the journal of financial data science
VOLUME 1, NUMBER 1 WINTER 2019

Until recently, the depth and breadth of datasets available to financial researchers was, to put it mildly, extremely shallow. Some exchanges did not record volume information until the early 2000s. The wide adoption of time stamping with millisecond resolution took even longer. Outside exchange trade records and infrequent government statistics, alternative data sources were rare. The implication is that financial researchers conducted the large majority of their analyses on daily price series. This state of data paucity set a hard limit on the sophistication of the techniques that financial researchers could use. In that financial paleo-data age, the linear regression method was a reasonable choice, even though most of us suspected that the linearity assumption may not provide a realistic representation of a system as complex and dynamic as modern financial markets.

Today, we live in a different era, the age of financial Big Data. Researchers have at their disposal datasets that only a few years ago were unimaginable: Satellite images, credit card transactions, sensor data, web scrapes, sentiment from news and tweets, recordings from speeches, geolocation of cargos crossing the oceans, web searches, supply-chain statistics, and the like. The size, quality, and variety of these sources of information, combined with the power of modern computers, allow us to apply more sophisticated mathematical techniques.

However, the adoption of these new techniques is not straightforward. It requires researchers to abandon the comfort of closed-form solutions and embrace the flexibility of numerical and nonparametric methods. The goal of this journal is to facilitate this transition among academics and practitioners. We, the editors, felt that the established journals were not ready to serve this goal for multiple reasons. Our readers will find in this journal high-quality academic articles that are applicable to the practical problems faced by asset managers. These articles present fresh ideas that challenge the traditional way of thinking about finance, the economy, and investing. Through case studies, we offer a front-row view of the cutting-edge of empirical research in financial economics.

In the first article, two of the co-editors, Joseph Simonian and Frank J. Fabozzi, position financial data science within the broader history of econometrics. They explain why its ascendance marks a reorientation of the field toward a more empirical and pragmatic stance, and that due to the unique nature of financial information, financial data science should be considered a field in its own right and not just an application of data science methods to finance.

Ashby Monk, Marcel Prins, and Dane Rook explain how in finance as alternative data become mainstream, institutional investors

may benefit from rethinking how they engage with alternative datasets. By rethinking their approaches to alternative data as the authors suggest, institutional investors can select alternative datasets that better align with their organizational resources and contexts.

As a remedy to the shortcomings of traditional factor models, Joseph Simonian, Chenwei Wu, Daniel Itano, and Vyshaal Narayanam describe a machine learning approach to factor modeling based on the random forests algorithm. As a case study, the authors apply random forests to the well-known Fama-French-Carhart factors and analyze the major equity sectors, showing that compared to a traditional regression-based factor analysis, the random forests algorithm provides significantly higher explanatory power, as well as the ability to account for factors' nonlinear behavior and interaction effects. In addition to providing evidence that the random forests framework can enhance ex post risk analysis, the authors also demonstrate that combining the random forest algorithm with another machine learning framework, association rule learning, can also help produce useful ex ante trading signals.

It is well-known that the classic mean-variance portfolio framework generates weights for the optimized portfolios that are directly proportional to the inverse of the asset correlation matrix. However, most of contemporary portfolio optimization research focuses on optimizing the correlation matrix itself, and not its inverse. Irene Aldridge demonstrates that this is a mistake, specifically from a Big Data perspective. She demonstrates that the inverse of the correlation matrix is much more unstable and sensitive to random perturbations than the correlation matrix itself. The results she reports are novel in the Data Science space, extending far beyond financial data, and are applicable to any data correlation matrices and their inverses.

Although machine learning offers a set of powerful tools for asset managers, one crucial limitation involves data availability. Because machine learning applications typically require far more data than are available, especially for longer-horizon investing, it is important for asset managers to select the right application before applying the tools. Rob Arnott, Campbell Harvey, and Harry Markowitz provide a research checklist that can be used by asset managers and quantitative analysts to select the appropriate machine learning applications as well as, more generally, providing a framework for best practices in quantitative investment research.

Applying a machine learning technique that is new to finance called independent Bayesian classifier combination, David Bew, Campbell Harvey, Anthony Ledford, Sam Radnor, and Andrew Sinclair test whether valuable information can be extracted from analysts' recommendations of stock performance. The technique provides a way to weight analysts forecasts based on their performance in rating a particular stock as well as their performance rating other stocks. Their results show that a combination of their machine learning recommendations along with the analysts' ratings leads to excess returns in their sample suggesting this new technique could be useful for active investors.

Thousands of journal articles have claimed to have discovered a wide range of risk premia. Most of these discoveries are false, as a result of selection bias under multiple testing. Using a combination of extreme value theory and unsupervised learning, Marcos López de Prado proposes a practical method to discount the inflationary effect that selection bias has on a particular discovery.

Ananth Madhavan and Aleksander Sobczyk employ data science to create an investible, dynamic portfolio to mimic the factor characteristics of private equity. Using textual analysis, they first identify firms taken private and then use a multifactor model to measure the cross-sectional factor exposures of firms immediately prior to the announcement that they were being acquired by a private equity firm. Then the authors use holdings-based optimization to build a liquid, investible, long-only portfolio that dynamically mimics the factor characteristics of the portfolio of stocks that were taken private.

Julia Klevak, Joshua Livnat, and Kate Suslava illustrate how the utilization of text mining and scoring of an unstructured data can add information to investors beyond structured data. They demonstrate how the application to the analysis of earnings conference call transcripts produces a signal that is incrementally additive to earnings surprises and the short-term returns around the earnings announcement.

In their article, Sidney C. Porter and Sheridan Porter contribute two new fundamental properties of

indexes—similarity and stability—to indexing theory, made practical by advances in data science technology. In the application of the theory, they introduce a framework for a repeatable decomposition of private equity returns that disambiguates the quantification of manager skill.

A graph-theoretic framework for monitoring system-wide risk by extending methods widely deployed in social networks is provided by Sanjiv R. Das, Seoyoung Kim, and Daniel N. Ostrov. They introduce desired properties for any systemic risk measure and provide a novel extension of the well-known Merton credit risk model to a generalized stochastic network-based framework across large financial institutions.

The problem of optimally hedging an options book in a practical setting, where trading decisions are discrete and trading costs can be nonlinear and difficult to model. Using reinforcement learning, a well-established machine learning technique, Petter Kolm and Gordon Ritter propose a flexible, accurate and very promising model for solving this problem.

Frank J. Fabozzi
Editor