the journal of financial data science

VOLUME 2, NUMBER 3 SUMMER 2020

In the lead article in this issue, Dilip B. Madan and Yazid M. Sharaiha explain the linkages that may exist between automated trading and financial markets, both in theoretical and practical contexts. As they explain in their article, "Machine Trading: *Theory, Advances, and Applications*," dynamic contributions to trading are evaluated over some investment horizon using covariations between asset position and price changes. They compare machine learning strategies based on a Gaussian process regression with a least squares regression. The authors generalize these two regression methods by invoking conservative valuation schemes, leading to the study of conservative conditional expectations modeled by distorted expectations which, in turn, lead to the development of distorted least squares and distorted Gaussian process regression as the associated estimation or prediction schemes. The authors analyze trading strategies based on these four regression methods—Gaussian process regression, distorted Gaussian process regression, least squares regression, and distorted least squares—using a common database of securities and set of factor (predictor) variables. The regression methods are used to trigger entry into trades, with exits conducted on a comparable basis for all of them. Trading strategies are executed for nine sectors of the US economy using 14 different predictive factor sets. Results reported by the authors indicate improvements are made by Gaussian process regression over the least squares regression, and distorted least squares regression, with the distorted regression also favorably affecting the drawdowns.

Many financial models and applications rely on the ability to quickly determine which hidden state a new time-series observation belongs to. In "Greedy Online Classification of Persistent Market States Using Realized Intraday Volatility Features," Peter Nystrup, Petter N. Kolm, and Erik Lindström develop a greedy online classifier for classifying time-series data without the need to parse historical data. While it is difficult to maintain persistence when classifying observations online, their classification methodology is based on clustering temporal features and explicitly penalizing jumps between states with a fixed-cost regularization term. A series of simulations show that the new classifier results in a higher accuracy and increased robustness to misspecification than the correctly specified maximum likelihood estimator.

Hyperparameter optimization has become increasingly important as the number of machine-learning approaches to portfolio selection continues to increase. In "Hyperparameter Optimization for Portfolio Selection," Peter Nystrup, Erik Lindström, and Henrik Madsen propose a systematic approach to hyperparameter optimization by leveraging recent advances in automated machine learning and multi-objective optimization. They establish a connection between forecast uncertainty and holding- and trading-cost parameters. Nystrup, Lindström, and Madsen argue that these should be considered regularization parameters that can be adjusted to achieve optimal performance. In particular, they show that when the signal-to-noise ratio of return forecasts deteriorates, the optimal level of the transaction-cost parameters increases to prevent excessive trading. Additionally, the authors show that hyperparameter optimization using multi-objective optimization produces solutions with better risk-return tradeoffs than that of manual, grid, and random search. At the same time, the solutions found were more stable across in-sample training and out-of-sample testing, indicating that improvements were not a result of overfitting.

There is an extensive literature that has investigated hedge funds and hedge fund interconnectedness on an index level. However, what ultimately determines fund interconnectedness is single hedge fund strategies and their underlying risk factors. Yet, few studies have examined this interconnectedness. In "A Network Approach to Analyzing Hedge Fund Connectivity," Gueorgui S. Konstantinov and Joseph Simonian investigate this issue. Specifically, they investigate the hedge fund market as a network of interacting individual funds. Identifying and analyzing the most important hedge fund styles that could both affect the market and transmit systemwide shocks to other funds, individual asset classes, and beyond, they find that the most connected hedge fund database categories are global macro and equity long–short funds. Applying clustering algorithms to classify and evaluate hedge fund styles as well as individual hedge funds, they present empirical evidence that hedge fund networks are found to have different dynamics across different stages of the business cycle. Classifying funds using clustering, in which seemingly different funds are shown to cluster based on their shared factor exposures, is a major finding of their study. Rather than database classifications, when measuring hedge fund risk across the business cycle, this finding of Konstantinov and Simonian suggests that investors should consider fund connectivity and their attendant importance scores. Moreover, they propose a forecasting framework that can be used to predict hedge fund network behavior and the impact of individual factors on the network.

Financial data distributions change over time due to short-lived market dynamics and exogenous drivers among others. This is often manifested through nonlinear patterns that are known to be challenging to track by conventional models. In "Derivation of a Dynamic Market Risk Signal Using Kernel PCA and Machine Learning," Alireza Yazdani applies Kernel PCA on the spot return covariance matrix of G10 currencies, alongside other known drivers of the foreign-exchange market, to derive an aggregate market signal that can be used as an input feature in currency prediction models. Yazdani demonstrates that the Kernel PCA signal enhances in-sample and out-of-sample risk-adjusted performance across a range of machine learning strategies. He attributes the performance to the market "information conveyance" and "information compression" properties of the Kernel PCA method as well as the ability to track short- and mid-term market moves. This is particularly useful as a predictor of extreme events when utilized in a machine learning framework.

While the predictability of financial market returns is studied extensively in the literature, many studies focus on individual return factors and linear models. Such studies tend to make three strong parametric assumptions that undermine predictive performance: (1) a simple functional form between returns and predictors, (2) no interactions within the features, and (3) the feature set usually includes less than a handful of predictors. In "The Cross Section of Commodity Returns: *A Nonparametric Approach*", Clemens Struck and Enoch Cheng compare the predictability of commodity returns using standard linear approaches with that of tree-based methods which overcome the three parametric assumptions listed above. Struck and Cheng find in their out-of-sample analysis that up to 3.74% of returns are predictable using tree-based methods versus 0.33% using the standard linear approaches. Portfolios formulated using tree-based methods outperform their linear counterparts on both an absolute and risk-adjusted basis.

Multiple machine learning models have been found to enhance predictive performance. Combining models is commonly referred to as "committees". Rather than use the traditional approach to portfolio selection, Tsung-wu Ho proposes a way that the committee approach can be applied in portfolio selection, the most popular optimization frameworks for portfolio selection being risk-diversification, risk parity, and risk optimal. The basic methodology Ho uses in "Portfolio Selection Using Portfolio Committees," is as follows. Since each optimal portfolio is a combination of three basic elements (strategy, covariance matrix, and risk type), the combination is augmented to 250 optimal portfolios at each estimation period. Then a score as an information criterion is defined by Ho to select the best portfolio to hold in the next period. The author finds that the committee approach to portfolio selection provides superior performance. Moreover, it easy to implement.

The hierarchical risk parity approach introduced in 2016 by Marcos López de Prado elegantly addresses the weaknesses of portfolio models and establishes this approach as the cutting-edge application of machine learning to portfolio management. In "A Modified Hierarchical Risk Parity Framework for Portfolio Management," Marat Molyboga extends the original hierarchical risk parity approach by incorporating three intuitive elements commonly used by practitioners: (1) an exponentially weighted covariance matrix with Ledoit-Wolf shrinkage instead of the sample covariance matrix, (2) an equal volatility, rather than an inverse variance, allocation approach to diversify across portfolio constituents both within and across clusters, and (3) portfolio volatility targeting to diversify across time. Examining the impact of the enhancements on portfolios of Commodity Trading Advisors within a large-scale Monte-Carlo simulation framework that accounts for the realistic constraints of institutional investors, the author reports a striking improvement in the out-of-sample Sharpe ratio of 50% (on average), as well as a reduction in downside risk. Molyboga argues that the hierarchical risk parity framework as modified in the article has broad applications for portfolios of traditional and alternative investments.

Francesco A. Fabozzi
Managing Editor