ANALIZA KURSÓW AKCJI Z WYKORZYSTANIEM METODY ICA

Agnieszka Pasztyła
StatSoft Polska Sp. z o.o.; Akademia Ekonomiczna w Krakowie, Katedra Statystyki

Artykuł przedstawia nowe spojrzenie na zastosowanie analizy korelacji w praktyce inwestowania długoterminowego. Metoda składowych głównych, której podstawą jest analiza korelacji i metoda składowych niezależnych, będąca jej rozwinięciem, zostały zastosowane w celu wyodrębnienia czynników, które jednocześnie kształtują stopy zwrotu wybranych grup spółek. Analiza ta może być wykorzystana do tworzenia portfeli inwestycyjnych, w celu minimalizacji ryzyka, a także do prognozowania stóp zwrotu portfeli i spółek bez konieczności estymacji rozkładu zmiennych.

Analiza korelacji w konstruowaniu portfela inwestycyjnego

Badanie korelacji stóp zwrotu instrumentów finansowych jest jednym z podstawowych narzędzi analitycznych wykorzystywanych w konstruowaniu portfeli inwestycyjnych. Gdy zmienne, np. notowania akcji, są wzajemnie skorelowane, możemy mówić o współzależności tych zmiennych. Określenie kierunku i siły współzależności między szeregami czasowymi umożliwia taki wybór spółek, który umożliwi uzyskanie wymaganego poziomu ryzyka portfela. Miarą, która jest najczęściej wykorzystywana do ilościowego przedstawienia współzmienności szeregów, jest współczynnik liniowej korelacji Pearsona - ***r***. Obliczamy go, korzystając ze wzoru:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y},$$

gdzie:

$x_i, y_i, \bar{x}, \bar{y}$ – analizowane zmienne (np. stopy zwrotu), ich wartości i wartości średnie,

s_X, s_Y – odchylenie standardowe X i Y.

Współczynnik korelacji r, zgodnie z powyższym wzorem, przyjmuje wartości z przedziału <-1;1>. Im większa wartość bezwzględna r, tym silniejsza współzmienność. Znak współczynnika korelacji informuje o kierunku zmian. Dodatnia wartość mówi o jednokierunkowych odchyleniach szeregów, czyli równoczesnym spadku lub wzroście notowań spółek,

natomiast wartość ujemna świadczy o zmianach o przeciwnym kierunku. Jak wynika ze wzoru, współczynnik korelacji Pearsona dotyczy dwóch zmiennych. Aby analizować większą liczbę zmiennych, warto korzystać z macierzy korelacji, z której możemy odczytać siłę i kierunek współzmienności dowolnej pary zmiennych.

Współczynnik korelacji jest ważny w teorii portfela papierów wartościowych, ponieważ umożliwia dobór papierów wartościowych, który minimalizuje ryzyko portfela.

Istotnym mankamentem *r* jest fakt, że mierzy on tylko najprostszą, liniową zależność dwóch zmiennych. W sytuacji, gdy mamy do czynienia z nieliniowymi powiązaniami między zmianami notowań lub stóp zwrotu, mogą one nie zostać wykryte za pomocą współczynnika korelacji liniowej.

Wyodrębnianie wspólnych przyczyn wahań kursów spółek

Poza klasycznym zastosowaniem korelacji w zarządzaniu ryzykiem portfeli, opisanym powyżej, warto wziąć pod uwagę nieco inne spojrzenie na analizę współzmienności cech. Mianowicie, jeśli stopy zwrotu dwóch lub więcej akcji są ze sobą skorelowane, można mieć podejrzenie, że zmiany te mają wspólne przyczyny. Innymi słowy możemy powiedzieć, że skorelowane wzajemnie szeregi czasowe niosą, w stopniu określonym przez wartość *r*, te same informacje. Warto się zatem pokusić o wyodrębnienie części wspólnej analizowanych szeregów. Możemy to zrobić, korzystając np. z metody składowych głównych (ang. *Principal Component Analisys*).

Metoda składowych głównych ma swój początek w pracach Karla Pearsona, z przełomu dziewiętnastego i dwudziestego wieku, a następnie była rozwijana w latach trzydziestych przez H. Hotellinga. Jej znaczenie wzrosło wraz z możliwością wykorzystania do obliczeń komputerów i od tego czasu jest stale udoskonalana. Jest to metoda czysto matematyczna i nie wymaga żadnych założeń dotyczących rozkładu zmiennych ani tym bardziej reszt (nie buduje się modelu statystycznego, więc nie ma reszt). Koncepcja metody opiera się na określeniu stopnia współzależności zmiennych za pomocą współczynnika korelacji liniowej lub kowariancji, a następnie na wyodrębnieniu nowych, nieskorelowanych zmiennych, określanych jako składowe główne, które odpowiadają za część zmienności grup zmiennych lub nawet za zmienność całych grup. Nowo powstałe zmienne są liniowymi kombinacjami pierwotnych zmiennych i kolejne składowe mają za zadanie ujmować jak najwięcej informacji zawartych w oryginalnych danych. Istotne jest tutaj założenie, że pierwsza ze składowych powinna zawierać jak największą porcję informacji, natomiast kolejne coraz mniej. Miarą informacji w metodzie głównych składowych jest wariancja, czyli miara zmienności cechy, stąd kolejne składowe powinny charakteryzować się coraz mniejszą wariancją.

Aby wyznaczyć nowe zmienne, które mają być składowymi głównymi, korzystamy z własności macierzy korelacji lub kowariancji. Można bowiem dowieść, że kolejnymi składowymi głównymi są wektory własne macierzy korelacji (np. por. [1], s.58-62), które obliczamy metodami algebry liniowej.

Załóżmy, że dysponujemy zbiorem danych, w którym mamy *p* zmiennych, będących szeregami czasowymi stóp zwrotu dla *p* spółek, oznaczonych za pomocą X_1, X_2, X_3, ..., X_p. Przyjmijmy również, że poszczególne zmienne są skorelowane wzajemnie. W oparciu o macierz korelacji zmiennych znajdujemy za pomocą metody głównych składowych nowy zestaw zmiennych, np. Y_1, Y_2, ..., Y_p, które są nieskorelowane i których wariancja maleje dla kolejnych Y_j. Każda nowa zmienna Y_j będzie liniową kombinacją oryginalnych zmiennych X_i:

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + ... + a_{pj}X_p.$$

W ten sposób możemy wyodrębnić czynniki, które leżą u podstaw analizowanych szeregów, a które często są niemierzalne, a nawet niemożliwe jest ich określenie w sposób jednoznaczny. Poszczególne składowe Y_j mogą odpowiadać za wpływ różnych zjawisk, których efektem może być np. trend wzrostowy, trend spadkowy, wahania określonego typu itp. Metoda ta nie pozwala na interpretację wyizolowanych efektów. Wyjaśnienie poszczególnych tendencji jest możliwe tylko w oparciu o wiedzę merytoryczną i znajomość badanej problematyki.

W metodzie składowych głównych poszczególne składowe są nieskorelowane ze sobą. Oznacza to, że nie ma liniowego związku między nimi. Możemy mieć jednak do czynienia z zależnością nieliniową. Wówczas kowariancja lub współczynnik korelacji będą równe zeru, a pomimo tego zmienne będą zależne, czyli dalej będzie istnieć część wariancji, której przyczyna będzie wspólna dla obu zmiennych. Wyjątkiem jest sytuacja, gdy obie zmienne podlegają dwuwymiarowemu rozkładowi normalnemu, wówczas możemy mówić o niezależności zmiennych, jeśli wiemy, że korelacja jest zerowa. Można więc powiedzieć, że metoda składowych głównych najlepiej sprawdza się w przypadku zmiennych podlegających wielowymiarowemu rozkładowi normalnemu lub rozkładów zbliżonych do normalnego. Jednak, jak wiadomo, stopień dopasowania rozkładu normalnego do empirycznych danych giełdowych nie jest wysoki.

W odpowiedzi na ten istotny mankament rozpoczęto pracę nad metodą, która pozwoliłaby na zmianę założenia o braku korelacji składowych na założenie silniejsze, które dopuszczać będzie składowe niezależne statystycznie. W ciągu ostatnich kilkunastu lat zaproponowano wiele algorytmów, opartych m. in. na minimalizacji momentów centralnych wyższych rzędów (Cardoso, Soloumniac, 1993), minimalizacji wzajemnej informacji składowych lub maksymalizacji entropii składowych (Bell, Sejnowski, 1995) oraz minimalizacji odległości Kullbacka-Leiblera między łącznym i sumą brzegowych rozkładów poszczególnych składowych (Amari et. al, 1996). Wszystkie te rozwiązania noszą wspólną nazwę metody składowych niezależnych. Poniżej zostaną przedstawione te, które wykorzystują entropię.

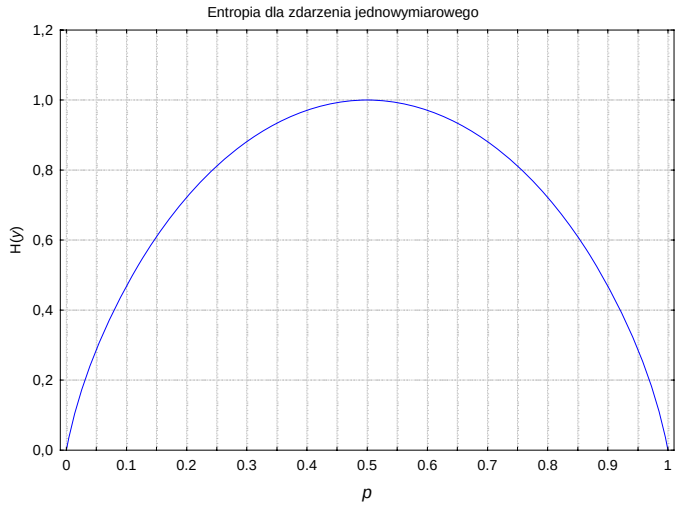
Entropia jako miara informacji zawartej w zmiennej

Ponieważ współczynnik korelacji sprawdza się jako miara niezależności składowych tylko w przypadku wielowymiarowego rozkładu normalnego, na początku lat dziewięćdziesiątych zaproponowano wykorzystanie w metodzie składowych niezależnych miar opartych

na entropii do badania niezależności składowych. W teorii informacji *entropia* jest miarą informacji zawartej w sygnale, natomiast w statystyce interpretuje się ją jako średnią wartość funkcji określonej na zbiorze prawdopodobieństw wszystkich możliwych realizacji pewnego doświadczenia. Funkcja ta określa ilość informacji, jaką niesie pojedyncze zdarzenie. Oznaczmy przez $H(Y)$ entropię zmiennej Y, wówczas

$$H(Y) = -\sum_{i=1}^{k} p_i \log_2 p_i,$$

gdzie p_i to prawdopodobieństwo wystąpienia zdarzenia y_i. Entropia jest zawsze nieujemna i równa zeru tylko w takim przypadku, gdy jedno zdarzenie występuje z prawdopodobieństwem równym jedności, a pozostałe mają prawdopodobieństwa równe zeru. Natomiast osiąga wartość maksymalną w przypadku, gdy prawdopodobieństwa wszystkich zdarzeń są równe. Jeśli wiadomo, że wystąpienie analizowanego zdarzenia jest pewne, to doniesienie o tym, że miało ono miejsce, nie dostarcza nam żadnej informacji. Np. jeśli w gorący letni dzień słyszymy, że temperatura wynosi powyżej 25°*C*, to taki komunikat wnosi znikomą ilość informacji. Jeśli natomiast lato jest deszczowe i zimne, to prawdopodobieństwo, że temperatura wzrośnie następnego dnia do 30°*C* jest różna od zera i jedności, i taka wiadomość wnosi znaczną ilość informacji. Możemy również powiedzieć, że entropia jest miarą nieokreśloności. W pierwszej sytuacji stopień nieokreśloności jest bliski zeru, natomiast w drugim dosyć duży. Dla pojedynczego zdarzenia, które może przyjąć dwa stany A i B, z prawdopodobieństwem p i q ($p + q = 1$), można wykreślić entropię jako funkcję prawdopodobieństwa, np. p.



Maksimum entropii (a więc nieokreśloności) jest osiągane wówczas, gdy $p = q = 0{,}5$, tzn. gdy oba stany są jednakowo prawdopodobne. Z kolei entropia jest minimalna, gdy istnieje pewność (p=1 lub q=1), że zdarzenie przyjmie wartość odpowiednio A lub B.

Aby określić stopień zależności między zmiennymi, konstruuje się miarę określaną jako *wzajemna informacja* (ang. *mutual information*) $I(Y)$, której podstawą jest entropia poszczególnych zmiennych. Wzajemna informacja obliczana jest jako suma różnic między entropią gęstości rozkładów brzegowych zmiennej (niezależnych statystycznie), $H(Y_j)$, i entropią gęstości rozkładu analizowanej zmiennej $H(Y)$.

$$I(Y) = \sum_{j=1}^{p} (H(Y_j) - H(Y)).$$

Miara ta jest modyfikacją funkcji *odległości Kullbacka–Leiblera* dla dwóch rozkładów prawdopodobieństwa, w której nie jest wykorzystywana entropia.

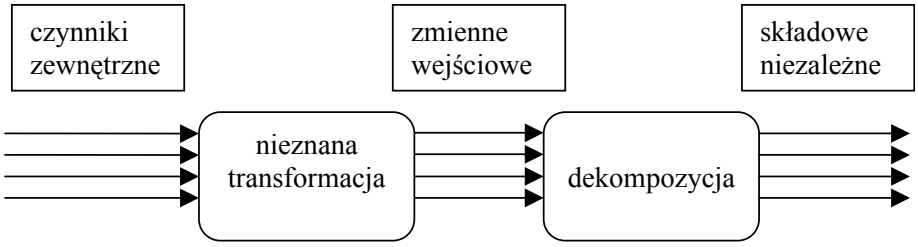
W praktyce do mierzenia zależności zmiennych zamiast entropii stosuje się *negentropię*, $J(Y_j)$, czyli miarę, która określa, jak bardzo różni się rozproszenie i koncentracja cechy o dowolnym rozkładzie od cechy o takiej samej wariancji, ale podlegającej rozkładowi normalnemu. Podstawą porównania jest rozkład normalny, ponieważ zmienna podlegająca temu rozkładowi charakteryzuje się największą entropią.

$$J(Y_j) = H(Z_j) - H(Y_j)$$

Z_j jest losową zmienną podlegającą rozkładowi normalnemu o takiej samej wariancji jak Y_j. Negentropia jest nieujemna i mierzy odległość rozkładu składowej Y_j od rozkładu normalnego.

Metoda składowych niezależnych

Metoda składowych niezależnych pozwala na dekompozycję, czyli rozkład zmiennych wejściowych na statystycznie niezależne składowe w oparciu o miary niezależności cech opisane powyżej.



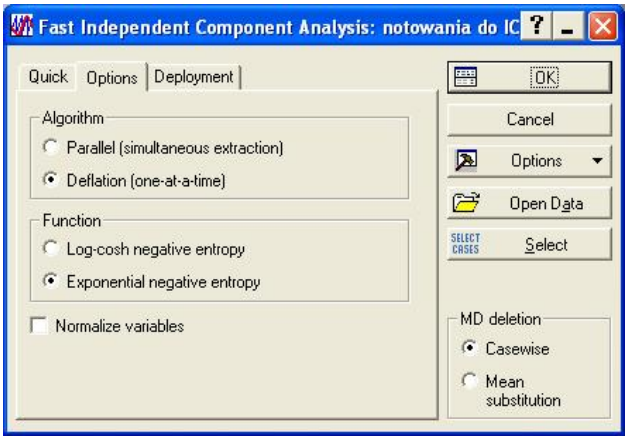
W przypadku giełdy czynniki zewnętrzne, takie jak: ożywienie rynku, ogólna kondycja gospodarki, wielkość deficytu budżetowego, stabilność rządu i inne, kształtują w dużym stopniu, obok informacji wewnętrznych pochodzących z samej spółki, wahania kursów i w efekcie stopy zwrotu z inwestycji. Wykazanie zależności między wybranymi czynnikami zewnętrznymi a notowaniami spółek jest utrudnione ze względu na problemy z przedstawieniem badanych efektów w postaci ilościowej. Pomocne tutaj może być wykorzystanie metody składowych niezależnych w celu wyodrębnienia wspólnych

przyczyn wahań notowań akcji. Niemierzalne czynniki zewnętrzne są estymowane za pomocą wyizolowanych składowych niezależnych. Dużą zaletą analizy jest brak konieczności estymacji rozkładu zmiennych. Jest to efekt zastosowania entropii jako podstawy miar niezależności zmiennych.

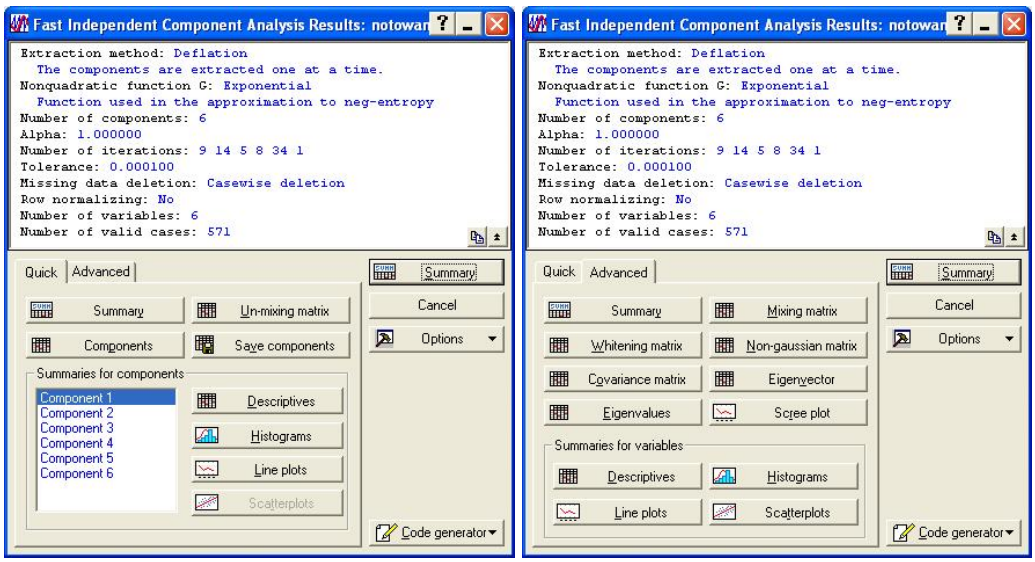
Przykład

W przykładzie jako zmienne wejściowe wzięto dzienne notowania spółek sektora informatycznego. Zbiór zawiera stopy zwrotu sześciu spółek (Comarch, Netia, Optimus, Interia, Prokom, TP SA) za okres od 28.01.2002 do 14.05.2004, obliczone na podstawie kursów zamknięcia. Pytanie, jakie możemy postawić na początku analizy, to np. czy możemy wyodrębnić wspólne źródła zmienności (wahań) kursów akcji spółek w ramach jednego sektora w długim okresie czasu (blisko 2,5 roku) oraz jaki procent zmienności ogółu cech może być wyjaśniony przez wyodrębnione czynniki.

W celu przeprowadzenia obliczeń wykorzystano algorytm zaimplementowany w programie *STATISTICA Data Miner*. W pierwszej kolejności określamy zmienne, które będą podlegać analizie (stopy zwrotu spółek) oraz liczbę składowych, które chcemy wyodrębnić. Liczba ta nie może przekraczać liczby zmiennych uwzględnianych w analizie. Dla potrzeb analizy ustalmy maksymalną liczbę składowych, czyli 6.



W programie *STATISTICA Data Miner* zostały zaimplementowane dwie wersje algorytmu: za pomocą opcji *Parallel* wyznaczamy równocześnie kilka składowych, natomiast opcja *Deflation* pozwala na kolejne wyznaczenie składowych. Szacowanie negentropii dla zmiennych odbywa się za pomocą funkcji *log-cosh* lub wykładniczej. Funkcjonalność programu obejmuje również normalizację zmiennych przed rozpoczęciem analizy. Ponieważ wykorzystywane przez nas stopy zwrotu przyjmują wartości z przedziału <-1;1>, nie jest konieczna ich normalizacja. Karty z wynikami zostały przedstawione poniżej.



Korzystając z karty *Qiuck*, otrzymujemy krótkie podsumowanie wartości parametrów i opcji wykorzystanych w analizie (*Summary*).

Specification	Summary of Fast Ir **Value**
Extraction method	Deflation
Nonquadratic function G	Exponential
Number of components	6
Alpha	1.000000
Number of iterations	9 14 5 8 34 1
Tolerance	0.000100
Missing data deletion	Casewise deletion
Row normalizing	No
Number of variables	6
Number of valid cases	571

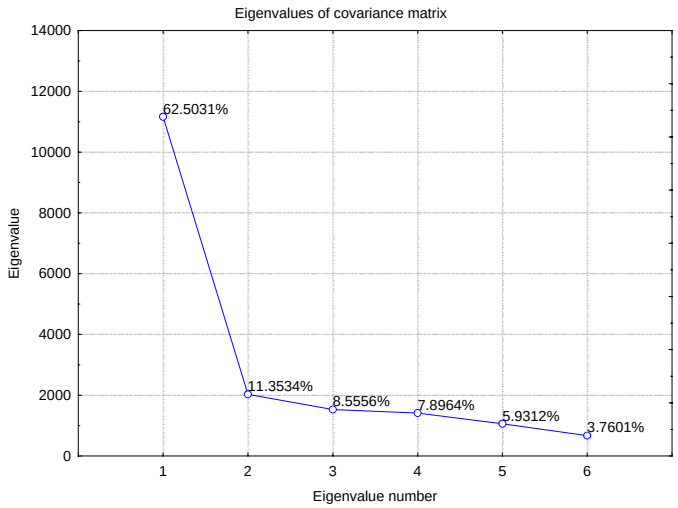
Otrzymujemy również tablicę z wartościami, jakie przyjmują poszczególne składowe (*Components*) oraz współczynniki przy zmiennych w kombinacjach liniowych, będących składowymi (*Un-mixing matrix*). Macierz ta spełnia równanie: $S = A^{-1}X$, gdzie S oznacza składowe główne, a A^{-1} to macierz współczynników przy zmiennych X_j. Fragment macierzy współczynników znajduje się poniżej.

	Un-mixing matrix (notowania do ICA) Estimated un-mixing matrix computed from centered data.					
	comarch	netia	onet	optimus	tpsa	interia
Component 1	0.005841	0.004944	0.007849	0.000084	0.003538	0.024619
Component 2	-0.021773	-0.006179	-0.011522	0.015014	-0.009398	0.008982
Component 3	-0.001219	-0.001746	0.006134	-0.009532	0.003191	-0.000654
Component 4	0.006355	-0.017290	0.001683	0.011821	0.006963	0.001542
Component 5	-0.006670	-0.000121	0.016370	0.018141	-0.018262	-0.004624
Component 6	-0.012395	0.007106	0.007019	0.013990	0.021772	-0.002400

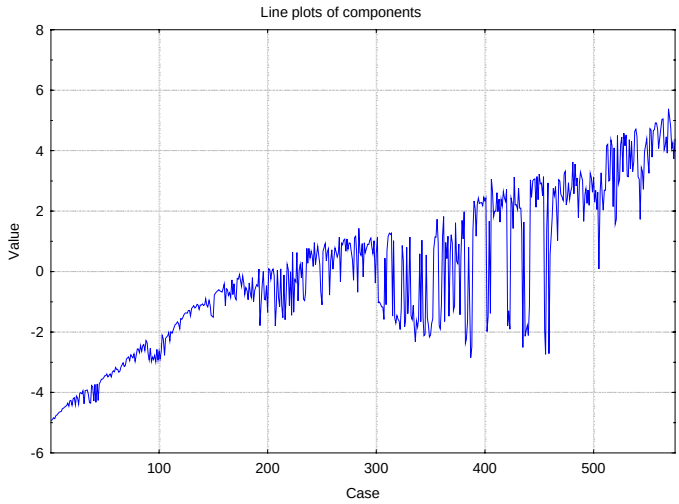
Macierz odwrotna, czyli A, spełnia następujące równanie: $X = AS$ i jest określana jako *mixing matrix*. Fragment macierzy A przedstawiony jest na poniższym rysunku.

Variable	Mixing matrix (notowania do ICA) Estimated mixing matrix computed from centered data. Component 1	Component 2	Component 3	Component 4	Component 5	Component 6
comarch	7.29658	-29.9723	-51.4494	13.0640	-1.5034	-12.0222
netia	8.34123	-14.9581	-48.7898	-39.5830	1.9709	13.6507
onet	10.58687	-10.7148	49.5329	4.5784	26.4037	7.0777
optimus	2.15427	-3.0416	-55.5566	12.1737	11.2413	12.0149
tpsa	0.35794	-5.1978	6.9598	11.2490	-17.8870	24.0078
interia	33.77875	14.2881	5.4031	1.7321	-5.9242	-5.6372

Aby odpowiedzieć na zadane na początku pytanie, czy możemy wyodrębnić czynniki, które kształtują stopy zwrotu wybranych spółek, skorzystamy z karty *Advanced* i tablicy zawierającej wartości własne macierzy kowariancji oraz z wykresu osypiska (*Scree plot*).

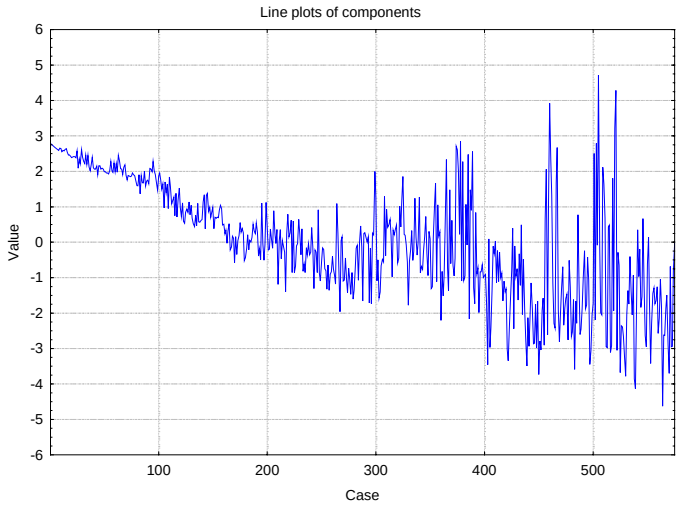


Na podstawie przedstawionych powyżej wyników możemy stwierdzić, że istnieje czynnik, który w bardzo dużym stopniu (62,5%) wyjaśnia zmienność ogółu analizowanych akcji.



Kształtowanie się tego czynnika w czasie możemy zobaczyć na powyższym wykresie (karta *Quick*, *Line plots*).

Poniżej znajduje się wykres drugiego czynnika (składowej), który niesie ok. 11% informacji wspólnych dla analizowanych spółek.



Podsumowanie

Wykorzystanie analizy składowych niezależnych może być różnorodne. Z pewnością wniesie istotną informację na temat badanego zjawiska, a w tym przypadku pozwoli na dokładniejsze poznanie przyczyn kształtujących analizowane stopy zwrotu. Podsumowując, metoda niezależnych składowych pozwala na oszacowanie, jaka część zmienności wybranych kursów niesie wspólną informację i na tej podstawie umożliwia planowanie długoterminowych inwestycji. Ponadto, stanowi podstawę do prognozowania wartości stóp procentowych.

Literatura

1. Chatfield, Ch., Collins, J. (1983), Introduction to Multivariate Analysis, Chapman and Hall Ltd, London.
2. Jolliffe, I. T. (2002), Principal Component Analysis, Springer.
3. Hastie, T., Tibshirani, R., Friedman, J. (2001), The Elements of Statistical Learning. Data Mining, Inference and Prediction, Springer.
4. Krawiec, K., Stefanowski, J. (2003), Uczenie maszynowe i sieci neuronowe, Wydawnictwo Politechniki Poznańskiej.

5. Kunysz, K. (1990), Elementy teorii informacji, Wydawnictwo Politechniki Rzeszowskiej.
6. Zeliaś, A., Pawełek, B., Wanat, S. (2000), Metody statystyczne, PWE, Warszawa.