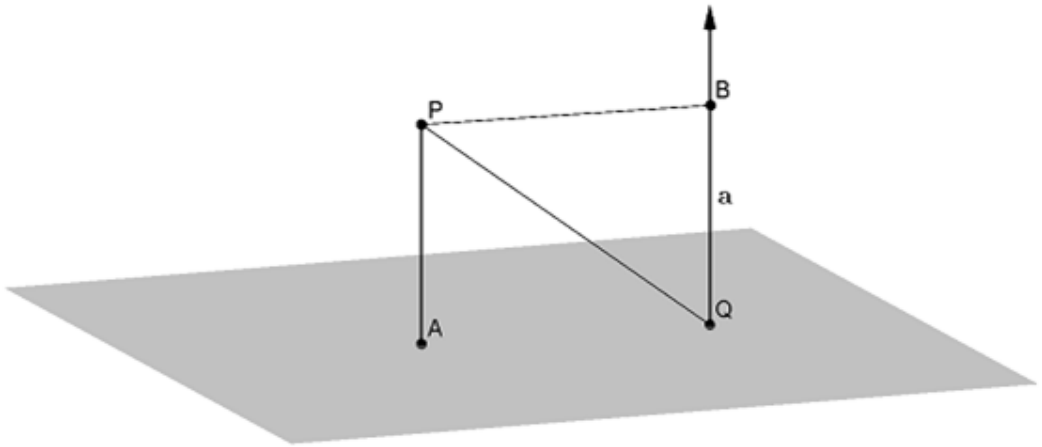
$$\{\mathbf{a}_{mk} : \mathbf{X}_{mk}\mathbf{a}_{mk} = \mathbf{1}_p\} \tag{23}$$

gdzie

- $\mathbf{X}_{mk}$ to macierz o wymiarach $p \times p$ składająca się z ułożonych wierszami p-elementowych wektorów (obserwacji) należących do $\mathbf{X}$
- $\mathbf{1}_p$ to p-wymiarowy wektor kolumnowy złożony wyłącznie z jedynek
- $\mathbf{a}_{mk}$ to wektor normalny do $p-1$wymiarowej hiperpłaszczyzny, do której należą punkty z $\mathbf{X}_{mk}$ (rozpinanej przez punkty z $\mathbf{X}_{mk}$)

Rozpatruje się odległość wektora $\mathbf{x}_i$ od hiperpłaszczyzny z wektorem normalnym $\mathbf{a}_{mk}$ (której to odległości odpowiada długość rzutu ortogonalnego wektora $\mathbf{x}_i$ na kierunek wektora $\mathbf{a}_{mk}$, o początku należącym do hiperpłaszczyzny), której kwadrat wyrażony jest przez:

$$d^2_{P,i}(\mathbf{a}_{mk}) = \frac{(\mathbf{x}'_i\mathbf{a}_{mk} - 1)^2}{||\mathbf{a}_{mk}||^2} \tag{24}$$



Rysunek 1: Odległość punktu P od płaszczyzny z wektorem normalnym $\mathbf{a}$, wykorzystywana w ramach procedury wyznaczania wartości estymatora PCS (punkt B, jest rzutem prostopadłym punktu P na wektor normalny $\mathbf{a}$, punkt Q jest dowolnym punktem płaszczyzny, punkt A jest rzutem prostopadłym punktu P na płaszczyznę, stąd rozważana odległość wynosi $|PA| = |BQ|$) (opracowanie własne)

gdzie $|d_{P,i}(\mathbf{a}_{mk})|$ to prostopadła odległość wektora $\mathbf{x}_i$ od $p-1$- wymiarowej hiperpłaszczyzny o wektorze normalnym $\mathbf{a}_{mk}$
Niech

- H_m oznacza podzbiór indeksów h obserwacji (punktów przestrzeni $\mathbb{R}^p$) ze zbioru $\mathbf{X}$, spośród których p punktów rozpina hiperpłaszczyznę o wektorze normalnym $\mathbf{a}_{mk}$: $H_m \subset \{1,\ldots,n\}$, $\#H_m = h \geqslant \lfloor \frac{n+p+1}{2} \rfloor$,
- $m = 1,\ldots,M$ to subskrypt numerujący h-elementowe podzbiory indeksów obserwacji z $\mathbf{X}$, przy czym $M = \left(\begin{array}{c} n \\ h \end{array} \right)$ (jednak w praktyce liczba rozważanych h-elementowych podzbiorów jest $< \left(\begin{array}{c} n \\ h \end{array} \right)$),
- H_{mk} to podzbiór indeksów h obserwacji (punktów) z całego zbioru $\mathbf{X}$, o najmniejszych kwadratach odległości od hiperpłaszczyzny z wektorem normalnym $\mathbf{a}_{mk}$:

$$H_{mk} = \left\{ i \in \{1,\ldots,n\} : d^2_{P,i}(\mathbf{a}_{mk}) \leqslant d^2_{P,(h)}(\mathbf{a}_{mk}) \right\} \tag{25}$$

  gdzie $d^2_{P,(h)}$ jest h-tą statystyką porządkową kwadratu odległości od hiperpłaszczyzny z wektorem normalnym $\mathbf{a}_{mk}$, dla wszystkich obserwacji $\mathbf{x}_i, i = 1,\ldots,n$ ze zbioru $\mathbf{X}$.

Wskaźnik inkongruencji H_m wzdłuż kierunku $\mathbf{a}_{mk}$ (ang. H_m *incongruence index along* $\mathbf{a}_{mk}$) wyrażony jest następująco:

$$I(H_m, \mathbf{a}_{mk}) = \log \left(\frac{\text{ave}_{i \in H_m} d^2_{P,i}(\mathbf{a}_{mk})}{\text{ave}_{i \in H_{mk}} d^2_{P,i}(\mathbf{a}_{mk})} \right) \tag{26}$$

Wskaźnik $I(H_m, \mathbf{a}_{mk})$ zawsze przyjmuje wartości dodatnie. Jeżeli rzuty punktów o indeksach ze zbioru H_m na kierunek $\mathbf{a}_{mk}$ (przy czym punkt początkowy wektora $\mathbf{a}_{mk}$ należy do hiperpłaszczyzny), w dużej części pokrywają się z rzutami punktów ze zbioru H_{mk} na ten kierunek, wartości indeksu będą niskie. Innymi słowy, wskaźnik inkongruencji $I(H_m, \mathbf{a}_{mk})$ przyjmuje niskie wartości, gdy h punktów o indeksach w zbiorze H_m skupia się w pobliżu hiperpłaszczyzny z wektorem

normalnym $\mathbf{a}_{mk}$ (rozpinanej przez podzbiór p punktów, spośród h o indeksach w H_m), czyli ich odległość od tej hiperpłaszczyzny[2] jest mniejsza w porównaniu z odległościami pozostałych punktów z $\mathbf{X}$, o indeksach nie należących do H_m.
Wartość $I(H_m, \mathbf{a}_{mk})$ wskaźnika inkongruencji H_m wzdłuż kierunku $\mathbf{a}_{mk}$, można także rozważać jako miernik stopnia, w jakim pokrywają się zbiory H_m (zawierający indeksy rozważanej h-elementowej podpróby $\mathbf{X}$) oraz H_{mk} (zawierający indeksy h punktów z całego zbioru $\mathbf{X}$ położonych najbliżej hiperpłaszczyzzny z wektorem normalnym $\mathbf{a}_{mk}$), uwzględniający równocześnie położenie względem rozważanej hiperpłaszczyzny punktów odnoszących się do zestawianych podzbiorów.
Punkty położone najbliżej hiperpłaszczyzny z wektorem normalnym $\mathbf{a}_{mk}$, nie związane ze zbiorem H_m, czyli te o indeksach w zbiorze $H_{mk} - H_m$, wpływają (poprzez swoje odległości) na obniżenie w $I(H_m, \mathbf{a}_{mk})$ wartości mianownika ułamka pod logarytmem, nie mając jednocześnie wpływu na wartość jego licznika. W przypadku, gdy zbiór indeksów H_m odnosi się do spójnego („niezanieczyszczonego") h-elementowego podzbioru punktów z $\mathbf{X}$, punkty o indeksach w H_m koncentrują się w pobliżu hiperpłaszczyzn (związanych z wektorami normalnymi $\mathbf{a}_{mk}$), rozpinanych przez kombinacje p punktów z tego h-elementowego podzbioru, co odzwierciedla się w niskich wartościach wskaźnika inkongruencji.
Natomiast w sytuacji, gdy indeksy w H_m odnoszą się do h obiektów z niespójnej podpróby, zawierającej oprócz obserwacji pochodzących z głównego rozkładu, także obserwacje odstające (wynikające z działania mechanizmu zakłócającego), zbiór H_m będzie miał mniej liczną część wspólną ze zbiorem H_{mk}, indeksów h punktów położonych najbliżej hiperpłaszczyzny rozpinanej przez kombinacje p-elementowe punktów z niespójnego podzbioru. Przekładać będzie się to na wyższe wartości wskaźnika inkongruencji dla takich podzbiorów.
W celu usunięcia wpływu konkretnego kierunku $\mathbf{a}_{mk}$ na wartość indeksu inkongruencji wyznaczanego dla zbioru H_m, rozważa się średnią względem wielu kierunków.
Wskaźnik inkongruencji dla H_m definiuje się jako następującą średnią:

$$I(H_m) = \text{ave}_{\mathbf{a}_{mk} \in B(H_m)} I(H_m, \mathbf{a}_{mk}) \tag{27}$$

gdzie $B(H_m)$ jest zbiorem wszystkich kierunków $\mathbf{a}_{mk}$ ortogonalnych do hiperpłaszczyzn rozpinanych przez różne kombinacje p punktów (obserwacji) należących do h-elementowego podzbioru punktów (obserwacji) o indeksach w H_m, $\#B(H_m) = \begin{pmatrix} h \\ p \end{pmatrix}$.
W praktyce w celu ograniczenia liczby wykonywanych operacji obliczeniowych nie rozpatruje się wszystkich $\#B(H_m)$ kierunków $\mathbf{a}_{mk}$, ale ich losowo wybrany K-elemetowy podzbiór $\tilde{B}(H_m)$.
Podzbiór H^* o minimalnej wartości $I(H_m)$ nazywa się *projection congruent subset*:

$$H^* = \arg \min_{\{H_m\}_{m=1}^{M}} I(H_m) \tag{28}$$

Wartość estymatora PCS wielowymiarowego położenia i skali, odpowiada wektorowej średniej arytmetycznej oraz macierz kowariancji, wyznaczonej w oparciu o obserwacje o indeksach ze zbioru H^*:

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = (\text{ave}_{i \in H^*} \mathbf{x}_i, \text{cov}_{i \in H^*} \mathbf{x}_i) \tag{29}$$

Wybrane własności estymatora PCS
Dowód afinicznej ekwiwariantności estymatora PCS przedstawiono w pracy [Schmitt i in., 2014].
W tej samej pracy wykazano, iż punkt załamania próby skończonej (FBP) dla estymatora PCS, w przypadku, gdy $n > p+1 > 2$ oraz $\mathbf{X}$ znajduje się w położeniu ogólnym wynosi $\varepsilon_n^*(\widehat{\boldsymbol{\Sigma}}, \mathbf{X}) = \frac{n-h+1}{n}$ i równy jest maksymalnej możliwej wartości dla estymatorów afinicznie ekwiwariantnych, gdy $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$.
Ocena stopnia obciążenia estymatora PCS wielowymiarowego rozrzutu, dla przypadku skończonej próby wygenerowanej przez rozkład główny F z mechanizmem (rozkładem) G zakłócającym w stopniu ε, tzn. $F_\varepsilon = (1-\varepsilon)F(\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F) + \varepsilon G(\boldsymbol{\mu}_G, \boldsymbol{\Sigma}_G)$, ma charakter empiryczny (symulacyjny). Dla zanieczyszczonej próby obciążenie estymatora PCS, będzie zależeć od wymiaru przestrzeni danych, stopnia zanieczyszczenia próby oraz położenia punktów odstających generowanych przez mechanizm zakłócający.
Niech $\mathbf{X}_\varepsilon$ oznacza n-elementową próbę (zbiór), w której odsetek obserwacji odstających, będących wynikiem działania mechanizmu zakłócającego G wynosi ε. Próbę $\mathbf{X}_\varepsilon$ można wyrazić jako sumę $\mathbf{X}_\varepsilon = \mathbf{X}_F \cup \mathbf{X}_G$, gdzie $\mathbf{X}_F$ jest zbiorem obserwacji pochodzących z rozkładu głównego F, a $\mathbf{X}_G$

[2] Odległość p punktów rozpinających hiperpłaszczyznę jest zerowa, gdyż punkty te przynależą do niej.

grupuje obserwacje odstające wygenerowane przez G. Ponadto I_F, I_G to zbiory obejmuje indeksy elementów próby należących odpowiednio do $\mathbf{X}_F$, $\mathbf{X}_G$.
W pracy [Vakili, Schmitt, 2014] wskazano dla estymatorów afinicznie ekwiwariantnych wielowymiarowego rozrzutu sytuacje, skutkujące przy określonych założeniach dotyczących obserwacji odstających, najwyższym możliwym poziomem obciążenia[3] .
Niech $\nu = \min_{i \in I_G} \sqrt{\frac{d^2_{MD,i}(\widehat{\boldsymbol{\mu}}_F, \widehat{\boldsymbol{\Sigma}}_F)}{\chi^2_{0,99,p}}}$ będzie miernikiem odległości pomiędzy obserwacjami z rozkładu głównego a obserwacjami odstającymi.
Rozpatrzono trzy przypadki dla afinicznie ekwiwariantnych estymatorów wielowymiarowego rozrzutu:

- rozkład zakłócający przesunięty względem głównego:

Przy założeniu ustalonej odległości ν oraz $|\boldsymbol{\Sigma}_G| \geqslant |\boldsymbol{\Sigma}_F|$, estymator afinicznie ekwiwariantny jest najbardziej obciążony, gdy $|\boldsymbol{\Sigma}_G| = |\boldsymbol{\Sigma}_F|$, a $\boldsymbol{\mu}_G$ przyjmuje wartość, przy której odległość wynosi ν.

- rozkład zakłócający skoncentrowany w punkcie (*point-mass configuration*):

znosząc założenie $|\boldsymbol{\Sigma}_G| \geqslant |\boldsymbol{\Sigma}_F|$, estymator afinicznie ekwiwariantny będzie charakteryzował się największym obciążeniem, gdy $|\boldsymbol{\Sigma}_G| = 0$, tzn. rozkład zakłócający jest skoncentrowany w punkcie,

- znosząc dodatkowo założenie ustalonego ν, największe obciążenie estymatora będzie obserwowane przy $\boldsymbol{\mu}_G = \boldsymbol{\mu}_F$, przykładem tutaj jest przypadek zanieczyszczenia typu „Barrow wheel" (*Barrow wheel contamination*), wprowadzonego w funkcji `rbwheel` pakietu R o nazwie `robustX` [Stahel, Maechler, 2009].

Symulacyjne badanie wpływu na uzyskane wartości estymatora PCS, przytoczonych typów zakłóceń, w zależności od liczebności n próby oraz wymiaru p przestrzeni zmiennych można znaleźć w rzeczonej pracy [Vakili, Schmitt, 2014].

3.3 Ponownie ważone estymatory (*reweighted estimates*) MCD i PCS

Często stosowanym zabiegiem względem uzyskanych wartości estymatorów MCD lub PCS jest zastosowanie procedury ponownego ważenia (ang. *reweighting*) obserwacji.
W ramach procedury ponownego ważenia tworzony jest system wag określony jako funkcja kwadratów odległości Mahalanobisa obiektów z $\mathbf{X}$, przy parametrach równych oszacowaniom $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ MCD bądź PCS. System wag wykorzystuje się przy wyznaczaniu ważonej wektorowej średniej oraz ważonej macierzy kowariancji, które przyjmuje się jako wartości ponownie ważonych estymatorów MCD bądź PCS.
Najczęściej stosuje się dwa systemy wag:
– *hard-rejection weight system*,
– *soft-rejection weight system*.
Funkcję wag w ramach systemu *hard-rejection* wyraża się następująco:

$$w_{HR} : \mathbb{R}_+ \to \mathbb{R}_+$$

$$w_{HR}(z) = I(z \leqslant k) \tag{30}$$

gdzie I jest funkcją wskaźnikową.
Za argument z funkcji wag w_{HR} przyjmuje się kwadrat odległości Mahalanobisa $d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$. Natomiast jako wartość progową k przyjmuje się:

$$k = \hat{c} \cdot \chi^2_{1-\beta,p} = \frac{\text{med}^n_{i=1} d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})}{\chi^2_{0,5,p}} \chi^2_{1-\beta,p} \tag{31}$$

gdzie $\hat{c}$ jest korektą na zgodność asymptotyczną przy wielowymiarowym rozkładzie normalnym, $\chi^2_{0,5,p}$ medianą rozkładu chi-kwadrat z p stopniami swobody, $\chi^2_{1-\beta,p}$ kwantylem $1-\beta$ tegoż rozkładu, zazwyczaj przyjmuje się dla β wartości 0,05 lub 0,025, a med jest empiryczną medianą spośród kwadratów odległości Mahalanobisa $d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}), i = 1, \ldots, n$

[3] Mierzonym za pomocą przytoczonego w rozdziale 2. pracy miernika bias$(\widehat{\boldsymbol{\Sigma}}, \mathbf{X}_\varepsilon)$ wykorzystującego wartości własne macierzy $\widehat{\boldsymbol{\Sigma}}$.

W sytuacji, gdy obserwacje z $\mathbf{X}$ generowane są przez niezakłócony p-wymiarowy rozkład normalny, można się spodziewać, że powyżej tak wyznaczonego progu dla odległości k, będzie znajdować się w przybliżeniu $n\beta$ obserwacji z $\mathbf{X}$.
Ponownie ważone wartości $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ afinicznie ekwiwariantnych estymatorów MCD bądź PCS wyznacza się według wzorów:

$$\tilde{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n} w\left(d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})\right) \mathbf{x}_i}{\sum_{i=1}^{n} w\left(d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})\right)} \tag{32}$$

$$\tilde{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{n} w\left(d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})\right) (\mathbf{x}_i - \tilde{\boldsymbol{\mu}}) (\mathbf{x}_i - \tilde{\boldsymbol{\mu}})'}{\sum_{i=1}^{n} w\left(d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})\right)} \tag{33}$$

Tym samym w przypadku systemu wag typu *hard-rejection* wyznaczenie ważonej wektorowej średniej sprowadza się do wyznaczenia próbkowej wektorowej średniej oraz kowariancji, w oparciu o obserwacje z $\mathbf{X}$, których indeksy należą do zbioru:

$$\tilde{H} = \left\{ i \in \{1, \ldots, n\} : d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) \leqslant \frac{\mathrm{med}^n_{i=1} d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})}{\chi^2_{0,5,p}} \chi^2_{1-\beta,p} \right\} \tag{34}$$

Drugi rozważany tutaj system typu *soft-rejection,* wagi reprezentuje jako funkcję:

$$w_{SR} : \mathbb{R}_+ \to \mathbb{R}_+$$

$$w_{SR}(z) = \min\left\{1, \frac{k}{z}\right\} \tag{35}$$

gdzie jako argument z przyjmuje się kwadrat odległości Mahalanobisa $d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, a za k najczęściej tą samą wartość progową rozważaną w ramach systemu *hard-rejection.*
W przypadku systemu *soft-rejection* obserwacje z $\mathbf{X}$, których odległości przekraczają próg k, przy wyznaczaniu ważonej wektorowej średniej i kowariancji, są uwzględniane z wagami dodatnimi, odwrotnie proporcjonalnymi do odległości $d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, a nie zerowymi jak to było w przypadku systemu *hard-rejection.*
Ponownie ważone estymatory MCD i PCS zachowują własność afinicznej ekwiwariantności (wynika to z afinicznej niezmienniczości odległości Mahalanobisa) oraz wysoki punkt załamania „wyjściowych" estymatorów. Wysoki punkt załamania zostaje zachowany, gdyż w sytuacji, gdy liczba obserwacji skutkujących niedopuszczalnymi wartościami estymatorów (nieskończonymi bądź należącymi do brzegu przestrzeni parametrów) nie przekracza wartości $n-h$ założonej dla „wyjściowego" estymatora MCD lub PCS, obserwacje te nie są uwzględnione przy wyznaczaniu wartości $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, tym samym odległości Mahalanobisa $d^2_{MD,i}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, wyznaczone dla niepożądanych obserwacji dążą do nieskończoności, co w przyjętym systemie skutkuje zmierzającymi do zera wagami.
W przypadku, gdy obserwacje z $\mathbf{X}$ generowane są przez niezakłócony wielowymiarowy model normalny, zastosowanie ponownego ważenia estymatora skutkuje zwiększeniem efektywności estymatora w szczególności, gdy dysponuje się mało liczebną próbą $\mathbf{X}$.
Niestety w przypadku występowania mechanizmu zakłócającego główny model generujący dane, wpływ ponownego ważenia na efektywność i obciążoność estymatora, zależeć będzie od charakteru mechanizmu zakłócającego oraz położenia generowanych przez niego punktów odstających [Maronna, Martin, 2006, s. 193].

4 Szkic algorytmów wyznaczania wartości estymatorów MCD i PCS

4.1 Dobór podzbiorów i punktów początkowych dla algorytmów

Punktem wyjściowym procedur wyznaczania wartości estymatorów MCD i PCS, jest określenie wstępnych podzbiorów obserwacji, które dodatkowo w procedurze wyznaczania wartości MCD są podstawą do wyznaczenia początkowych przybliżeń wartości estymatorów $\left(\widehat{\boldsymbol{\mu}}^{(0)}, \widehat{\boldsymbol{\Sigma}}^{(0)}\right)$.

W rozważanych procedurach FastMCD oraz FastPCS, jako wyjściowe (początkowe) rozpatruje się $p+1$-elementowe podzbiory n-elementowego zbioru $\mathbf{X}$.
W przypadku dużych zbiorów danych rozważenie wszystkich możliwych $\begin{pmatrix} n \\ p+1 \end{pmatrix}$ podzbiorów okazuje się być zbyt czasochłonne, stąd też przyjęto dobierać losowo mniejszą liczbę spośród nich i przyjmować je za podzbiory wyjściowe.
Losowo dobiera się $p+1$-elementowe podzbiory, gdyż w przypadku takiej ich liczebności, prawdopodobieństwo uzyskania próbki nie zawierającej w ogóle wartości odstających, wynosi $\phi = (1-\varepsilon)^{p+1}$, przy czym $\varepsilon \in (0,1)$ jest odsetkiem obserwacji odstających.
Natomiast prawdopodobieństwo uzyskania wśród M_{p+1} próbek $p+1$-elementowych, przynajmniej jednej próbki pozbawionej wartości odstających wynosi $1-(1-\phi)^{M_{p+1}}$ i jest ono dodatnie, przy dowolnym n, także $n \to \infty$, gdyż zależy ono wyłącznie od wartości $p+1$.
Wspomniane prawdopodobieństwo jest wyższe niż w przypadku wyboru próbek h-elementowych w sytuacji, gdy $h = \lfloor \frac{n+p+1}{2} \rfloor$, ponieważ prawdopodobieństwo $\phi_h = (1-\varepsilon)^h = (1-\varepsilon)^{\lfloor \frac{n+p+1}{2} \rfloor}$, przy $n \to \infty$ dąży do 0, tym samym także prawdopodobieństwo $1-(1-\phi_h)^{M_h}$ dąży do 0.
W celu uzyskania przynajmniej jednej „niezanieczyszczonej" $p+1$-elementowej podpróby z prawdopodobieństwem na poziomie powyżej $1-\psi$ (np. $\psi = 0,01$) potrzeba, aby
$\log \psi \geqslant M_{p+1} \log(1-\phi) \approx -M_{p+1} \cdot \phi$, stąd przyjęta liczba losowań M_{p+1} powinna być niemniejsza niż $M_{p+1} \geqslant \frac{|\log \psi|}{|\log(1-(1-\varepsilon)^{p+1})|} \approx \frac{|\log \psi|}{(1-\varepsilon)^{p+1}}$, przy czym przybliżenie jest właściwe dla odpowiednio wysokich wartości $p+1$.

4.2 Algorytm FastMCD

Niech zbiór $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^p,\ i = 1, \ldots, n\}$ reprezentuje n-elementową próbę p-wymiarowych ($p \geqslant 2$) obserwacji. Ponadto h to liczba obserwacji uwzględnianych przy wyznaczaniu próbkowych wektorowych średnich i macierzy kowariancji. Domyślnie przyjmuje się, że $h = \lfloor \frac{n+p+1}{2} \rfloor$ (wybór dający maksymalny FBP), ale można przyjąć dowolne h, takie, że $\lfloor \frac{n+p+1}{2} \rfloor \leqslant h < n$. Natomiast M_{p+1} jest liczbą rozpatrywanych w algorytmie początkowych podzbiorów $p+1$-elementowych.

Algorytm FastMCD – wariant dla $n \leqslant 600$ [Rousseeuw, Driessen, 1999]:

1. Dla $m = 1, 2, \ldots, M_{p+1}$:
   1.1. Losowo dobrać podzbiór $p+1$ elementów ze zbioru $\mathbf{X}$, których indeksy utworzą podzbiór $H_m^{(0)} \subset \{1, \ldots, n\}$, gdzie $\#H_m^{(0)} = p+1$. W oparciu o zbiór $H_m^{(0)}$ wyznaczyć wartości początkowe dla procedury iteracyjnej: $\left(\widehat{\boldsymbol{\mu}}_m^{(0)}, \widehat{\boldsymbol{\Sigma}}_m^{(0)}\right)$.
   1.2. Wykonać dwa kroki koncentracji $l = 1, 2$, uzyskując kolejno h-elementowe podzbiory $H_m^{(1)}$ i $H_m^{(2)}$ oraz odpowiadające im wartości $\left(\widehat{\boldsymbol{\mu}}_m^{(l)}, \widehat{\boldsymbol{\Sigma}}_m^{(l)}\right), l = 1, 2$.
2. Przyjmując, że $\mathcal{M} = \left\{m \in \{1, \ldots, M_{p+1}\} : \left|\widehat{\boldsymbol{\Sigma}}_m^{(2)}\right| \leqslant \left|\widehat{\boldsymbol{\Sigma}}_{(10)}^{(2)}\right|\right\}$, gdzie $\left|\widehat{\boldsymbol{\Sigma}}_{(r)}^{(2)}\right|$ jest r-tą statystyką porządkową dla wyznacznika macierzy kowariancji.
   Dla każdego $m \in \mathcal{M}$ (czyli dla 10 podzbiorów $H_m^{(2)}$, związanych z najniższymi wartościami $\left|\widehat{\boldsymbol{\Sigma}}_m^{(2)}\right|$), wychodząc od $\left(\widehat{\boldsymbol{\mu}}_m^{(2)}, \widehat{\boldsymbol{\Sigma}}_m^{(2)}\right)$ wykonywać kroki koncentracji (*C-steps*), aż do uzyskania zbieżności w kroku L_m, w którym wartości $\left(\widehat{\boldsymbol{\mu}}_m^{(L_m)}, \widehat{\boldsymbol{\Sigma}}_m^{(L_m)}\right)$ wyznacza się w oparciu o obserwacje z h-elementowego podzbioru $H_m^{(L_m)}$.
3. Jako wartości estymatora MCD przyjąć wartości $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = (\mathrm{ave}_{i \in H^*} \mathbf{x}_i, \mathrm{cov}_{i \in H^*} \mathbf{x}_i)$, wyznaczone w oparciu o ten spośród podzbiorów $H_m^{(L_m)}, m \in \mathcal{M}$, oznaczany symbolem H^*, dla którego wartość $\left|\widehat{\boldsymbol{\Sigma}}_m^{(L_m)}\right|$ jest najniższa : $H^* = \mathrm{argmin}_{\left\{H_m^{(L_m)}\right\}_{m \in \mathcal{M}}} |\widehat{\boldsymbol{\Sigma}}_m^{(L_m)}|$.

Wraz ze wzrostem liczebności próby n, wzrasta czas wykonywania obliczeń, w szczególności w wyniku konieczności wyznaczenia w każdym kroku algorytmu n odległości Mahalanobisa dla elementów $\mathbf{X}$.

W wariancie algorytmu FastMCD [Rousseeuw, Driessen, 1999] dla przypadku, gdy wielkość próby wynosi $n > 600$ obserwacji, w celu skrócenia czasu wykonywania obliczeń, dokonuje się podziału całości próby (w przypadku, gdy $n > 1500$, liczebność próby uprzednio ogranicza się poprzez bezzwrotne losowanie do 1500), na co najwyżej 5 możliwie równolicznych, rozłącznych podzbiorów (podprób), dla których osobno wykonywane są wstępne dwa kroki koncentracji przedstawionego algorytmu. Dla każdej z 5 podprób zachowuje się 10 najlepszych rozwiązań, po czym łączy się podpróby, a 50 zachowanych rozwiązań służy jako rozwiązania początkowe. Następnie w oparciu o te rozwiązania początkowe oraz połączoną próbę dokonuje się po dwa kroki koncentracji oraz zachowuje się 10 najlepszych rozwiązań i te stanowią rozwiązania początkowe w kolejnym etapie. Dla całości próby wychodząc od rozwiązań początkowych dokonuje się kroków koncentracji, aż do osiągnięcia zbieżności .
Jako, że przedstawiona procedura iteracyjna wyznaczania wartości estymatora MCD jest algorytmem o lokalnej zbieżności do minimum globalnego, uzyskana wartość estymatora MCD zależy od przyjętej wartości początkowej, dobieranej tutaj w sposób losowy. W celu uniezależnienia wyników procedury iteracyjnej od losowego doboru, w modyfikacji estymatora MCD nazwanej detMCD (*Deterministic MCD*, [Hubert i in., 2012]) wartości początkowe dla procedury, wyznacza się w oparciu o wartości kilku wariantów estymatorów wielowymiarowej skali (przyjmujących dla danego zbioru zawsze te same wartości), dla zestandaryzowanych danych, w tym m.in. macierzy korelacji rang Spearmana, czy też estymatora zortogonalizowanego Gnanadesikana-Kettenringa (OGK).

4.3 Algorytm FastPCS

Dla metody PCS w przypadku, gdy $n > p+1 > 2$ oraz $\mathbf{X}$ znajduje się w położeniu ogólnym skończonopróbkowy punkt załamania wynosi $\varepsilon_n^* = \frac{n-h+1}{n}$. W przypadku przyjęcia $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$ uzyskuje się maksymalny możliwy FBP dla afinicznie ekwiwariantnych estymatorów rozrzutu, taką też wartość h przyjmuje się w przedstawionym poniżej algorytmie FastPCS.
Niech $\mathbf{X}$, n, p, h oraz M_{p+1} mają takie samo znaczenie jak w przypadku zaprezentowanego wcześniej algorytmu.
W prezentowanym algorytmie FastPCS przyjęto natomiast, że $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$.

Algorytm FastPCS [Vakili, Schmitt, 2014]:

1. Dla $m = 1, \ldots, M_{p+1}$:
   1.1. Losowo dobrać $p+1$ elementów ze zbioru $\mathbf{X}$, których indeksy utworzą podzbiór $H_m^{(0)} \subset \{1, \ldots, n\}$, gdzie $\#H_m^{(0)} = p+1$.
   1.2. Dla $l = 1, \ldots, L$:
   $D_i(H_m^{(l)}) = \text{ave}_{k=1}^{K} \frac{d_{P,i}^2(\mathbf{a}_{mk})}{\text{ave}_{j \in H_m^{(l-1)}} d_{P,j}^2(\mathbf{a}_{mk})}$, dla każdego $i = 1, \ldots, n$
   $q = \left\lfloor (n-p-1)\frac{l}{2L} \right\rfloor + p + 1$
   $H_m^{(l)} = \{i : D_i(H_m^{(l)}) \leqslant D_{(q)}(H_m^{(l)})\}$ (krok koncentracji)
   1.3. $H_m \overset{ozn.}{=} H_m^{(L)}$
   1.4. $I(H_m) = \text{ave}_{k=1}^{K} I(H_m, \mathbf{a}_{mk})$ (indeks inkongruencji dla podzbioru H_m)
2. $H^* = \text{argmin}_{\{H_m\}_{m=1}^{M_{p+1}}} I(H_m)$. Jako wartości estymatora PCS przyjąć $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = (\text{ave}_{i \in H^*} \mathbf{x}_i, \text{cov}_{i \in H^*} \mathbf{x}_i)$.
3. (opcjonalny) W wariancie algorytmu za elementy podzbioru H^{**} przyjąć indeksy h obserwacji, dla których kwadraty odległości Mahalanobisa wyznaczonych przy $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ są najmniejsze: $H^{**} = \left\{ i \in \{1, \ldots, n\} : d_{MD,i}^2(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) \leqslant d_{MD,(h)}^2(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) \right\}$. Jako wartości estymatora przyjąć $(\widehat{\boldsymbol{\mu}}^{**}, \widehat{\boldsymbol{\Sigma}}^{**}) = (\text{ave}_{i \in H^{**}} \mathbf{x}_i, \text{cov}_{i \in H^{**}} \mathbf{x}_i)$.

W ramach punktu 1.2. algorytm FastPCS stopniowo zwiększa w kolejnych L krokach koncentracji rozmiary podzbiorów $H_m^{(l)}$ od $p+1$ dla $H_m^{(0)}$ do ostatecznego rozmiaru $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$ dla $H_m^{(L)}$.

Inaczej jest w przypadku algorytmu FastMCD , który już w pierwszym kroku koncentracji przechodzi od $p+1$-elementowego pozbioru $H_m^{(0)}$ do h-elementowego podzbioru $H_m^{(1)}$. Zabieg stopniowego zwiększania liczebności podzbiorów $H_m^{(l)}$ zastosowany w FastPCS, umożliwia zwiększenia odporności algorytmu w przypadku, gdy obserwacje odstające wynikające z działania mechanizmu zakłócającego, położone są w pobliżu „dobrych" danych, wygenerowanych przez proces główny.
W l-tym powtórzeniu punktu 1.2. algorytmu FastPCS, w ramach tworzenia podzbioru $H_m^{(l)}$, dla kolejnych punktów $\mathbf{x}_i, i = 1, \ldots, n$ ze zbioru $\mathbf{X}$, rozpatruje się ich odległości od hiperpłaszczyzny z wektorem normalnym a_{mk}, w relacji do przeciętnej odległości od tej hiperpłaszczyzny dla tych z obserwacji ze zbioru X, których indeksy należą do $H_m^{(l-1)}$. W celu uniezależnienia się od konkretnego kierunku $\mathbf{a}_{mk}$, wspomniane relatywne odległości dla punktów $x_i, i = 1, \ldots, n$ ze zbioru X wyznacza się dla wszystkich rozważanych kierunków a_{mk}, $k = 1, \ldots, K$, po czym uśrednia się je po k, uzyskując wartości $D_i(H_m^{(l)}), i = 1, \ldots, n$. W kroku koncentracji podzbiór $H_m^{(l)}$ tworzą indeksy $q = \left\lfloor (n-p-1)\frac{l}{2L} \right\rfloor + p + 1$ obserwacji spośród n obserwacji ze zbioru $\mathbf{X}$, dla których uśrednione relatywne odległości są najmniejsze.
Wychodząc od m-tego podzbioru startowego $H_m^{(0)}$, po wykonaniu L-krotnie kroku 1.2. uzyskuje się zbiór H_m, dla którego wyznacza się wartości wskaźników inkongruencji wzdłuż kierunków $a_{mk}, k = 1, \ldots, K$ oraz wskaźnik inkogruencji podzbioru H_m, będący ich uśrednieniem po k. Jako zbiór H^*, zawierający indeksy obserwacji będących podstawą wyznaczania wartości estymatorów, przyjmuje się ten spośród zbiorów H_m, $m = 1, \ldots, M_{p+1}$, dla którego wskaźnik inkongruencji $I(H_m)$ jest najmniejszy. Wartości estymatora PCS wyznacza się jako próbkową wektorową średnią oraz macierz kowariancji, w oparciu o nerwacje o indeksach w H^*.

5 Estymatory MCD i PCS – podsumowanie

Niech $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, n\}$ będzie n elementową próbą p-wymiarowych obserwacji, natomiast h liczebnością podpróby ze zbioru $\mathbf{X}$. Skrótowe podsumowanie własności estymatorów wielowymiarowego rozrzutu MCD i PCS oraz algorytmów wyznaczania ich wartości przedstawiono w tabeli.

Aspekt	**Estymator MCD**	**Estymator PCS**
Wyrażenie problemu optymalizacyjnego definiującego wartość estymatora dla zbioru $\mathbf{X}$	Minimalizacja z ograniczeniami funkcji kryterium $\lvert\boldsymbol{\Sigma}\rvert$, zależnej bezpośrednio od parametru $\boldsymbol{\Sigma}$	Poszukiwanie najbardziej jednorodnego podzbioru zbioru $\mathbf{X}$, niezawierającego wartości odstających, charakteryzującego się minimalną wartością indeksu inkongruencji, który nie jest wyrażony poprzez bezpośrednią zależność od parametru $\boldsymbol{\Sigma}$
Zgodność estymatora w przypadku wielowymiarowego rozkładu normalnego bez mechanizmu zakłócającego	niezgodny asymptotycznie (wprowadza się korektę poprzez mnożnik c dla $\widehat{\boldsymbol{\Sigma}}$), niezgodny w sensie Fishera (wprowadza się korektę poprzez mnożnik[4] c_γ)	
Nazwa algorytmu poszukiwania wartości estymatora	FastMCD [Rousseeuw, Driessen, 1999]	FastPCS [Vakili, Schmitt, 2014]
Pakiet R /funkcja/ z implementacją procedury wyznaczania wartości estymatora	`robustbase` /`covMcd`/ [Rousseeuw i in., 2015] `rrcov` /`CovMcd`/ [Todorov i in., 2009]	`FastPCS` /`FastPCS`/ [Schmitt i in., 2014]
Miara odległości dla elementów zbioru $\mathbf{X}$, wykorzystywana w ramach algorytmu	kwadrat odległość Mahalanobisa $d^2_{MD,i}$	kwadrat odległość punktu od hiperpłaszczyzny $d^2_{P,i}$
Podzbiór zbioru $\mathbf{X}$, będący podstawą wyznaczania wartości estymatora	h-elementowy podzbiór, dla którego próbkowa macierz kowariancji ma minimalny wyznacznik: $\lvert\widehat{\boldsymbol{\Sigma}}\rvert$	h-elementowy podzbiór dla którego wartość wskaźnika inkongruencji jest minimalna: $I(H^*)$
Skończonopróbkowy punkt załamania estymatora	$\frac{n-h+1}{n} \leqslant \frac{1}{n}\left\lfloor\frac{n-p+1}{2}\right\rfloor$ gdzie n – liczebność próby, p – wymiar przestrzeni	$\frac{n-h+1}{n} \leqslant \frac{1}{n}\left\lfloor\frac{n-p+1}{2}\right\rfloor$ gdzie n – liczebność próby, p – wymiar przestrzeni
Złożoność obliczeniowa algorytmu (dla każdego kolejnego $p+1$ elementowego zbioru początkowego, spośród M_{p+1} rozpatrywanych)	FastMCD: $O(p^3 + np^2)$	FastPCS: $O(p^3 + np)$
Możliwość ponownego ważenia (*reweighting*)	system wag jako funkcja kwadratów odległości Mahalanobisa, przy wartościach oszacowań „wyjściowego" estymatora	system wag jako funkcja kwadratów odległości Mahalanobisa, przy wartościach oszacowań „wyjściowego" estymatora
Inne		mniejszy wpływ stopnia koncentracji obserwacji odstających, na wynik działania algorytmu

Tabela 1: Sumaryczne porównanie estymatorów MCD i PCS (opracowanie własne)

6 Przykład empiryczny zastosowania odpornych estymatorów rozrzutu MCD i PCS – odporna analiza portfelowa

W celu praktycznego rozwiązania problemów optymalizacyjnych w ramach analizy portfelowej wykorzystuje się oszacowania parametrów odpowiedniego wielowymiarowego rozkładu stóp zwrotu. W ramach odpornej analizy portfelowej postuluje się zastąpienie klasycznych estymatorów parametrów ich odpornymi odpowiednikami, w celu przeciwdziałania znaczącym odstępstwom wartości oszacowań od prawdziwych wartości parametrów, wywoływanym występowaniem odstających realizacji stóp zwrotu.

Za parametry w funkcji kryterium, optymalizowanej w ramach analizy portfelowej podstawiane są punktowe oszacowania parametrów wartości oczekiwanych stóp zwrotu oraz ich macierzy kowariancji (estymatory klasyczne bądź odporne), które obarczone są niepewnością.
Ze względu na wspomnianą niepewność wnioskowania o parametrach, nawet niewielkie odchylenia oszacowań od nieznanych prawdziwych parametrów mogą skutkować dosyć odmiennymi rozwiązaniami problemów optymalizacyjnych w ramach analizy portfelowej.
Ponieważ wpływ na wyniki optymalizacji (uzyskaną strukturę wag portfela) w relacji do stopnia zmienności związanej z niepewnością w przypadku oszacowań wektora oczekiwanych stóp zwrotu jest większy niż ma to miejsce w przypadku oszacowań macierzy kowariancji [Pfaff, 2012, s. 160], w ramach odpornych technik analizy portfelowej, przyjęło się, żeby bezpośrednio uwzględniać niepewność oszacowań wektora wartości oczekiwanych, przy jednoczesnym traktowaniu parametru kowariancji jako ustalonego na poziomie oszacowania punktowego.
Problem optymalizacyjny w ramach poszukiwania portfela o minimalnej zmienności, nie wykorzystuje w swej konstrukcji parametrów oczekiwanej stopy zwrotu, a jedynie parametry składowe macierzy kowariancji, co pozwala na uniknięcie konieczności rozważania niepewności oraz umożliwia bardziej bezpośrednią ocenę wpływu przyjętego estymatora macierzy kowariancji na wyniki analizy portfelowej.
W przypadku problemu poszukiwania portfela o minimalnej zmienności, warunkowej minimalizacji (założenie braku dźwigni finansowej oraz braku możliwości zajmowania krótkich pozycji na instrumentach składowych portfela) podlega następująca funkcja kryterium:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \sigma_P = \sigma(\mathbf{w}) = \sqrt{\mathbf{w}'\mathbf{\Sigma}\mathbf{w}} \tag{36}$$

pod warunkiem

$$\mathbf{1}'\mathbf{w} = 1 \wedge \mathbf{w} \geqslant \mathbf{0} \tag{37}$$

przy czym $\mathbf{w}$ jest poszukiwanym p-wymiarowym wektorem kolumnowym wag, $\mathbf{\Sigma}$ jest macierzą kowariancji (w praktycznej analizie zastępowany wybranym jej oszacowaniem), $\mathbf{1}$ jest p-wymiarowym wektorem kolumnowym złożonym z jedynek
Innym przykładem problemu optymalizacyjnego w ramach analizy portfelowej, który wykorzystuje w funkcji kryterium wyłącznie parametry z macierzy kowariancji jest portfel ERC (*Equal Risk Contribution*).
W portfelu ERC strukturę wag określa się w taki sposób, aby krańcowy udział w ryzyku portfela (mierzonego jego zmiennością $\sigma_P = \sqrt{\mathbf{w}'\mathbf{\Sigma}\mathbf{w}}$) dla każdego instrumentu składowego portfela był równy.
Krańcowy wkład i-tej składowej portfela w jego łączne ryzyko wyraża się jako $\sigma_i(\mathbf{w}) = w_i \cdot \frac{\partial\sigma(\mathbf{w})}{\partial w_i}$, gdzie $\frac{\partial\sigma(\mathbf{w})}{\partial w_i} = \frac{w_i\sigma_i^2+\sum_{j\neq i} w_j\sigma_{ij}}{\sigma(\mathbf{w})}$ oraz σ_i^2 jest i-tym elementem diagonalnym macierzy kowariancji stóp zwrotu, natomiast σ_{ij} to element na przecięciu i-tego wiersza i j-tej kolumny tej macierzy. Jako, że funkcja $\sigma(\mathbf{w})$ wyrażająca ryzyko (zmienność) portfela σ_P, jest funkcją jednorodną w stopniu 1, stąd na mocy twierdzenia Eulera $\sigma_P = \sigma(\mathbf{w}) = \sum_{i=1}^{p} \sigma_i(\mathbf{w}) = \sum_{i=1}^{p} w_i \cdot \frac{\partial\sigma(\mathbf{w})}{\partial w_i}$, czyli łączne ryzyko portfela można wyrazić jako sumę iloczynów pochodnych cząstkowych względem kolejnych zmiennych $\frac{\partial\sigma(\mathbf{w})}{\partial w_i}$ przez wartości tych zmiennych w_i, co odpowiada sumie krańcowych wkładów poszczególnych instrumentów składowych portfela w jego zmienność.
W przypadku portfela ERC jednym z możliwych sposobów wyrażenia problemu optymalizacyjnego z warunkami ograniczającymi (założenie braku możliwości zajmowania pozycji krótkich na aktywach portfela) jest:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \sum_{i=1}^{p} \left(\frac{\sigma_i(\mathbf{w})}{\sigma(\mathbf{w})} - \frac{1}{p} \right)^2 \tag{38}$$

pod warunkiem

$$\mathbf{1}'\mathbf{w} = 1 \wedge \mathbf{w} \geqslant \mathbf{0} \tag{39}$$

Zasadnym krokiem w ramach analizy portfelowej jest uwzględnienie wpływu na rozkład stóp zwrotu, ich realizacji w najbliższej przeszłości, co prowadzi do dynamicznej analizy portfelowej.

W dynamicznej analizie portfelowej zakładającej codzienny *rebalancing*[5], bardzo ważnym aspektem jest trafne oszacowanie i prognozowanie warunkowej kowariancji. Jako alternatywę dla parametrycznych modeli MGARCH czy MSV (szerokie omówienie tych klas modeli zawiera praca [Pajor, 2010]), zaproponowano nieparametryczne ruchome estymatory warunkowej kowariancji [Fleming i in., 2001, Pooter i in., 2008].
W przypadku wykorzystania danych dziennych wskazuje się następującą konstrukcję ruchomego estymatora macierzy kowariancji warunkowej względem przeszłości procesu:

$$\mathbf{S}_t = \exp(-\alpha)\mathbf{S}_{t-1} + \alpha \exp(-\alpha)\mathbf{r}_{t-1}\mathbf{r}_{t-1}^{'} \tag{40}$$

przy czym

- $\mathbf{S}_t$ – oszacowanie dziennej macierzy kowariancji warunkowej w dniu sesyjnym t
- α – parametr zanikania (*decay rate*),
- $\mathbf{r}_t$ – dzienna stopa zwrotu w dniu sesyjnym t, rozumiana jako logarytmiczny przyrost pomiędzy cenami na zamknięcie w dniu t oraz $t-1$

Analizy [Barndorff-Nielsen, Shephard, 2004] pokazują, że bardziej efektywne prognozy warunkowej kowariancji uzyskuje się, wykorzystując w ich konstrukcji macierz zrealizowanej kowariancji wyznaczoną w oparciu o wewnątrzdzienne stopy zwrotu:

$$\mathbf{S}_t = \exp(-\alpha)\mathbf{S}_{t-1} + \alpha \exp(-\alpha)(\mathbf{V}_{t-1} + \boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^{'}) \tag{41}$$

przy czym

- $\mathbf{V}_t$ – macierz zrealizowanej kowariancji (*ex post*) w dniu sesyjnym t
- $\boldsymbol{\eta}_t$– stopa zwrotu „przez noc" (*overnight*), tzn. logarytmiczny przyrost pomiędzy ceną na otwarcie w dniu t a ceną na zamknięcie w dniu $t-1$
- pozostałe symbole rozumiane są jak wyżej

Wartość parametru zanikania α można dobrać arbitralnie bądź też można dokonać jego estymacji za pomocą metody quasi-MNW, przy założeniu $\mathbf{r}_t = \mathbf{S}_t\boldsymbol{\varepsilon}_t$, $\{\boldsymbol{\varepsilon}_t\} \sim iid(\mathbf{0}, \mathbf{I}_p)$. Inną propozycję procedury doboru parametru α przedstawiono w pracy [Fleming i in., 2003].
Przedstawiony model ruchomego estymatora warunkowej kowariancji, można rozpatrywać jako MGARCH z ograniczeniami nałożonymi na parametry. W ramach analiz [Fleming i in., 2003] pokazano, że pomimo, iż model MGARCH daje lepsze dopasowanie do danych, zastosowanie ruchomego oszacowania warunkowej kowariancji daje lepsze wyniki finansowe w analizie portfelowej, co przypisuje się bardziej gładkiemu przebiegowi w czasie tak szacowanej macierzy warunkowej kowariancji.
Przyjmując, że przedział $[0, 1]$ odpowiada przedziałowi czasowemu dla pierwszego rozważanego dnia sesyjnego, dla t-tego dnia sesyjnego jest to odpowiednio przedział $[t-1, t]$, macierz zrealizowanej kowariancji $\mathrm{RCov}_{t,\Delta}$ definiuje się z wykorzystaniem wewnątrzdziennych stóp zwrotu za podokresy o długości Δ, przy czym $\Delta < 1$, stąd każdy t-ty dzień sesyjny obejmuje $\lfloor \frac{1}{\Delta} \rfloor$ podokresów oraz związanych z nimi wewnątrzdziennych stóp zwrotu.
Miernik zrealizowanej kowariancji RCov (ang. *Realized Covariation*) wyraża się następująco:

$$\mathrm{RCov}_{t,\Delta} = \sum_{i=(t-1)\cdot\lfloor 1/\Delta \rfloor + 1}^{t\cdot\lfloor 1/\Delta \rfloor} \mathbf{r}_{i,\Delta}\mathbf{r}_{i,\Delta}^{'} \tag{42}$$

przy czym

- $\mathrm{RCov}_{t,\Delta}$ – zrealizowana kowariancja w dniu sesyjnym t
- $\mathbf{r}_{i,\Delta}$ – wewnątrzdzienna logarytmiczna stopa zwrotu za i-ty okres o długości Δ , gdzie $i = (t-1)\cdot\lfloor 1/\Delta \rfloor + 1, \ldots, t\cdot\lfloor 1/\Delta \rfloor$, przebiega indeksy wewnątrzdziennych stóp zwrotu w ramach t-tego dnia sesyjnego

[5] Wybór optymalnych wag portfela w oparciu o parametry rozkładu warunkowego względem przeszłości, w każdym kolejnym dniu sesyjnym.

W przypadku wykorzystania wewnątrzdziennych stóp zwrotu 1–, 5–, 10–, 15– minutowych, przyjmując, że mamy do czynienia z akcjami notowanymi w systemie ciągłym na GPW, liczba okresów $\lfloor 1/\Delta \rfloor$ w ramach pojedynczego dnia sesyjnego wynosi odpowiednio: 480, 96, 48, 32.
Niemniej jednak miernik $\mathrm{RCov}_{t,\Delta}$ nie jest odporny na występowanie addytywnych skoków nakładających się na proces kształtujący ceny rozważanych aktywów.
Wpływ skoków na kształtowanie się mierników zrealizowanej kowariancji, można rozważyć przyjmując, że dynamikę logarytmów cen instrumentu opisuje p-wymiarowy proces dyfuzyjny ze skokami o skończonej aktywności BSMFAJ (*Brownian Semimartingale with Finite Activity Jumps*), zadany przez:

$$\mathrm{d}\mathbf{p}(s) = \boldsymbol{\mu}(s)\,\mathrm{d}s + \boldsymbol{\Omega}(s)\,\mathrm{d}\mathbf{W}(s) + \mathbf{K}(s)\,\mathrm{d}\mathbf{q}(s) \tag{43}$$

przy czym

- $\{\mathbf{p}(s)\}_{s\geqslant 0}$ – p-wymiarowy proces z czasem ciągłym $s \geqslant 0$
- $\mathbf{p}(s)$ – p-wymiarowy wektor wartości logarytmów cen instrumentów w momencie s
- $\boldsymbol{\mu}(s)\,\mathrm{d}s$ – p-wymiarowa deterministyczna składowa procesu, opisująca dynamikę oczekiwanej wartości logarytmów cen instrumentów
- $\boldsymbol{\Omega}(s)\,\mathrm{d}\mathbf{W}(s)$ – p-wymiarowa stochastyczna składowa procesu opisująca dynamikę zmienności
- $\{\mathbf{W}(s)\}_{s\geqslant 0}$ – p-wymiarowy standardowy proces Wienera
- $\boldsymbol{\Omega}(s)$ – zmienność chwilowa (*spot volatility*) w momencie s (macierz $p \times p$)
- $\boldsymbol{\Sigma}(s) = \boldsymbol{\Omega}(s)\boldsymbol{\Omega}'(s)$ – chwilowa macierz kowariancji w momencie s (macierz klasy $PDS(p)$)
- $\mathbf{K}(s)\,\mathrm{d}q(s)$ – składowa obejmująca addytywny p-wymiarowy proces skoków
- $\{\mathbf{q}(s)\}_{s\geqslant 0}$ – proces liczący dla skoków ze skończoną aktywnością (*finite activity jumps*) zadany przez p-wymiarowy proces Poissona
- $\mathbf{K}(s)$ – macierz o wymiarach $p \times p$, opisująca poziom natężenia skoków oraz sposób transmisji ich wpływu pomiędzy poszczeglnymi wymiarami procesu w momencie s

W przypadku pominięcia składowej związanej z działaniem skoków, tzn. $\{\mathbf{K}(s) \equiv 0\}_{s\geqslant 0}$, proces BSMFAJ sprowadza się do procesu BSM (*Brownian Semimartingale*). Przykład zastosowania procesów dyfuzyjnych ze skokami w ekonometrii finansowej wraz z opisem procedury bayesowskiego wnioskowania o ich parametrach można znaleźć w pracy [Kostrzewski, 2014].
Zintegrowaną macierz kowariancji $\mathrm{ICov}_{[0,1]}$ procesu, dla przedziału $[0, 1]$, będącą miernikiem poziomu zmienności w ramach (pierwszego rozpatrywanego) dnia sesyjnego, definiuje się jako:

$$\mathrm{ICov}_{[0,1]} = \int_0^1 \boldsymbol{\Sigma}(s)\,\mathrm{d}s \tag{44}$$

Logarytmiczne stopy zwrotu za okres Δ, można rozważać jako dyskretne przyrosty w ramach rozważanego procesu z czasem ciągłym, generującego logarytmy cen:

$$\mathbf{r}_{i,\Delta} = \mathbf{p}(i\Delta) - \mathbf{p}\left((i-1)\Delta\right) \tag{45}$$

W przypadku krótkich okresów Δ dla uproszczenia ignoruje się część procesu dyfuzyjnego związaną ze średnią, przyjmując, że $\int_{(i-1)\Delta}^{i\Delta} \boldsymbol{\mu}(s)\,\mathrm{d}s \approx 0$, stąd w przypadku braku występowania skoków w rozważanym przedziale czasowym o długości Δ przyjmuje się, że $E\left(\mathbf{r}_{i,\Delta}\right) = \mathbf{0}$.
Macierz kowariancji zrealizowanej $\mathrm{RCov}_{t,\Delta}$ dla $t = 1$, w przypadku procesu BSMFAJ, nie jest zgodnym estymatorem macierzy zintegrowanej kowariancji $\mathrm{ICov}_{[0,1]}$, gdyż:

$$\mathrm{plim}_{\Delta\to 0}\mathrm{RCov}_{1,\Delta} = \mathrm{ICov}_{[0,1]} + \sum_{j=1}^{J} \boldsymbol{\kappa}_j \boldsymbol{\kappa}_j^{'} \tag{46}$$

przy czym j to indeks liczący kolejne wystąpienia skoków w dowolnym wymiarze addytywnego p-wymiarowego procesu skoków na przedziale czasowym $[0, 1]$, ponadto p-wymiarowy wektor κ_j mierzy wpływ j-tego skoku na rozważany p-wymiarowy proces logarytmów cen, natomiast łączną

liczbę wystąpień skoków w przedziale czasowym $[0,1]$ w dowolnym z wymiarów addytywnego procesu skoków wyznacza się jako: $J = \int_0^1 \mathrm{d}q^*(s)$, gdzie $q^*(s) = \mathbf{1}'\mathbf{q}(s)$, natomiast $\mathbf{1}$ to p-wymiarowy wektor kolumnowy jedynek
[Boudt i in., 2011] zaproponowali odporny na występowanie skoków (rozpatrywanych jako wielowymiarowe obserwacje odstające) miernik kowariancji zrealizowanej ROWCov (ang. *Realized Outlyingness Weighted Covariation*), umożliwiający przy założeniu procesu BSMFAJ, zgodną estymację wartości zintegrowanej macierzy kowariancji ICov, dla kolejnych dni sesyjnych.
ROWCov wykorzystuje w swej pierwotnej konstrukcji odporny estymator wielowymiarowego rozrzutu MCD, niemniej jednak jako jego alternatywę można także zastosować estymator PCS. Wartość miernika ROWCov wyznacza się jako sumę ważoną wewnątrzdziennych logarytmicznych stóp zwrotu, dla których wagi przypisywane są z wykorzystaniem odległości Mahalanobisa (będącej miernikiem stopnia odstawania obserwacji), wykorzystującej odporne oszacowanie rozrzutu okresowych (wewnątrzdziennych) stóp zwrotu w ramach tzw. lokalnego okna.
W ramach procedury tworzenia miernika ROWCov dokonuje się podziału dziedziny czasu na lokalne okna o długości λ (przy czym $\lambda \leqslant 1$ oraz $\lambda > \Delta$), w ramach których to przedziałów czasowych zakłada się stały poziom (brak znaczących zmian) zmienności chwilowej, a tym samym kowariancji chwilowej, tzn. $\mathbf{\Omega}(s) = \mathbf{\Omega}((l-1)\lambda)$, dla $s \in [(l-1)\lambda, l\lambda)$, $l \in \mathbb{Z}$, stąd też dla tego przedziału $\mathbf{\Sigma}(s) = \mathbf{\Sigma}((l-1)\lambda)$, dla uproszczenia przyjmuje się oznaczenie $\mathbf{\Sigma}_l = \mathbf{\Sigma}((l-1)\lambda)$.
W ramach lokalnego okna można wyróżnić $\lfloor \frac{\lambda}{\Delta} \rfloor$ podokresów o długości Δ. Zbiór indeksów stóp zwrotu z okresów o długości Δ, należących do tego samego okna lokalnego co stopa zwrotu $\mathbf{r}_{i,\Delta}$ wyraża się jako $N_i = \left\{ j = (l-1)\lfloor \frac{\lambda}{\Delta} \rfloor + 1, \ldots, l \lfloor \frac{\lambda}{\Delta} \rfloor : \ \ l = \lceil \frac{i\Delta}{\lambda} \rceil \right\}$.
I tak wartość odpornego estymatora rozrzutu MCD albo PCS $\widehat{\mathbf{\Sigma}}_l$, będącą szacunkiem chwilowej macierzy kowariancji w l-tym lokalnym oknie, które obejmuje stopę $\mathbf{r}_{i,\Delta}$, czyli $l = \lceil \frac{i\Delta}{\lambda} \rceil$, wyznacza się w oparciu o standaryzowane ze względu na długość okresu Δ stopy zwrotu $\mathbf{r}_{j,\Delta} \cdot \Delta^{-\frac{1}{2}}$, $j \in N_i$.
Przy określaniu odległości Mahalanobisa względem p-wymiarowego wektora zer dla stopy zwrotu $\mathbf{r}_{i,\Delta}$, wykorzystuje się wartość odpornego estymatora rozrzutu, dla lokalnego okna $l = \lceil \frac{i\Delta}{\lambda} \rceil$, do którego przynależy i-ta okresowa stopa zwrotu: $\widehat{\mathbf{\Sigma}}_{i,\Delta}\Delta \equiv \widehat{\mathbf{\Sigma}}_l\Delta$.
Kwadrat odległości Mahalanobisa pomiędzy $\mathbf{r}_{i,\Delta}$ a wektorem zerowym, przy macierzy kowariancji stóp zwrotu za okres Δ, wyrażonej przez $\widehat{\mathbf{\Sigma}}_{i,\Delta}\Delta$, dany jest następująco:

$$d^2_{i,\Delta} = \frac{\mathbf{r}'_{i,\Delta}\widehat{\mathbf{\Sigma}}^{-1}_{i,\Delta}\mathbf{r}_{i,\Delta}}{\Delta} \tag{47}$$

Tak zdefiniowane kwadraty odległości $d^2_{i,\Delta}$ pełnią rolę argumentu dla funkcji wag.
Jako funkcję wag $w : \mathbb{R}_+ \to \mathbb{R}_+$ przyjmuje się funkcję ciągłą, dla której wartości $w(z)z$ są ograniczone.
Przykładem takiej funkcji jest funkcja wagowa *Hard Rejection* (HR):

$$w_{HR}(z) = I(z \leqslant k) \tag{48}$$

gdzie I jest funkcją wskaźnikową
oraz funkcja wagowa *Soft Rejection* (SR):

$$w_{SR}(z) = \min\left\{1, \frac{k}{z}\right\} \tag{49}$$

gdzie $0 < k < \infty$ jest ustalonym parametrem.
Przy założeniu, że logarytmy cen instrumentów generowane są przez p-wymiarowy proces BSM, $d^2_{i,\Delta}$ ma rozkład chi-kwadrat z p stopniami swobody, wtedy za wartość k, przyjmuje się wartość kwantyla $1-\beta$ (częstym wyborem jest $\beta = 0,001$ albo $0,005$) wspomnianego rozkładu.
Po wyborze funkcji wagowej wartość miernika $\mathrm{ROWCov}_{t,\Delta}$ dla t-tego dnia sesyjnego, wyznacza się następująco:

$$\mathrm{ROWCov}_{t,\Delta} = c_w \sum_{i=(t-1)\cdot\lfloor 1/\Delta \rfloor + 1}^{t\cdot\lfloor 1/\Delta \rfloor} w(d^2_{i,\Delta})\mathbf{r}_{i,\Delta}\mathbf{r}'_{i,\Delta} \tag{50}$$

przy czym

- $\mathbf{r}_{i,\Delta}$ – wewnątrzdzienna logarytmiczna stopa zwrotu za i-ty okres o długości Δ

- $w(d^2_{i,\Delta})$ – waga odpowiadająca i-tej stopie zwrotu, której wartość wyznaczana jest w oparciu o kwadrat odległości Mahalanobisa wektora

- $c_w = \frac{p}{E[w(z)z]}$ – poprawka na zgodność ROWCov jako estymatora ICov dla dnia sesyjnego, w przypadku, gdy dynamikę logarytmów cen opisuje proces BSM, gdzie p jest wymiarem procesu, a z jest zmienną losową o rozkładzie chi-kwadrat z p stopniami swobody

Należy nadmienić, iż ROWCov ma własność afinicznej ekwiwariantności (dowód tej własności można znaleźć w pracy [Boudt i in., 2011]).
Do odpornej dynamicznej analizy portfelowej wybrano akcje trzech notowanych na GPW spółek KGHM, PEKAO oraz PKOBP, które charakteryzowały się wysoką płynnością (mierzoną przeciętnym czasem pomiędzy kolejnymi transakcjami). Jako zakres czasowy analizy przyjęto przedział od 2.07.2014 do 5.11.2015, co odpowiada $t \in \{0, 1, \ldots, T = 340\}$ (próba objęła 341 dni sesyjne, w związku z brakiem danych dla stopy zwrotu „przez noc" dla daty 4.07.2014).
Dane dzienne[6] oraz dane transakcyjne[7] dotyczące notowań przywołanych spółek pobrano z repozytorium Domu Maklerskiego BOŚ SA.
Częstym wyborem w analizach jest przyjęcie długości lokalnego okna odpowiadającej pojedynczemu dniu sesyjnemu, tzn. $\lambda = 1$ oraz długości okresów Δ odpowiadających 1–, 5–, 10– lub 15–minutowym okresom (dla notowań akcji GPW byłoby to Δ równe odpowiednio 1/480, 1/96, 1/48, 1/32). Biorąc pod uwagę płynność przyjętych do analizy akcji, zdecydowano się przyjąć $\Delta = 1/32$.
Podjęto się budowy portfela o minimalnej zmienności z codziennym *rebalancingiem*, tzn. w celu określenia struktury wag dla każdego kolejnego dnia sesyjnego $t = 1, \ldots, T$, w minimalizowanej funkcji kryterium za macierz kowariancji przyjmuje się dla t-tego dnia sesyjnego, prognozę dotyczącą macierzy kowariancji warunkowej względem przeszłości procesu.
Do oszacowania i prognozowania warunkowej względem przeszłości procesu macierzy kowariancji $\mathbf{S}_t$, przyjęto podejście nieparametryczne wykorzystujące ruchomy estymator, przyjmując za $\mathbf{V}_t$, oparte na 15–minutowych stopach zwrotu ($\Delta = 1/32$) mierniki kowariancji zrealizowanej $\mathrm{RCov}_{t,\Delta}$oraz $\mathrm{ROWCov}_{t,\Delta}$, bazujące na odpornych estymatorach rozrzutu MCD i PCS, przyjmując lokalne okno $\lambda = 1$ (dla przywołanych mierników przyjęto symbole $\mathrm{ROWCov}^{MCD}_{t,\Delta}$ oraz $\mathrm{ROWCov}^{PCS}_{t,\Delta}$):

$$\mathbf{S}_t = \exp(-\alpha)\mathbf{S}_{t-1} + \alpha \exp(-\alpha)(\mathbf{V}_{t-1} + \boldsymbol{\eta}_{t-1}\boldsymbol{\eta}'_{t-1}) \tag{51}$$

Jak wcześniej wspominano $\boldsymbol{\eta}_{t-1}$ to stopa zwrotu przez noc (*overnight*) w dniu sesyjnym $t-1$.
Im wyższa wartość parametru zanikania α tym wyższy udział w $\mathbf{S}_t$ ostatnich poziomów mierników zrealizowanej kowariancji i mniej gładki przebieg w czasie wartości oszacowań warunkowej macierzy kowariancji.
W pracy [Pooter i in., 2008] proponuje się rozważenie wartości parametru zanikania $\alpha = 0,05$, $0,1$ oraz $0,2$.
Jako rozwiązania problemu optymalizacyjnego uzyskuje się wagi portfela (w tym przypadku trójwymiarowe wektory wag – $p = 3$) dla każdego kolejnego t-tego dnia sesyjnego, w zależności od przyjętego za V_{t-1} miernika zrealizowanej kowariancji $\mathrm{RCov}_{t-1,\Delta}$, $\mathrm{ROWCov}^{MCD}_{t-1,\Delta}$ oraz $\mathrm{ROWCov}^{PCS}_{t-1,\Delta}$, które oznaczane będą odpowiednio symbolami: $\mathbf{w}_t^{\mathrm{RCov}}$, $\mathbf{w}_t^{\mathrm{ROWCov,\ MCD}}$ oraz $\mathbf{w}_t^{\mathrm{ROWCov,\ PCS}}$. Wektory wag pochodzące z sekwencji rozwiązań problemów optymalizacyjnych utworzą odpowiednie szeregi czasowe.
Stopę zwrotu portfela w t-tym dniu sesyjnym wyznacza się jako:

$$R^m_{P,t} = (\mathbf{w}^m_t)'\mathbf{R}_t \tag{52}$$

przy czym

- $R^m_{P,t}$ – dzienna prosta stopa zwrotu w t-tym dniu sesyjnym dla portfela o strukturze wag $\mathbf{w}^m_t$, gdzie w zależności od portfela m zastępuje się RCov, ROWCov − MCD albo ROWCov − PCS

- $\mathbf{R}_t$ – wektor dziennych prostych stóp zwrotu aktywów portfela w dniu sesyjnym t

[6] `http://bossa.pl/pub/ciagle/mstock/mstcgl.zip`
[7] `http://bossa.pl/index.jsp?layout=intraday&page=1&news_cat_id=875&dirpath=/mstock/cgl/`

W celu oceny poziomu zmienności portfeli w ramach tzw. *backtestingu* wyznacza się wartości odchylenia standardowego stóp zwrotu rozważanych portfeli $R_{P,t}^{\mathrm{RCov}}$, $R_{P,t}^{\mathrm{ROWCov-\ MCD}}$ oraz $R_{P,t}^{\mathrm{ROWCov-PCS}}$, w przywołanym wcześniej przedziale czasowym, z pominięciem pierwszych 30 dni sesyjnych (ze względu na konstrukcję ruchomego estymatora warunkowej kowariancji, który potrzebuje tzw. okresu „zapłonu" <ang. *burn-in period*>), tzn. analiza objęła dni sesyjne $t = 31, \ldots, 340$.
Ponadto wyznacza się wartości pozostałych statystyk opisowych dla stóp zwrotu rozważanych portfeli.
W celu porównania zachowania się portfeli wykorzystujących w konstrukcji prognoz odporne mierniki zrealizowanej kowariancji ROWCov (bazujące na MCD oraz PCS) z portfelem wykorzystującym w tym miejscu miernik RCov, wyznacza się dla kolejnych dni sesyjnych nadwyżki stóp zwrotu dwóch pierwszych portfeli nad stopą zwrotu portfela ostatniego (odpowiadają one stopom zwrotu z inwestycji polegającej na zajęciu pozycji długiej na porównywanym portfelu oraz pozycji krótkiej na portfelu z wagami $\mathbf{w}_t^{\mathrm{RCov}}$).
Dodatkowo w celu oceny trafności przewidywań dotyczących warunkowej kowariancji oraz jej wpływu na zmienność portfela (która nie jest bezpośrednio obserwowalna), wyznacza się dla każdego t-tego dnia sesyjnego, zmienność portfeli wykorzystując ocenę *ex post* zrealizowanej kowariancji:

$$\mathrm{sd}_{P,t}^{m} = \sqrt{(\mathbf{w}_t^m)' \mathbf{V}_t \mathbf{w}_t^m} \tag{53}$$

Dla każdego t-tego dnia sesyjnego dla każdego z trzech rozważanych portfeli (tzn. o strukturach wag $\mathbf{w}_t^{\mathrm{RCov}}$, $\mathbf{w}_t^{\mathrm{ROWCov-\ MCD}}$ oraz $\mathbf{w}_t^{\mathrm{ROWCov-\ PCS}}$), dokonuje się pomiaru zmienności sd_t^m z wykorzystaniem macierzy kowariancji zrealizowanej *ex post* przyjmując za $\mathbf{V}_t$ trzy jej warianty $\mathrm{RCov}_{t,\Delta}$, $\mathrm{ROWCov}_{t,\Delta}^{MCD}$ oraz $\mathrm{ROWCov}_{t,\Delta}^{PCS}$, niezależnie od tego który typ miernika został wykorzystany przy tworzeniu prognoz macierzy kowariancji warunkowej, będącej podstawą wyznaczania wag.
We wszystkich powyżej wskazanych porównaniach zestawia się dodatkowo wyniki dla portfela z równymi wagami dla wszystkich instrumentów: $\mathbf{w}_t^{eq} = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]'$. Dla stóp zwrotu portfela z równymi wagami przyjęto symbol $R_{P,t}^{\mathrm{eq}}$.

Portfel o minimalnej zmienności

Sekwencyjną analizę dla portfela o minimalnej zmienności, wykonano dla trzech wartości parametru $\alpha = 0,05$, $0,1$ oraz $0,2$, dla ruchomego oszacowania macierzy kowariancji warunkowej. W każdej z powyższych prób, dla wykorzystywanych w konstrukcji ROWCov estymatorów odpornych MCD i PCS, przyjmując, że parametr$h = \lfloor \gamma \cdot (\#N_i) \rfloor$, gdzie $\#N_i$ jest mocą zbioru N_i (obejmującego indeksy okresowych stóp zwrotu w tym samym oknie lokalnym co $r_{i,\Delta}$), rozważono dwa poziomy $\gamma = 0,5$ oraz $\gamma = 0,75$. Jako funkcję wagową używaną przy określaniu wartości ROWCov przyjęto funkcję typu HR (*Hard Rejection*), a parametr k ustalono, jako wartość kwantyla $1 - \beta$ rozkładu chi-kwadrat z $p = 3$ stopniami swobody, przyjmując $\beta = 0,001$.
W ramach tworzenia kodu w języku R umożliwiającego implementację przyjętej procedury budowy portfeli o minimalnej zmienności korzystano z funkcji następujących pakietów:
`xts` – [Ryan, Ulrich, 2011], `highfrequency` (funkcje `rCov`, `rOWCov`) – [Boudt i in., 2012], `fPortfolio` (funkcja `minvariancePortfolio`) – [Würtz i in., 2009].
Dla każdej z trzech rozważanych wartości parametru zanikania α, wyniki dotyczące portfela minimalnej zmienności uzyskane dla ruchomego oszacowania warunkowej kowariancji były bardzo zbliżone. Nieznacznie lepsze wyniki uzyskiwano przyjmując w ramach konstrukcji miernika ROWCov, dla estymatora MCD oraz PCS wartość parametru $h = \lfloor 0,75 \cdot (\#N_i) \rfloor$.
W związku z powyższymi ustaleniami, zgodnie z arbitralną decyzją, zaprezentowane zostaną wyniki dla konfiguracji $\alpha = 0,05$ i $h = \lfloor 0,75 \cdot (\#N_i) \rfloor$, odnoszące się do wyrażonych procentowo (zwielokrotnionych 100-krotnie) stóp zwrotu portfela oraz zmienności portfela.
W celu oceny jakości konstruowanych sekwencyjnie portfeli, dla których warunkowej minimalizacji ze względu na wartości wag podlega funkcja opisująca zmienność stóp zwrotu portfela, mierzoną odchyleniem standardowym, jednym z głównych aspektów jest zbadanie poziomu zmienności stóp zwrotu w próbie w ramach *backtestingu*. Dokonuje się tego poprzez wyznaczenie odchylenia standardowego dla stóp zwrotu portfela zaobserwowanych w próbie obejmującej zakres czasowy *backtestingu*. Ważne uzupełnienie tej oceny, polega na wyznaczeniu dla każdego dnia sesyjnego miary zmienności indukowanej przez wagi portfela oraz określony *ex post* poziom kowariancji zrealizowanej, wyznaczonej w oparciu o wewnątrzdzienne stopy zwrotu składowych portfela, której miernikiem jest na przykład RCov, czy też jego odporny odpowiednik ROWCov. Zmienność portfela w ujęciu dziennym nie jest bezpośrednio obserwowalna, tym samym poziom zmienności zrealizowanej $\mathrm{sd}_{P,t}^m$ indukowany przez mierniki kowariancji zrealizowanej dla kolejnych dni sesyjnych,

umożliwia badanie jego zmian w ujęciu dynamicznym. Sumaryczną informację o zmienności portfela $\mathrm{sd}_{P,t}^{m}$ można uzyskać poprzez wyznaczenie dla niej statystyk opisowych w ramach przyjętego przedziału czasowego (domyślnie takiego samego jak w ramach *backtestingu*).
W *backtestingu* dla dni sesyjnych $t = 31, \ldots, 340$ odchylenie standardowe stóp zwrotu portfeli, dla których przyjęto szeregi wag $\mathbf{w}_t^{\mathrm{RCov}}$, $\mathbf{w}_t^{\mathrm{ROWCov-\ MCD}}$, $\mathbf{w}_t^{\mathrm{ROWCov-PCS}}$ oraz $\mathbf{w}_t^{eq}$, przedstawiało się następująco:

m	RCov	ROWCov-MCD	ROWCov-PCS	eq
odchylenie standardowe dla próby $R_{P,t}^{m}$, $t = 31, \ldots, 340$	1.3914	1.3893	1.3849	1.4350

Tabela 2: Odchylenie standardowe stóp zwrotu portfeli minimalnej zmienności (opracowanie własne)

Wartość odchylenia standardowego wyznaczonego w próbie dla stóp zwrotu portfeli (uzyskanych poprzez optymalizację wag w ramach problemu portfela o minimalnej zmienności), było zbliżone niezależnie od tego czy w konstrukcji prognozy warunkowej kowariancji wykorzystano miarę RCov (odchylenie standardowe portfela wyniosło 1,3914), czy jej odporne odpowiedniki ROWCov-MCD i ROWCov-PCS (odchylenie standardowe odpowiednio 1,3893 i 1,3849). Biorąc pod uwagę wyniki *backtestingu*, wpływ wspomnianych odpornych odpowiedników na zmniejszenie poziomu zmienności wartości portfela można uznać za marginalny, jednocześnie zaobserwowany poziom zmienności portfeli minimalnego ryzyka był zauważalnie niższy od tego dla portfela z równymi wagami.
Statystyki opisowe dla stóp zwrotu portfeli z wagami $\mathbf{w}_t^{\mathrm{RCov}}$, $\mathbf{w}_t^{\mathrm{ROWCov-\ MCD}}$, $\mathbf{w}_t^{\mathrm{ROWCov-PCS}}$ oraz $\mathbf{w}_t^{eq}$ $(t = 31, \ldots, 340)$, przedstawiały się następująco:

m	$R_{P,t}^{m}$ średnia arytm.	$R_{P,t}^{m}$ mediana	$R_{P,t}^{m}$Q1	$R_{P,t}^{m}$Q3	$R_{P,t}^{m}$min	$R_{P,t}^{m}$ max	$R_{P,t}^{m}$ rozstęp
RCov	-0.0890	-0.0863	-0.8530	0.7852	-6.2455	3.9933	10.2388
ROWCov-MCD	-0.0931	-0.0882	-0.8519	0.7462	-6.4611	3.9850	10.4461
ROWCov-PCS	-0.0908	-0.0904	-0.8616	0.7601	-6.0896	3.9753	10.0649
eq	-0.0660	-0.0248	-0.8601	0.7891	-6.6375	4.7572	11.3947

Tabela 3: Statystyki opisowe stóp zwrotu portfeli minimalnej zmienności (opracowanie własne)

W rozważanym przypadku maksymalizacja oczekiwanej stopy zwrotu portfela nie jest zadana w ramach problemu optymalizacyjnego, co odzwierciedliło się w umiarkowanie niższych poziomach średnich stóp zwrotu portfeli z wagami $\mathbf{w}_t^{\mathrm{ROWCov-\ MCD}}$, $\mathbf{w}_t^{\mathrm{ROWCov-PCS}}$ oraz $\mathbf{w}_t^{\mathrm{RCov}}$ (średnia wartość stopy zwrotu portfeli odpowiednio na poziomie -0,0931, -0,0908 oraz -0,0890), względem portfela o równych wagach $\mathbf{w}_t^{eq}$(średnia: -0,0660). Wartość minimalna stopy zwrotu $R_{P,t}^{\mathrm{ROWCov-PCS}}$ dla portfela wykorzystującego w konstrukcji prognoz ROWCov-PCS, jest wyższa od odpowiedników w pozostałych portfelach, w tym zauważalnie wyższa w relacji do minimum $R_{P,t}^{\mathrm{eq}}$. Najniższy rozstęp stóp zwrotu zaobserwowano dla portfela z prognozami skonstruowanymi z wykorzystaniem ROWCov-PCS.
Statystyki dotyczące mierników zrealizowanej zmienności portfela $\mathrm{sd}_{P,t}^{m}$ (mierniki te wyznaczono z wykorzystaniem trzech mierników zrealizowanej kowariancji $\mathbf{V}_t$: $\mathrm{RCov}_{t,\Delta}$, $\mathrm{ROWCov}_{t,\Delta}^{MCD}$, $\mathrm{ROWCov}_{t,\Delta}^{PCS}$), dla $t = 31, \ldots, 340$ przedstawiały się następująco:

$\mathbf{V}_t$ (miernik zrealizowanej kowariancji *ex post*)	m	$\mathrm{sd}^m_{P,t}$ średnia arytm.	$\mathrm{sd}^m_{P,t}$ mediana	$\mathrm{sd}^m_{P,t}$ Q1	$\mathrm{sd}^m_{P,t}$ Q3	$\mathrm{sd}^m_{P,t}$ min	$\mathrm{sd}^m_{P,t}$ max	$\mathrm{sd}^m_{P,t}$ rozstęp
$\mathrm{RCov}_{t,\Delta}$	RCov	1.0382	0.9734	0.7714	1.1865	0.3465	3.7506	3.4041
	ROWCov-MCD	1.0369	0.9717	0.7773	1.1946	0.3471	3.8488	3.5017
	ROWCov-PCS	1.0392	0.9696	0.7755	1.1984	0.3468	3.6829	3.3361
	eq	1.0363	0.9791	0.7752	1.2082	0.3516	3.9078	3.5562
$\mathrm{ROWCov}^{MCD}_{t,\Delta}$	RCov	0.9928	0.9446	0.7390	1.1509	0.3474	3.5836	3.2362
	ROWCov-MCD	0.9921	0.9450	0.7342	1.1550	0.3480	3.6617	3.3137
	ROWCov-PCS	0.9936	0.9499	0.7335	1.1547	0.3477	3.5285	3.1808
	eq	0.9899	0.9286	0.7415	1.1528	0.3526	3.7292	3.3766
$\mathrm{ROWCov}^{PCS}_{t,\Delta}$	RCov	1.0044	0.9508	0.7558	1.1656	0.3474	3.7602	3.4128
	ROWCov-MCD	1.0037	0.9540	0.7553	1.1631	0.3480	3.8586	3.5106
	ROWCov-PCS	1.0053	0.9551	0.7489	1.1739	0.3477	3.6924	3.3447
	eq	1.0024	0.9400	0.7522	1.1805	0.3526	3.9180	3.5654

Tabela 4: Mierniki zrealizowanej zmienności portfeli minimalnej zmienności (opracowanie własne)

Przeciętne poziomy zrealizowanej zmienności *ex post* zbudowanych portfeli (bez względu na podstawę określenia struktury wag: $\mathbf{w}_t^{\mathrm{RCov}}$, $\mathbf{w}_t^{\mathrm{ROWCov-\,MCD}}$, $\mathbf{w}_t^{\mathrm{ROWCov-PCS}}$ oraz $\mathbf{w}_t^{eq}$), były wyższe (1,03-1,04), gdy podstawą wyznaczania miernika zmienności portfela była macierz RCov, niż miało to miejsce w przypadku zastosowania jej odpornych odpowiedników ROWCov-MCD (0,99) oraz ROWCov-PCS (1,00–1,01).
Różnice w przeciętnych poziomach zmienności zrealizowanej *ex post* – niezależnie od rozważanej miary, są dla portfeli o różnych strukturach wag bardzo niewielkie. Niekorzystnym wydaje się być fakt, iż przeciętny poziom zrealizowanej zmienności portfeli *ex post*, które w problemie optymalizacyjnym zakładały minimalizację ryzyka portfela (mierzonego w oparciu o prognozy warunkowej względem przeszłości kowariancji stóp zwrotu) nie był znaczący niższy niż ten dla portfeli o równych wagach. Tylko w przypadku miernika zrealizowanej zmienności portfela opartego na ROWCov-MCD, najniższym średnim jej poziomem charakteryzował się portfel o strukturze wag $\mathbf{w}_t^{\mathrm{ROWCov-\,MCD}}$, dla którego prognozy warunkowej kowariancji (podstawiane jako oszacowania parametrów w problemie optymalizacyjnym) wykorzystywały w swej konstrukcji tę samą miarę zrealizowanej kowariancji. Także w tym przypadku różnica w rozważanym poziomie zmienności zrealizowanej względem pozostałych portfeli była znikoma.
Statystyki dotyczące nadwyżki stopy zwrotu $R^m_{P,t}$ zestawianego portfela nad stopą zwrotu portfela o strukturze wag $\mathbf{w}_t^{\mathrm{RCov}}$, $t = 31, \ldots, 340$, przedstawiają się następująco:

m	$R^m_{P,t} - R^{RCov}_{P,t}$ średnia arytm.	$R^m_{P,t} - R^{RCov}_{P,t}$ mediana	$R^m_{P,t} - R^{RCov}_{P,t}$ odch. stand.	$R^m_{P,t} - R^{RCov}_{P,t}$ Q1	$R^m_{P,t} - R^{RCov}_{P,t}$ Q3	$R^m_{P,t} - R^{RCov}_{P,t}$ min	$R^m_{P,t} - R^{RCov}_{P,t}$ max
ROWCov-MCD	-0.0040	0.0017	0.0602	-0.0209	0.0213	-0.6301	0.2122
ROWCov-PCS	-0.0017	0.0002	0.0440	-0.0203	0.0208	-0.2654	0.1559
eq	0.0230	0.0029	0.2491	-0.0608	0.0764	-0.9943	1.1772

Tabela 5: Statystyki nadwyżek stóp zwrotu odpornych portfeli minimalnej zmienności (opracowanie własne)

Statystyki dotyczące kształtowania się wskaźnika Giniego mierzącego nierównomierność rozkładu wag portfela, w przedziale czasowym $t = 31, \ldots, 340$:

Struktura wag portfela	Gini - wagi średnia arytm.	Gini - wagi mediana	Gini - wagi odch. stand.	Gini - wagi Q1	Gini - wagi Q3	Gini - wagi min	Gini - wagi max
$\mathbf{w}_t^{\text{RCov}}$	0.1056	0.0820	0.0670	0.0561	0.1398	0.0049	0.2709
$\mathbf{w}_t^{\text{ROWCov- MCD}}$	0.1039	0.0826	0.0701	0.0516	0.1657	0.0047	0.2608
$\mathbf{w}_t^{\text{ROWCov-PCS}}$	0.1081	0.0913	0.0705	0.0504	0.1564	0.0015	0.2638
$\mathbf{w}_t^{eq}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Tabela 6: Wskaźniki koncentracji rozkładu wag portfeli minimalnej zmienności (opracowanie własne)

Przeciętny poziom koncentracji (nierównomierności) wag był dla portfeli o strukturze $\mathbf{w}_t^{\text{RCov}}$, $\mathbf{w}_t^{\text{ROWCov- MCD}}$, $\mathbf{w}_t^{\text{ROWCov-PCS}}$ zbliżony i niewielki, zawierał się w przedziale 0,10–0,11 (mediany: 0,08–0,09). W przypadku portfela z równymi wagami $\mathbf{w}_t^{eq}$, w związku z definicją wskaźnika Giniego, poziom nierówności jest niezmiennie zerowy.
Statystyki dotyczące kształtowania się poziomu koncentracji mierzonej wskaźnikiem Giniego, dotyczącej krańcowego wkładu poszczególnych aktywów w łączne ryzyko portfela, wyrażane przez $\text{sd}_{P,t}^m = \sqrt{(\mathbf{w}_t^m)' \mathbf{V}_t \mathbf{w}_t^m}$
(za macierz zrealizowanej kowariancji *ex post* $\mathbf{V}_t$ przyjęto $\text{RCov}_{t,\Delta}$ – ze względu na podobieństwo wyników pomija się prezentację wyników dla $\mathbf{V}_t$ równego $\text{ROWCov}_{t,\Delta}^{MCD}$, oraz $\text{ROWCov}_{t,\Delta}^{PCS}$), w przedziale czasowym $t = 31, \ldots, 340$:

Struktura wag portfela	Gini – wkład w ryzyko średnia arytm.	Gini – wkład w ryzyko mediana	Gini – wkład w ryzyko odch. stand.	Gini – wkład w ryzyko Q1	Gini – wkład w ryzyko Q3	Gini – wkład w ryzyko min	Gini – wkład w ryzyko max
$\mathbf{w}_t^{\text{RCov}}$	0.1955	0.1856	0.0971	0.1222	0.2552	0.0176	0.5718
$\mathbf{w}_t^{\text{ROWCov- MCD}}$	0.1950	0.1903	0.0970	0.1216	0.2583	0.0095	0.5744
$\mathbf{w}_t^{\text{ROWCov-PCS}}$	0.1991	0.1953	0.0991	0.1294	0.2650	0.0066	0.5756
$\mathbf{w}_t^{eq}$	0.1442	0.1313	0.0802	0.0895	0.1883	0.0145	0.5758

Tabela 7: Wskaźniki koncentracji wkładów aktywów w łączne ryzyko portfeli minimalnej (opracowanie własne)

Przeciętny poziom koncentracji krańcowych wkładów aktywów w ryzyko portfela (indukowane przez wagi portfela oraz zrealizowaną macierz kowariancji $\text{RCov}_{t,\Delta}$), był zbliżony dla portfeli optymalizowanych w kierunku minimalnej zmienności (tzn. o strukturze wag $\mathbf{w}_t^{\text{RCov}}$,$\mathbf{w}_t^{\text{ROWCov- MCD}}$, $\mathbf{w}_t^{\text{ROWCov-PCS}}$) i zawierał się on w przedziale 0,19–0,20. Średni poziom koncentracji ryzyka tych portfeli był wyższy w porównaniu do portfela o równych wagach $\mathbf{w}_t^{eq}$, dla którego wyniósł 0,14.

Portfel ERC (*Equal Risk Contribution*)

W przypadku sekwencyjnej konstrukcji portfeli ERC problem optymalizacyjny sprowadza się do określenia przy odpowiednich ograniczeniach wag w taki sposób, aby krańcowy udział każdej składowej portfela w jego ryzyku był (możliwie jak najbardziej) równy. Ryzyko portfela określone jest przez $\sigma_P = \sigma(\mathbf{w}) = \sqrt{\mathbf{w'\Sigma w}}$, przy krańcowym wkładzie i-tej składowej portfela definiowanej przez $\sigma_i(\mathbf{w}) = w_i \cdot \frac{\partial \sigma(\mathbf{w})}{\partial w_i}$, dla każdego dnia sesyjnego parametr kowariancji zastępuje się prognozą $\mathbf{S}_t$ dotycząca warunkowej względem przeszłości macierzy kowariancji.
Jako narzędzie prognoz wykorzystano ruchomy model warunkowej kowariancji. W celu dokonania porównania wyników z tymi prezentowanymi wcześniej dla portfela o minimalnej wariancji, przyjęto jako podstawę prognoz ten sam model. Dla przypomnienia parametr zanikania α ustalono na poziomie 0,05, za $\mathbf{V}_{t-1}$, przyjmowano odpowiednio $\text{RCov}_{t-1,\Delta}$, $\text{ROWCov}_{t-1,\Delta}^{MCD}$ oraz $\text{ROWCov}_{t-1,\Delta}^{PCS}$, w przypadku tych dwóch ostatnich przy wyznaczaniu ich wartości wykorzystano estymatory MCD i PCS z $h = \lfloor 0,75 \cdot (\#N_i) \rfloor$, system wag HR z parametrem $k = \chi^2_{1-\beta,p}$, $\beta = 0,001$, gdzie $\chi^2_{1-\beta,p}$ oznacza kwantyl $1-\beta$ rozkładu chi-kwadrat z p stopniami swobody. Wektory wag

$\mathbf{w}_t^{\text{RCov}}$, $\mathbf{w}_t^{\text{ROWCov- MCD}}$, $\mathbf{w}_t^{\text{ROWCov-PCS}}$ odpowiadające rozwiązaniom problemu optymalizacyjnego, w którym przyjmowano za parametry kolejne prognozy warunkowej kowariancji oparte na danym modelu, tworzą szeregi czasowe.
W ramach tworzenia kodu w języku R umożliwiającego implementację przyjętej procedury budowy portfeli ERC poza wcześniej wymienianymi pakietami wykorzystano funkcję `PERC` pakietu `FRAPO` – Pfaff, 2011.
W ramach badania jakości tworzonych portfeli ERC, ciężar przenosi się z oceny poziomu ich zmienności $\text{sd}_{P,t}^m$, czyli ryzyka, na pomiar równomierności krańcowych wkładów składowych portfela w to ryzyko. Zarówno ryzyko portfela jak i krańcowy wkład w nie, dla kolejnych dni sesyjnych wyznacza się z wykorzystaniem wyznaczonej *ex post* (w oparciu o wewnątrzdzienne stopy zwrotu aktywów) macierzy zrealizowanej kowariancji (w wariancie RCov, ROWCov-MCD, ROWCov-PCS).
Pomiaru stopnia koncentracji (nierówności) krańcowych wkładów w ryzyko portfeli, dla kolejnych dni sesyjnych dokonuje się z wykorzystaniem wskaźnika Giniego.
Analizą w tym względzie, tak jak w poprzednim przypadku objęto przedział czasowy $t = 31, \ldots, 340$.
Dodatkowo przywołane zostaną podstawowe statystyki opisowe stóp zwrotu portfelu w ramach *backtestingu*.
W *backtestingu* dla dni sesyjnych $t = 31, \ldots, 340$ odchylenie standardowe stóp zwrotu portfeli ERC, dla których przyjęto szeregi wag $\mathbf{w}_t^{\text{RCov}}$, $\mathbf{w}_t^{\text{ROWCov- MCD}}$, $\mathbf{w}_t^{\text{ROWCov-PCS}}$ oraz $\mathbf{w}_t^{eq}$, przedstawiało się następująco:

m	RCov	ROWCov-MCD	ROWCov-PCS	eq
odchylenie standardowe dla próby $R_{P,t}^m$, $t = 31, \ldots, 340$	1.4162	1.4171	1.4136	1.4350

Tabela 8: Odchylenie standardowe stóp zwrotu portfeli ERC (opracowanie własne)

Wartość odchylenia standardowego wyznaczona dla stóp zwrotu portfeli ERC w przedziale czasowym objętym *backtestingiem,* kształtowała się na bardzo zbliżonym poziomie powyżej 1,41, który był wyższy niż w przypadku portfeli minimalnej zmienności – 1,38–1,39 (w przypadku portfeli ERC problem optymalizacyjny nie obejmuje minimalizacji ich zmienności), a tym samym niższy niż takowy dla portfeli o równych wagach – 1,44.
Zanim przejdzie się do omówienia stopnia koncentracji krańcowych wkładów w ryzyko portfeli ERC, warto poddać ocenie stopień koncentracji rozkładu wag.
Statystyki dotyczące kształtowania się wskaźnika Giniego mierzącego nierównomierność rozkładu wag portfeli ERC, w przedziale czasowym $t = 31, \ldots, 340$:

Struktura wag portfela	Gini - wagi średnia arytm.	Gini - wagi mediana	Gini - wagi odch. stand.	Gini - wagi Q1	Gini - wagi Q3	Gini - wagi min	Gini - wagi max
$\mathbf{w}_t^{\text{RCov}}$	0.0327	0.0267	0.0212	0.0182	0.0407	0.0017	0.0900
$\mathbf{w}_t^{\text{ROWCov- MCD}}$	0.0307	0.0262	0.0202	0.0164	0.0420	0.0017	0.0826
$\mathbf{w}_t^{\text{ROWCov-PCS}}$	0.0330	0.0299	0.0218	0.0162	0.0432	0.0004	0.0867
$\mathbf{w}_t^{eq}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Tabela 9: Wskaźniki koncentracji rozkładu wag portfeli ERC (opracowanie własne)

Przeciętny poziom nierównomierności rozkładu wag dla portfeli ERC był bardzo niewielki – wartość wskaźnika Giniego równa 0,03. Oznacza to, że przeciętnie struktura portfela ERC była zbliżona do portfela z równymi wagami. Nawet maksymalna odnotowana wartość wskaźnika Giniego dla wag była niewielka i wynosiła 0,08–0,09.
Jak już wspomniano głównym kryterium oceny jakości zbudowanych portfeli ERC jest stopień równomierności krańcowych wpływów składowych portfela na jego łączne ryzyko, które to mierniki wyznaczane są *ex post* dla kolejnych dni sesyjnych, w oparciu o trzy warianty kowariancji zrealizowanej: RCov, ROWCov-MCD, ROWCov-PCS. Jego oceny dokonuje się w oparciu o szeregi czasowe wartości wskaźników Giniego.

Statystyki dotyczące kształtowania się poziomu koncentracji mierzonej wskaźnikiem Giniego, dotyczącej krańcowego wkładu poszczególnych aktywów w łączne ryzyko portfeli ERC, wyrażane przez $\mathrm{sd}_{P,t}^{m} = \sqrt{(\mathbf{w}_t^m)' \mathbf{V}_t \mathbf{w}_t^m}$
(za macierz zrealizowanej kowariancji *ex post* $\mathbf{V}_t$ przyjęto $\mathrm{RCov}_{t,\Delta}$ – ze względu na podobieństwo wyników pomija się prezentację wyników dla $\mathbf{V}_t$ równego $\mathrm{ROWCov}_{t,\Delta}^{MCD}$, oraz $\mathrm{ROWCov}_{t,\Delta}^{PCS}$), w przedziale czasowym $t = 31, \ldots, 340$:

Struktura wag portfela	Gini – wkład w ryzyko średnia arytm.	Gini – wkład w ryzyko mediana	Gini – wkład w ryzyko odch. stand.	Gini – wkład w ryzyko Q1	Gini – wkład w ryzyko Q3	Gini – wkład w ryzyko min	Gini – wkład w ryzyko max
$\mathbf{w}_t^{\mathrm{RCov}}$	0.1423	0.1319	0.0819	0.0865	0.1843	0.0050	0.5745
$\mathbf{w}_t^{\mathrm{ROWCov-\ MCD}}$	0.1418	0.1310	0.0806	0.0861	0.1875	0.0029	0.5753
$\mathbf{w}_t^{\mathrm{ROWCov-PCS}}$	0.1428	0.1314	0.0813	0.0881	0.1868	0.0067	0.5757
$\mathbf{w}_t^{eq}$	0.1442	0.1313	0.0802	0.0895	0.1883	0.0145	0.5758

Tabela 10: Wskaźniki koncentracji wkładów aktywów w łczne ryzyko portfeli ERC (opracowanie własne)

Dla portfeli ERC przeciętny stopień nierównomierności wkładów w ryzyko aktywów odzwierciedlający się wartością wskaźnika Giniego na poziomie 0,14 zdaje się być wysoki, tym bardziej, że dla tych portfeli problem optymalizacyjny, wykorzystujący w swej konstrukcji prognozy warunkowej kowariancji, zakłada minimalizację odstępstw pomiędzy indywidualnymi udziałami w łącznym ryzyku portfela. Wspomniany poziom jest prawie identyczny z tym dla portfeli o równych wagach, co ma związek z ogólną tendencją do równomiernego rozkładu wag w skonstruowanych portfelach ERC. Maksymalna wartość wskaźnika Giniego dla wkładów w ryzyko portfeli ERC, była bardzo wysoka i wynosiła 0,57–0,58, przy górnym kwartylu na poziomie 0,18–0,19. Co prawda przeciętny poziom nierównomierności krańcowych wkładów w ryzyko portfeli ERC był niższy niż takowy dla portfeli minimalnej zmienności – przeciętna wartość wskaźnika Giniego na poziomie 0,19–0,20.
Wyniki dla portfeli ERC zdają się być niezadowalające, bez względu na to, które mierniki zrealizowanej kowariancji wykorzystano w ramach konstrukcji prognoz warunkowej kowariancji – klasyczny ROWCov, czy też odporne ROWCov-MCD i ROWCov-PCS.

7 Podsumowanie

Biorąc pod uwagę wykonane analizy, wpływ zastosowania w ramach tworzenia prognoz warunkowej kowariancji, zamiast klasycznego miernika zrealizowanej kowariancji RCov, jego odpornych odpowiedników ROWCov (które w swej konstrukcji wykorzystują wartości odpornych estymatorów rozrzutu MCD oraz PCS dla wewnątrzdziennych stóp zwrotu w ramach lokalnych okien), na poprawę wyników konstruowanych portfeli okazał się być niewielki. Zarówno w przypadku portfeli, dla których celem było uzyskanie minimalnej zmienności (patrząc na wartość odchylenia standardowego stóp zwrotu portfela w ramach *backtestingu* oraz kształtowanie się poziom jego zrealizowanej zmienności – ryzyka), jak też portfeli ERC, w których dąży się do uzyskania identycznego wpływu poszczególnych aktywów na łączne ryzyko (patrząc na kształtowania się poziom wskaźnika Giniego dla krańcowych wkładów w zrealizowane ryzyko portfela), zastosowanie podejścia odpornego nie dało znaczącej poprawy wartości najważniejszych kryteriów oceny w danej kategorii portfeli.
Niemniej jednak w niepublikowanych badaniach symulacyjnych autorów z wykorzystaniem pakietu R `DepthProc` [Kosiorowski, Zawadzki, 2016], w których obserwacje odstające wykrywane były w oparciu o odporny wariant funkcji głębi Mahalanobisa (próbkową macierz kowariancji zastępuje się odpornym estymatorem rozrzutu), zastosowanie estymatora PCS dawało lepsze wyniki niż MCD. W ramach dalszych badań autorzy planują ocenę możliwości zastosowania odpornych estymatorów rozrzutu w konstrukcji wielowymiarowych kart kontrolnych (m.in. karty T^2 Hotellinga).

Literatura

[Barndorff-Nielsen, Shephard, 2004] Barndorff-Nielsen O. E., Shephard N. (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72(3):885–925.

[Boudt i in., 2012] Boudt K., Cornelissen J., Payseur S. (2012). Highfrequency: Toolkit for the analysis of highfrequency financial data in r. *URL: http://cran. r-project. org/web/packages/highfrequency/vignettes/highfrequency. pdf*.

[Boudt i in., 2011] Boudt K., Croux C., Laurent S. (2011). Outlyingness weighted covariation. *Journal of Financial Econometrics*, 9(4):657–684.

[Butler i in., 1993] Butler R., Davies P., Jhun M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, pages 1385–1400.

[Croux, Haesbroeck, 1999] Croux C., Haesbroeck G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71(2):161–190.

[Davies, 1987] Davies P. L. (1987). Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, pages 1269–1292.

[Donoho, 1982] Donoho D. L. (1982). Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston. URL http://www-stat. stanford. edu/˜ donoho/Reports/Oldies/BPMLE. pdf.

[Fleming i in., 2001] Fleming J., Kirby C., Ostdiek B. (2001). The economic value of volatility timing. *The Journal of Finance*, 56(1):329–352.

[Fleming i in., 2003] Fleming J., Kirby C., Ostdiek B. (2003). The economic value of volatility timing using realized volatility. *Journal of Financial Economics*, 67(3):473–509.

[Hubert i in., 2012] Hubert M., Rousseeuw P. J., Verdonck T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21(3):618–637.

[Jensen i in., 2007] Jensen W. A., Birch J. B., Woodall W. H. (2007). High breakdown estimation methods for phase i multivariate control charts. *Quality and Reliability Engineering International*, 23(5):615–629.

[Kosiorowski, Zawadzki, 2016] Kosiorowski D., Zawadzki Z. (2016). *DepthProc An R Package for Robust Exploration of Multidimensional Economic Phenomena.*

[Kostrzewski, 2014] Kostrzewski M. (2014). Bayesian inference for the jump-diffusion model with m jumps. *Communications in Statistics-Theory and Methods*, 43(18):3955–3985.

[Maronna, Martin, 2006] Maronna R. A., Martin D. (2006). Yohai. robust statistics. *Wiley Series in Probability and Statistics. John Wiley and Sons*, 2:3.

[Maronna i in., 1992] Maronna R. A., Stahel W. A., Yohai V. J. (1992). Bias-robust estimators of multivariate scatter based on projections. *Journal of Multivariate Analysis*, 42(1):141–161.

[Pajor, 2010] Pajor A. (2010). Wielowymiarowe procesy wariancji stochastycznej w ekonometrii finansowej: ujęcie bayesowskie. *Zeszyty Naukowe/Uniwersytet Ekonomiczny w Krakowie. Seria Specjalna, Monografie*, (195).

[Pfaff, 2012] Pfaff B. (2012). *Financial risk modelling and portfolio optimization with R.* John Wiley & Sons.

[Pison i in., 2002] Pison G., Van Aelst S., Willems G. (2002). Small sample corrections for lts and mcd. *Metrika*, 55(1-2):111–123.

[Pooter i in., 2008] Pooter M. d., Martens M., Dijk D. v. (2008). Predicting the daily covariance matrix for s&p 100 stocks using intraday databut which frequency to use? *Econometric Reviews*, 27(1-3):199–229.

[Rousseeuw i in., 2015] Rousseeuw P., Croux C., Todorov V., Ruckstuhl A., Salibian-Barrera M., Verbeke T., Koller M., Maechler M. (2015). robustbase: Basic robust statistics. r package version 0.92-3.

[Rousseeuw, 1984] Rousseeuw P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880.

[Rousseeuw, Driessen, 1999] Rousseeuw P. J., Driessen K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

[Ryan, Ulrich, 2011] Ryan J. A., Ulrich J. M. (2011). xts: Extensible time series. *R package version 0.8-2*.

[Schmitt i in., 2014] Schmitt E., Öllerer V., Vakili K. (2014). The finite sample breakdown point of pcs. *Statistics & Probability Letters*, 94:214–220.

[Stahel, Maechler, 2009] Stahel W., Maechler M. (2009). robustx: experimental extraneous extraordinary. *Functionality for Robust Statistics. R package version*, pages 1–1.

[Stahel, 1981] Stahel W. A. (1981). *Breakdown of covariance estimators*. Fachgruppe für Statistik, Eidgenössische Techn. Hochsch.

[Steiner i in., 2009] Steiner S. H., Variyath A. M., et al. (2009). A multivariate robust control chart for individual observations. *Journal of Quality Technology*, 41(3):259.

[Todorov i in., 2009] Todorov V., Filzmoser P., et al. (2009). *An object-oriented framework for robust multivariate analysis*. Citeseer.

[Vakili, Schmitt, 2014] Vakili K., Schmitt E. (2014). Finding multivariate outliers with fastpcs. *Computational Statistics & Data Analysis*, 69:54–66.

[Würtz i in., 2009] Würtz D., Chalabi Y., Chen W., Ellis A. (2009). *Portfolio optimization with R/RMetrics*. Rmetrics.