the journal of
financial data science

VOLUME 1, NUMBER 2 SPRING 2019

After the well-received launch of the inaugural issue, we have received a good number of excellent manuscripts for publication consideration. In this second issue of JFDS, eight of the top submissions are included. We believe these articles and those we plan for the future will continue expanding the frontiers of data-driven investing.

In the lead article, Sanjiv R. Das, Seoyoung Kim, and Bhushan Kothari demonstrate how natural language processing—a fast-growing field within financial data science concerned with the interactions between human language and computers—can be applied to develop an early warning system for identifying corporate failure. Potential indicators within employees' sent email content and sender/recipient networks are explored to determine whether they can effectively predict changes in a firm's risk and subsequent financial performance. Sentiment-based indicators in the message bodies and nontextual structural characteristics (e.g., the number of emails sent, the average email length, and the shifting sender/recipient networks within the company over time) are explored and found to be useful indicators. Based on their analysis, a RegTech expert system solution to systematically and effectively detect escalating risk or potential malaise is proposed by the authors. This solution does not require the manual reading of individual employee emails and suggests that RegTech is a promising avenue for computational linguistics.

Koushik Balasubramanian, Harish Sundaresh, Diqing Wu, and Kevin Kearns provide a complex network approach to analyze the dynamics of returns in the US stock market. They begin with a description of a machine learning technique to learn a latent space representation of the financial instruments using a complex network approach to analyze the return dynamics of the stock market. The authors then demonstrate that the clustering of co-moving assets and breaking of already existing clusters have interesting connections to regime changes in the market. Some preliminary ideas on incorporating the impact of news and announcements in the dynamics of the assets are provided.

In today's stock market, a high percentage of trade orders are driven by trading systems using computational finance to exploit the inefficiencies or inaccuracies of the market. These systems analyze historical data and detect and seek to capitalize on time-and-asset localized opportunities for managing risk in uncertain scenarios by dynamically adapting portfolio weights. A novel approach for generating virtual scenarios of multivariate financial data of arbitrary length and composition of assets is proposed in the article by Javier Franco-Pedroso, Joaquin Gonzalez-Rodriguez, Jorge Cubero, Maria Planas, Rafael Cobo, and Fernando Pablos. Their proposed approach generates artificial asset returns that behave much like those observed for historical returns. Virtual scenarios involving decades of trading days for hundreds of assets within a given market can be simulated at low computational cost. These virtual scenarios produce diverse scenarios that can be used to test and improve quantitative investment strategies. Using an extensive set of measurements, the authors validate their simulations and find a significant degree of agreement with the reference performance of real financial series which they find is superior than that obtained with other classical and state-of-the-art approaches.

Portfolio selection based on pattern matching has shown great potential. Pattern matching involves selecting trading periods in history that are like the present trading period. The similarity between two trading periods is measured by the distance between their status vectors. Applying capital growth theory, Yang Wang and Dong Wang demonstrate that a pattern-matching approach can be derived from a symmetric market perspective, where the relationship between market status and optimal portfolio is quantitatively defined in terms of Pearson correlation. The authors' perspective, motivated by a revised pattern-matching algorithm (symmetric CORN-K), involves selecting the portfolio that simultaneously maximizes the returns of similar periods and minimizes the returns of dissimilar periods. The algorithm was further extended by the authors to a general symmetry-based pattern-matching algorithm (functional CORN-K) that uses the symmetry property in a principled way. The authors' experiments demonstrate that their new algorithms have the potential to generate superior returns, larger Sharpe ratios, and lower maximum drawdown.

Regime-switching models are popular in academia but have generally shown themselves to be dubious investment tools, a drawback compounded by their often-convoluted formal structure. As an alternative to traditional regime-switching models, Joseph Simonian and Chenwei Wu present a new mathematical basis for quant-macro investing based on spectral clustering, a graph theoretic approach to classifying data. Inspired by the ideas of Hyman Minsky and John Geanakoplos, the authors present a macro framework driven by the interaction of growth, inflation, and leverage signals, and show how it can be used to produce tradable information. In doing so, the authors show that spectral clustering can provide both an elegant formal characterization of the leverage cycle and a sound, reliable framework for quant-macro investing.

Current methodologies for constructing credit default swap (CDS) proxy rates attempt to directly construct such rates based on region, sector, and rating data, separately for each maturity. By treating all counterparties in a given region, sector, and rating bucket in the same way, these methodologies have several limitations because they: (1) ignore the counterparty-specific component of default risk, (2) fail one of the regulatory criteria for sound CDS proxy-rate construction, (3) introduce model arbitrage in the form of arbitrage opportunities between constructed CDS rates of different maturities, and (4) are highly dependent on the quality of credit ratings (which are sometimes observed to be inconsistent with the default risks implied by the CDS rates). In the first of two articles dealing with CDS, Raymond Brummelhuis and Zhongmin Luo propose a methodology that instead of constructing proxy rates, constructs proxy names in the sense that one maps a given non-observable to an observable that, based on sets of financial market data that have been shown to possess predictive power for counterparts default, best resembles the non-observable. The liquidly quoted CDS rates of this proxy name can then be used as proxy rates for the non-observable. By definition, these proxy rates are market rates and consequently there is no risk of introducing artificial arbitrage opportunities. Moreover, because these names can be traded, they can be employed for hedging or speculation. Because the construction of such proxies involves solving a classification problem, the authors use and test a variety of machine learning techniques, going beyond the regression approach in finance. Testing 126 classifiers coming from the eight most popular algorithms, the authors investigate performance variations among and within the classifiers and rank-order their performance. In the first of the two articles, after presenting their methodology, the authors report on cross-classifier performance. In their second article, they focus on parametrization and intra-classifier performance, investigate correlation effects, and perform a benchmarking exercise. The two articles are the first systematic study of CDS proxy construction by machine learning techniques and a first classifier comparison study entirely based on financial market data. The techniques should be of interest for financial institutions seeking proxies for CDS rates or other financial variables.

The limitations of the capital asset pricing model (CAPM) have been well documented since its introduction in the 1960s. There are theoretical challenges and empirical findings that are inconsistent with what the CAPM would suggest. With respect to the contrary empirical findings reported by numerous studies, Kekoura Sakouvogui and William Nganje investigate how estimation errors of the CAPM can be minimized using the cross-validation technique. This technique is widely applied in machine learning (ML-CAPM). Using data from the S&P 500 and the Dow Jones Industrial Average, Sakouvogui and Nganje apply their approach to test the assumption that the CAPM is a well-diversified portfolio model. The results from the ML-CAPM validate both market indexes as well diversified, with statistically insignificant variation in unsystematic risks during and after the 2007 financial crisis. Furthermore, the ML-CAPM provides smaller root mean square error and mean absolute deviations compared to the traditional CAPM.

Frank J. Fabozzi,
Marcos López de Prado,
Joseph Simonian
Editors