Forecasting Realized Covariance via Big Data and Machine Learning

Sophia Zhengzi Li and Yushan Tang

Abstract

This project aims to develop a new generation of machine learning-based automatic system for predicting realized covariance via big data. We consider five types of features: 1) lagged realized covariance measures based on intraday data from the NYSE TAQ database; 2) market condition variables such as market returns, VIX, and VVIX; 3) firm economic links including text-based product similarity scores and customer-supplier relationship; 4) firm information links such as common institutional ownership, common analyst coverage, and common geographic location; and 5) firm characteristics. By taking advantage of novel machine learning techniques and aforementioned big data, the system is expected to: 1) limits human intervention in feature engineering and algorithm selection, 2) incorporates many features while preventing overfitting noises; 3) utilizes more flexible and state-of-the-art learning algorithms; 4) achieves good and consistent out-of-sample performance; 5) sheds lights on feature contributions to realized covariance forecasts; and 6) facilitates portfolio management with accurate predictions.

Research Design

Forecasting asset volatility and co-movement is central to portfolio construction and risk management. Existing literature on realized covariance forecasting typically focuses on a small set of predictors based on market data within the classical OLS-based framework. In this project, our objective is to build an ambitious and powerful automatic forecasting system to predict realized covariance. Our system consists of both feature engineering and algorithm fitting. In the feature engineering step, we consider hundreds of features including lagged realized covariance measures (Bollerslev, Li, Patton, and Quaedvlieg, 2020), market condition

variables that drive up/down asset correlations, economic and information links of firms, and firm characteristics. In the algorithm fitting step, both sparse linear models (LASSO, Principal Component Regression) and nonlinear algorithms (Random Forest, Gradient Boosted Regression Tree, Neural Network) will be implemented to prevent overfitting and to capture nonlinear relation between features and the response variable. Following Gu, Kelly, and Xiu (2020) and Li and Tang (2020), we will first use OOS R2 to gauge the model performance and the Diebold-Mariano test for model comparison. Next, we aim to shed lights on model interpretation by providing detailed discussions on feature contribution based on population variable importance measures such as Shapley values suggested by Lundberg and Lee (2017) and Williamson and Feng (2020). Lastly, we will demonstrate that the improvement in OOS R2 can translate into meaningful economic gains in portfolio management.

Reference

[1] Bollerslev, T., Li, J., Patton, A. J., Quaedvlieg, R., 2020. Realized semicovariances. Econometrica 88(4), 1515-1551.

[2] Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Review of Financial Studies 33, 2223–2273.

[3] Li, S. Z., Tang, Y., 2021. Forecasting realized volatility: an automatic system using many features and many machine learning algorithms. Working paper.

[4] Lundberg, S. M., Lee, S., 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30, 4768-4777.

[5] Williamson, B. D., Feng, J., 2020. Efficient nonparametric statistical inference on population feature importance using Shapley values. Proceedings of the Thirty-seventh International Conference on Machine Learning.