Zastosowanie metod data mining w strategii arbitrażu statystycznego na rynku kapitałowym

Przemysław Jaśko

Słowa kluczowe: arbitraż statystyczny, pairs trading, ilościowe strategie inwestycyjne, test arbitrażu statystycznego JTTW, losowe układy dynamiczne, losowe rozmaitości niezmiennicze, losowe foliacje, mutliplikatywne twierdzenie ergodyczne, przestrzenie Oseledeca, spektrum Lapunowa, norma adaptowana do losowego układu dynamicznego, luka spektralna, skala przestrzeni Banacha, twierdzenie Hartmana-Grobmana, postać normalna losowego układu dynamicznego, prawie-tożsamościowe przekształcenie współrzędnych, równania kohomologiczne, kointegracja, test kointegracji liniowej Breitunga, test kointegracji w informacji Aparicio i Escribano, test zliczający ekstrema (RCC) Escribano dla nieliniowej kointegracji, rangowy test Breitunga dla kointegracji monotonicznej, VECM-MGARCH, model rzadkiej kointegracji, CECM, losowe układy dynamiczne typu wolny-szybki, kointegracja PCA, kointegracja ICA, uczenie rozmaitości, mapy dyfuzji, bayesowski model MS-VECM, bayesowski model TVP-VECM-SV z klasycznymi rozkładami *a priori*, bayesowski model TVP-VECM-SV ze skurczającymi rozkładami *a priori* i przerzedzaniem *ex post* metodą SAVS, metacechy szeregów czasowych, data mining, analiza szeregów czasowych, ekonometria finansowa, statystyka obliczeniowa, finanse obliczeniowe, WIG20, GPW, R, Python, C++, REDUCE, system algebry komputerowej

Streszczenie

Arbitraż statystyczny jest ilościową strategią inwestycyjną, rynkowo neutralną, typu długo-krótko, dla której przedstawiamy matematyczną definicję oraz procedurę weryfikacji czy ma się z nią do czynienia, za pomocą formalnego testu arbitrażu statystycznego JTTW (Jarrow, Teo, Tse, Warachka), opartego o proces stochastyczny wartości strategii inwestycyjnej.

Pierwszym celem pracy jest matematyczne określenie struktury losowych układów dynamicznych i ich generatorów (w postaci stochastycznych równań różnicowych), reprezentujących zmiany (logarytmów) cen akcji, które to struktury umożliwiałyby budowę strategii arbitrażu statystycznego, bazującej na modelu dynamiki cen powiązanych aktywów.

Drugim, empirycznym celem pracy jest znalezienie wielowymiarowych procesów stochastycznych (logarytmów) cen akcji, które utworzą losowy układ dynamiczny, którego własności pozwolą na budowę strategii arbitrażu statystycznego.

Pierwszy cel pracy realizujemy wykorzystując teorię losowych układów dynamicznych, obejmującą takie obiekty matematyczne jak losowe rozmaitości niezmiennicze (na których ruchy cen charakteryzują się określoną dynamiką, tzn. wykazują określoną zbieżność wprzód bądź wstecz w czasie, do losowych punktów z rozkładu implikowanego przez miarę niezmienniczą) oraz losowe foliacje, a także normalną postać losowego układu dynamicznego (która jest najprostszą możliwą postacią stochastycznych równań różnicowych reprezentujących dynamikę cen i może być wyznaczona z wykorzystaniem losowego, nieliniowego przekształcenia układu współrzędnych, które w ogólności zmienia się w czasie). W oparciu o postać normalną losowego układu dynamicznego możemy podać wzory dla losowych rozmaitości niezmienniczych i losowych foliacji względem nich oraz wzory dotyczące dynamiki układu w ramach wspomnianych obiektów, które można wykorzystać w budowie portfeli w ramach strategii arbitrażu statystycznego.

Jako generatory losowych układów dynamicznych, umożliwiających konstrukcję strategii arbitrażu statystycznego, rozpatrujemy statystyczne modele kointegracji z parametrami stałymi w czasie, takie jak VECM-MGARCH, model rzadkiej kointegracji CECM (warunkowy model ECM) oraz losowe równania różnicowe dla zmiennych wolnych i szybkich. Modele te reprezentują dynamikę (logarytmów) cen akcji.

Rozważamy również metody wydobywania przestrzeni kointegracyjnej takie jak PCA i ICA oraz metodę uczenia rozmaitości, nazywaną mapą dyfuzji, pozwalającą na wydobycie wolnych i szybkich rozmaitości dla losowych układów dynamicznych.

Jako narzędzia umożliwiające wykrywanie powiązanych (za pomocą stałej w czasie kointegracji) procesów logarytmów cen, przedstawiamy następujące testy statystyczne: test kointegracji liniowej Johansena i Breitunga oraz test kointegracji w informacji i test RCC zliczający ekstrema dla (ogólnego typu) nieliniowej kointegracji, a także rangowy test Breitunga dla kointegracji monotonicznej.

Biorąc pod uwagę, że parametry modeli mogą ulegać zmianom w czasie, rozważamy również bayesowskie modele z parametrami zmieniającymi się w czasie, takie jak MS-VECM (model VECM z przełączeniami typu Markowa), TVP-VECM-SV (model VECM z parametrami zmieniającymi się w czasie i stochastyczną zmiennością) z klasycznymi rozkładami *a priori* oraz TVP-VECM-SV ze skurczającymi rozkładami *a priori* i przerzedzaniem *ex post* metodą SAVS (ang. *Signal Adaptive Variable Selector*). Ten ostatni model umożliwia jednoczesne stwierdzenie, które procesy są skointegrowane oraz określenie czy kointegracja ta jest stała, czy zmieniająca się w czasie. Przedstawiamy również procedury MCMC dla bayesowskiej estymacji wspomnianych modeli.

W części empirycznej pracy wykorzystujemy metodologię badawczą rozwiniętą na gruncie data mining. Oznacza to, że najpierw wykonujemy analizę eksploracyjną, której wyniki wykorzystywane są do postawienia precyzyjnych hipotez statystycznych, które są następnie na etapie konfirmacyjnym testowane w ramach procedury specyfikacji i weryfikacji modeli statystycznych.

Zbiór dla analiz empirycznych obejmuje 21 szeregów (logarytmów) cen na zamknięcie sesji indeksu WIG20 oraz 20 jego składowych. Długość szeregów czasowych to $T = 643$.

Na etapie eksploracyjnym wyznaczamy 184 metacechy dla jednowymiarowych szeregów czasowych logarytmów cen i logarytmicznych stóp zwrotu, pozwalające na ocenę ich struktury. W ramach tego etapu przeprowadzamy również testy kointegracji (test liniowej kointegracji Breitunga oraz dla kointegracji nieliniowej test RCC zliczający ekstrema Escribano oraz rangowy test Breitunga) w oparciu o 21 szeregów (logarytmów) cen, w celu znalezienia potencjalnie powiązanych procesów cen akcji, w ramach dziedziny obejmującej WIG20 oraz 20 jego składowych, których dynamikę można by wykorzystać do budowy strategii arbitrażu statystycznego.

Testy kointegracji, które wykorzystujemy sugerują, że powiązane mogą być procesy logarytmów cen aktywów takich jak m.in. ALIOR-SANPL, CCC-JSW oraz DINOPL-PGE-PZU. Dla (logarytmów) cen przywołanych trzech podzbiorów akcji budujemy oddzielne modele TVP-VECM-SV ze skurczającymi rozkładami *a priori* (w dwóch wariantach: rozkłady *a priori* normalne-gamma oraz rozkłady *a priori* regresji grzbietowej) oraz przerzedzaniem *ex post* metodą SAVS macierzy kointegracji. Taka struktura modelu pozwala nam na jednoczesne stwierdzenie czy zachodzi kointegracja oraz określenie (w przypadku jej istnienia) czy zmienia się ona w czasie (z możliwymi podprzedziałami czasu, w których kointegracja zanika, co można stwierdzić z wykorzystaniem procedury przerzedzania *ex post* macierzy kointegracji), czy też jest stała w czasie.

Wnioski ze zbudowanych modeli TVP-VECM-SV są takie, że dla par ALIOR-SANPL, CCC-JSW, nie występuje kointegracja w okresie czasu objętym analizą. Uniemożliwia to więc dla wspomnianych par budowę strategii arbitrażu statystycznego opartej o modele dynamiki dla logarytmów ich cen. Dla trójki DINOPL-PGE-PZU kointegracja zmieniająca się w czasie, występowała w krótkich podprzedziałach czasu: 4% przedziału czasu objętego analizą dla modelu z rozkładami *a priori* normalnymi-gamma oraz 18% przedziału czasu objętego analizą dla modelu z rozkładami *a priori* regresji grzbietowej. Znacznie ogranicza to możliwość budowy strategii arbitrażu statystycznego dla wspomnianej trójki akcji.

W części empirycznej dodatkowo implementujemy (z wykorzystaniem systemu algebry komputerowej) procedurę określenia postaci normalnej losowego układu dynamicznego generowanego przez układ stochastycznych równań różniczkowych, reprezentujący możliwą dynamikę (logarytmów) cen akcji. Postać normalna umożliwia nam wyprowadzenie wzorów dla klasycznych losowych rozmaitości niezmienniczych oraz dla dynamiki wykazywanej na nich przez rozważany układ, które mogą posłużyć do budowy strategii arbitrażu statystycznego.