the journal of
financial data science
VOLUME 1, NUMBER 3 SUMMER 2019

In this third issue of the journal, there are eight articles that cover various applications of machine learning techniques to asset management.

A good number of studies have analyzed industry return predictability. These studies typically rely on popular predictor variables from the literature on aggregate market return predictability, such as the aggregate dividend yield, nominal yields, and yield spreads. In "Industry Return Predictability: *A Machine Learning Approach*," David E. Rapach, Jack K. Strauss, Jun Tu, and Guofu Zho use least absolute shrinkage and selection operator (LASSO), a powerful and popular technique in machine learning, to analyze return predictability for 30 US industries based on the information in lagged industry returns. After controlling for post-selection inference and multiple testing, they report significant in-sample evidence of industry return predictability. Lagged returns for the financial sector and commodity- and material-producing industries exhibit widespread predictive ability. This finding is consistent with the gradual diffusion of information across economically linked industries. Out-of-sample industry return forecasts that incorporate the information in lagged industry returns are economically valuable: controlling for systematic risk using leading multi-factor models, an industry-rotation portfolio that goes long (short), industries with the highest (lowest) forecasted returns delivering an annualized alpha of over 8%. The industry-rotation portfolio also generates substantial gains during economic downturns, including the Great Recession. Their findings offer machine-learning-based evidence against the weak form of market efficiency.

The article "Pest Control: *Eliminating Nuisance Allocations through Empirical Asset Class Identification*" demonstrates how two unsupervised machine learning techniques—exploratory factor analysis and hierarchical cluster analysis—help asset managers to identify empirical asset classes. The authors—Chao Ma, Brian Jacobsen, and Wai Lee—then apply a regression tree, which is a supervised machine learning technique to identify the most important feature of the assets within a particular asset class when allocating within that asset class. Machine learning techniques also help asset managers to assess whether a finer partitioning of the investment universe is worth the effort by quantifying the reduction in variance or increase in information from finer partitions. The authors report that the resulting asset classes satisfy desirable properties of practical asset classes: they are consistent over time, economically sensible, and statistically significant.

In their article "Machine Learning for Recession Prediction and Dynamic Asset Allocation," Alexander James, Yaser S. Abu-Mostafa, and Xiao Qiao employ Support Vector Machines (SVM), a machine learning algorithm, to determine the beginning and end of recessions in real time. More specifically, they examine two questions. The first is whether it is possible using SVM to identify business cycle turning points in a timelier manner. The second question is whether recession prediction could help investors navigate the changing macroeconomic environment. That is, can an investor with knowledge of how stocks and bonds perform in recessions and expansions use this information to dynamically adjust a portfolio to reflect the prevailing macroeconomic conditions? The authors show that SVM has excellent predictive performance for recessions, capturing all six occurrences from 1973 to 2018 while providing the signal with minimal lag. Using the timeliness of SVM signals to test a dynamic asset allocation strategy between stocks and bonds, the authors find that superior returns are generated relative to an equal-risk contribution portfolio without increasing tail risk.

In their article "When More or Less Is Less: *Managers' Clichés*," Julia Klevak, Joshua Livnat, and Kate Suslava document the most commonly used clichés in earnings conference calls and construct a dictionary of these expressions. They use natural language processing (NLP) software to detect clichés and use the findings to signal potential negative tone. Controlling for other characteristics and signals conveyed in earnings calls that are most correlated with excessive use of clichés, they find using NLP that managers' excessive use of clichés is negatively correlated with excess stock returns. The excessive use of clichés, the authors find, is associated with negative earnings growth and prior stock returns. Managers appear to use clichés to "soften the blow" associated with bad news. The evidence suggests that (1) market returns around earnings calls are incrementally and significantly more negative after controlling for the surprise in earnings and the tone of the conference call itself and (2) there are economic and statistically significant returns in a hedge portfolio holding long positions in all the companies that used no clichés in their earnings conference calls and short positions in companies that used at least four clichés.

Simulation methods are often employed in various ways in asset management, such as asset and derivative pricing. "The ETS Challenges: *A Machine Learning Approach to the Evaluation of Simulated Financial Time Series for Improving Generation Processes*" provides an evaluation framework for quantifying the degree of realism of simulated financial time series. The objective of the author team of Javier Franco-Pedroso, Joaquin Gonzalez-Rodriguez, Maria Planas, Jorge Cubero, Rafael Cobo, and Fernando Pablos is to discover and improve unknown characteristics that are not being properly reproduced by simulation methods. Since this can be dealt with as a binary classification problem for distinguishing between two classes (i.e., real and simulated financial time series), the authors employ machine learning techniques to automatically extract features and train classifiers on large datasets, with the aim of properly modeling the differences between classes based on these features. Their findings reveal some interesting properties of financial data and resulted in substantial improvements in the simulation methods in their study.

In transaction cost analysis, asset managers measure implementation shortfall, defined as the difference in performance between a paper portfolio and a real portfolio. This measure is then decomposed as a sum of execution cost and opportunity cost. In "Computation of Implementation Shortfall for Algorithmic Trading by Sequence Alignment," Raymond Chan, Kelvin Kan, and Alfred Ma explain why the original framework for measuring implementation shortfall is not directly applicable to algorithmic trading and propose an effective and objective two-stage framework for computing the decomposition of implementation shortfall. Inspired by DNA sequence alignment techniques, in the first stage the trade records from a paper portfolio and a real portfolio are aligned based on sequence alignment techniques. In the second stage, the implementation shortfall is decomposed as delay cost, market impact, over-trade cost, and under-trade cost. The proposed two-stage framework is simple, objective, and computationally efficient—the complexity only grows linearly with respect to the number of trades for the paper and real portfolios. Hence the proposed framework is applicable to high-frequency trading data.

Using an extended linear clones method and a new sequential oscillating selection method, David Byrd, Sourabh Bajaj, and Tucker Hybinette Balch investigate the problem of inferring the number and identity of constituents of an unknown portfolio given a time series of the portfolio's aggregate value. The approach they suggest in their article, "Fund Asset Inference Using Machine Learning Methods: *What's in That Portfolio?*" has several applications in asset management. One application is to detect "window dressing," in which a fund manager might rebalance a portfolio prior to a reporting deadline to show that it holds reputable stocks when in reality it had been holding risky stocks. Another application of the approach is identifying whether an asset manager has taken market positions in advance of a significant fund disclosure date for the purpose of earning a profit from those pursuing a replication strategy after disclosure. The authors test their algorithm to infer the constituents of exchange-traded funds.

In "From Risk Factors to Networks: *A Case Study on Interconnectedness Using Currency Funds*," Gueorgui Konstantinov and Jonas Rebmann introduce an approach using methods from network science and risk factor analysis for investigating currency funds. They document a positive relationship between currency funds style exposure, fund age, size, and connectedness, providing both economically and statistically significant results. They find that the most important funds in the network can influence the currency market with significant exposure to the risks factors carry, value, and trend. Their approach helps investors identify market interconnectedness, explains how risk can be transmitted, and highlights the factors that could represent significant idiosyncratic, systematic, and systemic economic risk. The authors argue that the interconnectedness of currency funds is both return- and factor-based. The approach proposed by Konstantinov and Rebmann in this article provides asset managers with a new framework for investigating global currency portfolio risk. That is, a network's future fund factor exposure can be identified and evaluated by an asset manager.

Francesco A. Fabozzi
Managing Editor