An Application of Data Mining Methods in a Statistical Arbitrage Strategy on the Stock Market

Przemysław Jaśko

Keywords: statistical arbitrage, pairs trading, quantitative trading strategies, JTTW statistical arbitrage test, random dynamical systems, random invariant manifolds, random foliations, multiplicative ergodic theorem, Oseledec's spaces, Lyapunov spectrum, norm adapted to random dynamical system, spectral gap, scale of Banach spaces, Hartman-Grobman theorem, normal form of random dynamical system, near-identity transform of coordinates, cohomological equations, cointegration, Breitung linear cointegration test, Aparicio-Escribano cointegration-in-information test, Escribano RCC nonlinear cointegration tests, Breitung rank test for monotonic cointegration, VECM-MGARCH, sparse cointegration model, CECM, random dynamical system with slow and fast variables, PCA cointegration, ICA cointegration, manifold learning, diffusion map, Bayesian MS-VECM, Bayesian TVP-VECM-SV with classical priors, Bayesian TVP-VECM-SV with shrinkage priors and SAVS postsparsification, time series metafeatures, data mining, time series analysis, financial econometrics, computational statistics, computational finance, WIG20, Warsaw Stock Exchange, R, Python, C++, REDUCE, Computer Algebra System

Abstract

A statistical arbitrage is a long-short, market neutral, quantitative trading strategy. We present mathematical definition of a statistical arbitrage and the verification procedure if considered trading strategy is of this kind, employing formal JTTW (Jarrow, Teo, Tse, Warachka) test of statistical arbitrage, based on stochastic process of strategy value.

The first aim of the work is to mathematically establish structures of random dynamical systems and their generators (in a form of stochastic difference equations) representing movement of (log) prices of assets, which enable us to pursue statistical arbitrage strategy based on modeled dynamics of prices of the related stocks.

The second aim is to empirically find multivariate stochastic processes of related stock (log) prices, that will form a random dynamical system, whose properties will allow us to pursue a statistical arbitrage strategy based on it.

The first aim is realized on a ground of random dynamical models theory, which involves such mathematical objects as random invariant manifolds (on which price movement is characterized by a particular dynamics, i.e. it have particular convergence, forward or backward in time, to random points from the distribution implied by an invariant measure), random foliations and normal form of random dynamical systems (which provides the easiest possible form of stochastic difference equations representing price dynamics, and which can be derived by the random, nonlinear transformation of coordinate system which generally can change in time). From the normal form of a random dynamical system, we can extract the formulas for random invariant manifolds and random foliations, which can be used in forming of a statistical arbitrage strategy portfolios.

As generators of random dynamical systems useful for establishing a statistical arbitrage strategy, we formally consider frequentist statistical models of cointegration with parameters constant in time, such as VECM-MGARCH, sparse cointegration model, namely CECM (Conditional ECM), and random difference equations for slow and fast variables. These models represent dynamics of stock (log) prices.

We also consider methods of extracting cointegration space such as PCA, ICA, and manifold learning method called diffusion map, for extracting slow and fast manifolds for random dynamical system. As tools to find related processes (in the form of time constant cointegration) of asset (log) prices we present following statistical tests: Johansen's and Breitung's tests for linear cointegration; cointegration-in-information and Record Counting Cointegration (RCC) tests for (general type) nonlinear cointegration,

and Breitung's rank test for monotonic cointegration.

Noting that parameters of the models could change in time, we also consider time varying cointegration Bayesian models such as MS-VECM (Markov Switching VECM), TVP-VECM-SV (Time Varying Parameter VECM-SV) with classical priors, and TVP-VECM-SV with shrinkage priors and SAVS (Signal Adaptive Variable Selector) postsparsification (which enables to simultaneously establish which stock prices processes are cointegrated and is this cointegration constant or time varying). We also present MCMC procedures for Bayesian estimation of considered models.

We achieve second, empirical aim of the work by using research methodology developed in the data mining field. It means that first we use exploratory analysis, and its results are used to state precise statistical hypotheses, which in the subsequent confirmatory analysis phase are statistically tested, within a procedure of specification and verification of statistical models. The dataset for the empirical research encompasses 21 time series of (logarithms of) closing prices of WIG20 and its 20 constituent stocks, of length $T = 643$.

At the level of exploratory data analysis we calculate 184 metafeatures for univariate time series of logarithmic prices and logarithmic returns, these metafeatures give information about the structure of time series. At this phase, we also conduct cointegration tests (Breitung's linear cointegration test and for a nonlinear cointegartion, Escribano's RCC and Breitung's rank tests) for 21 time series of (log) prices, to find possibly related processes of stock prices in a domain of WIG20 index and its 20 constituents, whose dynamics can be used to form a statistical arbitrage trading strategy.

Cointegration tests we use point out that the (log) prices process of the following assets could be related: ALIOR-SANPL, CCC-JSW, and DINOPL-PGE-PZU (among others). For (log) prices of the three stated subsets of assets, we build separate time varying parameter Bayesian dynamical models, namely TVP-VECM-SV with shrinkage priors (in two variants: normal gamma priors and ridge regression priors) and SAVS postsparsfication of cointegration matrix. Such structure of a model enables us to simultaneously test if cointegration is present, and when this is true, is it time varying (with possible time subperiods in which cointegration disappears, which can be easily established using postsparsification procedure for cointegration matrix) or time constant.

Conclusions from TVP-VECM-SV models we construct, is that for the pairs ALIOR-SANPL, CCC-JSW there is no cointegration during all the analyzed time period. This situation excludes for these two pairs, construction of statistical arbitrage strategy, based on their models of (log) price dynamics. For the triplet DINOPL-PGE-PZU time varying cointegration was present for short subperiods of time: 4% of analyzed time period, for model with normal-gamma priors, and 18% of considered time period, for model with ridge regression priors. So it restricts an ability to establish the statistical arbitrage strategy for the triplet.

In the empirical part, we additionally employ (using Computer Algebra System) a procedure for the derivation of random normal form of stochastic differential equations representing possible nonlinear dynamics of (log) prices. Normal form enables us to extract formulas for classical random invariant manifolds and dynamics on them, which can be used to establish a statistical arbitrage strategy.