Data-driven Methods for Simulation and Forecasting of Financial Time Series

Chao Zhang

The Queen's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2023

Abstract

This thesis develops data-driven methods for the simulation and forecasting of financial time series. The contributions are structured into four main components.

In the first part, we propose Tail-GAN, a novel nonparametric approach that combines a Generative Adversarial Network (GAN) with the joint elicitability property of Value-at-Risk (VaR) and Expected Shortfall (ES) for learning to simulate price scenarios that preserve tail risk features for a set of benchmark trading strategies.

In the second part, we investigate the impact of order flow imbalance (OFI) on price movements in equity markets in a multi-asset setting. Our results show that, once the information from multiple levels is integrated into the OFI, multi-asset models with cross-impact do not provide additional explanatory power for contemporaneous impact compared to a sparse model without the cross-impact terms. We show however that cross-asset OFIs do improve the forecasting of future returns.

In the third part, we apply machine learning models to forecast intraday realized volatility (RV), by exploiting commonality in intraday volatility by pooling stocks together, and by incorporating a proxy for market volatility. Neural networks dominate linear regression and tree-based models in terms of performance, and remain robust and competitive on unseen stocks not included in the training set, thus providing new empirical evidence for a universal volatility mechanism among stocks. We also propose a new approach to forecasting one-day-ahead RVs using past intraday RVs as predictors, and expose interesting time-of-day effects that aid the forecasting mechanism.

In the last part, we develop a method for forecasting the realized covariance matrix of asset returns in the U.S. equity market by exploiting the predictive information of graphs in volatility and correlation. Specifically, we augment the Heterogeneous Autoregressive (HAR) model via neighborhood aggregation on these graphs. The results generally suggest that the augmented model incorporating graph information yields both statistically and economically significant improvements for out-of-sample performance over the traditional models.

Data-driven Methods for Simulation and Forecasting of Financial Time Series

Chao Zhang

The Queen's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisors, Rama Cont and Mihai Cucuringu, for their invaluable guidance, continuous support, and generous encouragement throughout my DPhil journey. Their insightful feedback, vision, and motivation have been instrumental in sharpening my thinking and elevating my work to new heights.

I would like to extend my sincere appreciation to my viva examiners, Álvaro Cartea and Dacheng Xiu, for their immense expertise, meticulous examination, and valuable feedback. I am also deeply indebted to the faculty members, Zhongmin Qian, Hanqing Jin, Samuel Cohen, Justin Sirignano, and Blanka Horvath, for providing their valuable feedback and suggestions throughout my studies.

Then, I would like to thank my outstanding research collaborators, including Darwin Choi, Wenxi Jiang, Renyuan Xu, Yihuang Zhang, Xingyue Pu, Zihao Zhang, Xiaowen Dong, and Stefan Zohren for the stimulating discussions and valuable cooperation. I would also like to extend my sincere gratitude to all my lab mates, MCF, and OMI members for those inspiring discussions.

My heartfelt thanks to the poets Su Shi and Li Jian, whose philosophical poems have inspired and motivated me to persist in pursuing the research I am passionate about. I would also like to thank the Clarendon Fund and the Queen's College for sponsoring my studies and all other support.

Finally, I am extremely thankful to my parents for their unwavering support and love throughout my academic pursuits. Thank you all for being an integral part of this journey.

Abstract

This thesis develops data-driven methods for the simulation and forecasting of financial time series. The contributions are structured into four main components.

In the first part, we propose Tail-GAN, a novel nonparametric approach that combines a Generative Adversarial Network (GAN) with the joint elicitability property of Value-at-Risk (VaR) and Expected Shortfall (ES) for learning to simulate price scenarios that preserve tail risk features for a set of benchmark trading strategies.

In the second part, we investigate the impact of order flow imbalance (OFI) on price movements in equity markets in a multi-asset setting. Our results show that, once the information from multiple levels is integrated into the OFI, multi-asset models with cross-impact do not provide additional explanatory power for contemporaneous impact compared to a sparse model without the cross-impact terms. We show however that cross-asset OFIs do improve the forecasting of future returns.

In the third part, we apply machine learning models to forecast intraday realized volatility (RV), by exploiting commonality in intraday volatility by pooling stocks together, and by incorporating a proxy for market volatility. Neural networks dominate linear regression and tree-based models in terms of performance, and remain robust and competitive on unseen stocks not included in the training set, thus providing new empirical evidence for a universal volatility mechanism among stocks. We also propose a new approach to forecasting one-day-ahead RVs using past intraday RVs as predictors, and expose interesting time-of-day effects that aid the forecasting mechanism.

In the last part, we develop a method for forecasting the realized covariance matrix of asset returns in the U.S. equity market by exploiting the predictive information of graphs in volatility and correlation. Specifically, we augment the Heterogeneous Autoregressive (HAR) model via neighborhood aggregation on these graphs. The results generally suggest that the augmented model incorporating graph information yields both statistically and economically significant improvements for out-of-sample performance over the traditional models.

Contents

List of Figures

1 Introduction

Contents

The advent of "big data" has fundamentally reshaped the financial sector, due to the availability of high-frequency data and alternative data. Big data in finance is characterized by three key properties: large size, high dimension, and complex structure (Goldstein, Spatt, and Ye [115]). Consequently, traditional econometrics techniques have struggled in dealing with these challenges, particularly for unstructured data. On the other hand, recent years have seen significant developments in data-driven methods such as machine learning (ML), which can identify patterns and relationships between variables and provide powerful tools for making informed decisions based on data insights.

One essential component of financial data is the time series structure, which plays a crucial role in analyzing how the market and economy change over time. With the rise of high-frequency and machine-based trading, there is a growing need for advanced data-driven methods to simulate and forecast financial time series.

For example, financial institutions are required by the Basel Committee's Fundamental Review of the Trading Book (FRTB) to estimate the risk of portfolios and trading strategies. Data-driven methods, such as nonparametric scenario generation, can contribute to this task, leading to better decision-making and a lower likelihood of financial losses. Additionally, data-driven methods can offer more accurate forecasts of financial time series by analyzing vast amounts of data and identifying complex patterns.

1.1 Thesis outline

This thesis summarizes my work on the modeling and forecasting of financial time series using machine learning and graph-based methods. We focus on four problems: the design of market simulators (Chapter 2), the study of cross-impact in equity markets (Chapter 3), the forecasting of (intraday) volatility of financial assets (Chapter 4), and the analysis of spillover effects and covariance matrix forecasting (Chapter 5). Fig 1.1 illustrates the connections between the chapters and the core theme, namely, the data-driven analysis of financial time series. We now further describe these topics and our main contributions.

1.1.1 Simulation of market scenarios

The estimation of loss distributions for dynamic portfolios requires the simulation of scenarios representing realistic joint dynamics of their components, with particular importance devoted to the simulation of *tail risk* scenarios. In Chapter 2, we propose Tail-GAN, a novel data-driven approach for simulating tail risk scenarios. Our approach uses a Generative Adversarial Network (GAN) which exploits the joint elicitability property of Value-at-Risk (VaR) and Expected Shortfall (ES) for learning to simulate price scenarios that preserve tail risk features for benchmark trading strategies, including consistent statistics such as VaR and ES.

We prove a universal approximation theorem for our generator for a broad class of risk measures. In addition, we show that the training of Tail-GAN may be formulated as a max-min game, leading to a more effective approach for training.

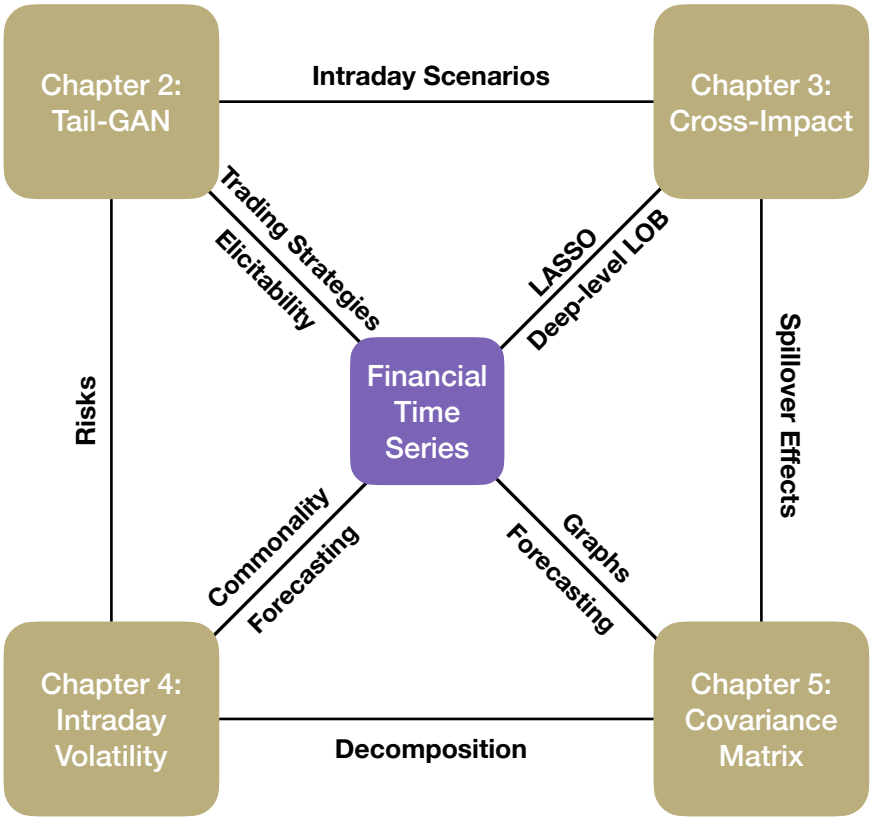


Figure 1.1: Thesis structure.

Our numerical experiments show that, in contrast to other data-driven scenario generators, our proposed scenario simulation method correctly captures tail risk for both static and dynamic portfolios.

1.1.2 Market impact and cross-impact in equity markets

Accurate estimation and forecasting of the impact of trading behavior of market participants on the price movements of assets carries practical implications for both practitioners and academics, such as trading cost analysis and optimal execution models.

Chapter 3 studies the impact of order flow imbalance (OFI) on price movements in equity markets in a multi-asset setting. First, we propose a systematic approach for combining OFIs at the top levels of the limit order book (LOB) into an integrated OFI variable which better explains price impact, compared to the

best-level OFI. Equipped with this, we show that in the contemporaneous cross-impact setting, once the information from multiple levels is integrated into the OFI, multi-asset models with cross-impact do not provide additional explanatory power for contemporaneous impact compared to a parsimonious model without the cross-impact terms. Furthermore, we demonstrate the opposite effect in the forward-looking cross-impact setting, namely, cross-asset OFIs do improve the forecast of future returns, both in terms of R^2 and economic incentives, thus providing evidence of cross-impact. We also establish that this cross-impact mainly manifests at short-term horizons and decays rapidly in time.

1.1.3 Modeling and forecasting of volatility

Forecasting and modeling stock return volatility has been of interest to both academics and practitioners over the past years. Recent advances in high-frequency trading (HFT) highlight the need for robust and accurate intraday volatility forecasts.

In Chapter 4, we apply machine learning models to model and forecast intraday realized volatility (RV), by exploiting commonality in intraday volatility by pooling data from different stocks, and by incorporating a proxy for the market volatility. Neural networks dominate linear regressions and tree models in terms of performance, due to their ability to uncover and model complex latent interactions among variables. Our findings remain robust when we apply trained models to new stocks that have not been included in the training set, thus providing new empirical evidence for a universal volatility mechanism among stocks. Finally, we propose a new approach to forecasting one-day-ahead RVs using past intraday RVs as predictors, and highlight interesting time-of-day effects that aid the forecasting mechanism. The results demonstrate that the proposed methodology yields superior out-of-sample forecasts over a strong set of traditional baselines that only rely on past daily RVs.

1.1.4 Covariance matrix forecasting and spillover effects

There is substantial empirical evidence to support the idea that the covariance of asset returns varies over time. Practitioners continually monitor and predict

covariances as they evolve, and adjust their portfolios in response, in order to maximize return while limiting risk.

In Chapter 5, we forecast the realized covariance matrix of asset returns in the U.S. equity market by exploiting the predictive information of graphs in volatility and correlation. Specifically, we augment the Heterogeneous Autoregressive (HAR) model via neighborhood aggregation on these graphs. Our proposed method allows for the modeling of interdependence in volatility (also known as spillover effects) and correlation, while maintaining parsimony and interpretability. We explore various graph construction methods, including sector membership and graphical LASSO (for modeling volatility), and line graph (for modeling correlation). The results generally suggest that the augmented model incorporating graph information yields both statistically and economically significant improvements for out-of-sample performance over the traditional models. Such improvements remain significant over horizons up to one week ahead, but decay in time. The robustness tests demonstrate that the forecast improvements are obtained consistently over the different out-of-sample sub-periods, and are insensitive to measurement errors of volatilities.

2

Tail-GAN: Learning to Simulate Tail Risk Scenarios

Contents

2.1 Introduction

Scenario simulation is extensively used in finance for evaluating the loss distribution of portfolios and trading strategies, often with a focus on the estimation of risk measures such as Value-at-Risk and Expected Shortfall (Glasserman [113]). The estimation of such risk measures for static and dynamic portfolios involves the simulation of scenarios representing realistic joint dynamics of their components. This requires both a realistic representation of the temporal dynamics of individual assets (*temporal dependence*), as well as an adequate representation of their co-movements (*cross-asset dependence*).

Risk estimation has become increasingly important in financial applications in recent years, in light of the Basel Committee's Fundamental Review of the Trading Book (FRTB), an international standard that regulates the amount of capital banks ought to hold against market risk exposures (see Bank for International Settlements [22]). FRTB particularly revisits and emphasizes the use of Value-at-Risk vs Expected Shortfall (Du and Escanciano [90]) as a measure of risk under stress, thus ensuring that banks appropriately capture tail risk events. In addition, FRTB requires banks to develop clear methodologies to specify how various extreme scenarios are simulated, and how the stress scenario risk measures are constructed using these scenarios. This suite of capital rules has taken effect January 2022 to strengthen the financial system, with an eye towards capturing tail risk events that came to light during the 2007-2008 financial crisis.

A common approach in scenario simulation is to use parametric models. The specification and estimation of such parametric models poses challenges in situations in which one is interested in heterogeneous portfolios or intraday dynamics. As a result of these issues, and along with the scalability constraints inherent in nonlinear models, many applications in finance have focused on Gaussian factor models for scenario generation, even though they fail to capture many stylized features of market data (Cont [72]).

Generative Adversarial Networks (GANs) (Goodfellow et al. [116]) have emerged in recent years as an efficient alternative to parametric models for the simulation of

patterns whose features are extracted from complex and high-dimensional data sets. A GAN is composed of a pair of neural networks: a *generator* network G, which generates random scenarios, and a *discriminator* network D which aims to identify whether the generated sample comes from the desired distribution. The generator G is then trained to output samples which closely reproduce properties of the training data set under a certain criterion. GANs have been successfully applied for the generation of images (Goodfellow et al. [116] and Radford, Metz, and Chintala [193]), audio (Donahue, McAuley, and Puckette [88] and Oord et al. [184]), and text (Fedus, Goodfellow, and Dai [99] and Zhang et al. [237]), which can be further combined with downstream tasks such as image reconstruction (Zhou et al. [238]), facial recognition (Huang et al. [141]) and anomaly detection (Cao, Guo, and Wang [52]).

GANs have been recently used in several instances for the simulation of financial market scenarios. In particular, Takahashi, Chen, and Tanaka-Ishii [211] used GAN to generate one-dimensional financial time series and observed that GANs are able to capture certain stylized facts of univariate price returns, such as heavy-tailed return distribution and volatility clustering. Wiese et al. [226] introduced the Quant-GAN architecture, where the generator utilizes the temporal convolutional network (TCN), first proposed in Oord et al. [184], to capture long-range dependencies in financial data. Marti [174] applied a convolutional-network-based GAN framework (denoted as DCGAN) to simulate empirical correlation matrices of asset returns. However, no dynamic patterns such as autocorrelation could be captured in the framework. The Conditional-GAN (CGAN) architecture, first introduced in Mirza and Osindero [177], and its variants were proposed to simulate financial data or time series in a line of works (see Fu et al. [108], Koshiyama, Firoozye, and Treleaven [155], Li et al. [166], Ni et al. [180], and Vuletić, Prenzel, and Cucuringu [219]). Compared to the classic GAN architecture, CGAN has an additional input variable for both the generator and discriminator, in order to incorporate certain structural information into the training stage, for example, the lag information inherent in the time series. Yoon, Jarrett, and Schaar [232] introduced a stepwise supervised loss to learn from the transition dynamics of the market data for producing realistic multivariate time series.

Unlike generative models for images, which can be validated by visual inspection, *model validation* for such data-driven market generators remains a challenging question. In particular, it is not clear whether the scenarios simulated by such "market generators" are sufficiently accurate to be useful for various applications in risk management. We argue in fact that the model validation criterion and the training objective should not be chosen independently, but rather target a *specific use case* for the output scenarios. In the present work, we illustrate this idea in a specific application, namely the design and performance validation of generative models that could correctly quantify the *tail risk* of a set of user-specified benchmark strategies.

2.1.1 Main contributions

We propose Tail-GAN, a novel approach for multi-asset market scenario simulation that focuses on generating tail risk scenarios for a user-specified class of trading strategies. In contrast to previous GAN-based market generators, which are trained using cross-entropy or Wasserstein loss functions, Tail-GAN starts from a set of *benchmark* trading strategies and uses a bespoke loss function to accurately capture the tail risk of these benchmark portfolios, as measured by their Value-at-Risk (VaR) and Expected Shortfall (ES). This is achieved by exploiting the joint *elicitability* property of VaR and ES (Acerbi and Szekely [3] and Fissler, Ziegel, et al. [100]).

From a theoretical perspective, we provide two contributions. First, we prove a universal approximation theorem for our generator under a broad class of risk measures: given a tail risk measure (VaR, ES, or any spectral risk measures that are Hölder continuous) and any tolerance level $\varepsilon > 0$, there exists a fully connected generator network whose outputs lead to $(1 - \varepsilon)$-accurate tail risk measures for the benchmark strategies. The second theoretical result is related to the training of the generator and the discriminator, which is traditionally formulated as a bi-level optimization problem (Goodfellow et al. [116]). We prove that, in our method, the bi-level optimization is equivalent to a max-min game, leading to a more effective and practical formulation for training.

From the perspective of applications, our extensive numerical experiments, using synthetic and market data, show that TAIL-GAN provides accurate tail risk estimates and is able to capture certain stylized features observed in financial time series, such as heavy tails, and complex temporal and cross-asset dependence patterns. Our results also show that including dynamic portfolios in the training set of benchmark portfolios is crucial for learning the temporal features of the underlying time series. Last but not least, we show that combining TAIL-GAN with Principal Component Analysis (PCA) enables the design of scenario generators that are scalable to a large number of heterogeneous assets.

2.1.2 Related literature

The idea of incorporating *quantile* properties into the simulation model has been explored in Ostrovski, Dabney, and Munos [186], which introduced an autoregressive implicit quantile network (AIQN). Their goal is to train a simulator via supervised learning so that the quantile divergence between the empirical distributions of the training data and the generated data is minimized. However, the quantile divergence adopted in AIQN is an *average performance* across all quantiles, which provides no guarantees for the tail risks. In addition, the simulator trained with supervised learning may suffer from accuracy issues and the lack of generalization power (see Section 2.5.3 for a detailed discussion).

Bhatia, Jain, and Hooi [32] employed GANs conditioned on the statistics of extreme events to generate samples using Extreme Value Theory (EVT). By contrast, our approach is fully non-parametric and does not rely on parametrization of tail probabilities.

The idea of exploiting input price scenarios (or simulated price scenarios) via non-linear functionals has also been proposed in recent studies via the notion of *signature*. Buehler et al. [45, 46] developed a generative model based on Variational Autoencoder (VAE) and signatures of price series. The concept of signature tensor has been used for time series generation using GANs (see Ni et al. [180]).

The remainder of this research is structured as follows. Section 2.2 discusses tail risk measures, the concept of elicitability, and properties of score functions for tail risk measures. Section 2.3 introduces the Tail-GAN framework, including the generator, the discriminator, and the loss function. Section 2.4 explains the methodology and criteria that we use to evaluate the performance. Section 2.5 demonstrates the numerical performance on synthetic scenarios for model validation, which also includes a numerical validation of generalization power and scalability of Tail-GAN. Section 2.6 discusses the performance of Tail-GAN on real-world intraday scenarios.

2.2 Tail risk measures and score functions

Most GAN frameworks use divergence measures such as cross-entropy (Chen et al. [62] and Goodfellow et al. [116]) or Wasserstein distance (Arjovsky, Chintala, and Bottou [16]) to measure the similarity of the generated data to the input data. However, using such divergence measures as objective functions in GAN training may result in poor performance if one is primarily interested in *tail properties* of the distribution. But tail risk measures such as quantile or Expected Shortfall, may fail to be continuous for these divergence measures: two probability distributions may have arbitrarily small cross-entropy of Wasserstein divergence yet widely different tail risk measures.

To overcome this difficulty, we consider an alternative divergence measure which quantifies closeness of the tails of two distributions. Inspired by recent work on elicitability of risk measures (Acerbi and Szekely [3] and Fissler, Ziegel, and Gneiting [101]), we design a score function related to tail risk measures used in financial risk management. We define these tail risk measures and the associated score functions in Section 2.2.1 and discuss some of their properties in Section 2.2.2.

2.2.1 Tail risk measures

Tail risk refers to the risk of large portfolio losses. *Value at Risk (VaR)* and *Expected Shortfall (ES)* are commonly used statistics for measuring the tail risk of portfolios.

The gain of a portfolio at a certain horizon may be represented as a random variable $X : \Omega \to \mathbb{R}$ on the set Ω of market scenarios. Given a probabilistic model, represented by a probability measure μ on Ω, the VaR at confidence level $0 < \alpha < 1$ is defined as the α-quantile of X under μ:

$$\mathrm{VaR}_\alpha(\mu) := \inf\{x \in \mathbb{R} : \mu(X \leq x) \geq \alpha\}.$$

We will consider such tail risk measures under different probabilistic models, each represented by a probability measure μ on the space Ω of market scenarios, and the notation above emphasizes the dependence on μ.

ES is an alternative to VaR which is sensitive to the tail of the loss distribution:

$$\mathrm{ES}_\alpha(\mu) := \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\beta(\mu)\mathrm{d}\beta.$$

Elicitability and score functions. A statistical functional is *elicitable* if it admits a consistent M-estimator (Gneiting [114], Lambert, Pennock, and Shoham [161], and Osband [185]). More specifically, a statistical functional $T : \mathcal{F} \mapsto \mathbb{R}$ defined on a set of distributions $\mathcal{F}$ on $\mathbb{R}^d$ is *elicitable* if there is a score function $S(x, y)$ such that

$$T(\mu) = \arg\min_x \int S(x, y)\mu(\mathrm{d}y),$$

for any $\mu \in \mathcal{F}$. Examples of elicitable statistical functionals and their strictly consistent score functions include the mean $T(\mu) = \int x\,\mu(\mathrm{d}x)$ with $S(x, y) = (x-y)^2$, and the median $T(\mu) = \inf\{x \in \mathbb{R} : \mu(X \leq x) \geq 0.5\}$ with $S(x, y) = |x - y|$.

It has been shown in Gneiting [114] and Weber [225] that ES is not elicitable, whereas VaR at level $\alpha \in (0, 1)$ is elicitable for random variables with a unique α-quantile. However, it turns out that ES is elicitable of *higher order*, in the sense that the pair $(\mathrm{VaR}_\alpha(\mu), \mathrm{ES}_\alpha(\mu))$ is jointly elicitable. In particular, the following result in [100, Theorem 5.2] gives a family of score functions which are strictly consistent for $(\mathrm{VaR}_\alpha(\mu), \mathrm{ES}_\alpha(\mu))$.

Proposition 2.2.1. [100, Theorem 5.2] *Assume* $\int |x|\mu(\mathrm{d}x) < \infty$. *If* $H_2 : \mathbb{R} \to \mathbb{R}$ *is strictly convex and* $H_1 : \mathbb{R} \to \mathbb{R}$ *is such that*

$$v \mapsto R_\alpha(v, e) := \frac{1}{\alpha}vH_2'(e) + H_1(v), \tag{2.1}$$

is strictly increasing for each $e \in \mathbb{R}$*, then the score function*

$$\begin{aligned} S_\alpha(v,e,x) &= (1_{\{x\le v\}}-\alpha)(H_1(v)-H_1(x)) \\ &\quad +\frac{1}{\alpha}H_2'(e)1_{\{x\le v\}}(v-x)+H_2'(e)(e-v)-H_2(e), \end{aligned} \tag{2.2}$$

is strictly consistent for $(\mathrm{VaR}_\alpha(\mu),\mathrm{ES}_\alpha(\mu))$*, i.e.*

$$(\mathrm{VaR}_\alpha(\mu),\mathrm{ES}_\alpha(\mu)) = \arg\min_{(v,e)\in\mathbb{R}^2}\int S_\alpha(v,e,x)\mu(\mathrm{d}x). \tag{2.3}$$

2.2.2 Properties of score functions for tail risk measures

The computation of the estimator (2.3) involves the optimization of

$$s_\alpha(v,e) := \int S_\alpha(v,e,x)\mu(\mathrm{d}x), \tag{2.4}$$

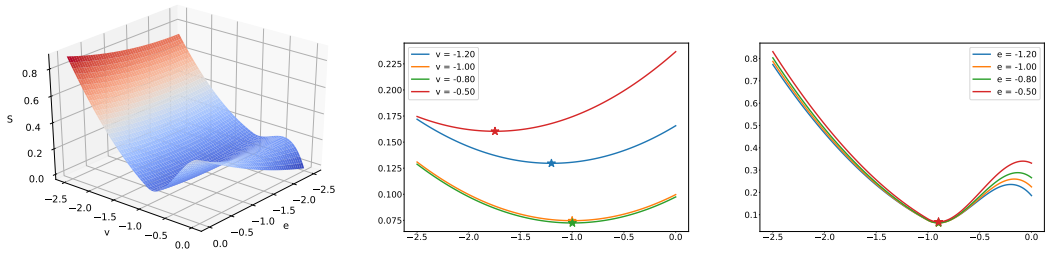
for a given one-dimensional distribution μ. While any choice of H_1, H_2 satisfying the conditions of Proposition 2.2.1 theoretically leads to consistent estimators in (2.3), different choices of H_1 and H_2 lead to optimization problems with different landscapes, with some being easier to optimize than others. We use a specific form of the score function, proposed by Acerbi and Szekely [3], which has been adopted by practitioners for backtesting purposes:

$$S_\alpha(v,e,x) = \frac{W_\alpha}{2}(1_{\{x\le v\}}-\alpha)(x^2-v^2)+1_{\{x\le v\}}e(v-x)+\alpha e\left(\frac{e}{2}-v\right), \text{with } \frac{\mathrm{ES}_\alpha(\mu)}{\mathrm{VaR}_\alpha(\mu)} \ge W_\alpha \ge 1. \tag{2.5}$$

This choice is special case of (2.2), where H_1 and H_2 are given by

$$H_1(v) = -\frac{W_\alpha}{2}v^2,\ H_2(e) = \frac{\alpha}{2}e^2, \quad \text{with} \quad \frac{\mathrm{ES}_\alpha(\mu)}{\mathrm{VaR}_\alpha(\mu)} \ge W_\alpha \ge 1.$$

It is easy to check that (2.5) satisfies the conditions in Proposition 2.2.1 on the subspace $\{(v,e) \in \mathbb{R}^2 \mid W_\alpha\, v \le e \le v \le 0\}$.



(a) $s_\alpha(v, e)$. (b) $s_\alpha(v, e)$ as a function of e with given value of v. (c) $s_\alpha(v, e)$ as a function of v with given value of e.

Figure 2.1: Landscape of $s_\alpha(v, e)$.

Note: This is based on (2.5) with $\alpha = 0.05$ for the uniform distribution on $[-1, 1]$

Next, we provide the following theoretical guarantee for the well-behaved optimization landscape of the score function (2.5); also, see Figure 2.1 for a visualization.

Proposition 2.2.2. *(1) Assume* $\mathrm{VaR}_\alpha(\mu) < 0$, *for* $\alpha < 1/2$. *Then the score* $s_\alpha(v, e)$ *based on* (2.5) *is strictly consistent for* $(\mathrm{VaR}_\alpha(\mu), \mathrm{ES}_\alpha(\mu))$ *and the Hessian of* $s_\alpha(v, e)$ *is positive semi-definite on the region*

$$\mathcal{B} = \{(v, e) \mid v \le \mathrm{VaR}_\alpha(\mu), \text{ and } W_\alpha v \le e \le v \le 0\}.$$

(2) In addition, we assume there exist $\delta_\alpha \in (0, 1)$, $\varepsilon_\alpha \in \left(0, \frac{1}{2} - \alpha\right)$, $z_\alpha \in \left(0, \frac{1}{2} - \alpha\right)$, *and* $W_\alpha > \frac{1}{\sqrt{\alpha}}$ *such that*

$$\frac{\mu(\mathrm{d}x)}{\mathrm{d}x} \ge \delta_\alpha \text{ for } x \in [\mathrm{VaR}_\alpha(\mu), \mathrm{VaR}_{\alpha+\varepsilon_\alpha}(\mu)] \quad \text{and} \quad \mathrm{ES}_\alpha(\mu) \ge W_\alpha \mathrm{Var}_\alpha(\mu) + z_\alpha \quad (2.6)$$

Then the Hessian of $s_\alpha(v, e)$ *is positive semi-definite on the region*

$$\widetilde{\mathcal{B}} = \{(v, e) \mid v \le \mathrm{VaR}_{\alpha+\beta_\alpha}(\mu), \text{ and } W_\alpha v + z_\alpha \le e \le v \le 0\},$$

where $\beta_\alpha = \min\left\{\varepsilon_\alpha, \frac{z_\alpha \delta_\alpha}{2 W_\alpha}\right\}$

See the proof of Proposition 2.2.2 in Appendix A.1.

Example 2.2.3 (Example for condition (2.6))**.** *Condition* (2.6) *holds when* X *has a strictly positive density under measure* μ. *Take an example where* X *follows the standard normal distribution. Denote* $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ *as the density function, and* $F(y) = \int_{-\infty}^{y} f(x)\mathrm{d}x$ *as the cumulative density function for* X. *Then we have*

$\mathrm{VaR}_\alpha(\mu) = F^{-1}(\alpha)$ *and* $\mathrm{ES}_\alpha(\mu) = -\frac{f(F^{-1}(\alpha))}{\alpha}$. *Setting* $\alpha = 0.05$ *and* $\varepsilon_\alpha = 0.05$, *we have* $\mathrm{VaR}_{0.05}(\mu) \approx -1.64$ *and* $\mathrm{ES}_{0.05}(\mu) \approx -2.06$ *by direct calculation. Then we can set* $\delta_\alpha = f(F^{-1}(0.05)) \approx 0.103$, $W_\alpha = 5$ *and* $z_\alpha = \frac{1}{4}$. *Hence* (2.6) *holds for* $\beta_\alpha = \min\{\varepsilon_\alpha, \frac{z_\alpha \delta_\alpha}{2W_\alpha}\} \approx 0.0025$.

Proposition 2.2.2 implies that $s_\alpha(v, e)$ has a well-behaved optimization landscape on regions $\mathcal{B}$ and $\widetilde{\mathcal{B}}$ if the corresponding conditions are satisfied. In particular, the minimizer of $s_\alpha(v, e)$, i.e., $(\mathrm{VaR}_\alpha(\mu), \mathrm{ES}_\alpha(\mu))$, is on the boundary of region $\mathcal{B}$. $\widetilde{\mathcal{B}}$ contains an open ball with center $(\mathrm{VaR}_\alpha(\mu), \mathrm{ES}_\alpha(\mu))$.

In summary, $s_\alpha(v, e)$ has a positive semi-definite Hessian in a neighborhood of the minimum, which leads to desirable properties for convergence. There are other score functions with different choices of H_1 and H_2, but some of them may have undesirable properties (Fissler, Ziegel, et al. [100] and Fissler, Ziegel, and Gneiting [101]) as the following example shows.

Example 2.2.4. *Let* X *be uniformly distributed on* $[-1, 1]$ *and* $\alpha = 0.05$. *When* $H_2(x) = \exp(x)$, *there does not exist a constant* $c < 0$ *such that*

$$\frac{\partial s_\alpha^2}{\partial e^2} \geq 0 \text{ on } (-\infty, c]. \tag{2.7}$$

(2.7) *implies that the objective function* s_α *is not convex which renders the optimization problem difficult.*

Proof of (2.7). If X is uniformly distributed on $[-1, 1]$ and $\alpha = 0.05$, we have

$$\mu(X \leq v)v - \int_{-\infty}^{v} x\mu(x) + (e - v)\alpha = \frac{1}{4}v^2 + \left(\frac{1}{2} - 0.05\right)v + \frac{1}{4} + 0.05e,$$

which yields

$$\frac{\partial^2 s_\alpha}{\partial e^2} = \frac{\exp(e)}{\alpha}\left[\frac{1}{4}(v + 0.9)^2 + 0.0475 + \alpha + 0.05e\right].$$

Letting $v = -0.9$, we arrive at $\frac{\partial^2 s_\alpha}{\partial e^2}|_{v=-0.9} < 0$ for all $e < -1.95$.

The above analysis on the optimization landscapes for different score functions supports the choice of (2.5), as proposed by Acerbi and Szekely [3] in our learning objective function (see next section).

2.3 Learning to generate tail scenarios

We now introduce the *Tail Generative Adversarial Network* (TAIL-GAN) for simulating multivariate price scenarios. Given a set of input price scenarios as training data, TAIL-GAN learns to simulate new price scenarios which lead to accurate *tail risk statistics* for a set of benchmark strategies, by solving a max-min game between a generator and a discriminator. The generator creates samples that are intended to approximate the tail distribution of the training data. The discriminator evaluates the quality of the simulated samples using *tail risk measures*, namely VaR and ES, across a set of benchmark trading strategies, including static portfolios and dynamic trading strategies. Static portfolios and dynamic strategies capture properties of the price scenarios from different perspectives: static portfolios explore the correlation structure among the assets, while dynamic trading strategies, such as mean-reversion and trend-following strategies, discover temporal properties. To train the generator and the discriminator, we use an objective function that leverages the elicitability of VaR and ES, and guarantees the consistency of the estimator.

2.3.1 Discriminator

Given measurable spaces (X_1, Σ_1) and (X_2, Σ_2), a measurable mapping $\Phi\colon X_1 \to X_2$ and a measure $\mu\colon \Sigma_1 \to [0, +\infty]$, the pushforward of μ is defined to be the measure $\Phi\#\mu$ given by, for any $B \in \Sigma_2$,

$$\Phi\#\mu(B) = \mu\Big(\Phi^{-1}(B)\Big). \tag{2.8}$$

In addition, we consider M assets and K different trading strategies of interest. Denote $\mathbf{p} = \{(p_{m,t})_{t=1}^{T}\}_{m=1}^{M}$ as the matrix in $\mathbb{R}^{M\times T}$ that records the prices of each asset m, at the beginning of consecutive time intervals with duration Δ. That is, each price scenario $\mathbf{p}$ contains the price information of M assets over a total period of length $\Delta \times T$. Depending on the purpose of the simulator, Δ could range from milliseconds to minutes, and even to days. Each strategy is allocated the same initial capital, and trading decisions are made at discrete timestamps $t = 1, 2, \ldots, T$.

In order for the strategies to be self-financed, there is no exogenous injection or withdrawal of capital. We consider K *benchmark strategies* to discriminate the performance of the generator. For each strategy $k = 1, 2, \ldots, K$, a mapping $\Pi^k : \mathbb{R}^{M\times T} \to \mathbb{R}$ is defined to map the price scenarios $\mathbf{p}$ to the final PnL x^k at terminal time T, that is, $\Pi^k(\mathbf{p}) = x^k$. Finally, we use $\mathbf{\Pi} := (\Pi^1, \ldots, \Pi^K)$ to define the mapping of all benchmark strategies.

Ideally, the discriminator $\overline{D}$ takes strategy PnL distributions as inputs, and outputs two values for each of the K strategies, aiming to provide the correct $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$. Mathematically, this amounts to

$$\overline{D}^* \in \arg\min_{\overline{D}} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[S_\alpha\left(\overbrace{\overline{D}(\underbrace{\Pi^k \# \mathbb{P}_r}_{\text{strategy PnL distribution}})}^{\text{VaR and ES prediction from } \overline{D}}; \mathbf{p} \right)\right]. \tag{2.9}$$

However, it is impossible to access the true distribution of $\mathbf{p}$, denoted as $\mathbb{P}_r$, in practice. Therefore we consider a sample-based version of the discriminator. More specifically, we consider PnL samples $\{\mathbf{p}_i\}_{i=1}^n$ with a fixed size n as the input of the discriminator. Mathematically, we write

$$D^* \in \arg\min_{D} \frac{1}{K}\sum_{k=1}^{K} \frac{1}{n}\sum_{i=1}^{n}\left[S_\alpha\left(\overbrace{D(\underbrace{\Pi^k(\mathbf{p}_j), j \in [n]}_{\text{strategy PnL samples}})}^{\text{VaR and ES prediction from } D}; \mathbf{p}_i \right)\right]. \tag{2.10}$$

Here the expectation in (2.9) is replaced by the empirical mean using finite samples $\{\mathbf{p}_i\}_{i=1}^n$. In (2.10), we search the discriminator D over all Lipschitz functions parameterized by the neural network architecture. Specifically, the discriminator adopts a neural network architecture with $\tilde{L}$ layers, and the input dimension is $\tilde{n}_1 := n$ and the output dimension is $\tilde{n}_{\tilde{L}} := 2$. Note that the α-VaR of a distribution can be approximated by the $\lfloor \alpha n \rfloor^{th}$ smallest value in a sample of size n from this distribution, which is permutation-invariant to the ordering of the samples. Given that the goal of the discriminator is to predict the α-VaR and α-ES, including a sorting function in our architecture design could potentially improve the stability of the discriminator. We denote this (differentiable) neural sorting function as $\widetilde{\Gamma}$ (Grover et al. [118]), with details deferred to Appendix A.2.3.

In summary, the discriminator is given by

$$D(\mathbf{x}^k;\delta) = \widetilde{\mathbf{W}}_{\tilde{L}} \cdot \sigma\left(\widetilde{\mathbf{W}}_{\tilde{L}-1} \ldots \sigma(\widetilde{\mathbf{W}}_1 \widetilde{\Gamma}(\mathbf{x}^k) + \widetilde{\mathbf{b}}_1) \ldots + \widetilde{\mathbf{b}}_{\tilde{L}-1}\right) + \widetilde{\mathbf{b}}_{\tilde{L}}, \tag{2.11}$$

where $\delta = (\widetilde{\mathbf{W}}, \widetilde{\mathbf{b}})$ represent all the parameters in the neural network. Here we have $\widetilde{\mathbf{W}} = (\widetilde{\mathbf{W}}_1, \widetilde{\mathbf{W}}_2, \ldots, \widetilde{\mathbf{W}}_{\tilde{L}})$ and $\widetilde{\mathbf{b}} = (\widetilde{\mathbf{b}}_1, \widetilde{\mathbf{b}}_2, \ldots, \widetilde{\mathbf{b}}_{\tilde{L}})$ with $\widetilde{\mathbf{W}}_l \in \mathbb{R}^{n_l \times n_{l-1}}$, $\widetilde{\mathbf{b}}_l \in \mathbb{R}^{n_l \times 1}$ for $l = 1, 2, \ldots, \tilde{L}$. In the neural network literature, the $\widetilde{\mathbf{W}}_l$'s are often called the *weight* matrices, the $\widetilde{\mathbf{b}}_l$'s are called *bias* vectors. The outputs of the discriminator are two values for each of the K strategies, (hopefully) representing the α-VaR and α-ES. The operator $\sigma(\cdot)$ takes a vector of any dimension as input, and applies a function component-wise. $\sigma(\cdot)$ is referred to as the *activation function*. Specifically, for any $q \in \mathbb{Z}^+$ and any vector $\mathbf{u} = (u_1, u_2, \ldots, u_q)^\top \in \mathbb{R}^q$, we have that $\sigma(\mathbf{u}) = (\sigma(u_1), \sigma(u_2), \ldots, \sigma(u_q))^\top$. Several popular choices for the activation function include ReLU with $\sigma(u) = \max(u, 0)$, Leaky ReLU with $\sigma(u) = a_1 \max(u, 0) - a_2 \max(-u, 0)$ and $a_1, a_2 > 0$, and smooth functions such as $\sigma(\cdot) = \tanh(\cdot)$. We sometimes use the abbreviation D_δ or D instead of $D(\cdot;\delta)$ for notation simplicity.

Accordingly, we define $\mathcal{D}$ as a class of discriminators

$$\mathcal{D}(\tilde{L}, \tilde{n}_1, \ldots, \tilde{n}_{\tilde{L}}) = \Big\{ D : \mathbb{R}^n \to \mathbb{R}^2 \;\Big|\; D \text{ takes the form in (2.11) with } \tilde{L} \text{ layers and } \tilde{n}_l \text{ as the width of each layer, } \|\widetilde{\mathbf{W}}_l\|_\infty, \|\widetilde{\mathbf{b}}_l\|_\infty < \infty \text{ for } l = 1, 2, \ldots, \tilde{L} \Big\}, \tag{2.12}$$

where $\|\cdot\|_\infty$ denotes the max-norm that takes the max absolute value of all elements in the input matrix or vector.

2.3.2 Generator

For the generator, we use a neural network with $L \in \mathbb{Z}^+$ layers. Denoting by n_l the width of the l-th layer, the functional form of the generator is given by

$$G(\mathbf{z};\gamma) = \mathbf{W}_L \cdot \sigma\left(\mathbf{W}_{L-1} \ldots \sigma(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1) \ldots + \mathbf{b}_{L-1}\right) + \mathbf{b}_L, \tag{2.13}$$

in which $\gamma := (\mathbf{W}, \mathbf{b})$ represents the parameters in the neural network, with $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_L)$ and $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_L)$. Here $\mathbf{W}_l \in \mathbb{R}^{n_l \times n_{l-1}}$ and $\mathbf{b}_l \in \mathbb{R}^{n_l \times 1}$ for $l = 1, 2, \ldots, L$, where $n_0 = N_z$ is the dimension of the input variable.

We define $\mathcal{G}$ as a class of generators that satisfy given regularity conditions on the neural network parameters

$$\mathcal{G}(L, n_1, n_2, \ldots, n_L) = \Big\{ G : \mathbb{R}^{N_z} \to \mathbb{R}^{M\times T} \;\Big|\; G \text{ takes the form in (2.13) with } L \text{ layers and } n_l \text{ as the width of each layer, } \|\mathbf{W}_l\|_\infty, \|\mathbf{b}_l\|_\infty < \infty \text{ for } l = 1, 2, \ldots, L \Big\}. \tag{2.14}$$

To ease the notation, we may use the abbreviation $G_\gamma(\cdot)$ or drop the dependency of $G(\cdot\,;\gamma)$ on the neural network parameters γ and conveniently write $G(\cdot)$. We further denote $\mathbb{P}_G$ as the distribution of price series generated by G under the initial distribution $\mathbf{z} \sim \mathbb{P}_{\mathbf{z}}$.

Universal approximation property of the generator. We first demonstrate the universal approximation power of the generator under the VaR and ES criteria, and then we provide a similar result for more general risk measures that satisfy certain Hölder regularity property.

Given two probability measures μ and ν on $\mathbb{R}^d$, recall in (2.8) that a transport map Φ between μ and ν is a measurable map $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ such that $\nu = \Phi\#\mu$. We denote by $\Gamma(\mu, \nu)$ the set of transport plans between μ and ν which consists of all coupling measures γ of μ and ν, i.e., $\gamma(A \times \mathbb{R}^d) = \mu(A)$ and $\gamma(\mathbb{R}^d \times B) = \gamma(B)$ for any measurable $A, B \subset \mathbb{R}^d$.

We assume that the portfolio values are Lipschitz-continuous with respect to price paths:

Assumption 2.3.1 (Lipschitz Continuity of the Portfolio Value)**.** *For $k = 1, 2, \cdots, K$,*

$$\exists \ell_k > 0, \qquad |\Pi^k(x) - \Pi^k(y)| \le \ell_k \|x - y\|, \quad \forall x, y \in \mathbb{R}^{M\times T}.$$

Recall $\mathbb{P}_r$ and $\mathbb{P}_z$ are respectively the target distribution and the distribution of the input noise.

Assumption 2.3.2 (Noise Distribution and Target Distribution)**.** *$\mathbb{P}_r$ and $\mathbb{P}_z$ are probability measures on $\mathbb{R}^{M\times T}$ (i.e. $N_z = M \times T$) satisfying the following conditions:*

- *$\mathbb{P}_z \in \mathcal{P}^2(\mathbb{R}^{M\times T})$ is absolutely continuous with respect to the Lebesgue measure.*

- $\mathbb{P}_r$ *has a bounded moment of order* $\beta > 1$*:* $\int \|x\|^\beta \mathbb{P}_r(dx) < \infty$,.

Theorem 2.3.3 (Universal Approximation under VaR and ES Criteria)**.** *Under Assumptions 2.3.1-2.3.2 and the additional assumption*

A3 *The random variables* $\Pi^k(X)$ *have continuous densities* f_k *under* $\mathbb{P}_r$ *and there exist* $\eta_k, \delta_k > 0$ *such that* $f_k(x) > \eta_k$ *for all* $x \in [\mathrm{VaR}_\alpha(\Pi^k \# \mathbb{P}_r) - \frac{\delta_k}{2}, \mathrm{VaR}_\alpha(\Pi^k \# \mathbb{P}_r) + \frac{\delta_k}{2}]$.

we have, for any $\varepsilon > 0$

- *there exists a positive integer* $n_1 = \mathcal{O}(\varepsilon^{-2})$*, and a feed-forward neural network* G_1 *with* $L = \lceil \log n_1 \rceil$ *fully connected layers of equal width* $N = 2^L$ *and ReLU activation, such that*

$$\left| \mathrm{VaR}_\alpha\Big(\Pi^k \# \Big((\nabla G_1) \# \mathbb{P}_z\Big)\Big) - \mathrm{VaR}_\alpha\Big(\Pi^k \# \mathbb{P}_r\Big) \right| < \varepsilon;$$

- *there exists a positive integer* $n_2 = \mathcal{O}(\varepsilon^{-\frac{\beta}{\beta-1}})$*, and a feed-forward neural network* G_2 *with* $L = \lceil \log n_2 \rceil$ *fully connected layers of equal width* $N = 2^L$ *and ReLU activation, such that*

$$\left| \mathrm{ES}_\alpha\Big(\Pi^k \# \Big((\nabla G_2) \# \mathbb{P}_z\Big)\Big) - \mathrm{ES}_\alpha\Big(\Pi^k \# \mathbb{P}_r\Big) \right| < \varepsilon.$$

Theorem 2.3.3 implies that a feed-forward neural network with fully connected layers of equal-width neural network is capable of generating scenarios which reproduce the tail risk properties (VaR and ES) for the benchmark strategies with arbitrary accuracy. This justifies the use of this simple network architecture for TAIL-GAN. The size of the network, namely the width and the length, depends on the tolerance of the error ε and depends on β in the case of ES.

The proof (given in Appendix A.1) consists in using the theory of (semi-discrete) optimal transport to build a transport map of the form $\Phi = \nabla\psi$ which pushes the source distribution $\mathbb{P}_z$ to the empirical distribution $\mathbb{P}_r^{(n)}$. The potential ψ has an explicit form in terms of the maximum of finitely many affine functions. Such

an explicit structure enables the representation of ψ with a finite deep neural network (Lu and Lu [172]).

We now provide a universal approximation result for more general law-invariant risk measures. Consider a risk measure, defined as a map $\rho : \mathcal{L} \to \mathbb{R}$ on the set $\mathcal{L}$ of probability measures on $\mathbb{R}$. VaR, ES, and spectral risk measures are examples of such risk measures (Cont, Deguest, and He [73]). With a slight abuse of notation, for a real-valued random variable X with CDF F, we will write $\rho(X) = \rho(F)$.

Denote by

$$\mathcal{W}_p(\mu,\nu) = \inf_{\xi\in\mathcal{C}(\mu,\nu)} \Big[\mathbb{E}_{(X,Y)\sim\xi}\|X-Y\|^p\Big]^{1/p},$$

the Wasserstein distance of order $p \in [1,\infty)$ between two probability measures μ and ν on $\mathbb{R}^d$, where $\mathcal{C}(\mu,\nu)$ denotes the collection of all distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginal distributions μ and ν.

Theorem 2.3.4 (Universal Approximation under General Risk Measure)**.** *Under Assumptions 2.3.1-2.3.2 and the additional assumption that*

A4 *ρ is Hölder continuous for the Wasserstein-1 metric:*

$$\exists L>0, \quad \exists \kappa \in (0,1], \quad \Big|\rho(\Pi^k\#\mu) - \rho(\Pi^k\#\nu))\Big| \le L\Big(\mathcal{W}_1\Big(\Pi^k\#\mu, \Pi^k\#\nu\Big)\Big)^{\kappa} \tag{2.15}$$

for any ε, there exists a positive integer n_3, and a feed-forward neural network G_3 with $L = \lceil \log n_3 \rceil$ fully connected layers with equal width $N = 2^L$ and ReLU activation such that

$$\Big|\rho\Big(\Pi^k\#\Big(\nabla G_3\#\mathbb{P}_z\Big)\Big) - \rho\Big(\Pi^k\#\mathbb{P}_r\Big)\Big| < \varepsilon.$$

Furthermore,

(1) $n_3 = \mathcal{O}\left(\varepsilon^{-\frac{\beta}{\kappa(\beta-1)}}\right)$ *when* $M = T = 1$ *and* $1 < \beta \le 2$*;*

(2) $n_3 = \mathcal{O}\left(\varepsilon^{-\frac{2}{\kappa}}\right)$ *when* $M = T = 1$ *and* $\beta \ge 2$*;*

(3) $n_3 = \mathcal{O}\left(\varepsilon^{-\frac{M\times T}{\kappa}}\right)$ *when* $M \times T \ge 2$ *and* $\frac{1}{M\times T} + \frac{1}{\beta} < 1$*;*

(4) $n_3 = \mathcal{O}\left(\varepsilon^{-\frac{\beta}{\kappa(\beta-1)}}\right)$ *when* $M \times T \geq 2$ *and* $\frac{1}{M\times T} + \frac{1}{\beta} \geq 1$.

Remark 2.3.5. *Examples of risk measures that are Hölder continuous, i.e. which satisfy Assumption A4, include the optimized certainty equivalent (Ben-Tal and Teboulle [28, 29]), spectral risk measures (Acerbi [2]), and utility-based shortfall risk (Föllmer and Schied [103]).*

As suggested in Theorem 2.3.4, the depth of the neural network depends on β, $M \times T$, and κ. $M = T = 1$ corresponds to the simulation of a single price value of a single asset, which is not much of an interesting case. When $M \times T \geq 2$ and $\beta > \frac{M\times T}{M\times T-1}$, the complexity of the neural network, characterized by n_3, depends on the ratio between the dimension of the price scenario $M \times T$ and the Lipschitz exponent κ. When $\mathbb{P}_r$ is heavy-tailed in the sense that $1 < \beta < \frac{M\times T}{M\times T-1}$, n_3 is determined by $\frac{\beta}{(\beta-1)\kappa}$. The proof of Theorem 2.3.4 is deferred to Appendix A.1.

2.3.3 Loss function

We now design a loss function to train both the generator and the discriminator together. We start with a bi-level optimization formulation and then introduce the max-min game formulation by relaxing the lower-level optimization problem. We also show that these two formulations are equivalent under mild conditions.

Bi-level optimization problem. We first start with a theoretical version of the objective function to introduce some insights and then provide the practical sample-based version for training. Given two classes of functions $\overline{\mathcal{G}} := \{\overline{G} : \mathbb{R}^{N_z} \to \mathbb{R}^{M\times T}\}$ and $\overline{\mathcal{D}} := \{\overline{D} : \mathcal{P}(\mathbb{R}) \to \mathbb{R}^2\}$, our goal is to find a generator $\overline{G}^* \in \overline{\mathcal{G}}$ and a discriminator $\overline{D}^* \in \overline{\mathcal{D}}$ via the following bi-level (or constrained) optimization problem

$$\overline{G}^* \in \arg\min_{\overline{G}\in\overline{\mathcal{G}}} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r} \left[S_\alpha\Big(\overline{D}^*\Big(\Pi^k \# \mathbb{P}_{\overline{G}}\Big),\ \Pi^k(\mathbf{p})\Big)\right], \tag{2.16}$$

where $\mathbb{P}_{\overline{G}} \in \mathcal{P}(\mathbb{R}^{M\times T})$ is the distribution of the samples from $\overline{G}$ and

$$\overline{D}^* \in \arg\min_{\overline{D}\in\overline{\mathcal{D}}} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r} \left[S_\alpha\Big(\overline{D}\Big(\Pi^k \# \mathbb{P}_r\Big),\ \Pi^k(\mathbf{p})\Big)\right]. \tag{2.17}$$

In the bi-level optimization problem (2.16)-(2.17), the discriminator $\overline{D}^*$ aims to map the PnL distribution $\Pi^k \# \mathbb{P}_r$ to the associated α-VaR and α-ES values. Given the definition of the score function and the joint elicitability property of VaR and ES, we have $\overline{D}^*(\cdot) := (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)(\cdot)$ according to (2.3). Assume $\overline{D}^*$ solves (2.17), the simulator $\overline{G}^* \in \overline{\mathcal{G}}$ in (2.16) aims to map the noise input to a price scenario that has *consistent* VaR and ES values of the strategy PnLs applied to $\mathbb{P}_r$. Note our current formulation (2.16)-(2.17) implicitly assumes that the user of our Tail-GAN framework assigns equal importance to the losses of all trading strategies of interest. It can be adapted to account for different weights if some of the strategies are more important than others.

From bi-level optimization to max-min game. In practice, constrained optimization problems are difficult to solve, and one can instead relax the constraint by applying the Lagrangian relaxation method with a dual parameter $\lambda > 0$, leading to a max-min game between two neural networks $\overline{D}$ and $\overline{G}$,

$$\max_{\overline{D} \in \overline{\mathcal{D}}_0} \min_{\overline{G} \in \overline{\mathcal{G}}} \frac{1}{K} \sum_{k=1}^{K} \left[\mathbb{E}_{\mathbf{p} \sim \mathbb{P}_r} \left[S_\alpha\Big(\overline{D}\Big(\Pi^k \# \mathbb{P}_{\overline{G}}\Big), \Pi^k(\mathbf{p})\Big)\right] - \lambda\, \mathbb{E}_{\mathbf{p} \sim \mathbb{P}_r} \left[S_\alpha\Big(\overline{D}(\Pi^k \# \mathbb{P}_r),\ \Pi^k(\mathbf{p})\Big)\right] \right] \tag{2.18}$$

where

$$\overline{\mathcal{D}}_0 := \Big\{ \overline{D} :\ \ \mathcal{P}(\mathbb{R}) \to \mathbb{R}^2 \ \ \textit{and}\ \ \exists \mu \in \mathcal{P}(\mathbb{R}^{M \times T}) \text{ with a finite first moment } \text{ s.t.} \\ \overline{D}(\Pi^k \# \mu) = (\mathrm{VaR}_\alpha(\Pi^k \# \mathbb{P}_r), \mathrm{ES}_\alpha(\Pi^k \# \mathbb{P}_r)), k = 1, 2, \cdots, K \Big\} \tag{2.19}$$

is a smaller set of discriminators such that $\overline{\mathcal{D}}_0 \subseteq \mathcal{D}$. Note that the condition in (2.19) implies that $\overline{D}$ is sensitive to the input distribution, leading to a non-degenerate mapping. This is not a restrictive condition to impose. Intuitively, if the discriminator detects that the samples come from the true distribution, it will provide a pair of values that is close to the minimizer of the score function, namely the VaR and ES values of the true distribution, in accordance with the elicitability property.

Theorem 2.3.6 (Equivalence of the Formulations)**.** *Set $N_z = M \times T$. Assume that $\mathbb{P}_z$ has a finite first moment and is absolutely continuous with respect to the*

Lebesgue measure. Then the max-min game (2.18) *with* $\overline{\mathcal{D}}_0$ *is equivalent to the bi-level optimization problem* (2.16)-(2.17) *for any* $\lambda > 0$.

The proof of Theorem 2.3.6 is deferred to Appendix A.1.

A sample-based version of the loss function. As we explained in Section 2.3.1, it is impossible to access the true distribution $\mathbb{P}_r$ in practice. Therefore we consider a sample-based version of the discriminator. This leads to the following formulation of the loss function

$$\max_{D\in\mathcal{D}}\min_{G\in\mathcal{G}} \frac{1}{Kn}\sum_{k=1}^{K}\sum_{j=1}^{n}\Big[S_\alpha\Big(D\big(\Pi^k(\mathbf{q}_i); i\in[n]\big), \Pi^k(\mathbf{p}_j)\Big) - \lambda\, S_\alpha\Big(D\big(\Pi^k(\mathbf{p}_i); i\in[n]\big),\ \Pi^k(\mathbf{p}_j)\Big)\Big] \tag{2.20}$$

where $\mathbf{p}_i, \mathbf{p}_j \sim \mathbb{P}_r$ and $\mathbf{q}_i \sim \mathbb{P}_G$ $(i, j = 1, 2, \ldots, n)$. The discriminator D takes n PnL samples as the input and aims to provide the VaR and ES values of the sample distribution as the output. (See Section 2.3.1 for the design of the discriminator.) The score function S_α is defined in (2.5). In addition, the max-min structure of (2.18) encourages the exploration of the generator to simulate scenarios that are not exactly the same as what is observed in the input price scenarios, but are equivalent under the criterion of the score function, hence improving generalization. We refer the readers to Section 2.5.3 for a comparison between TAIL-GAN and supervised learning methods and a demonstration of the generalization power of TAIL-GAN.

Compared to the Jensen-Shannon divergence used in the loss function of standard GAN architectures (Goodfellow et al. [116]) and Wasserstein distance in WGAN (Arjovsky, Chintala, and Bottou [16]), the objective function in (2.18) with the score function S_α is more sensitive to tail risk and leads to an output which better approximates the α-ES and α-VaR values. The architecture of TAIL-GAN is depicted in Figure 2.2.

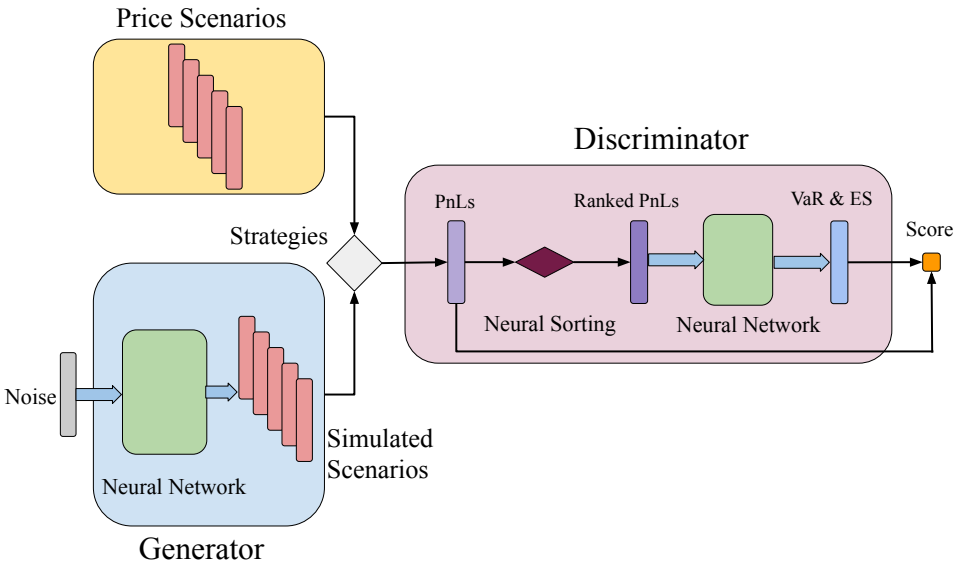


Figure 2.2: Architecture of TAIL-GAN.

Note: The (blue) thick arrows represent calculations with learnable parameters, and the (black) thin arrows represent calculations with fixed parameters.

2.4 Numerical experiments: methodology and performance evaluation

Before we proceed with the numerical experiments on both synthetic data sets and real financial data sets, we first describe the methodologies, performance evaluation criterion, and several baseline models for comparison.

2.4.1 Methodology

Algorithm 1 provides a detailed description of the training procedure of TAIL-GAN, which allows us to train the simulator with different benchmark trading strategies.

For ease of exposition, unless otherwise stated, we denote TAIL-GAN as the model (introduced in Section 2.3) trained with both **static and dynamic** trading strategies across multiple assets, as this is a typical set-up used by asset managers.[1]

[1]Here the static trading strategy refers to the buy-and-hold strategy and dynamic trading strategies are the portfolios with time-dependent and history-dependent weights. For static multi-asset portfolios, the weights are generated randomly before the trading period. It is not our main goal to run a horse race between TAIL-GAN variants trained with various strategies. Therefore, regarding dynamic trading strategies, we only consider the practically popular mean-reversion strategies and trend-following strategies. It is easy to incorporate more complex strategies, such

Algorithm 1 TAIL-GAN.

Input:
• Price scenarios $\mathbf{p}_1, \ldots, \mathbf{p}_N \in \mathbb{R}^{M\times T}$.
• Description of trading strategies: $\mathbf{\Pi} = (\Pi^1, \ldots, \Pi^K) : \mathbb{R}^{M\times T} \to \mathbb{R}^K$.
• Hyperparameters: learning rate l_D for discriminator and l_G for generator; number of training epochs; batch size N_B; dual parameter λ.

for number of epochs **do**
  for $j = 1 \to \lfloor N/N_B \rfloor$ **do**
    Generate N_B IID noise samples $\{\mathbf{z}_i, i \in [N_B]\} \sim \mathbb{P}_{\mathbf{z}}$.
    Sample a batch $\mathcal{B}_j \subset \{1, \ldots, N\}$ of size N_B from the input data $\{\mathbf{p}_i, i = 1, \ldots, N\}$.
    Compute the loss of the discriminator on the batch $\mathcal{B}_j$

$$\mathcal{L}_D(\delta) = \frac{1}{K\,N_B} \sum_{k=1}^{K} \sum_{n\in\mathcal{B}_j} \Big[S_\alpha\Big(D_\delta(\Pi^k(G_\gamma(\mathbf{z}_i)), i \in [N_B]), \Pi^k(\mathbf{p}_n)\Big) - \lambda S_\alpha\Big(D_\delta(\Pi^k(\mathbf{p}_i), i \in \mathcal{B}_j), \Pi^k(\mathbf{p}_n)\Big)\Big].$$

    Update the discriminator

$$\delta \leftarrow \delta + l_D \nabla \mathcal{L}_D(\delta).$$

    Generate N_B IID noise samples $\{\tilde{\mathbf{z}}_i, i \in [N_B]\} \sim \mathbb{P}_{\mathbf{z}}$.
    Compute the loss of the generator

$$\mathcal{L}_G(\gamma) = \frac{1}{K\,N_B} \sum_{k=1}^{K} \sum_{n\in\mathcal{B}_j} S_\alpha\Big(D_\delta(\Pi^k(G_\gamma(\tilde{\mathbf{z}}_i)), i \in [N_B]), \Pi^k(\mathbf{p}_n)\Big).$$

    Update the generator

$$\gamma \leftarrow \gamma - l_G \nabla \mathcal{L}_G(\gamma).$$

  end for
end for
$\gamma^* = \gamma, \qquad \delta^* = \delta.$

Outputs:
• δ^*: trained discriminator weights; γ^*: trained generator weights.
• Simulated scenarios: $G_{\gamma^*}(\mathbf{z}_i)$ where $\mathbf{z}_i \sim \mathbb{P}_{\mathbf{z}}$ IID.

We compare Tail-GAN with four benchmark models:

(1) Tail-GAN-Raw: Tail-GAN trained (only) with (static) buy-and-hold strategies on *individual asset*.

(2) Tail-GAN-Static: Tail-GAN trained (only) with static multi-asset portfolios.

(3) Historical Simulation Method (HSM): using VaR and ES computed from historical data as the prediction for VaR and ES of future data.

(4) Wasserstein GAN (WGAN): trained on asset return data with Wasserstein distance as the loss function. We refer to Arjovsky, Chintala, and Bottou [16] for more details on WGAN.

As Tail-GAN-Raw is trained with the PnLs of single-asset portfolios, it shares the same spirit as Quant-GAN (Wiese et al. [226]) as well as many GAN-based generators for financial time series (such as Koshiyama, Firoozye, and Treleaven [155], Li et al. [166], Marti [174], Ni et al. [180], and Takahashi, Chen, and Tanaka-Ishii [211]). Tail-GAN-Static is trained with the PnLs of multi-asset portfolios, thus more flexible than Tail-GAN-Raw by allowing different capital allocations among different assets. This could potentially capture the correlation patterns among different assets. In addition to the static portfolios, Tail-GAN also includes dynamic strategies to capture temporal dependence information in financial time series.

2.4.2 Performance evaluation criteria

We introduce the following criteria to compare the scenarios simulated using Tail-GAN with other simulation models: (1) tail behavior comparison; (2) structural characterizations such as correlation and auto-correlation; and (3) model verification via (statistical) hypothesis testing. The first two evaluation criteria are applied throughout the numerical analysis for both synthetic and real financial data, while

as statistical arbitrage.

the hypothesis tests are only for synthetic data as they require the knowledge of "oracle" estimates of the true data generating process.

Tail behavior. To evaluate how closely the VaR (ES) of strategy PnLs, computed under the generated scenarios, match the ground-truth VaR (ES) of the strategy PnLs computed under input scenarios, we employ both quantitative and qualitative assessment methods to gain a thorough understanding.

The quantitative performance measure is *relative error* of VaR and ES. For any strategy k ($1 \leq k \leq K$), the relative error of VaR is defined as $\frac{\left|\mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_G^{(\mathfrak{N})}\right) - \mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|}{\left|\mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|}$, where $\mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)$ is the α-VaR of the PnL for strategy k evaluated under $\mathbb{P}_r$ and $\mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_G^{(\mathfrak{N})}\right)$ is the empirical estimate of VaR for strategy k evaluated with $\mathfrak{N}$ samples under $\mathbb{P}_G$. Similarly, for the estimates $\mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_G^{(\mathfrak{N})}\right)$, we define the relative error as $\frac{\left|\mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_G^{(\mathfrak{N})}\right) - \mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|}{\left|\mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|}$ for the ES of strategy k. We then use the following *average relative errors* of VaR and ES as the overall measure of model performance

$$\mathrm{RE}(\mathfrak{N}) = \frac{1}{2K} \sum_{k=1}^{K} \left(\frac{\left|\mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_G^{(\mathfrak{N})}\right) - \mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|}{\left|\mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|} + \frac{\left|\mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_G^{(\mathfrak{N})}\right) - \mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|}{\left|\mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|} \right).$$

One useful benchmark is to compare the above relative error with the *sampling error* below, when only using a finite number of real samples to calculate VaR and ES

$$\mathrm{SE}(\mathfrak{N}) = \frac{1}{2K} \sum_{k=1}^{K} \left(\frac{\left|\mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_r^{(\mathfrak{N})}\right) - \mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|}{\left|\mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|} + \frac{\left|\mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_r^{(\mathfrak{N})}\right) - \mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|}{\left|\mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_r\right)\right|} \right),$$

where $\mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_r^{(\mathfrak{N})}\right)$ and $\mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_r^{(\mathfrak{N})}\right)$ are the estimates for VaR and ES of strategy k using ground-truth data with sample size $\mathfrak{N}$.

GANs are usually trained with a finite number of samples, and thus it is difficult for the trained GAN to achieve a better sampling error of the training data with the same sample size. To this end, we may conclude that the trained GAN reaches its best finite-sample performance if $\mathrm{RE}(\mathfrak{N})$ is comparable to $\mathrm{SE}(\mathfrak{N})$. This benchmark comparison and model validation steps are important to evaluate the performance of GAN models. However, these steps are missing in the GAN literature for simulating financial time series (Wiese et al. [226]).

We also use *rank-frequency distribution* to visualize the tail behaviors of the simulated data versus the market data. Rank-frequency distribution is a discrete form of the quantile function, i.e., the inverse cumulative distribution, giving the size of the element at a given rank. By comparing the rank-frequency distribution of the market data and simulated data of different strategies, we gain an understanding of how good the simulated data is in terms of the risk measures of different strategies.

Structural characterization. We are also interested in testing whether Tail-GAN is capable of capturing structural properties, such as temporal and spatial correlations, of the input price scenarios. To do so, we calculate and compare the following statistics of the output price scenarios generated by each simulator: (1) the sum of the absolute difference between the *correlation* coefficients of the input price scenario and those of generated price scenario, and (2) the sum of the absolute difference between the *autocorrelation* coefficients (up to 10 lags) of the input price scenario and those of the generated price scenario.

Hypothesis testing for synthetic data. Given the benchmark strategies and a simulation model (generically referred to as $\mathcal{M}$), we are interested in testing (or rejecting) whether risk measures for benchmark strategies estimated from simulated scenarios $\mathcal{M}$ are as accurate as "oracle" estimates given knowledge of the true data generating process. Here, $\mathcal{M}$ may represent Tail-GAN, Tail-GAN-Raw, Tail-GAN-Static, HSM, and WGAN.

We explore two methods, the Score-based Test and the Coverage Test, to verify the relationship between the simulator $\mathcal{M}$ and the true model. We first introduce the *Score-based Test* to verify the hypothesis

$$\begin{aligned}\mathcal{H}_0: \quad & \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[S_\alpha\Big(\mathrm{VaR}_\alpha\left(\Pi^k\#\mathbb{P}_{\mathcal{M}}\right),\mathrm{ES}_\alpha\left(\Pi^k\#\mathbb{P}_{\mathcal{M}}\right),\Pi^k(\mathbf{p})\Big)\right] \\ & = \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[S_\alpha\Big(\mathrm{VaR}_\alpha\left(\Pi^k\#\mathbb{P}_r\right),\mathrm{ES}_\alpha\left(\Pi^k\#\mathbb{P}_r\right),\Pi^k(\mathbf{p})\Big)\right].\end{aligned}$$

By making use of the joint elicitability property of VaR and ES, Fissler, Ziegel, and Gneiting [101] proposed the following test statistic to verify $\mathcal{H}_0$

$$T^k = \frac{\bar{S}^k_{\mathcal{M}} - \bar{S}^k_{\text{Ground-Truth}}}{\hat{\sigma}^k},$$

where

$$\begin{aligned}
\bar{S}^k_{\mathcal{M}} &= \frac{1}{\mathfrak{N}}\sum_{i=1}^{\mathfrak{N}} S_\alpha\Big(\mathrm{VaR}_\alpha\left(\Pi^k\#\mathbb{P}_{\mathcal{M}}\right), \mathrm{ES}_\alpha\left(\Pi^k\#\mathbb{P}_{\mathcal{M}}\right), \Pi^k(\mathbf{p}_i)\Big),\\
\bar{S}^k_{\text{Ground-Truth}} &= \frac{1}{\mathfrak{N}}\sum_{i=1}^{\mathfrak{N}} S_\alpha\Big(\mathrm{VaR}_\alpha\left(\Pi^k\#\mathbb{P}_{r}\right), \mathrm{ES}_\alpha\left(\Pi^k\#\mathbb{P}_{r}\right), \Pi^k(\mathbf{p}_i)\Big),\\
\hat{\sigma}^k &= \sqrt{\frac{\hat{s}^2_{\mathcal{M}} + \hat{s}^2_{\text{Ground-Truth}}}{\mathfrak{N}}}.
\end{aligned}$$

Here $\{\mathbf{p}_i\}_{i=1}^{\mathfrak{N}}$ represents the observations from $\mathbb{P}_r$ and $\{\Pi^k(\mathbf{p}_i)\}_{i=1}^{\mathfrak{N}}$ represents the PnL observations of strategy k under $\mathbb{P}_r$. $\mathbb{P}_{\mathcal{M}}$ denotes the distribution of generated data from simulator $\mathcal{M}$. $\mathrm{VaR}_\alpha\left(\Pi^k\#\mathbb{P}_{\mathcal{M}}\right)$ and $\mathrm{ES}_\alpha\left(\Pi^k\#\mathbb{P}_{\mathcal{M}}\right)$ represent the estimates of VaR and ES for PnLs of strategy k evaluated under $\mathbb{P}_{\mathcal{M}}$. $\mathrm{VaR}_\alpha\left(\Pi^k\#\mathbb{P}_{r}\right)$ and $\mathrm{ES}_\alpha\left(\Pi^k\#\mathbb{P}_{r}\right)$ represent the ground-truth estimates of VaR and ES for PnLs of strategy k evaluated under $\mathbb{P}_r$. Furthermore, $\hat{s}^2_{\mathcal{M}}$ and $\hat{s}^2_{\text{Ground-Truth}}$ are the empirical variances of
$S_\alpha\Big(\mathrm{VaR}_\alpha\left(\Pi^k\#\mathbb{P}_{\mathcal{M}}\right), \mathrm{ES}_\alpha\left(\Pi^k\#\mathbb{P}_{\mathcal{M}}\right), \Pi^k(\mathbf{p})\Big)$ and $S_\alpha\Big(\mathrm{VaR}_\alpha\left(\Pi^k\#\mathbb{P}_{r}\right), \mathrm{ES}_\alpha\left(\Pi^k\#\mathbb{P}_{r}\right), \Pi^k(\mathbf{p})\Big)$, respectively. Under $\mathcal{H}_0$, the test statistic T^k has expected value equal to zero, and the asymptotic normality of the test statistics T^k can be similarly proved as in Diebold and Mariano [86].

The second test we explore is the *Coverage Test*, also known as Kupiec Test (Campbell [50] and Kupiec [157]). It measures the simulator performance by comparing the observed violation rate of estimates from a simulator with the expected violation rate. The null hypothesis of Coverage Test is

$$\mathcal{H}_0 : \mathbb{P}\left(\Pi^k(\mathbf{p}) < \mathrm{VaR}_\alpha\left(\Pi^k\#\mathbb{P}_{\mathcal{M}}\right)\right) = \alpha.$$

Here $\Pi^k(\mathbf{p})$ is a random variable which represents the PnL of strategy k under $\mathbb{P}_r$.

Kupiec [157] proposed the following statistics, which is a likelihood ratio between two Binomial likelihoods, to verify the null hypothesis. Furthermore, Kupiec [157] and Lehmann and Romano [163] proved that that $\mathrm{LR} \sim \chi^2_1$ under $\mathcal{H}_0$

$$\mathrm{LR} = -2\ln\left(\frac{(1-\alpha)^{\mathfrak{N}-C^k(\mathfrak{N})}\alpha^{C^k(\mathfrak{N})}}{\left(1-\frac{C^k(\mathfrak{N})}{\mathfrak{N}}\right)^{\mathfrak{N}-C^k(\mathfrak{N})}\left(\frac{C^k(\mathfrak{N})}{\mathfrak{N}}\right)^{C^k(\mathfrak{N})}}\right),$$

where $C^k(\mathfrak{N}) = \sum_{i=1}^{\mathfrak{N}} 1_{\{\Pi^k(\mathbf{p}_i) < \text{VaR}_\alpha(\Pi^k \# \mathbb{P}_\mathcal{M})\}}$ represents the number of violations observed in the estimates from simulator $\mathcal{M}$ and $\{\Pi^k(\mathbf{p}_i)\}_{i=1}^n$ represents the PnL observations of strategy k under $\mathbb{P}_r$.

In-sample test vs out-of-sample test. Throughout the experiments, both in-sample tests and out-of-sample tests are used to evaluate the trained simulators. In particular, the in-sample tests are performed on the training data, whereas the out-of-sample tests are performed on the testing data. For each Tail-GAN variant, the in-sample test uses the same set of strategies as in its loss function. For example, the in-sample test for Tail-GAN-Raw is performed with buy-and-hold strategies on individual assets. On the other hand, all benchmark strategies are used in the out-of-sample test for each simulator.

2.5 Numerical experiments with synthetic data

In this section, we test the performance of Tail-GAN on a synthetic data set, for which we can validate the performance of Tail-GAN by comparing to the true input price scenarios distribution. We divide the entire data set into two disjoint subsets, i.e. the training data and the testing data, with no overlap in time. The training data is used to estimate the model parameters, and the testing data is used to evaluate the out-of-sample (OOS) performance of different models. In this examination, 50,000 samples are used for training and 10,000 samples are used for performance evaluation.

The main takeaway from our comparison against benchmark simulation models is that the consistent tail-risk behavior is difficult to attain by *only* training on price sequences, without incorporating the dynamic trading strategies in the loss function, as we propose to do in our pipeline. As a consequence, if the user is indeed interested in including dynamic trading strategies in the portfolio, training a simulator on raw asset returns, as suggested by Wiese et al. [226], will be insufficient.

2.5.1 Multi-asset scenario

In this section, we simulate five financial instruments under a given correlation structure, with different temporal patterns and tail behaviors in the return distributions. The marginal distributions of these assets are: • Gaussian distribution, • AR(1) with autocorrelation $\phi_1 > 0$, • AR(1) with autocorrelation $\phi_2 < 0$, • GARCH(1, 1) with $t(\nu_1)$ noise and • GARCH(1, 1) with $t(\nu_2)$ noise. Here, $t(\nu)$ denotes the Student's t-distribution with ν degrees of freedom. The AR(1) models with positive and negative autocorrelations represent the trending scenario and mean-reversion scenario, respectively. The GARCH(1,1) models with noise from Student's t-distribution with different degrees of freedom provide us with heavy-tailed return distributions. We refer the reader to details of the simulation setup in Appendix A.2.1.

Here we examine the performance with one quantile value $\alpha = 0.05$. The architecture of the network configuration is summarized in Table A.2 in Appendix A.2.2. Experiments with other risk levels or multiple risk levels are demonstrated in Section 2.5.2.

Figure 2.3 reports the convergence of in-sample errors[2], and Table 2.1 summarizes the out-of-sample errors of TAIL-GAN-Raw, TAIL-GAN-Static, TAIL-GAN and WGAN.

Performance accuracy. We draw the following observations from Figure 2.3 and Table 2.1.

- For the evaluation criterion RE(1000) (see Figure 2.3), all three simulators, TAIL-GAN-Raw, TAIL-GAN-Static and TAIL-GAN, converge within 2000 epochs with errors smaller than 10%. This implies that all three generators are able to capture the static information contained in the market data.

[2]The in-sample error of WGAN is not reported in Figure 2.3 because WGAN uses a different training metric.

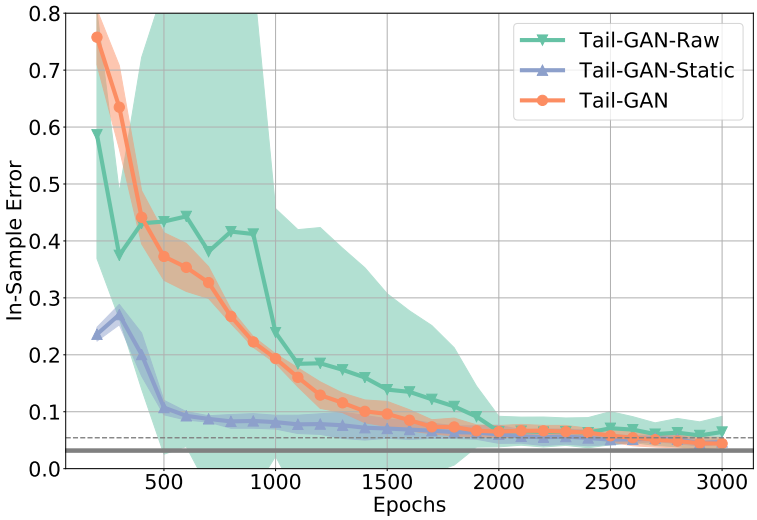


Figure 2.3: Training performance: relative error RE(1000) with 1000 samples.

Note: Grey horizontal line: average simulation error SE(1000). Dotted line: average simulation error plus one standard deviation. Each experiment is repeated five times with different random seeds. The performance is visualized with mean (solid lines) and standard deviation (shaded areas).

- For the evaluation criterion RE(1000), with both static portfolios and dynamic strategies on out-of-sample tests (see Table 2.1), only Tail-GAN converges to an error 4.6%, whereas the other two generators fail to capture the dynamic information in the market data.

- Compared to Tail-GAN-Raw and Tail-GAN-Static, Tail-GAN has the lowest training variance across multiple experiments (see standard deviations in Table 2.1). This implies that Tail-GAN has the most stable performance among all three simulators.

- Compared to Tail-GAN, WGAN yields a less competitive out-of-sample performance in terms of generating scenarios that have consistent tail risks of static portfolios and dynamic strategies. A possible explanation is that the objective function of WGAN focuses on the full distribution of raw returns, which does not guarantee the accuracy of tail risks of certain dynamic strategies.

Table 2.1: Mean and standard deviation (in parentheses) of relative errors for out-of-sample tests.

	SE(1000)	HSM	TAIL-GAN-Raw	TAIL-GAN-Static	TAIL-GAN	WGAN
OOS Error (%)	3.0	3.4	83.3	86.7	4.6	21.3
	(2.2)	(2.6)	(3.0)	(2.5)	(1.6)	(2.2)

Note: Each experiment is repeated five times with different random seeds.

Figure 2.4 shows the empirical quantile function of the strategy PnLs evaluated with price scenarios sampled from TAIL-GAN-Raw, TAIL-GAN-Static, TAIL-GAN, and WGAN. The testing strategies are, from left to right (in Figure 2.4), static single-asset portfolio (buy-and-hold strategy), single-asset mean-reversion strategy and single-asset trend-following strategy. We only demonstrate here the performance of the AR(1) model, and the results for other assets are provided in Figure A.2 in Appendix A.3.[3] In each subfigure, we compare the rank-frequency distribution of strategy PnLs evaluated with input price scenario (in blue), three TAIL-GAN simulators (in orange, red and green, respectively), and WGAN (in purple). Based on the results depicted in Figure 2.4, we conclude that

- All three variants of TAIL-GAN are able to capture the tail properties of the static single-asset portfolio at quantile levels above 1%, as shown in the first column of Figure 2.4.
- For the PnL distribution of the dynamic strategies, only TAIL-GAN is able to generate scenarios with compatible tail statistics of the PnL distribution, as shown in the second and third columns of Figure 2.4.
- TAIL-GAN-Raw and TAIL-GAN-Static underestimate the risk of the mean-reversion strategy at $\alpha = 5\%$ quantile level, and overestimate the risk of the trend-following strategy at $\alpha = 5\%$ quantile level, as illustrated in the second and third columns of Figure 2.4.

[3]We observe from Figure A.2 that for other input scenarios such Gaussian, AR(1), and GARCH(1,1), WGAN is able to generate scenarios that match the tail risk characteristics of benchmark trading strategies.

- It appears that WGAN fails to generate scenarios that retain consistent risk measures with the ground truth when the input market scenarios have heavy tails, such as AR(1) with autocorrelation 0.5, and GARCH(1,1) with $t(5)$ noise.

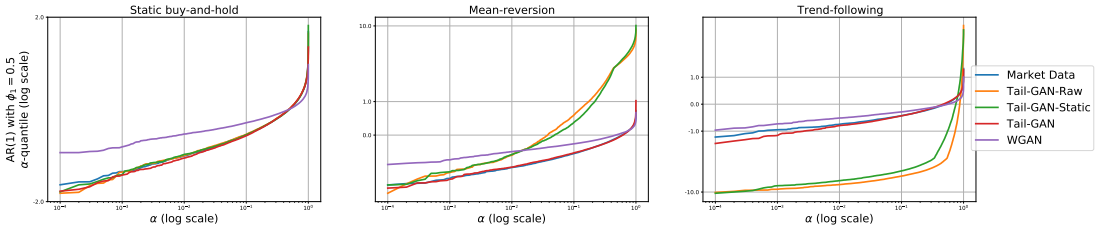
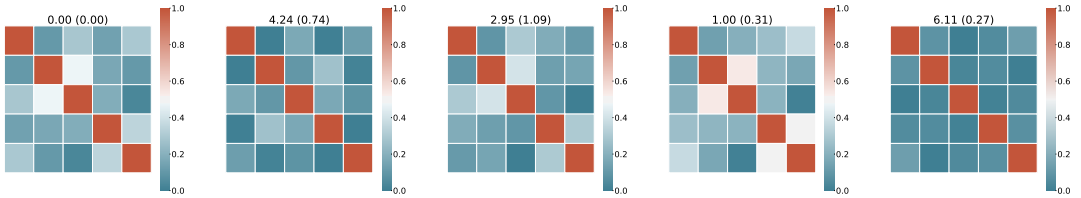


Figure 2.4: Tail behavior.

Note: This figure presents the empirical rank-frequency distribution of the strategy PnL (based on AR(1) with autocorrelation 0.5). The columns represent the strategy types.

Learning the temporal and correlation patterns. Figures 2.5 and 2.6 show the correlation and auto-correlation patterns of market data (Figures 2.5(a) and 2.6(a)) and simulated data from Tail-GAN-Raw (Figures 2.5(b) and 2.6(b)), Tail-GAN-Static (Figures 2.5(c) and 2.6(c)), Tail-GAN (Figures 2.5(d) and 2.6(d)), and WGAN (Figures 2.5(e) and 2.6(e)).

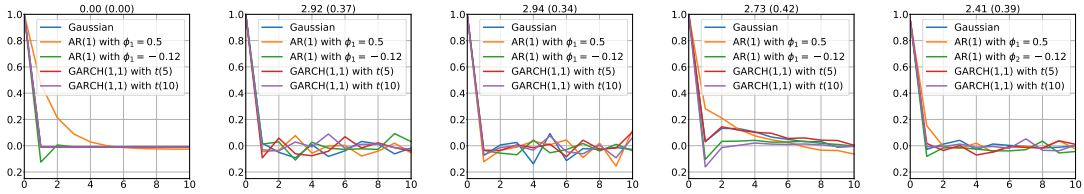


(a) Market data. (b) Tail-GAN-Raw. (c) Tail-GAN-Static. (d) Tail-GAN. (e) WGAN.

Figure 2.5: Correlations of the price increments from different trained GAN models.

Note: The numbers at the top of each plot denote the mean and standard deviation (in parentheses) of the sum of the absolute element-wise difference between the correlation matrices, computed with 10,000 training samples and 10,000 generated samples.

Figures 2.5 and 2.6 demonstrate that the auto-correlation and cross-correlations returns are best reproduced by Tail-GAN, trained on multi-asset dynamic portfolios. On the contrary, Tail-GAN-Raw and WGAN, trained on raw returns, have



(a) Market Data. (b) TAIL-GAN-Raw. (c) TAIL-GAN-Static. (d) TAIL-GAN. (e) WGAN.

Figure 2.6: Auto-correlations of the price increments from different trained GAN models.

Note: The numbers at the top of each plot denote the mean and standard deviation (in parentheses) of the sum of the absolute difference between the auto-correlation coefficients computed with 10,000 training samples and 10,000 generated samples.

the lowest accuracy in this respect. This illustrates the importance of training the algorithm on benchmark strategies instead of only raw returns.

Table 2.2: Average rejection rate of the null hypothesis in two tests.

	HSM	TAIL-GAN-Raw	TAIL-GAN-Static	TAIL-GAN	WGAN
Coverage Test (%)	17.9	53.6	22.9	17.1	44.9
Score-based Test (%)	0.00	21.3	15.4	0.00	11.4

Note: We use sample size $1,000$ and repeat the above experiment 100 times on testing data.

Statistical significance. Table 2.2 summarizes the statistical test results for Historical Simulation Method, TAIL-GAN-Raw, TAIL-GAN-Static, TAIL-GAN, and WGAN. Table 2.2 suggests that TAIL-GAN achieves the lowest average rejection rate of the null hypothesis described in Section 2.4.2. In other words, scenarios generated by TAIL-GAN have more consistent VaR and ES values for benchmark strategies compared to those of other simulators.

2.5.2 Discussion on the risk levels

In addition to the VaR and ES introduced in Section 2.2.1, the spectral risk measure (Kusuoka [158]) is another well-known risk measure commonly used in the literature. For a distribution μ, the spectral risk measure, denoted ρ^{ϕ}, is defined as

$$\rho^{\phi}(\mu) = \int_{[0,1]} \mathrm{VaR}_{\alpha}(\mu)\, \phi(\mathrm{d}\alpha), \tag{2.21}$$

Table 2.3: Mean and standard deviation (in parentheses) of relative errors for various risk levels.

OOS Error	SE(1000)	TAIL-GAN(5%)	TAIL-GAN(10%)	TAIL-GAN(1%&5%)
$\alpha = 1\%$	4.3 (3.0)	6.4 (2.7)	7.0 (3.1)	5.9 (2.6)
$\alpha = 5\%$	3.0 (2.2)	4.6 (1.6)	4.8 (1.6)	4.2 (1.8)
$\alpha = 10\%$	2.9 (2.1)	3.7 (1.5)	3.5 (1.5)	3.5 (1.7)

Note: The columns represent the models trained for certain risks. The rows represent the out-of-sample performance for different risk levels.

where the *spectrum* ϕ is a probability measure on $[0, 1]$. Theorem 5.2 in Fissler, Ziegel, et al. [100] shows that spectral risk measures under a spectral measure with finite support are jointly elicitable. We refer the reader to a class of score functions in Fissler, Ziegel, et al. [100]. This enables us to train TAIL-GAN with multiple quantile levels $\{\alpha_m\}_{m=1}^M$ simultaneously. The theoretical developments in Section 2.3 could be generalized accordingly.

To further examine the robustness of our framework and verify that TAIL-GAN is effective for not only one particular risk level, we evaluate the previously trained model TAIL-GAN(5%) at some different levels. Here TAIL-GAN(a%) represents TAIL-GAN model trained at risk level a. As shown in the second column of Table 2.3, the performance of TAIL-GAN(5%) is comparable to the baseline estimate SE(1000). We also train the model at a different level 10% (see Column TAIL-GAN(10%)). We observe that TAIL-GAN(10%) is slightly worse than TAIL-GAN(5%) in terms of generating scenarios to match the tail risks at levels 1% and 5%, but better in 10%. Furthermore, we investigate the performance of TAIL-GAN trained with multiple risk levels. In particular, the model TAIL-GAN(1%&5%) is trained with the spectral risk measure defined in (2.21). The results suggest that, in general, including multiple levels can further improve the simulation accuracy.

2.5.3 Generalization error

If the training and test errors closely follow one another, it is said that a learning algorithm has good generalization performance, see Arora et al. [17]. Generalization error quantifies the ability of machine learning models to capture certain inherent properties from the data or the ground-truth model. In general, machine learning models with a good generalization performance are meant to learn "underlying rules" associated with the data generation process, rather than only memorizing the training data, so that they are able to extrapolate learned rules from the training data to new unseen data. Thereby, the generalization error of a generator G can be measured as the difference between the empirical divergence of the training data $d(\mathbb{P}_r^{(n)}, \mathbb{P}_G^{(n)})$ and the ground-truth divergence $d(\mathbb{P}_r, \mathbb{P}_G)$.

To provide a systematic quantification of the generalization capability, we adopt the notion of generalization proposed in Arora et al. [17]. For a fixed sample size n, the generalization error of $\mathbb{P}_G$ under divergence $d(\cdot, \cdot)$ is defined as

$$\left| d(\mathbb{P}_r^{(n)}, \mathbb{P}_G^{(n)}) - d(\mathbb{P}_r, \mathbb{P}_G) \right|, \tag{2.22}$$

where $\mathbb{P}_r^{(n)}$ is the empirical distribution of $\mathbb{P}_r$ with n samples, i.e., the distribution of the training data, and $\mathbb{P}_G^{(n)}$ is the empirical distribution of $\mathbb{P}_G$ with n samples drawn after the generator G is trained. A small generalization error under definition (2.22) implies that GANs with good generalization property should have consistent performances with the empirical distributions (i.e., $\mathbb{P}_r^{(n)}$ and $\mathbb{P}_G^{(n)}$) and with the true distributions (i.e., $\mathbb{P}_r$ and $\mathbb{P}_G$). We consider two choices for the divergence function: d_q based on quantile divergence, and d_s based on the score function we use. See the mathematical definition in Appendix A.2.4.

To illustrate the generalization capabilities of TAIL-GAN, we compare it with a supervised learning benchmark using the same loss function. Given the optimization problem (2.16)-(2.17), one natural idea is to construct a simulator (or a generator) using empirical VaR and ES values in the evaluation. To this end, we consider the following optimization

$$\min_{G \in \mathcal{G}} \frac{1}{Kn} \sum_{k=1}^{K} \sum_{j=1}^{n} S_\alpha\Big(\Big(\mathrm{VaR}_\alpha\big(\Pi^k \# \mathbb{P}_G^{(n)}\big), \mathrm{ES}_\alpha\big(\Pi^k \# \mathbb{P}_G^{(n)}\big) \Big),\ \Pi^k(\mathbf{p}_j) \Big), \tag{2.23}$$

where $\mathbb{P}_G^{(n)}$ is the empirical measure of n samples drawn from $\mathbb{P}_G$, and $\mathbf{p}_j$ are samples under the measure $\mathbb{P}_r$ $(j = 1, 2, \ldots, n)$. The optimization problem (2.23) falls into the category of training simulators with supervised learning (Ostrovski, Dabney, and Munos [186]).

Compared with Tail-GAN, the presented supervised learning framework has several disadvantages, which we illustrate in a set of empirical studies. The first issue is the bottleneck in statistical accuracy. When using $\mathbb{P}_r^{(n)}$ as the guidance for supervised learning, as indicated in (2.23), it is not possible for the α-VaR and α-ES values of the simulated price scenarios $\mathbb{P}_G$ to improve on the sampling error of the empirical α-VaR and α-ES values estimated with the n samples. In particular, ES is very sensitive to tail events, and the empirical estimate of ES may not be stable even with 10,000 samples. The second issue concerns the limited ability in generalization. A generator constructed via supervised learning tends to mimic the exact patterns in the input financial scenarios $\mathbb{P}_r^{(n)}$, instead of generating new scenarios that are equally realistic compared to the input financial scenarios under the evaluation of the score function.

We proceed to compare the performance of Tail-GAN with that of a Generator-Only Model (GOM) according to (2.23), from the point of view of both statistical accuracy and generalization ability. The detailed setup is provided in Appendix A.2.4.

Performance accuracy of Tail-GAN vs GOM. Figure 2.7 reports the convergence of in-sample errors, and Table 2.4 summarizes the out-of-sample errors of GOM and Tail-GAN. From Table 2.4, we observe that the relative error of Tail-GAN is 4.6%, which is a 30% reduction compared to the relative error of GOM of around 7.2%. Compared to (2.23), the advantage of using neural networks to learn the VaR and ES values, as designed in Tail-GAN, is that it memorizes information in previous iterations during the training procedure, and therefore the statistical bottleneck with n samples can be overcome when the number of iterations increases. Therefore, we conclude that Tail-GAN outperforms GOM in terms of simulation accuracy, demonstrating the importance of the discriminator.

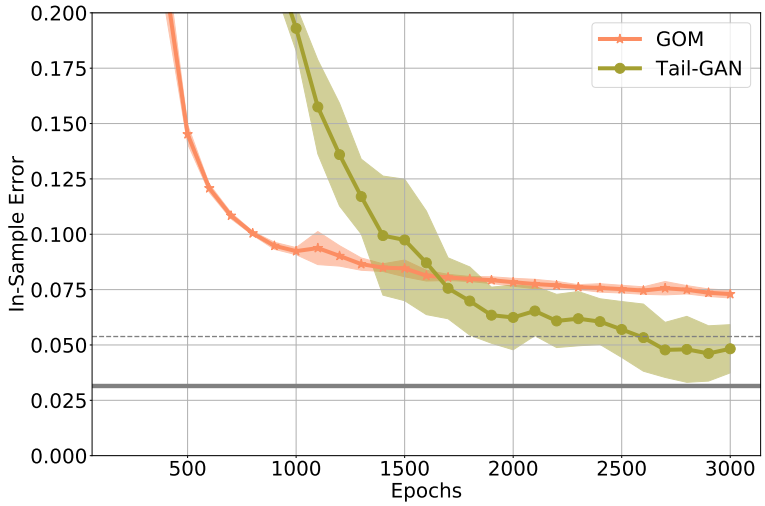


Figure 2.7: Training performance of GOM and TAIL-GAN.

Note: Grey horizontal line: average simulation error SE(1000). Dotted line: average simulation error plus one standard deviation. Each experiment is repeated five times with different random seeds. The performance is visualized with mean (solid lines) and standard deviation (shaded areas).

Table 2.4: Mean and standard deviation (in parentheses) of relative errors for out-of-sample tests.

	TAIL-GAN	GOM
OOS Error (%)	4.6	7.2
	(1.6)	(0.2)

Note: Each experiment is repeated five times with different random seeds.

Table 2.5 provides the generalization errors, under both d_q and d_s, for TAIL-GAN and GOM. We observe that under both criteria, the generalization error of GOM is twice that of TAIL-GAN, implying that TAIL-GAN has better generalization power, in addition to higher performance accuracy.

2.5.4 Scalability

In practice, most portfolios held by asset managers are constructed with 20-30 or more financial assets. In order to scale TAIL-GAN with a comparable number of assets, we use PCA-based eigenvectors. The resulting *eigenportfolios* (Avellaneda and Lee [18]) are uncorrelated and able to explain the most variation in the cross-section of returns with the smallest number of portfolios. This idea enables to train

Table 2.5: Mean (in percentage) and standard deviation (in parentheses) of generalization errors under both divergence functions.

Error metric	TAIL-GAN	GOM
d_q	0.214 (0.178)	0.581 (0.420)
d_s	0.017 (0.014)	0.032 (0.026)

Note: The mathematical formulations for divergences are deferred to (A.21) and (A.22)). Results are averaged over 10 repeated experiments (synthetic data sets).

TAIL-GAN with the minimum number of portfolios, hence rendering TAIL-GAN scalable to generate price scenarios with a large number of heterogeneous assets.

In this section, we train TAIL-GAN with eigenportfolios of 20 assets, and compare its performance with TAIL-GAN trained on 50 randomly generated portfolios. TAIL-GAN with the eigenportfolios shows dominating performance, which is also comparable to simulation error (with the same number of samples). The detailed steps of the eigenportfolio construction are deferred to Appendix A.2.5.

Data. To showcase the scalability of TAIL-GAN, we simulate the price scenarios of 20 financial assets for a given correlation matrix ρ, with different temporal patterns and tail behaviors in return distributions. Among these 20 financial assets, five of them follow Gaussian distributions, another five follow AR(1) models, another five of them follow GARCH(1, 1) with light-tailed noise, and the rest follow GARCH(1, 1) with heavy-tailed noise. Other settings are the same as in Section 2.5.1.

Results. Figure 2.8 shows the percentage of explained variance of the principal components. We observe that the first principal component accounts for more than 23% of the total variation across the 20 asset returns.

To identify and demonstrate the advantages of the eigenportfolios, we compare the following two TAIL-GAN architectures

(1) TAIL-GAN(Rand): GAN trained with 50 multi-asset portfolios and dynamic strategies,

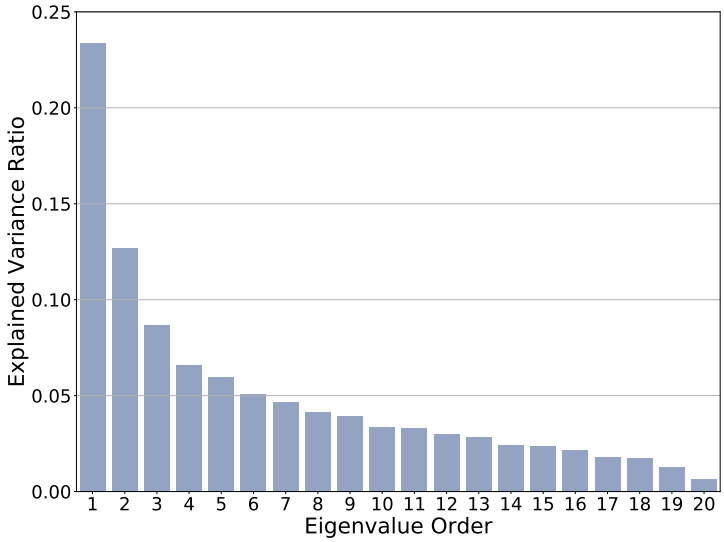


Figure 2.8: Explained variance ratios of eigenvalues.

(2) TAIL-GAN(Eig): GAN trained with 20 multi-asset eigenportfolios and dynamic strategies.

The weights of static portfolios in TAIL-GAN(Rand) are randomly generated such that the absolute values of the weights sum up to one. The out-of-sample test consists of $K = 90$ strategies, including 50 convex combinations of eigenportfolios (with weights randomly generated), 20 mean-reversion strategies, and 20 trend-following strategies.

Performance accuracy. Figure 2.9 reports the convergence of in-sample errors. Table 2.6 summarizes the out-of-sample errors and shows that TAIL-GAN(Eig) achieves better performance than TAIL-GAN(Rand) with fewer number of portfolios.

Table 2.6: Mean and standard deviation (in parentheses) for relative errors in out-of-sample tests.

	HSM	TAIL-GAN(Rand)	TAIL-GAN(Eig)
OOS Error (%)	3.5 (2.3)	10.4 (1.8)	6.9 (1.5)

Note: Each experiment is repeated five times with different random seeds.

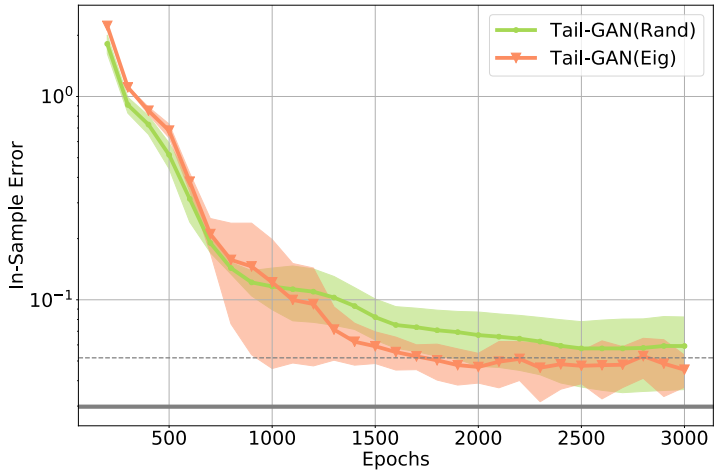


Figure 2.9: Training performance on 50 random portfolios vs 20 eigenportfolios.

Note: Grey horizontal line: average simulation error. Dotted line: average simulation error plus one standard deviation. Each experiment is repeated five times with different random seeds. The performance is reported with mean (solid lines) and standard deviation (shaded areas).

2.6 Application to simulation of intraday market scenarios

We use the Nasdaq ITCH data from LOBSTER[4] during the intraday time interval 10:00AM-3:30PM, for the period 2019-11-01 until 2019-12-06. The reason for excluding the first and last 30 minutes of the trading day stems from the increased volatility and volume inherent in the market following the opening session, and preceding the closing session. The Tail-GAN simulator is trained on the following five stocks: AAPL, AMZN, GOOG, JPM, QQQ.

The mid-prices (average of the best bid and ask prices) of these assets are sampled at a $\Delta = 9$-second frequency, with $T = 100$ for each price series representing a financial scenario during a 15-minute interval. We sample the 15-minute paths every one minute, leading to an overlap of 14 minutes between two adjacent paths[5]. The architecture and configurations are the same as those reported in Table A.2 in Appendix A.2.2, except that the training period here is from 2019-

[4]https://lobsterdata.com/

[5]Tail-GAN are also trained on market data with no time overlap and the conclusions are similar.

	Static buy-and-hold		Mean-reversion		Trend-following	
	VaR	ES	VaR	ES	VaR	ES
AAPL	-0.351	-0.548	-0.295	-0.479	-0.316	-0.485
AMZN	-0.460	-0.720	-0.398	-0.639	-0.399	-0.628
GOOG	-0.316	-0.481	-0.272	-0.426	-0.273	-0.419
JPM	-0.331	-0.480	-0.275	-0.419	-0.290	-0.427
QQQ	-0.254	-0.384	-0.202	-0.321	-0.210	-0.328

Table 2.7: Empirical VaR and ES values for trading strategies evaluated on the training data.

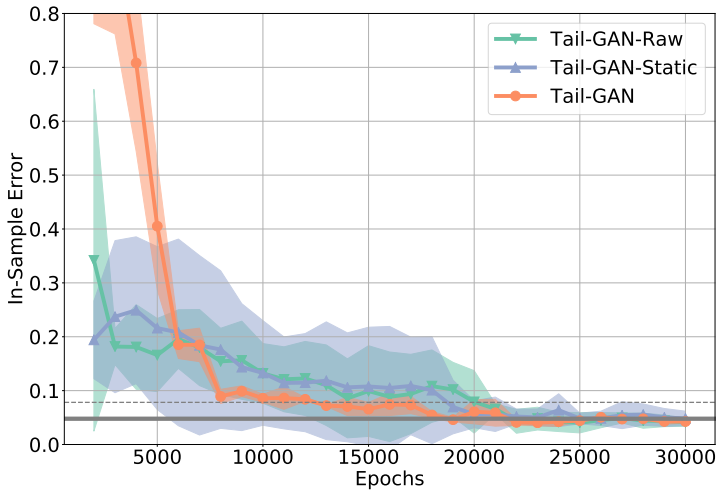


Figure 2.10: Training performance: relative error RE(1000) with 1000 samples.

Note: Grey horizontal line: average simulation error SE(1000). Dotted line: average simulation error plus one standard deviation. Each experiment is repeated 5 times with different random seeds. The performance is visualized with mean (solid lines) and standard deviation (shaded areas).

11-01 to 2019-11-30, and the testing period is the first week of 2019-12. Thus, the size of the training data is $N = 6300$. Table 2.7 reports the 5%-VaR and 5%-ES values of several strategies calculated with the market data of AAPL, AMZN, GOOG, JPM, and QQQ.

Performance accuracy. Figure 2.10 reports the convergence of in-sample errors and Table 2.8 summarizes the out-of-sample errors.

We draw the following conclusions from the results of Figure 2.10 and Table 2.8.

- For the evaluation criterion RE(1000) based on in-sample data (see Figure 2.10), all three GAN simulators, TAIL-GAN-Raw, TAIL-GAN-Static and

Table 2.8: Mean and standard deviation (in parentheses) for relative errors in out-of-sample tests.

	"Oracle"	HSM	Tail-GAN-Raw	Tail-GAN-Static	Tail-GAN	WGAN
OOS Error (%)	2.4 (1.6)	10.4 (3.6)	112.8 (7.8)	75.8 (8.0)	10.1 (1.1)	26.9 (1.7)

Note: "Oracle" represents the sampling error of the testing data. Each experiment is repeated five times with different random seeds.

Tail-GAN, converge within 20,000 epochs and reach in-sample errors smaller than 5%.

- For the evaluation criterion RE(1000), with both static portfolios and dynamic strategies based on out-of-sample data (Table 2.8), only Tail-GAN converges to an error of 10.1%, whereas the other two Tail-GAN variants fail to capture the temporal information in the input price scenarios.

- The HSM method comes close with an error of 10.4% and WGAN reaches an error of 26.9%. As expected, all methods attain higher errors than the sampling error of the testing data (denoted by "oracle" in Table 2.8).

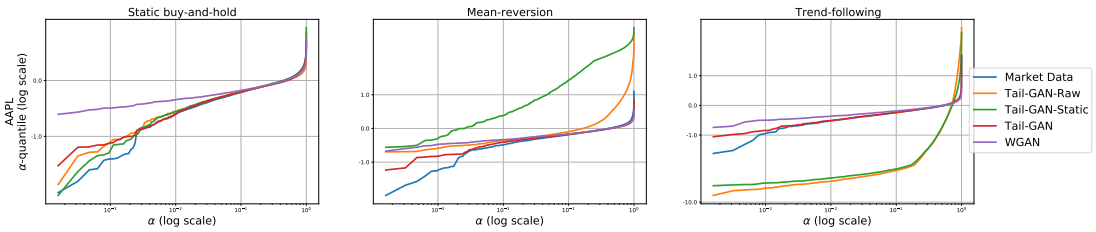


Figure 2.11: Tail behavior.

Note: This figure presents the empirical rank-frequency distribution of the strategy PnL (based on AAPL). The columns represent the strategy types.

To study the tail behavior of the intraday scenarios, we implement the same rank-frequency analysis as in Section 2.5.1. For the AAPL stock, we draw the following conclusions from Figure 2.11

- All three Tail-GAN simulators are able to capture the tail properties of static single-asset portfolio for quantile levels above 1%.

- For the PnL distribution of the dynamic strategies, only TAIL-GAN is able to generate scenarios with comparable (tail) PnL distribution. That is, only scenarios sampled from TAIL-GAN can correctly describe the risks of the trend-following and the mean-reversion strategies.
- TAIL-GAN-Raw and TAIL-GAN-Static underestimate the risk of loss from the mean-reversion strategy at the $\alpha = 5\%$ quantile level, and overestimate the risk of loss from the trend-following strategy at the $\alpha = 5\%$ quantile level.
- While WGAN can effectively generate scenarios that align with the bulk of PnL distributions (e.g. above 10%-quantile), it tends to fail in accurately capturing the tail parts, usually resulting in underestimation of risks).

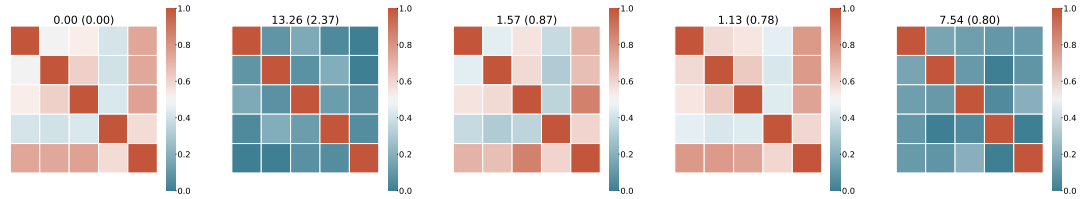
Note that some of the blue curves corresponding to the market data (almost) coincide with the red curves corresponding to TAIL-GAN, indicating a promising performance of TAIL-GAN to capture the tail risk of various trading strategies. See Figure A.3 in Appendix A.3 for the results for other assets.

Learning temporal and cross-correlation patterns. Figures 2.12 and 2.13 present the in-sample correlation and auto-correlation patterns of the market data (Figures 2.12(a) and 2.13(a)), and simulated data from TAIL-GAN-Raw (Figures 2.12(b) and 2.13(b)), TAIL-GAN-Static (Figures 2.12(c) and 2.13(c)), TAIL-GAN (Figures 2.12(d) and 2.13(d)) and WGAN (Figures 2.12(e) and 2.13(e)).

As shown in Figures 2.12 and 2.13, TAIL-GAN trained on dynamic strategies learns the information on cross-asset correlations more accurately than TAIL-GAN-Raw and WGAN, which are trained on raw returns.

Scalability. To test the scalability property of TAIL-GAN on realistic scenarios, we conduct a similar experiment as in Section 2.5.4. The stocks considered here include the top 20 stocks in the S&P500 index. The training period is between 2019-11-01 and 2019-11-30.

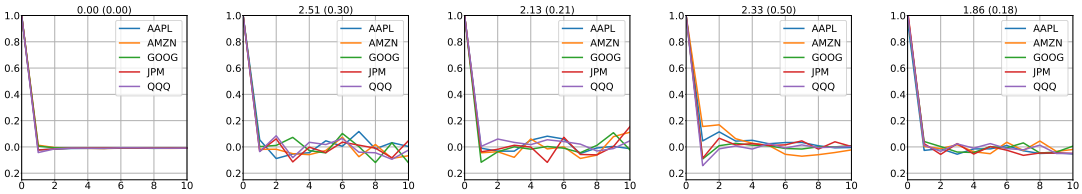
Figure 2.14 reports the convergence of in-sample errors, and Table 2.9 summarizes the out-of-sample errors of TAIL-GAN(Rand) and TAIL-GAN(Eig). Table 2.9



(a) Market data. (b) TAIL-GAN-Raw. (c) TAIL-GAN-Static. (d) TAIL-GAN. (e) WGAN.

Figure 2.12: Cross-asset correlations of the price increments in the market data and from different trained GAN models.

Note: Numbers on the top: mean and standard deviation (in parentheses) of the sum of the absolute difference between the correlation coefficients computed with all training samples and 1,000 generated samples.



(a) Market data. (b) TAIL-GAN-Raw. (c) TAIL-GAN-Static. (d) TAIL-GAN. (e) WGAN.

Figure 2.13: Auto-correlations of the price increments from different trained GAN models

Note: Numbers on the top: mean and standard deviation (in parentheses) of the sum of the absolute element-wise difference between auto-correlation coefficients computed with all training samples and 1,000 generated samples.

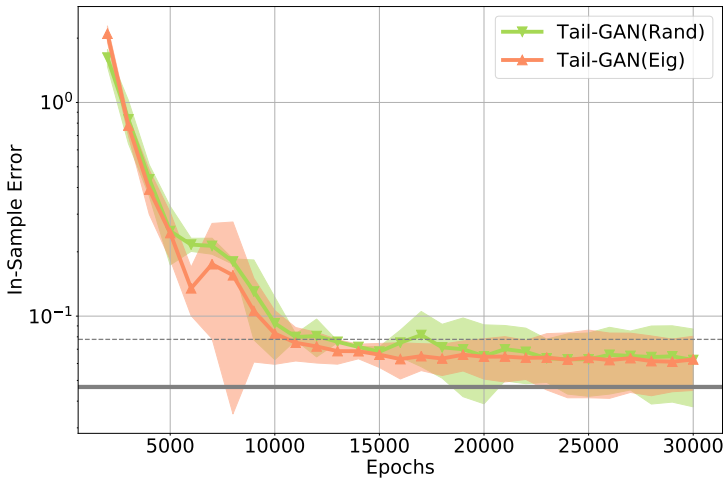


Figure 2.14: Training performance on 50 random portfolios vs 20 eigenportfolios

Note: Grey horizontal line: average simulation error. Dotted line: average simulation error plus one standard deviation. Each experiment is repeated 5 times with different random seeds.

shows that TAIL-GAN(Eig) achieves better performance than TAIL-GAN(Rand) with fewer training portfolios.

Table 2.9: Mean and standard deviation (in parentheses) for relative errors in out-of-sample tests.

	"Oracle"	HSM	TAIL-GAN(Rand)	TAIL-GAN(Eig)
OOS Error (%)	2.2	25.9	31.0	25.6
	(1.7)	(5.1)	(1.0)	(1.0)

Note: "Oracle" represents the sampling error of the testing data. Each experiment is repeated five times with different random seeds.

3

Cross Impact of Order Flow Imbalance in Equity Markets

Contents

3.1 Introduction

Accurately estimating and forecasting the impact of trading behavior of market participants on the price movements of assets carries practical implications for both practitioners and academics, such as trading cost analysis (Frazzini, Israel, and Moskowitz [105]) and optimal execution models (Cartea and Jaimungal [57] and Guo and Zervos [120]). It has been studied in the previous literature that the trade orders of an asset move its own price, also known as price-impact (Bouchaud [40], Cont, Kukanov, and Stoikov [74], and Lillo, Farmer, and Mantegna [169]). However, the impact of trading a given asset on the price of *other* assets, denoted as **cross-impact**, has been much less studied (see Benzaquen et al. [30], Capponi and Cont [53], and Pasquariello and Vega [188]). This phenomenon implies that looking solely at price-impact, without including cross-impact, amounts to disregarding an important component of trading costs when evaluating the performance of a portfolio-level trading strategy.

There are studies that investigate the *contemporaneous cross-impact* of returns and order flows by examining their cross-correlation structure. For example, Hasbrouck and Seppi [131] revealed that commonality in returns among Dow 30 stocks is mostly attributed to order flow commonality. Capponi and Cont [53] showed that the positive covariance between returns of a specific stock and order flow imbalances of other stocks does not necessarily constitute evidence of cross-impact. They further demonstrated that, as long as the common factor in order flow imbalances is taken into account, adding cross-impact terms only marginally improves model performance, and thus may be disregarded. Tomas, Mastromatteo, and Benzaquen [215] built a principled approach to choosing a cross-impact model for various markets. To the best of our knowledge, there have been no studies that examine the influence of order flows on price movements in a multi-asset setting, while also taking into account the deeper levels in the limit order book.[1]

[1] Xu, Gould, and Howison [231] studied the contemporaneous price-impact (not cross-impact) model by extending the model of Cont, Kukanov, and Stoikov [74] to multi-level order flow imbalance.

A more challenging problem than explaining contemporaneous returns is to examine the impact of trade orders on prices over future horizons, which has received a lot less attention in the literature, despite its important economic implications. Some studies have examined the relationship between order imbalances and *future daily* returns.[2] Chordia, Roll, and Subrahmanyam [67] revealed that daily stock market returns are strongly related to contemporaneous and lagged order imbalances. Chordia and Subrahmanyam [68] further found that there exists a positive relation between lagged order imbalances and daily individual stock returns. The authors also showed that imbalance-based trading strategies, i.e. buy if the previous day's imbalance is positive, and sell if the previous day's imbalance is negative, are able to yield statistically significant profits.[3] Pasquariello and Vega [188] provided empirical evidence of cross-asset informational effects in NYSE and NASDAQ stocks between 1993 and 2004, and demonstrated that the daily order flow imbalance in one stock, or across one industry, has a significant and persistent impact on daily returns of other stocks or industries. Rosenbaum and Tomas [198] provided a characterization of the class of cross-impact kernels for a market that employs Hawkes processes to model trades and applied their method to two instruments from E-Mini Futures.

Given the recent progress in high-frequency trading (HFT), it is increasingly crucial to obtain accurate estimations of the cross-impact on *future intraday* returns (Cartea, Gan, and Jaimungal [56]). Benzaquen et al. [30] introduced a multivariate linear propagator model (see Kyle [159]) to describe the structure of cross-impact and found that a significant fraction of the covariance of stock returns can be accounted for by this model. Wang, Neusüß, and Guhr [220] and Wang, Schäfer, and Guhr [221] empirically analyzed and discussed the impact of trading a specific stock on the average price change of the whole market or of individual sectors. Schneider

[2]Several studies, such as Buccheri, Corsi, and Peluso [42], Chinco, Clark-Joseph, and Ye [63], Hou [140], and Menzly and Ozbas [176], investigated the lead-lag effect in equity returns across various assets, but did not take into account order flows.

[3]Nonetheless, according to Madhavan, Richardson, and Roomans [173], short-term predictability in returns could stem from autocorrelation in order flows, limit orders, asymmetric information, and other microstructure effects. Meanwhile, Lillo and Farmer [168] and Doyne Farmer et al. [89] maintained that there is no possibility of arbitrage, possibly because of the fluctuating nature of asymmetric liquidity.

and Lillo [200] derived theoretical limits for the size and form of cross-impact and verified them on sovereign bonds data. However, when modeling cross-impact, these methods do not consider the possibility of high correlations between cross-asset order flows, which may result in overfitting issues. This is also evidenced by studies such as Benzaquen et al. [30] and Tomas, Mastromatteo, and Benzaquen [215]. Moreover, these studies mainly investigated the cross-impact coefficients for a fixed time period (i.e., in a static setting), ignoring the temporal dynamics of cross-impact.

In recent years, machine learning models including deep neural networks, have achieved substantial developments, leading to their applications in financial markets, especially for the task of modeling stock returns. For example, Huck [142] utilized state-of-the-art techniques, such as random forests, to construct a portfolio over a period of 22 years, and the results demonstrated the power of machine learning models to produce profitable trading signals. Krauss, Do, and Huck [156] applied a series of machine learning methods to forecast the probability of a stock outperforming the market index, and then constructed long-short portfolios from the predicted one-day-ahead trading signals. Gu, Kelly, and Xiu [119] employed a set of machine learning methods to make one-month-ahead return forecasts, and demonstrated the potential of machine learning approaches in empirical asset pricing, due to their ability to handle nonlinear interactions. Aït-Sahalia et al. [5] investigated the predictability of high-frequency stock returns and durations using LASSO and tree methods via many relevant predictors derived from returns and order flows. Tashiro et al. [212] and Kolm, Turiel, and Westray [154] applied deep neural networks with LOB-based features to predict high-frequency returns. Nonetheless, to the best of our knowledge, cross-asset order flow imbalances have not been considered as predictors for forecasting future high-frequency returns in the literature, which is one of the main directions we explore in the second half of this chapter.

3.1.1 Main contributions

The present study makes two main contributions to the literature regarding the *contemporaneous* and *predictive* cross-impact of order flow imbalances on price returns.

First, we revisit the significance of contemporaneous cross-impact by comparing it with the price-impact model. Instead of only looking at the best-level orders, we systematically examine the impact of **multi-level** order flows in a **cross-asset** setting. We find that the cross-impact model including the best-level orders of multiple assets as candidate features can provide small but significant additional explanatory power for price movements, compared to the price-impact model with only the best-level order information. Moreover, our results show that, once the information from multi-level orders is incorporated, cross-impact models do not provide additional explanatory power for contemporaneous impact, compared to a parsimonious model without the cross-impact terms. To the best of our knowledge, this is the first study to comprehensively analyze the relations between contemporaneous individual returns and multi-level orders in both single-asset and multi-asset settings.

In addition, we consider the challenging setting of *future cross-impact*, where we investigate the predictive power of the cross-asset order flows on future price returns. Specifically, we study the multi-period cross-impact model using lagged order flows to predict future price returns. Overall, our results suggest that cross-impact terms do provide significant information content for intraday forecasting of future returns over a short horizon of up to several minutes, but their predictability decays quickly through time.

The remainder of this chapter is structured as follows. In Section 3.2, we introduce the data and construct the variables employed throughout the chapter. Section 3.3 presents studies about the cross-impact model with multi-level order flows against contemporaneous returns. In Section 3.4, we first discuss the out-of-sample forecasting performance of price-impact and cross-impact models over 1-minute ahead horizon from two perspectives: R^2 values and economic gains, and then examine the predictability over longer horizons. Finally, we conclude our analysis in Section 3.5 and highlight potential future research directions.

3.2 Data and variables

3.2.1 Data

We use the Nasdaq ITCH data from LOBSTER to compute the independent and dependent variables. Our data includes the top 100 components of S&P 500 index, existing from 2017-01-01 to 2019-12-31.

3.2.2 Independent variables

Cont, Kukanov, and Stoikov [74] found that over short time intervals, price changes are mainly driven by the Order Flow Imbalance (henceforth denoted as OFI). Kolm, Turiel, and Westray [154] also demonstrated that forecasting deep learning models trained on OFIs significantly outperform most models trained directly on order books or returns. Therefore, we adopt the OFIs as features in our below analysis.

During the interval $(t-h, t]$, we enumerate the observations of all order book updates by n. Given two consecutive order book states for a given stock i at $n-1$ and n, we compute the bid order flows ($\mathrm{OF}_{i,n}^{m,b}$) and ask order flows ($\mathrm{OF}_{i,n}^{m,a}$) of stock i at level m at time n as

$$\mathrm{OF}_{i,n}^{m,b} := \begin{cases} q_{i,n}^{m,b}, & \text{if } P_{i,n}^{m,b} > P_{i,n-1}^{m,b}, \\ q_{i,n}^{m,b} - q_{i,n-1}^{m,b}, & \text{if } P_{i,n}^{m,b} = P_{i,n-1}^{m,b}, \\ -q_{i,n}^{m,b}, & \text{if } P_{i,n}^{m,b} < P_{i,n-1}^{m,b}, \end{cases}$$

$$\mathrm{OF}_{i,n}^{m,a} := \begin{cases} -q_{i,n}^{m,a}, & \text{if } P_{i,n}^{m,a} > P_{i,n-1}^{m,a}, \\ q_{i,n}^{m,a} - q_{i,n-1}^{m,a}, & \text{if } P_{i,n}^{m,a} = P_{i,n-1}^{m,a} \\ q_{i,n}^{m,a}, & \text{if } P_{i,n}^{m,a} < P_{i,n-1}^{m,a}, \end{cases}$$

where, $P_{i,n}^{m,b}$ and $q_{i,n}^{m,b}$ denote the bid price and size (in number of shares) of stock i at level m, respectively. Similarly, $P_{i,n}^{m,a}$ and $q_{i,n}^{m,a}$ denote the ask price and ask size at level m, respectively. Note that the variable $\mathrm{OF}_{i,t}^{m,b}$ is positive when (i) the bid price increase; (ii) the bid price remains the same and the bid size increases. $\mathrm{OF}_{i,t}^{m,b}$ is negative when (i) the bid price decreases; (ii) the bid price remains the same and the bid size decreases. One can perform an analogous analysis and interpretation for the ask order flows $\mathrm{OF}_{i,t}^{m,a}$.

Best-level OFI. It calculates the accumulative OFIs at the best bid/ask side during a given time interval (see Cont, Kukanov, and Stoikov [74] and Kolm, Turiel, and Westray [154]), and is defined as[4]

$$\mathrm{OFI}_{i,t}^{1,(h)} \coloneqq \sum_{n=N(t-h)+1}^{N(t)} \mathrm{OF}_{i,n}^{1,b} - \mathrm{OF}_{i,n}^{1,a}, \tag{3.1}$$

where $N(t-h)+1$ and $N(t)$ are the indexes of the first and the last order book event in the interval $(t-h, t]$.

Deeper-level OFI. A natural extension of the best-level OFI defined in Eqn (3.1) is deeper-level OFI (see Kolm, Turiel, and Westray [154] and Xu, Gould, and Howison [231]). We define OFI at level m ($m \geq 1$) as follows

$$\mathrm{OFI}_{i,t}^{m,(h)} \coloneqq \sum_{n=N(t-h)+1}^{N(t)} \mathrm{OF}_{i,n}^{m,b} - \mathrm{OF}_{i,n}^{m,a}, \tag{3.2}$$

Due to the intraday pattern in limit order depth, we use the average size to scale OFIs at the corresponding levels (consistent with Ahn, Bae, and Chan [4] and Harris and Panchapagesan [129]), and consider

$$\mathrm{ofi}_{i,t}^{m,(h)} = \frac{\mathrm{OFI}_{i,t}^{m,(h)}}{Q_{i,t}^{M,(h)}}, \tag{3.3}$$

where $Q_{i,t}^{M,(h)} = \frac{1}{M}\sum_{m=1}^{M} \frac{1}{2\Delta N(t)} \sum_{n=N(t-h)+1}^{N(t)} \left[q_{i,n}^{m,b} + q_{i,n}^{m,a}\right]$ is the average order book depth across the first M levels and $\Delta N(t) = N(t) - N(t-h)$ is the number of events during $(t-h, t]$. In this chapter, we consider the top $M = 10$ levels of LOB and denote the multi-level OFI vector as $\mathbf{ofi}_{i,t}^{(h)} = \left(\mathrm{ofi}_{i,t}^{1,(h)}, \cdots, \mathrm{ofi}_{i,t}^{10,(h)}\right)^T$.

[4] In Cont, Kukanov, and Stoikov [74], OFI was mathematically defined as $\mathrm{OFI}_{i,t}^{1,(h)} = L_{i,h}^{1,b} - C_{i,h}^{1,b} - M_{i,h}^{1,b} - L_{i,h}^{1,s} + C_{i,h}^{1,s} - M_{i,h}^{1,s}$, where $L_{i,h}^{1,b}$ denotes the total size of buy orders that arrived to the current best bid during the time interval $(t-h, t]$; $C_{i,h}^{1,b}$ denotes the total size of buy orders that canceled from the current best bid during the time interval $(t-h, t]$; $M_{i,h}^{1,b}$ denotes the total size of marketable buy orders that arrived to current best ask during the time interval $(t-h, t]$. The quantities $L_{i,h}^{1,s}, C_{i,h}^{1,s}, M_{i,h}^{1,s}$ for sell orders are defined analogously. However, in the empirical study of Cont, Kukanov, and Stoikov [74], the OFI was computed from fluctuations in best bid/ask prices and their sizes according to Eqn (3.1). The reason is that information about individual orders is not available in the data set. For better comparison, we employ the same formula, i.e. Eqn (3.1), to compute OFI.

Integrated OFI. Our following analysis in Section 3.2.4 will show that there exist strong correlations between multi-level OFIs, and that the first principal component can explain over 89% of the total variance among multi-level OFIs. In order to make use of the information embedded in multiple LOB levels and avoid overfitting, we propose an integrated version of OFIs via Principal Components Analysis (PCA) as shown in Eqn (3.4), which only preserves the first principal component.[5] We further normalize the first principal component by dividing by its l_1 norm so that the weights of multi-level OFIs in constructing integrated OFIs sum to 1, leading to

$$\text{ofi}_{i,t}^{I,(h)} = \frac{\boldsymbol{w}_1^T \mathbf{ofi}_{i,t}^{(h)}}{\|\boldsymbol{w}_1\|_1}, \tag{3.4}$$

where $\boldsymbol{w}_1$ is the first principal vector computed from historical data. To the best of our knowledge, this is the first work to *aggregate multi-level OFIs into a single variable.*

3.2.3 Dependent variables

In the present work, we are interested in the cross-impact of OFIs on the price returns over multiple horizons.

Logarithmic return. Our dependent variable is the logarithmic asset return. Specifically, we define the returns over the interval $(t-h, t]$ as follows:

$$r_{i,t}^{(h)} = \log\left(\frac{P_{i,t}}{P_{i,t-h}}\right), \tag{3.5}$$

where $P_{i,t}$ is the mid-price at time t, i.e. $P_{i,t} = \frac{P_{i,t}^{1,b} + P_{i,t}^{1,a}}{2}$.

3.2.4 Summary statistics

Table 3.1 presents the summary statistics of multi-level OFIs, integrated OFIs, and returns for the top 100 components of S&P 500 index. These descriptive

[5] A future research direction is to devise various weighting schemes that average the OFI information across the multiple levels, where the weights could be given, for example, by an inverse function of the distance of each price level to the mid price or applying tensor-SVD/PCA on this data.

Table 3.1: Summary statistics of OFIs and returns.

	Mean (bp)	Std (bp)	Skewness	Kurtosis	10% (bp)	25% (bp)	50% (bp)	75% (bp)	90% (bp)
$\text{ofi}^{1,(1m)}$	-0.01	6.26	-0.04	1.89	-7.97	-3.45	0.03	3.47	7.90
$\text{ofi}^{2,(1m)}$	0.01	6.86	-0.04	1.04	-8.86	-3.88	0.02	3.95	8.85
$\text{ofi}^{3,(1m)}$	-0.01	7.05	-0.04	0.71	-9.26	-4.08	0.01	4.11	9.19
$\text{ofi}^{4,(1m)}$	-0.02	7.22	-0.05	0.68	-9.50	-4.21	0.01	4.24	9.40
$\text{ofi}^{5,(1m)}$	-0.03	7.14	-0.05	0.79	-9.38	-4.14	0.01	4.15	9.25
$\text{ofi}^{6,(1m)}$	-0.03	6.87	-0.04	0.96	-8.98	-3.94	0.01	3.95	8.85
$\text{ofi}^{7,(1m)}$	-0.03	6.39	-0.05	1.29	-8.31	-3.59	0.01	3.59	8.16
$\text{ofi}^{8,(1m)}$	-0.03	6.03	-0.05	1.59	-7.80	-3.37	0.01	3.36	7.66
$\text{ofi}^{9,(1m)}$	-0.05	5.71	-0.05	1.96	-7.38	-3.18	0.01	3.14	7.19
$\text{ofi}^{10,(1m)}$	-0.05	5.38	-0.05	2.52	-6.92	-2.97	0.01	2.91	6.74
$\text{ofi}^{I,(1m)}$	0.01	6.53	-0.05	0.76	-8.52	-3.81	0.05	3.89	8.47
$r^{(1m)}$	0.02	4.81	-0.04	1.85	-6.22	-2.71	0.00	2.79	6.23

Note: These statistics are computed at the minute level across each stock and the full sample period. 1bp = 0.0001 = 0.01%.

Table 3.2: Average percentage and the standard deviation (in parentheses) of variance attributed to each principal component.

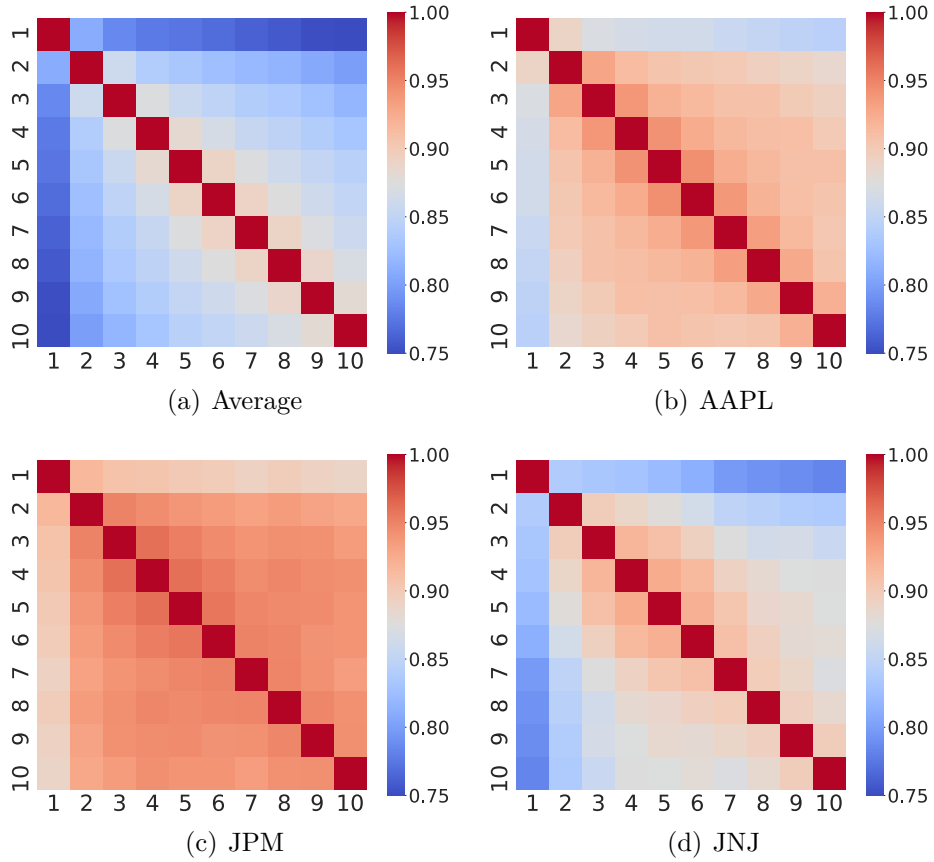
Principal Component	1	2	3	4	5	6	7	8	9	10
Explained Variance Ratio	89.06	4.99	2.28	1.28	0.80	0.54	0.39	0.29	0.21	0.15
	(6.12)	(3.52)	(1.26)	(0.74)	(0.48)	(0.34)	(0.25)	(0.19)	(0.15)	(0.11)

Note: The table reports the ratio (in percentage points) between the variance of each principal component and the total variance averaged across each stock and trading day.

statistics (e.g. mean, std, etc) are computed at the minute level and aggregated across trading days and stocks.

Figure 3.1 reveals that even though the correlation structure of multi-level OFIs may vary across stocks, they all show strong relationships (above 75%). It is worth pointing out that the best-level OFI exhibits the smallest correlation with any of the remaining nine levels, a pattern that persists across different stocks. Table 3.2 further reveals that the first principal component explains more than 89% of the total variance.

In Figure 3.2, we show statistics pertaining to the weights attributed to the top 10 levels in the first principal component. Plot 3.2(a) shows the average weights, and the one standard deviation bars, across all stocks in the universe. Plot 3.2(a) reveals that the best-level OFI has the smallest weight in the first principal component, but the highest standard deviation, hinting that it fluctuates significantly across stocks.



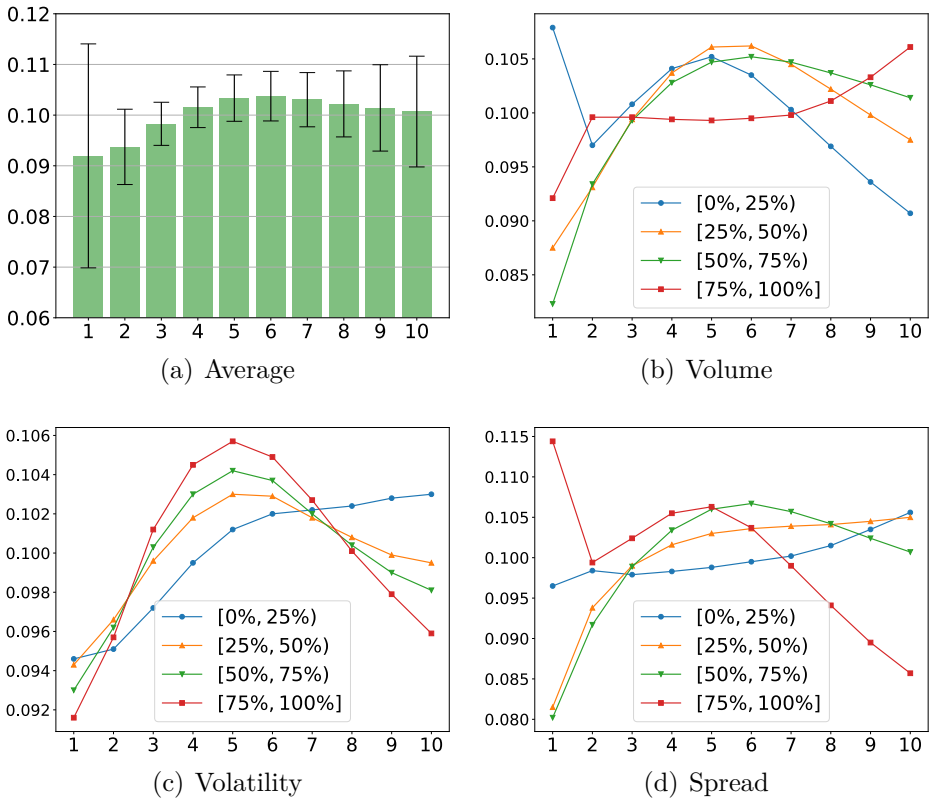
(a) Average (b) AAPL

(c) JPM (d) JNJ

Figure 3.1: Correlation matrix of multi-level OFIs.

Note: Plot (a) is averaged across each stock and each trading day, Plots (b)-(d): correlation matrix of Apple (AAPL), JPMorgan Chase (JPM), and Johnson & Johnson (JNJ) averaged across each trading day. The x-axis and y-axis represent different levels of OFIs.

Plots (b-d) show various patterns for the first principal component of multi-level OFIs, for each quantile bucket of various stock characteristics, in particular, for volume, volatility and spread. For instance, in Figure 3.2(b), the red curve shows the average weights in the first principal component for each of the 10 levels, where the average is taken over all the top 25% largest volume stocks. A striking pattern that emerges from this figure is that for *high-volume* (red line in 3.2(b)), and *low-volatility stocks* (blue line in 3.2(c)), OFIs deeper in the LOB receive more weights in the first component. However, for *low-volume* (blue line in 3.2(b)), and *large-spread stocks* (red line in 3.2(d)), the best-level OFIs account more than the deeper-level OFIs.



(a) Average (b) Volume

(c) Volatility (d) Spread

Figure 3.2: First principal component of multi-level OFIs, in quantile buckets for various stock characteristics.

Note: The x-axis indexes the top 10 levels of the OFIs. **Volume**: trading volume on the previous trading day. **Volatility**: volatility of 1-minute returns during the previous trading day. **Spread**: average bid-ask spread during the previous trading day. $[0\%, 25\%)$, respectively $[75\%, 100\%]$, denote the subset of stocks with the lowest, respectively highest, 25% values for a given stock characteristic.

3.3 Contemporaneous cross impact

In this section, we study the contemporaneous cross-impact model, i.e. how the price of a particular stock is related to the OFIs of other stocks. We examine the existence of contemporary cross-impact by comparing it with the price-impact model studied in Cont, Kukanov, and Stoikov [74].

3.3.1 Models

Price-impact of best-level OFIs. We first pay attention to the price-impact of best-level OFIs ($\text{ofi}_{i,t}^{1,(h)}$) on contemporaneous returns $r_{i,t}^{(h)}$ that materialize over

the same time bucket as the OFI, via the model

$$\mathrm{PI}^{[1]}: \quad r_{i,t}^{(h)} = \alpha_i^{[1]} + \beta_i^{[1]} \mathrm{ofi}_{i,t}^{1,(h)} + \epsilon_{i,t}^{[1]}. \tag{3.6}$$

Here, $\alpha_i^{[1]}$ and $\beta_i^{[1]}$ are the intercept and slope coefficients, respectively. $\epsilon_{i,t}^{[1]}$ is a noise term summarizing the influences of other factors, such as the OFIs at even deeper levels, and potentially the trading behaviors of other stocks. For the sake of simplicity, we refer to the above regression model as $\mathrm{PI}^{[1]}$ and use ordinary least squares (OLS) to estimate it.

Price-impact of integrated OFIs. The second model specification takes into account the impact of multi-level OFIs by leveraging the integrated OFIs, which we set up as follows and use OLS for estimation.

$$\mathrm{PI}^{I}: \quad r_{i,t}^{(h)} = \alpha_i^{I} + \beta_i^{I} \mathrm{ofi}_{i,t}^{I,(h)} + \epsilon_{i,t}^{I}. \tag{3.7}$$

Cross-impact of best-level OFIs. Assuming there are N stocks in the studied universe, we incorporate the multi-asset best-level OFIs, $\mathrm{ofi}_{j,t}^{1,(h)} (j = 1, \ldots, N)$, as candidate features to help fit the returns of the i-th stock $r_{i,t}^{(h)}$. For simplicity, we denote the impact from itself (stock i) as *Self* and that from other stocks as *Cross*, as shown below,

$$\mathrm{CI}^{[1]}: \quad r_{i,t}^{(h)} = \alpha_i^{[1]} + \underbrace{\beta_{i,i}^{[1]} \mathrm{ofi}_{i,t}^{1,(h)}}_{Self} + \sum_{j \neq i} \underbrace{\beta_{i,j}^{[1]} \mathrm{ofi}_{j,t}^{1,(h)}}_{Cross} + \eta_{i,t}^{[1]}. \tag{3.8}$$

Therefore, $\beta_{i,j}^{[1]}$ represents the influence of the j-th stock's best-level OFIs on the returns of stock i.

Cross-impact of integrated OFIs. Finally, we incorporate the cross-asset integrated OFIs to explore the impact of multi-level OFIs from other assets, resulting in the following CI^{I} model,

$$\mathrm{CI}^{I}: \quad r_{i,t}^{(h)} = \alpha_i^{I} + \underbrace{\beta_{i,i}^{I} \mathrm{ofi}_{i,t}^{I,(h)}}_{Self} + \sum_{j \neq i} \underbrace{\beta_{i,j}^{I} \mathrm{ofi}_{j,t}^{I,(h)}}_{Cross} + \eta_{i,t}^{I}. \tag{3.9}$$

Sparsity of cross-impact. As we are aware, OLS regression becomes ill-posed when there are fewer observations than parameters. Recall that we are now considering $N \approx 100$ independent variables in Eqns (3.8) and (3.9). Assuming the time interval is one minute and we are interested in estimating the intraday cross-impact models, e.g. relying on the 30-min estimation window and 1-min returns (as in Cont, Kukanov, and Stoikov [74]), then it seems inappropriate to estimate $\mathrm{CI}^{[1]}$ and CI^{I} for intraday scenarios using the OLS regression with more variables than observations. Moreover, the multicollinearity issue of features contradicts the necessary condition for a well-posed OLS. As displayed in Figure 3.3, a significant portion of the cross-asset correlations based on the best-level OFIs cannot be ignored. For example, approximately 10% of correlations are larger than 0.30. Last, Capponi and Cont [53] found that a certain number of cross-impact coefficients from their OLS regressions are not statistically significant at the 1% significance level.

With all the above considerations in mind, we assume that there is a small number of assets having a significant impact on a specific stock i, as opposed to the entire universe, in $\mathrm{CI}^{[1]}$ and $\mathrm{CI}^{[I]}$. To this end, we apply the Least Absolute Shrinkage and Selection Operator (LASSO)[6] to solve Eqns (3.8) and (3.9). The sparsity of cross-impact terms also facilitates the explanation of coefficients. Note that even though the sparsity of the cross-impact terms is not theoretically guaranteed, our empirical evidence confirms this modeling assumption.

[6]LASSO is a regression method that performs both variable selection and regularization, in order to enhance the prediction accuracy and interpretability of regression models (see more in Gu, Kelly, and Xiu [119], Hastie, Tibshirani, and Friedman [132], and Zhang, Ma, and Wang [236]). It can be formulated as a linear regression model and the objective function consists of two parts, i.e. the sum of squared residuals, and the l_1 constraint on the regression coefficients. In this work, we employ the cross-validation (see Shao [202]) to choose the l_1 penalty weight for each regression.

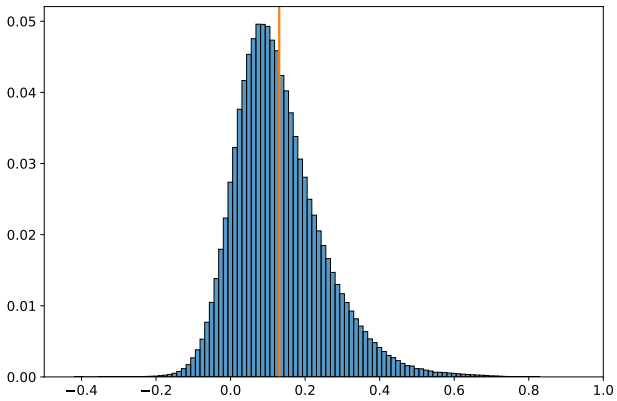


Figure 3.3: Distribution of correlations based on the best-level OFIs.

Note: The orange vertical line represents the average correlation.

3.3.2 Empirical results

For a more representative and fair comparison with previous studies, we apply a similar procedure described in Cont, Kukanov, and Stoikov [74] to our experiments. We exclude the first and last 30 minutes of the trading day due to the increased volatility near the opening and closing sessions, in line with Capponi and Cont [53], Chordia, Roll, and Subrahmanyam [67], Chordia and Subrahmanyam [68], Cont, Kukanov, and Stoikov [74], and Hasbrouck and Saar [130]. In particular, we use each non-overlapping 30-minute estimation window during the intraday time interval 10:00 am - 3:30 pm to estimate the regressions, namely Eqns (3.6), (3.7), (3.8), and (3.9). Within each window, returns and OFIs are computed for every minute.

In-sample performance

We first measure the model performance via in-sample adjusted-R^2, denoted as the **in-sample** R^2 or **IS** R^2. From Table 3.3, we first observe that $\text{PI}^{[1]}$ can explain 71.16% of the in-sample variation of a stock's contemporaneous returns, consistent with the findings of Cont, Kukanov, and Stoikov [74]. Meanwhile, PI^{I} displays higher and more consistent explanation power, with an average adjusted R^2 value of 87.14% and a standard deviation of 9.16%, indicating the effectiveness

Table 3.3: In-sample performance for contemporaneous returns.

	Best-level OFIs		Integrated OFIs	
	$\text{PI}^{[1]}$	$\text{CI}^{[1]}$	PI^{I}	CI^{I}
IS R^2	71.16	73.87	87.14	**87.85**
	(13.80)	(12.23)	(9.16)	(8.58)

Note: The table reports the mean values and standard deviations (in parentheses) of in-sample R^2 (in percentage points) of various models when modeling contemporaneous returns. The models include $\text{PI}^{[1]}$ (Eqn (3.6)), $\text{CI}^{[1]}$ (Eqn (3.8)), PI^{I} (Eqn (3.7)), and CI^{I} (Eqn (3.9)). These statistics are averaged across each stock and each regression window.

of our integrated OFIs.[7]

Table 3.3 also shows that the in-sample R^2 values increase as cross-asset OFIs are included as additional features, which is not surprising given that $\text{PI}^{[1]}$ (respectively, PI^{I}) is a nested model of $\text{CI}^{[1]}$ (respectively, CI^{I}). However, the increments of the in-sample R^2 are smaller when using integrated OFIs (87.85%-87.14%=0.71%), compared to the counterpart using best-level OFIs (73.87%-71.16%=2.71%). This indicates that cross-asset multi-level OFIs may not provide additional information on the variance in returns compared to the price-impact model with integrated OFIs.

Next, we take a closer look at the cross-impact coefficients based on either the best-level or integrated OFIs, i.e. $\beta_{i,j}^{[1]}$ and $\beta_{i,j}^{I}$ $(i, j = 1, \ldots, N)$. Table 3.4 reveals the frequency of self-impact and cross-impact variables selected by LASSO, i.e. the frequency of $\beta_{i,j}^{[1]} \neq 0$ (respectively, $\beta_{i,j}^{I} \neq 0$). We observe that self-impact variables are consistently chosen in both $\text{CI}^{[1]}$ and CI^{I}, as found in Bouchaud [40] and Cont, Kukanov, and Stoikov [74]. However, another interesting observation is that the frequency of a cross-asset integrated OFI variable selected by CI^{I} is around 1/2 of its counterpart in $\text{CI}^{[1]}$. When we turn to the size of the average regression coefficients as shown in Table 3.4, we obtain reasonably consistent results. The self-impact is much higher than the cross-impact in both the $\text{CI}^{[1]}$ and CI^{I} models, while the cross-impact coefficients in CI^{I} are about 1/3 in scale of their counterparts

[7]We also investigate the price-impact model with multi-level OFIs in Appendix B.1. The results demonstrate that the price-impact model using integrated OFIs outperforms those using multi-level OFIs in out-of-sample tests.

Table 3.4: Summary statistics of coefficients in the cross-impact models $\mathrm{CI}^{[1]}$ and CI^{I}.

	Frequency (%)		Magnitude	
	$\mathrm{CI}^{[1]}$	CI^{I}	$\mathrm{CI}^{[1]}$	CI^{I}
Self	99.85	99.96	1.02	1.24
	(0.34)	(0.18)	(0.31)	(0.34)
Cross	17.34	8.29	4.5×10^{-3}	1.6×10^{-3}
	(2.78)	(2.56)	(1.3×10^{-3})	(0.7×10^{-3})

Note: The table is calculated over each stock and each regression window. The first two columns describe the frequency of *Self* and *Cross* variables chosen by the corresponding model with a standard deviation (in parentheses); The last two columns describe the magnitude of *Self* and *Cross* coefficients in the corresponding model with a standard deviation (in parentheses).

in $\mathrm{CI}^{[1]}$. This suggests that once multi-level (or integrated) OFIs have been taken into account, the cross-impact terms might be negligible.

In addition, Figure 3.4 shows a comparison of the top 20 singular values of the coefficient matrices given by the best-level and integrated OFIs.[8] The relatively large singular values of the best-level OFI matrix are a consequence of the higher edge density, and thus average degree, of the network. Note that both networks exhibit a large top singular value of the adjacency matrix (akin to the usual *market mode* in Laloux et al. [160]), and the integrated OFI network has a faster decay of the spectrum, thus revealing its low-rank structure.

We visualize a network for each coefficient matrix, which only preserves the edges larger than a given threshold (following Curme et al. [78] and Kenett et al. [147]), as shown in Figure 3.5. We color stocks according to the GICS sector division, and sort them by their market capitalization within each sector.[9] As one can see from Figure 3.5(a), the cross-impact coefficient matrix $\left[\beta^{[1]}_{i,j}\right]_{j\neq i}$ displays a sectorial structure, in accordance with previous studies (e.g. Benzaquen et al. [30]). This behavior could be fueled by index arbitrage strategies, where traders may, for example, trade an entire basket of stocks coming from the same sector against an index.

[8]Here we only use the off-diagonal elements, i.e. $\left[\beta^{[1]}_{i,j}\right]_{i\neq j}$ and $\left[\beta^{I}_{i,j}\right]_{i\neq j}$.

[9]The Global Industry Classification Standard (GICS) is an industry taxonomy developed in 1999 by MSCI and Standard & Poor's (S&P) for use by the global financial community.

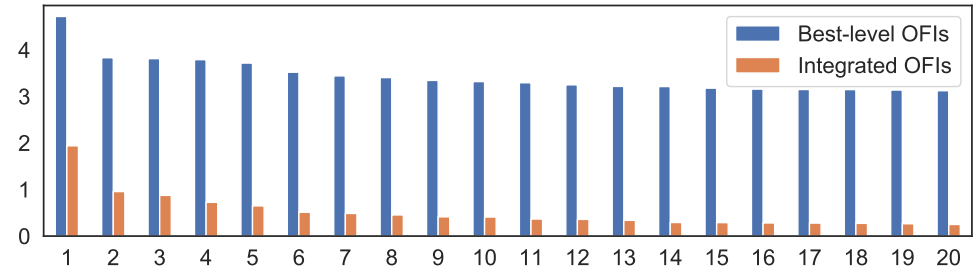


Figure 3.4: Barplot of singular values for the coefficient matrix in contemporaneous cross-impact models.

Note: We perform Singular Value Decomposition (SVD) on the coefficient matrix to obtain the singular values. Singular values are in descending order and the coefficients are averaged over each regression window between 2017–2019. The x-axis represents the singular value rank, and the y-axis represents the singular values.

Figure 3.5(b) presents the network of cross-impact coefficients based on integrated OFIs, i.e. $\left[\beta^I_{i,j}\right]_{j\neq i}$. Compared with Figure 3.5(a), the connections in Figure 3.5(b) are much weaker, implying that the cross-impact from stocks can be potentially explained by a stock's own multi-level OFIs, to a large extent. Note that there is only one connection from GOOGL to GOOG, as pointed out at the top of Figure 3.5(b). This stems from the fact that both stock ticker symbols pertain to Alphabet (Google). Our study also reveals that OFIs of GOOGL have more influence on the returns of GOOG, not the other way around. The main reason might be that GOOGL shares have voting rights, while GOOG shares do not.

In Figures 3.5(c) and 3.5(d), we set lower threshold values (75-th, respectively, 25-th percentile of coefficients) in order to promote more edges in the networks based on integrated OFIs. Interestingly, we observe only four connections in Figure 3.5(c). Except from bidirectional links between GOOGL and GOOG, there exists a one-way link from Cigna (CI) to Anthem (ANTM), and another one-way link from Duke Energy (DUK) to NextEra Energy (NEE). Anthem announced to acquire Cigna in 2015. After a prolonged breakup, this merger finally failed in 2020. Therefore, it is unsurprising that the OFIs of Cigna can affect the price movements of Anthem. Conversely, Anthem's OFIs also have an impact on the price movements of Cigna, but to a lesser extent. Further research should be undertaken to investigate this phenomenon. In terms of the second pair, Duke Energy rebuffed NextEra's

acquisition interest in 2020. Note that 2020 is not in our sample period. This finding hints that certain market participants may have noticed the special relationship between Duke Energy and NextEra Energy before this mega-merger was proposed.

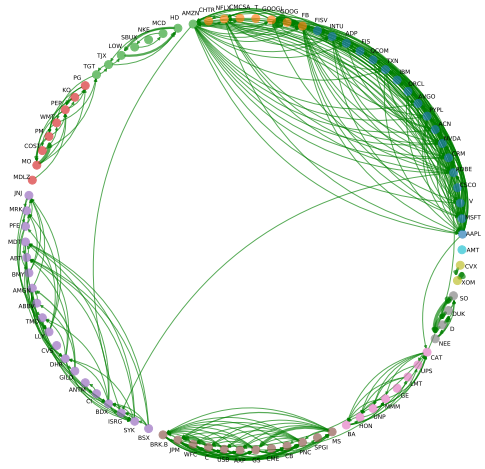
Out-of-sample performance

Although the in-sample estimation yields interesting findings, practitioners are eventually concerned about the out-of-sample estimation. Therefore, we propose to perform the following out-of-sample tests. We use the above fitted models to estimate returns on the following 30-minute data and compute the corresponding R^2, denoted as the **out-of-sample** R^2 or **OS** R^2.

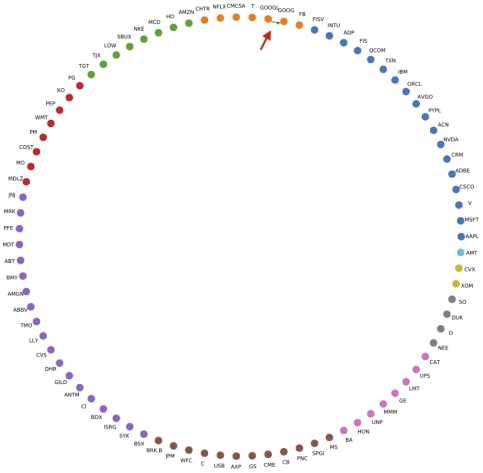
Previous studies either investigated the in-sample R^2 (including Capponi and Cont [53] and Cont, Kukanov, and Stoikov [74]), or adopted a cross-validation method (e.g. Xu, Gould, and Howison [231]). However, these works failed to consider the generalization error of their models or damaged the chronological order of the time-series data. In contrast, we obey the temporal ordering in our study. These matters are vital to practitioners, as only the historical data are accessible for the model fit in practice.

Table 3.5 reports the average values and their standard deviations of out-of-sample R^2 of $\text{PI}^{[1]}$, $\text{CI}^{[1]}$, PI^{I}, and CI^{I}. We first focus on the models using best-level OFIs. It appears $\text{CI}^{[1]}$ has a slight advantage compared with $\text{PI}^{[1]}$ for out-of-sample tests with an improvement of 1.39% (=66.03%-64.64%). However, when involving multi-level or integrated OFIs, the performance of CI^{I} is slightly worse than PI^{I}, indicating that the cross-impact model with integrated OFIs cannot provide extra explanatory power to the price-impact model with integrated OFIs. Overall, we observe that the models using integrated OFIs unveil significant and consistent improvements over those using only best-level OFIs.

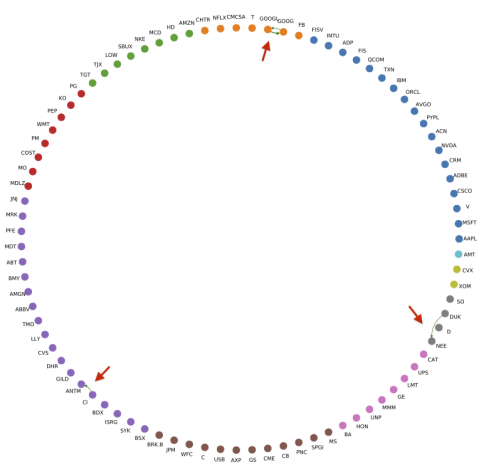
In general, we observe strong evidence implying $\text{CI}^{[1]}$ provides a better out-of-sample estimate than $\text{PI}^{[1]}$, while for PI^{I} and CI^{I}, the evidence is opposite. However, it is important to note that these conclusions are based on a point estimate and do not necessarily indicate statistical significance. Therefore, we



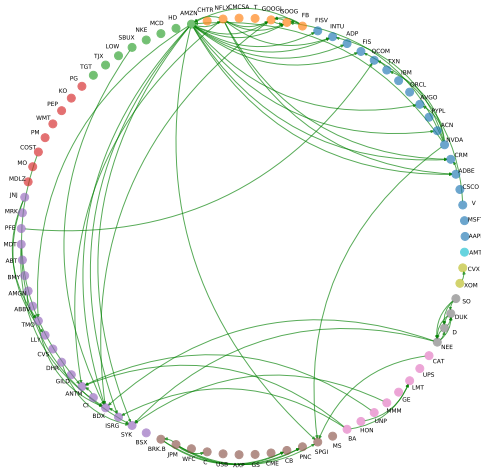
(a) Threshold = 95-th percentile, based on best-level OFIs



(b) Threshold = 95-th percentile, based on integrated OFIs



(c) Threshold = 75-th percentile, based on integrated OFIs



(d) Threshold = 25-th percentile, based on integrated OFIs

Figure 3.5: Illustrations of the coefficient networks constructed from contemporaneous cross-impact models.

Note: To render the networks more interpretable and for ease of visualization, we only plot the top 5% largest (a-b), or top 25% largest (c), or top 75% largest (d), in magnitude coefficients. The coefficients are averaged over each regression window between 2017–2019. Nodes are colored by the GICS structure and sorted by market capitalization. Green links represent positive values while black links represent negative values. The width of edges is proportional to the absolute values of their respective coefficients.

Table 3.5: Out-of-sample performance for contemporaneous returns.

	Best-level OFIs		Integrated OFIs	
	$\text{PI}^{[1]}$	$\text{CI}^{[1]}$	PI^{I}	CI^{I}
OS R^2	64.64	66.03	**83.83**	83.62
	(21.82)	(19.51)	(16.90)	(14.53)

Note: The table reports the mean values and standard deviations (in parentheses) of out-of-sample R^2 (in percentage points) of various models when modeling contemporaneous returns. The models include $\text{PI}^{[1]}$ (Eqn (3.6)), $\text{CI}^{[1]}$ (Eqn (3.8)), PI^{I} (Eqn (3.7)), and CI^{I} (Eqn (3.9)). These statistics are averaged across each stock and each regression window.

perform the following hypothesis test for each stock on the out-of-sample data to assess statistical significance,

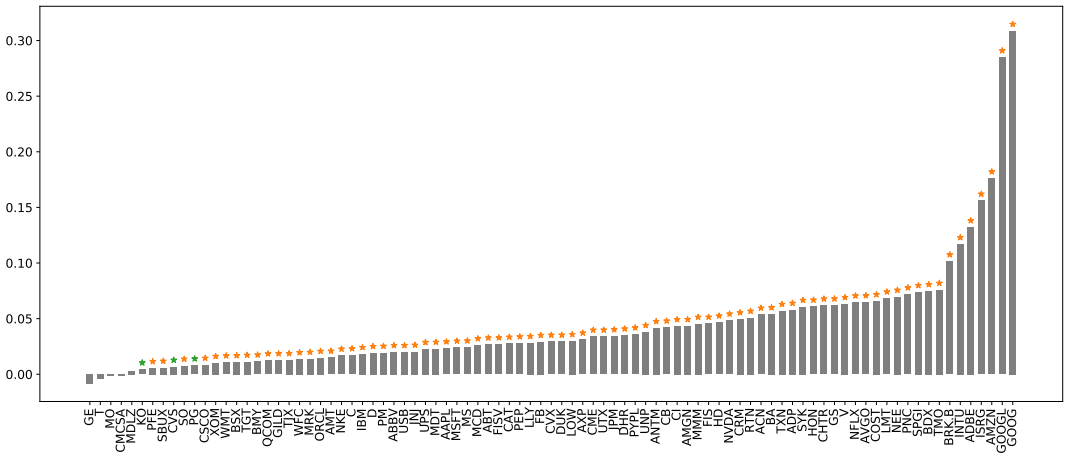
$$\mathcal{H}_0 : \mathbb{E}\left[R^2_{\text{OS}}\left(\text{CI}^{[1]}\right) - R^2_{\text{OS}}\left(\text{PI}^{[1]}\right)\right] \leq 0 \text{ vs. } \mathcal{H}_1 : \mathbb{E}\left[R^2_{\text{OS}}\left(\text{CI}^{[1]}\right) - R^2_{\text{OS}}\left(\text{PI}^{[1]}\right)\right] > 0.$$

We employ the approach from Giacomini and White [111] and Chinco, Clark-Joseph, and Ye [63] to assess statistical significance through a Wald-type test (see Ward and Ahlquist [224]). Theorem 1 in Giacomini and White [111] implies that we can use a standard t-test to evaluate the statistical significance of changes in R^2. A p-value less than a given significance level α rejects the null hypothesis in favor of the alternative at the $1-\alpha$ confidence level, implying $\text{CI}^{[1]}$ has significantly better estimation than $\text{PI}^{[1]}$. We also implement this test for the comparison between PI^{I} and CI^{I}.

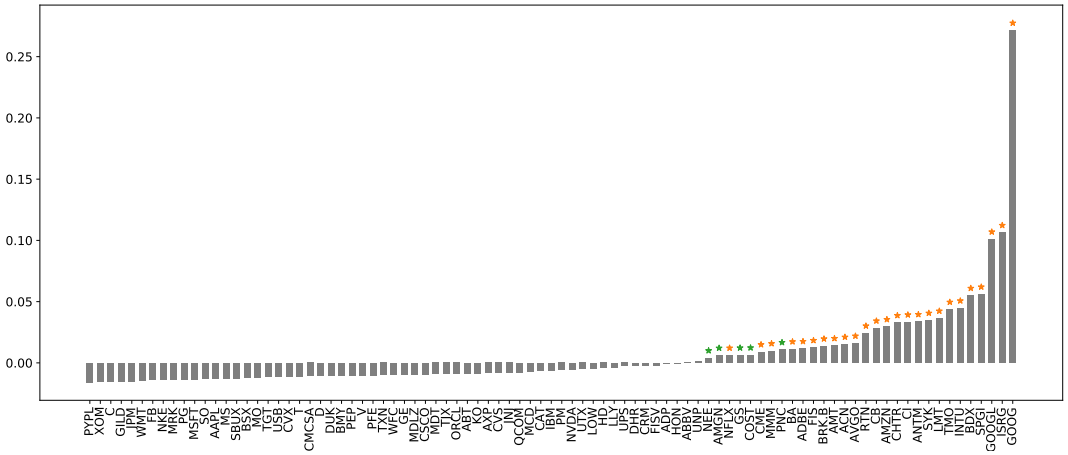
Figure 3.6 illustrates the main results from the above hypothesis tests. When using only the best-level OFIs, the cross-impact model is superior to the price-impact model for 91.0% (94.4%) of stocks, at the 1% (5%) confidence level. However, when examining the models using integrated OFIs, we reject the null hypothesis (i.e., in favor of the cross-impact model) only for 28.1% (33.7%) of stocks at the 1% (5%) confidence level. As expected, cross-impact terms can significantly improve the explanatory power of the price-impact model for GOOG and GOOGL.

3.3.3 Discussion about contemporaneous cross-impact

In summary, our previous results mainly show that when considering only the best-level OFI of a single stock, the addition of the best-level OFI from other



(a) $R^2_{\text{OS}}\left(\text{CI}^{[1]}\right) - R^2_{\text{OS}}\left(\text{PI}^{[1]}\right)$



(b) $R^2_{\text{OS}}\left(\text{CI}^{I}\right) - R^2_{\text{OS}}\left(\text{PI}^{I}\right)$

Figure 3.6: Mean differences of out-of-sample R^2 between CI and PI models.

Note: A positive (negative) number indicates superiority for the CI (PI) model. The y-axis represents the average difference of OS R^2 between CI and PI, while the x-axis lists the stock symbols. Stars indicate the p-values, with orange, green, and blue representing significance at the 1%, 5%, and 10% levels, respectively.

stocks slightly increases the explanatory power. On the other hand, when the information from multiple levels is integrated into the OFI, the improvement is negligible. In the meantime, it is unsurprising that taking into account more levels in the LOB (PI^I) could better explain price changes, compared to only considering best-level orders ($\text{PI}^{[1]}$).

After observing these results, several natural questions may arise: How can the

above facts be reconciled?[10] How do the cross-asset best-level OFIs interact with the multi-level OFIs, when modeling contemporaneous returns?

To address these questions, we consider the following scenario, also depicted in Figure 3.7. For simplicity, we denote the order from trading strategy A on stock i (respectively, j) as A_i (respectively, A_j). Analogously, we define orders from strategy B and S. Let us next consider the orders of stock i. There are three orders from different portfolios, given by A_i, B_i and S_i. A_i is at the third bid level of stock i and linked to an order at the best ask level of stock j, i.e. A_j. Also, B_i is at the best ask level of stock i and linked to an order at the best bid level of stock j, i.e. B_j. Finally, S_i is an individual bid order at the best level of stock i.

Now we observe that the best-level limit orders from stock j may be linked to price movements of stock i through paths $B_j \to B_i \to \text{ofi}_i^1 \to r_i$ and $A_j \to A_i \to \text{ofi}_i^3 \to r_i$. Thus the price-impact model for stock i which only utilizes its own best-level orders will ignore the information of A_i, while the cross-impact model can partially collect it along the path $A_j \to A_i$. This might illustrate why the best-level OFIs of multiple assets can provide slightly additional explanatory power to the price-impact model using only the best-level OFIs.

Nonetheless, if we can integrate multi-level OFIs in an efficient way (in our example, aggregate order imbalances caused by orders A_i, B_i and S_i), then there is no need to consider OFIs from other stocks for modeling price dynamics. For example, information hidden in the path $A_j \to A_i \to \text{ofi}_i^3 \to r_i$ can be leveraged as long as A_i is well absorbed into new integrated OFIs. In this sense, for stock i, cross-asset best-level OFIs (including A_j) are surrogates of its own OFIs at different levels (here A_i), to a certain extent. The likelihood of this relationship is attributed to massive portfolio trades that submit or cancel limit orders across a variety of assets

[10] One possible explanation for those facts is that the duration of the cross-impact terms might be shorter than the current time interval (30 minutes) used in our experiments, rendering the cross-impact terms vanish in out-of-sample tests. To verify this assertion, we implement additional experiments where models are updated more frequently. The results (deferred to Appendix B.3) reveal that even under higher-frequency updates (1-min) of the models, there is no benefit from introducing cross-impact terms to the price-impact model with integrated OFIs.

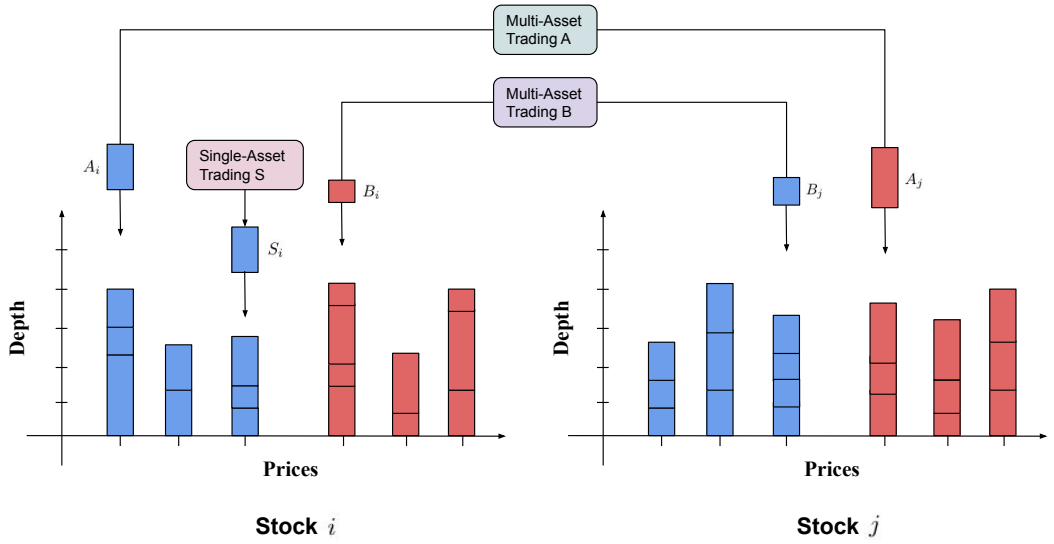


Figure 3.7: Illustration of the cross-impact model.

Note: The orders at different levels of each stock may come from single-asset and multi-asset trading strategies. The returns of stock i are potentially influenced by orders of stock j through the connections $B_j \to B_i \to \text{ofi}_i^1 \to r_i$ and $A_j \to A_i \to \text{ofi}_i^3 \to r_i$. Information along the path $A_j \to A_i \to \text{ofi}_i^3 \to r_i$ can be collected by the price-impact model with integrated OFIs but not by the price-impact model with only best-level OFIs.

at different levels.[11] We put forward this mechanism which potentially explains why the cross-impact model with integrated OFIs cannot provide additional explanatory power compared to the price-impact model with integrated OFIs.[12]

3.4 Forecasting future returns

In the previous section, the definitions of price-impact and cross-impact are based on contemporaneous OFIs and returns, meaning that both quantities pertain to the same bucket of time. In this section, we extend the above studies to future returns, and probe into the forward-looking price-impact and cross-impact models.

[11] Cao, Hansch, and Wang [51], Chakrabarty et al. [59], Hautsch and Huang [133], and Sirignano [207] showed that the depth of some deeper levels (such 2-3) is higher than the best level depth.

[12] It would be a very interesting research direction to derive testable predictions for this mechanism in future work. However, so far it hinges on the availability of client-ID based LOB data, i.e. knowing that the same market participant is behind the orders for two related instruments, released at approximately the same time, such as (A_i, A_j), (B_i, B_j).

3.4.1 Predictive models

We first propose the following forward-looking price-impact and cross-impact models, denoted as $\text{FPI}^{[1]}$, FPI^{I}, $\text{FCI}^{[1]}$, and FCI^{I}, respectively. $\text{FPI}^{[1]}$ (FPI^{I}) uses the lagged best-level (integrated) OFIs of stock i to predict its own future return $r^{(f)}_{i,t+f}$ during $(t, t+f]$, while $\text{FCI}^{[1]}$ (FCI^{I}) involves the lagged multi-asset best-level (integrated) OFIs. We employ OLS to fit the forward-looking price-impact models and LASSO to fit the cross-impact models.

$$\text{FPI}^{[1]}: \quad r^{(f)}_{i,t+f} = \alpha^{[1]}_i + \sum_{k\in L} \beta^{[1],k}_i \text{ofi}^{1,(kh)}_{i,t} + \epsilon^{[1]}_{i,t+f}, \tag{3.10}$$

$$\text{FCI}^{[1]}: \quad r^{(f)}_{i,t+f} = \alpha^{[1]}_i + \sum_{j=1}^{N}\sum_{k\in L} \beta^{[1],k}_{i,j} \text{ofi}^{1,(kh)}_{i,t} + \eta^{[1]}_{i,t+f}, \tag{3.11}$$

$$\text{FPI}^{I}: \quad r^{(f)}_{i,t+f} = \alpha^{I}_i + \sum_{k\in L} \beta^{I,k}_i \text{ofi}^{I,(kh)}_{i,t} + \epsilon^{I}_{i,t+f}, \tag{3.12}$$

$$\text{FCI}^{I}: \quad r^{(f)}_{i,t+f} = \alpha^{I}_i + \sum_{j=1}^{N}\sum_{k\in L} \beta^{I,k}_{i,j} \text{ofi}^{I,(kh)}_{i,t} + \eta^{I}_{i,t+f}, \tag{3.13}$$

where f is the forecasting horizon of future returns and $L = \{1, 2, 3, 5, 10, 20, 30\}$ represents the set of lags.

Furthermore, we compare OFI-based models with return-based models studied in previous works, where lagged returns are involved as predictors. AR (Eqn (3.14)) is an autoregressive (AR) model using various returns over different time horizons, inspired by Aït-Sahalia et al. [5] and Corsi [75]. CAR (Chinco, Clark-Joseph, and Ye [63]) uses the entire cross-section lagged returns as candidate predictors, as detailed in Eqn (3.15). We employ OLS to fit the ARs and LASSO to fit the CARs.

$$\text{AR}: \quad r^{(f)}_{i,t+f} = \alpha_i + \sum_{k\in L} \beta^{r,k}_i r^{(kh)}_{i,t} + \epsilon_{i,t+f} \tag{3.14}$$

$$\text{CAR}: \quad r^{(f)}_{i,t+f} = \alpha_i + \sum_{j=1}^{N}\sum_{k\in L} \beta^{r,k}_{i,j} r^{(kh)}_{i,t} + \eta_{i,t+f} \tag{3.15}$$

3.4.2 Empirical results

In this experiment, observations associated with returns and OFIs are computed minutely, i.e. $h = 1$ minute.[13] Following Chinco, Clark-Joseph, and Ye [63], we use

[13]Note that we choose to use the physical time as opposed to the trading time. This is because each stock has its own specific trading time, which is asynchronous with that of others. Thus it

data from the previous 30 minutes to estimate the model parameters and apply the fitted model to forecast future f-minute returns. We then move one minute forward and repeat this procedure to compute the rolling f-minute-ahead return forecasts. For all models, we initially focus on the 1-minute forecasting horizon. In Section 3.4.3, we consider return forecasts over longer horizons, including $f \in \{2, 3, 5, 10, 20, 30\}$ minutes, to assess the strength and duration of price-impact and cross-impact.

Following the analysis of Bollerslev et al. [35], Cartea, Donnelly, and Jaimungal [55], and Chinco, Clark-Joseph, and Ye [63], we demonstrate the effectiveness of the forward-looking price-impact and cross-impact models from two perspectives: (1) statistical performance, and (2) economic gain.

Statistical performance

Table 3.6 summarizes the out-of-sample R^2 values of the aforementioned predictive models when predicting the subsequent 1-minute returns, i.e. $f = 1$. It appears the cross-impact models FCI$^{[1]}$ (respectively, FCI^I, CAR) achieve higher out-of-sample R^2 statistics compared to the price-impact models FPI$^{[1]}$ (respectively, FPI^I, AR). We also implement the same hypothesis test described in Section 3.3 to investigate the statistical significance (unreported) of these results. We observe that the cross-impact models exhibit significantly superior performance than the price-impact models across all stocks, at the 1% confidence level.

Most of the empirical literature in return prediction focuses its evaluations on out-of-sample R^2. However, we remark that negative R^2 values do not imply that the forecasts are economically meaningless (see more discussions in Choi, Jiang, and Zhang [65] and Kelly, Malamud, and Zhou [146]).[14] To emphasize this point,

is difficult to work out the cross–impact between stocks on a trading time scale, also see Wang, Schäfer, and Guhr [221, 222]. The choice of a 1-minute bin size allows us to abstract away from microstructure effects which are not the focus of the present mesoscopic study, as is the case in Benzaquen et al. [30] and Chinco, Clark-Joseph, and Ye [63].

[14]A simple example can be framed as follows. Consider a model with one predictor and suppose that the estimated predictive coefficient is a significantly large multiple of the actual value. In this case, the R^2 will become negative. However, the predictions will be perfectly correlated with the true expected return, resulting in a positive expected return for our strategy. Proposition 4 in Kelly, Malamud, and Zhou [146] further proposed that we can worry less about the positivity of out-of-sample R^2 from a prediction model and focus more on the out-of-sample performance of specific trading strategies based on predicted returns.

we will incorporate these return forecasts into a forecast-based trading strategy, and showcase their profitability in the following subsection.

Table 3.6: Out-of-sample performance for one-minute-ahead returns.

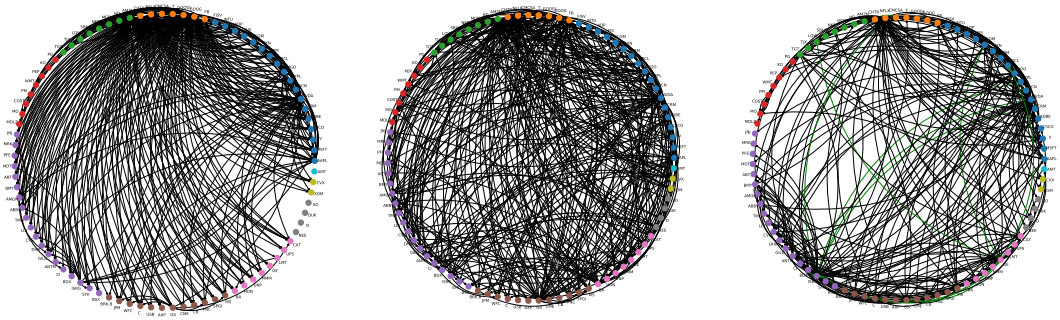
	Best-level OFIs		Integrated OFIs		Returns	
	$\text{FPI}^{[1]}$	$\text{FCI}^{[1]}$	FPI^{I}	FCI^{I}	AR	CAR
OS R^2	-0.37	-0.10	-0.36	-0.10	-0.36	-0.10
	(0.10)	(0.05)	(0.08)	(0.05)	(0.11)	(0.05)

Note: The table reports the mean values and standard deviations (in parentheses) of out-of-sample R^2 of various models when modeling one-minute-ahead returns. The predictive models include $\text{FPI}^{[1]}$ (Eqn (3.10)), $\text{FCI}^{[1]}$ (Eqn (3.11)), FPI^{I} (Eqn (3.12)), FCI^{I} (Eqn (3.13)), AR (Eqn (3.14)) and CAR (Eqn (3.15)). These statistics are averaged across each stock and each regression window.

Considering the different magnitudes of the OFIs and returns, we first normalize the coefficient matrix of each model by dividing by the average of the absolute coefficients. Figure B.4 (deferred to Appendix B.4) shows the average coefficient matrices of $\text{FCI}^{[1]}$, FCI^{I}, and CAR. For example, as revealed in Figure B.4(a) ($\text{FCI}^{[1]}$), for a specific stock, the main influence comes from its own OFI, i.e. the absolute values of diagonal elements are significantly larger than the off-diagonal ones. We observe that cross-impact is often negative, consistent with Pasquariello and Vega [188]. Except for the self-impact, most stocks are also influenced by stocks in Communication Services, Consumer Discretionary and Information Technology.

To better illustrate the interactions between different stocks, we construct a network for each normalized coefficient matrix and only preserve the cross-asset edges (i.e. off-diagonal elements) larger than the 95-th percentile of coefficients. Figure 3.8 illustrates some of the main characteristics of the coefficient networks for $\text{FCI}^{[1]}$, FCI^{I}, and CAR. For example, we again observe that there are more edges from Communication Services, Consumer Discretionary and Information Technology, indicating they may contain more predictive power for others.

To gain a better understanding of the structural properties of the resulting network, we aggregate node centrality measures (see Everett and Borgatti [97]) at the sector level, and also perform a spectral analysis of the adjacency matrix



(a) Based on best-level OFIs (b) Based on integrated OFIs (c) Based on returns

Figure 3.8: Network structure of the coefficient matrix constructed from forward-looking cross-impact models.

Note: The coefficients are averaged over 2017–2019. To render the networks more interpretable and for ease of visualization, we only plot the top 5% largest in magnitude coefficients. Nodes are colored by the GICS structure and sorted by market capitalization. Green links represent positive values while black links represent negative values. The width of edges is proportional to the absolute values of their respective coefficients.

Table 3.7: Group degree centrality for each GICS sector.

	Group In-degree Centrality			Group Out-degree Centrality		
	Best-level OFIs	Integrated OFIs	Returns	Best-level OFIs	Integrated OFIs	Returns
Information Technology	0.12	0.36	0.26	0.46	0.62	0.59
Communication Services	0.06	0.24	0.20	0.85	0.74	0.60
Consumer Discretionary	0.09	0.20	0.15	0.86	0.51	0.17
Consumer Staples	0.03	0.15	0.09	0.00	0.11	0.01
Health Care	0.10	0.37	0.19	0.12	0.22	0.59
Financials	0.12	0.21	0.17	0.03	0.41	0.08
Industrials	0.10	0.19	0.20	0.00	0.39	0.27
Utilities	0.00	0.07	0.04	0.00	0.06	0.00
Energy	0.06	0.07	0.04	0.00	0.14	0.00
Real Estate	0.00	0.05	0.02	0.00	0.00	0.01

Note: According to the threshold networks as shown in Figure 3.8, we compute the fraction of stocks outside of a specific sector connected to stocks in this specific sector. The color of each sector in this table corresponds to the color in Figure 3.8.

(see Kannan and Vempala [144] and Newman [179]). From Table 3.7, we observe that the out-degree centrality of Communication Services, Consumer Discretionary and Information Technology is significantly larger than that of others, consistent with previous findings. Figure 3.8(c) also shows that the network based on returns contains more inner-sector connections than the other two counterparts, thus implying a sectorial structure. Table 3.8 presents the top five most significant stocks in terms of out-degree centrality in each network, which exhibit more impact

Table 3.8: Top 5 stocks according to node out-degree centrality in threshold networks.

Best-level OFIs	Integrated OFIs	Returns
AMZN	NFLX	NVDA
GOOG	AMZN	NFLX
GOOGL	NVDA	ISRG
NVDA	GS	AVGO
NFLX	FB	GE

Note: The out-degree centrality for a node is the fraction of nodes its outgoing edges are connected to.

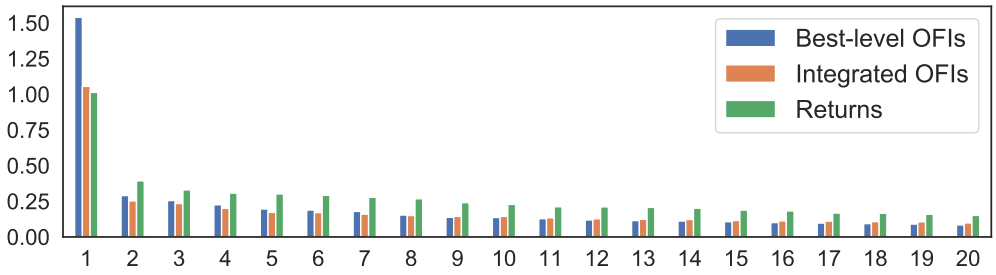


Figure 3.9: Barplot of normalized singular values for the average coefficient matrix in forward-looking cross-impact models.

Note: We perform Singular Value Decomposition (SVD) on the coefficient matrix and obtain the singular values. The x-axis represents the singular value rank, and the y-axis shows the normalized singular values. The coefficients are averaged over 2017–2019.

on the prices of other stocks.

Figure 3.9 shows a barplot with the average value for the top 20 largest singular values of the network adjacency matrix, for best-level OFIs, integrated OFIs, and returns, where the average is performed over all constructed networks. For ease of visualization and comparison, we first normalize the adjacency matrix before computing the top singular values, which exhibit a fast decay. In addition to the significantly large top singular value revealing that the networks have a strong rank-1 structure, the next 6-8 singular values are likely to correspond to the more prominent industry sectors.

Economic gains

On the basis of return forecasts, we employ a portfolio construction method, proposed by Chinco, Clark-Joseph, and Ye [63], to evaluate the economic gains

of the aforementioned predictive models.

Forecast-implied portfolio. For a specific forecasting model F, the motivations of portfolio construction can be summarized as follows.

- It only executes an order when the one-minute-ahead return forecast exceeds the bid-ask spread.
- It buys/sells more shares of the i-th stock when the absolute value of one-minute-ahead return forecast for i-th stock is higher.
- It buys/sells more shares of the i-th stock when the one-minute-ahead return forecasts for the i-th stock tend to be less volatile throughout the trading day.

This strategy allocates a fraction $w_{i,t}$ of its capital to the i-th stock

$$w_{i,t} \overset{\text{def}}{=} \frac{1_{\left\{\left|f^F_{i,t}\right|>sprd_{i,t}\right\}} \cdot f^F_{i,t} \,/\, \sigma^F_{i,t}}{\sum_{n=1}^{N} 1_{\left\{\left|f^F_{n,t}\right|>sprd_{n,t}\right\}} \cdot \left|f^F_{n,t}\right| \,/\, \sigma^F_{n,t}}, \tag{3.16}$$

where $f^F_{i,t}$ represents the one-minute-ahead return forecast according to model F for minute $(t+1)$, $sprd_{i,t}$ represents the relative bid-ask spread at time t, $\sigma^F_{i,t}$ represents the standard deviation of the model's one-minute-ahead return forecasts for the i-th stock during the previous 30 minutes of trading, i.e. the standard deviation of in-sample fits. The denominator is the total investment so that the strategy is self-financed. If there are no stocks with forecasts that exceed the spread in a given minute, then we set $w_{i,t} = 0, \forall i$.

Finally, we compute the *profit and loss (PnL)* of the resulting portfolios on each trading day by summing the strategy's minutely returns as in Chinco, Clark-Joseph, and Ye [63].

Table 3.9 compares the performance (annualized PnL) of the forecast-implied strategies, based on forecast returns from various predictive models. It is worth noting that in the following analysis, the strategy ignores trading costs, as this is not the focus of our chapter. Table 3.9 shows that portfolios based on forecasts

of the forward-looking cross-impact model outperform those based on forecasts of the forward-looking price-impact model.

Table 3.9: Economic performance of forecast-implied trading strategy.

	Best-level OFIs		Integrated OFIs		Returns	
	$\text{FPI}^{[1]}$	$\text{FCI}^{[1]}$	FPI^{I}	FCI^{I}	AR	CAR
PnL	0.21	0.43	0.23	0.39	0.23	0.40
	(0.12)	(0.17)	(0.13)	(0.19)	(0.13)	(0.18)

Note: The table reports the mean values and standard deviations (in parentheses) of annualized PnLs of forecast-implied trading strategy of various models for forecasting one-minute-ahead returns. The predictive models include $\text{FPI}^{[1]}$ (Eqn (3.10)), $\text{FCI}^{[1]}$ (Eqn (3.11)), FPI^{I} (Eqn (3.12)), FCI^{I} (Eqn (3.13)), AR (Eqn (3.14)) and CAR (Eqn (3.15)). These statistics are averaged over 2017-2019.

3.4.3 Longer forecasting horizons

One-minute-ahead return forecasts are not the only time horizon of interest to practitioners and academics. Additionally, we evaluate the performance of the above models and examine the forecasting ability of cross-impact terms over longer prediction horizons.

Figure 3.10 illustrates the model predictability from the perspective of raw annualized PnL across multiple horizons.[15] Due to the similar performance of $\text{FPI}^{[1]}$ and FPI^{I} (respectively, $\text{FCI}^{[1]}$ and FCI^{I}) over longer horizons, we only show the curves of $\text{FPI}^{[1]}$, FCI^{I}, AR, CAR, and a benchmark (S&P100 ETF). It appears that superior forecasting ability arises from cross-asset terms at short horizons. However, the PnL of cross-asset models declines more quickly over longer horizons. A further study with more focus on the reasons for the predictability of cross-asset OFIs over multiple horizons is therefore suggested. Finally, the models in which each stock only relies on its own returns/OFIs marginally outperform their counterparts which use the entire cross-sectional predictors.

[15]To plot this figure, we only accumulate the PnLs between [10:31, 15:30], which is the shared trading period for the studied forecasting horizons.

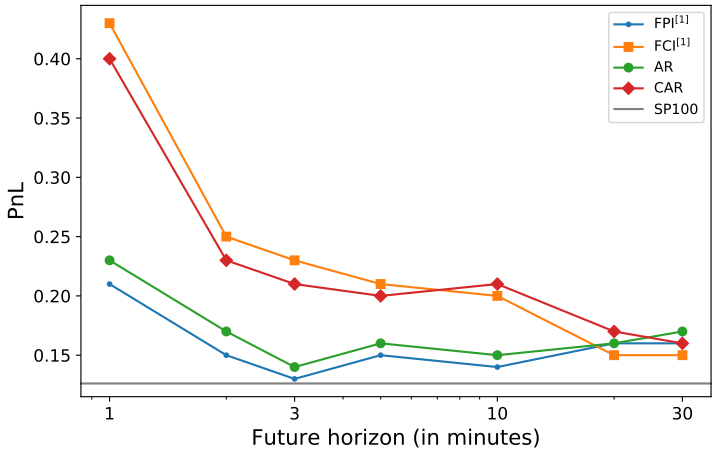


Figure 3.10: Annualized PnL as a function of the forecasting horizon.

Note: The x-axis represents the prediction horizon (in minutes), while the y-axis represents the annualized PnL. The grey horizontal line is the performance of the S&P100 ETF index.

3.4.4 Discussion about predictive cross-impact

Tables 3.6 and 3.9 reveal that, in contrast to the price-impact model, multi-asset OFIs can provide considerably more additional explanatory power for *future returns* compared to *contemporaneous returns*. A possible explanation for this asymmetric phenomenon is that there exists a time lag between when the OFIs of a given stock are formed (a so-called *flow formation period*) and the actual time when traders notice this change of flow and incorporate it into their trading model (see Buccheri, Corsi, and Peluso [42]).[16] For example, assume a trader submitted an unexpectedly large amount of buy limit orders of Apple (AAPL) at 10:00 am, at either the first level or potentially deeper in the book. Other traders may notice this anomaly and adjust their portfolios (including Apple) at a later time, for example, 10:01 am. In this case, the OFIs of Apple may indicate future price changes of other stocks.

[16]A closely-related phenomenon is the Epps effect documented by Epps [96], which showed that the empirical correlation estimates tend to decrease when the sampling is done at high frequencies. Previous research (including Renò [196], Tóth and Kertész [216], and Zhang [235]) demonstrated that the Epps effect might be explained by the non-synchronicity, the possible lead-lag relationship between stock returns, etc. Tóth and Kertész [216] also described that the Epps effect might be caused by the reaction time of traders to news and events, which is usually spread out over a time interval of a few minutes.

Consistent with our explanation, Hou [140] argued that the gradual diffusion of industry information is a leading cause of the lead-lag effect in stock returns. Cohen and Frazzini [71] found that certain stock prices do not promptly incorporate news pertaining to economically related firms, due to the presence of investors subject to attention constraints. Further research should be undertaken to investigate the origins of the success of multi-asset OFIs in predicting future returns.

It is also interesting to note that forward-looking models using integrated OFIs cannot significantly outperform models using the best-level OFIs. This phenomenon might stem from the fact that the integrated OFIs do not explicitly take into account the level information (distance of a given level to the best bid/ask) of multi-level OFIs, and are agnostic to different sizes resting at different levels on the bid and ask sides of the book. Previous studies (such as Cao, Hansch, and Wang [51], Cenesizoglu, Dionne, and Zhou [58], and Hasbrouck and Saar [130]) demonstrated that traders might strategically choose to place their orders in different levels of the book depending on various factors, therefore limit orders at different price levels may contain different information content with respect to predicting future returns. A further study with more focus on the impact of multi-level OFIs over different time horizons is suggested.

3.5 Conclusion

We have systematically examined the impact of OFIs from multiple perspectives. The main contributions can be summarized as follows.

First, we verify the effects of multi-level and cross-asset OFIs on contemporaneous price dynamics. We introduce a new procedure to examine the cross-impact on contemporaneous returns. Under the sparsity assumption of cross-impact coefficients, we use LASSO to describe such a structure and compare the performances with the price-impact model which only utilizes a stock's own OFIs. We implement models with the best-level OFIs and integrated OFIs, respectively. The results first demonstrate that our integrated OFIs provide a higher explanatory power for price movements than the widely-used best-level OFIs. More interestingly, in

comparison with the price-impact model using best-level OFIs, the cross-impact model exhibits additional explanatory power. However, the cross-impact model with integrated OFIs cannot provide extra explanatory power to the price-impact model with integrated OFIs, indicating the effectiveness of our integrated OFIs.

In addition, we apply the price-impact and cross-impact models to the challenging task of predicting future returns. The results reveal that involving cross-asset OFIs can increase out-of-sample R^2. We subsequently demonstrate that this increase in out-of-sample R^2 leads to additional economic profits, when incorporated in common trading strategies, thus providing evidence of cross-impact over short future horizons. We also find that predictability of cross-impact terms vanishes quickly over longer horizons.

Future research directions. There are a number of interesting avenues to explore in future research. One such direction pertains to the assessment of whether cross-asset **multi-level** OFIs can improve the forecast of future returns (in the present work, we only considered the best-level OFI and integrated OFI due to limited computing power). Another interesting direction pertains to performing a similar analysis as in the present chapter, but for the last 15-30 minutes of the trading day, where a significant fraction of the total daily trading volume occurs. For example, for the first few months of 2020 in the US equity market, about 23% of trading volume in the 3,000 largest stocks by market value has taken place after 3:30 pm, compared with about 4% from 12:30 pm to 1 pm (Banerji [21]). It would be an interesting study to explore the interplay between the OFI dynamics and this surge of trading activity at the end of U.S. market hours.

4

Volatility Forecasting with Machine Learning and Intraday Commonality

Contents

4.1 Introduction

Forecasting and modeling stock return volatility have been of interest to both academics and practitioners. Recent advances in high-frequency trading (HFT) highlight the need for robust and accurate intraday volatility forecasts. For example, Deutsche Börse, one of the world's leading data and technology service providers, launched the "Intraday Volatility Forecast" project in 2015 to provide intraday volatility forecasts up to 30-min for DAX, EURO STOXX 50 and Euro-Bund.

Engle and Sokalska [95] pointed out that intraday volatility forecasts are important for managing risk, pricing derivatives, and devising quantitative strategies, especially in HFT. Stroud and Johannes [208] also demonstrated that intraday measures are useful for market makers, HFT, and option traders. Specifically, intraday volatility forecasts may support traders in assessing the likelihood of price changes and therefore better understanding the risk involved in certain automated trading strategies (see Bates [27]). Screen traders could leverage volatility forecasts to support live trading and enhance the pre-trade transaction cost analysis from the risk assessment of price slippage. Intraday volatility forecasts are also helpful for practitioners screening for high-volatility opportunities and trading corresponding option strategies (see Ni, Pan, and Poteshman [181]).

To the best of our knowledge, unlike daily volatility forecasting, intraday volatility has not yet received much attention in the research literature. It is pointed out by Andersen and Bollerslev [10] that conventional parametric models, such as Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) and Stochastic Volatility (SV) models, may fail to reveal certain features of intraday returns. In Andersen et al. [13] and Corsi [75], high-frequency data are used to estimate daily realized volatility (RV) by summing squared intraday returns. Some related literature on this aspect will be reviewed briefly in the next section. These methods may reveal important information about characteristics of daily

returns and volatilities, but do not easily lend themselves applicable to the task of forecasting intraday volatility.

In the present chapter, we study several non-parametric machine learning (ML) models for forecasting *multi-asset intraday volatility* by leveraging high-frequency data from the U.S. equity market. We first propose a measure for evaluating the commonality in intraday volatility. The results demonstrate that, by taking advantage of commonality in intraday volatility, the forecasting performance of these ML models improves significantly. Neural networks (NNs) yield both statistically and economically significant improvements in out-of-sample performance over linear regressions and tree-based models, due to their ability to uncover the non-linearity and model complex latent interactions among variables. The improvements remain robust when we apply trained models to new stocks that have not been included in the training set, thus alleviating the overfitting concerns of NNs and providing new evidence toward a certain universality phenomenon in modeling volatility. In the end, our findings reveal that past intraday volatilities provide additional useful information for forecasting daily volatility, and reveal subtle time-of-day effects that aid the forecasting mechanism.

We augment our proposed methodology with a very thorough set of numerical experiments. The data covered in this work span the period from July 2011 to June 2021, and include the top 100 most liquid components of the Standard & Poor's (S&P) 500 index, and the 10-min, 30-min, 65-min, and daily (without overnight information) forecasting horizons are analyzed.

More specifically, a measure for evaluating the commonality in intraday volatility is proposed, which is the adjusted R^2 value from linear regressions of the RVs of a given stock against the market RVs. It is demonstrated that commonality over the daily horizon is turbulent over time, although the commonality in intraday RVs is strong and stable. The analysis of the high-frequency data from the real market reveals the following interesting phenomena. During a trading session, commonality achieves a peak near closing sessions, in contrast to the diurnal volatility pattern.

Second, in order to assess the benefits of incorporating commonality into models used to predict intraday volatility, multiple ML algorithms (including autoregressive integrated moving averages [ARIMA], heterogeneous autoregressive [HAR], ordinary least squares [OLS], least absolute shrinkage and selection operator [LASSO], XGBoost, multilayer perceptron [MLP], and long short-term memory [LSTM]) are implemented under three different schemes: (a) **Single**: training specific models for each asset; (b) **Universal**: training one model with pooled data for all assets; (c) **Augmented**: training one model using pooled data with an additional predictor which takes into account the impact of market RV. It is revealed that for most models, the incorporation of intraday commonality likely leads to better out-of-sample performance, based on the pooled data together with additional information of the market volatility.

The empirical results we present in the chapter demonstrate that NNs can be superior to other techniques. Empirical evidence is provided to demonstrate the capability of NNs for capturing complex interactions among predictors. Furthermore, to alleviate the concerns of over-fitting, a stringent out-of-sample test is conducted, where the trained models are evaluated on completely new stocks which have not been included in the training sample. Our results reveal that NNs can outperform other approaches. By comparing the result with the performance obtained by OLS models trained for each new stock, we show the validity of a universal volatility mechanism among stocks. Similar findings are reported in Sirignano and Cont [206] concerning universal features of price formation in equity markets.

We conclude the chapter by proposing a new approach for predicting daily volatility, in which the past intraday volatilities rather than the past daily volatilities are used as predictors. This approach fully utilizes the available high-frequency data, and therefore contributes to the improvement over traditional methods of modeling daily volatilities. The results presented in this chapter demonstrate that ML models, where past intraday volatilities are used as predictors, tend to outperform the traditional models with past daily volatilities (e.g., HAR of Corsi [75], SHAR of Patton and Sheppard [191], and HARQ of Bollerslev, Patton, and Quaedvlieg [38]).

We believe that our approach brings a novel perspective on research that studies the effectiveness of past intraday volatilities in forecasting future daily volatility, providing new insights into understanding of the volatility dynamics.

The remainder of this chapter is structured as follows. We begin with Section 4.2 by reviewing some closely related literature, which however should not be considered as a comprehensive survey of the subject. In Section 4.3, the data and the definition of realized volatility are described. In Section 4.4, we discuss the commonality in intraday volatility. Various ML models and three training schemes for predicting future intraday volatility are introduced in Section 4.5. Section 4.6 provides the forecasting results and discusses the empirical findings. In Section 4.7, a new approach to forecasting daily volatility using past intraday volatility as predictors is proposed. Finally, we summarize our study and discuss further avenues of investigation in Section 4.8.

4.2 Related literature

Our study is built upon several research streams proposed by various authors over recent years. The first stream is related to the research on the commonality in financial markets. Chordia, Roll, and Subrahmanyam [66] recognized the existence of commonality in liquidity, and Karolyi, Lee, and Van Dijk [145] suggested that commonality in liquidity is related to market volatility, in particular, the presence of international investors and trading activity. Dang, Moshirian, and Zhang [81] made an observation that the news commonality is associated with stock return co-movement and liquidity commonality.

The co-movement in daily volatility is well-known in the previous literature. Traditional GARCH and SV models (e.g. Andersen et al. [11] and Calvet, Fisher, and Thompson [49]) all make use of the volatility spillover effects. Herskovic et al. [136] provided empirical evidence of the co-movement in volatility across the equity market. Bollerslev et al. [35] observed strong similarities in daily RV and utilized them to forecast the daily RV. Engle and Sokalska [95] emphasized that pooled data are useful for intraday volatility forecasting and Herskovic et al. [135] reported that

volatilities co-move strongly over time. However, there is still a void of research related to commonality in intraday volatility and its implications for managing intraday risks, especially for forecasting purposes.

Second, there are numerous contributions in the existing literature on the topic of forecasting daily volatility. However, most methods proposed by various researchers for modeling and forecasting return volatility largely rely on the parametric GARCH or SV models, which provide forecasts of daily volatility from daily returns. As pointed out by Andersen et al. [11, 13] and Engle and Patton [94], these models employed to predict daily volatility cannot take advantage of high-frequency data, and suffer from the curse of high-dimensionality when dealing with multiple assets simultaneously. Due to the availability of high-frequency data, RV, computed from summing squared intraday returns, has gained popularity in recent years. Andersen et al. [13] proposed an Autoregressive Fractionally Integrated Moving Average (ARFIMA) model for forecasting daily RVs, which outperforms conventional GARCH and related approaches. Corsi [75] put forward a parsimonious AR-type model, termed HAR, for predicting daily RVs using various realized volatility components over different time horizons. Recently, Izzeldin et al. [143] made a comparison investigation for the forecasting performance of ARFIMA and HAR, and concluded that their performance is essentially indistinguishable. See Section 4.7 for more models to predict daily volatility.

Nonetheless, little attention has been paid to the role of forecasting intraday volatility. Taylor and Xu [213] proposed an hourly volatility model based on an ARCH specification and Engle and Sokalska [95] constructed a GARCH model for intraday financial returns, by specifying the variance as a product of daily, diurnal, and stochastic intraday components. These models, such as traditional GARCH and SV, are potentially restrictive due to their parametric nature, and are not able to effectively take into account the non-linear and highly complex relationships among different financial variables.

Third, ML models have demonstrated great potential in finance, such as their applications in asset pricing. The high-dimensional nature of ML methods allows

for better approximations of unknown and potentially complex data-generating processes, in contrast with traditional econometric models. Gu, Kelly, and Xiu [119] pointed out the superior performance of ML models for empirical asset pricing. Recently, Xiong, Nichols, and Shen [229] applied LSTMs to forecast S&P 500 volatility, with Google domestic trends as predictors, and Bucci [44] demonstrated that recurrent NNs (RNNs) are able to outperform all the traditional econometric methods in forecasting monthly volatility of the S&P index. Rahimikia and Poon [194] compared ML models with HAR models for forecasting daily RV by using variables extracted from limit order books and news. Li and Tang [167] proposed a simple average ensemble model combining multiple ML algorithms for forecasting daily (and monthly) RV, and Christensen, Siggaard, Veliyev, et al. [70] examined the performance of ML models in forecasting 1-day-ahead RV with firm-specific characteristics and macroeconomic indicators.

4.3 Data and RV

4.3.1 Data

We use the Nasdaq ITCH data from LOBSTER to compute intraday returns via mid-prices. We select the top 100 components of S&P 500 index, for the period between July 1, 2011 and June 30, 2021. After filtering out the stocks for which the dataset does not span the entire sample period, we are left with 93 stocks. Table 4.1 presents the number of stocks in each sector, according to the Global Industry Classification Standard (GICS) sector division.[1]

[1]The GICS is an industry taxonomy developed in 1999 by MSCI and S&P.

Table 4.1: Components in each sector.

Sector	Number	Tickers
Information Technology	20	AAPL ACN ADBE ADP AVGO CRM CSCO FIS FISV IBM INTC INTU MA MSFT MU NVDA ORCL QCOM TXN V
Health Care	19	ABT AMGN BDX BMY BSX CI CVS DHR GILD ISRG JNJ LLY MDT MRK PFE SYK TMO UNH VRTX
Financials	15	AXP BAC BLK BRK.B C CB CME GS JPM MMC MS PNC SCHW USB WFC
Industrials	9	BA CAT CSX GE HON LMT MMM UNP UPS
Consumer Discretionary	8	AMZN HD LOW MCD NKE SBUX TGT TJX
Consumer Staples	8	CL COST KO MO PEP PG PM WMT
Communication Services	6	CMCSA DIS GOOG NFLX T VZ
Others	8	AMT CCI COP CVX D DUK SO XOM

4.3.2 Realized volatility

In a general form, $P_{i,t}$ denotes the price process of a financial asset i and it follows

$$\mathrm{d}\log P_{i,t} = \mu_i \mathrm{d}t + \sigma_{i,t}\mathrm{d}W_t, \tag{4.1}$$

where μ_i is the drift, $\sigma_{i,t}$ is the instantaneous volatility, and W_t is the standard Brownian motion. The theoretical integrated variance (IV) of stock i during $(t-h,t]$ is estimated as

$$\mathrm{IV}_{i,t}(h) = \int_{t-h}^{t} \sigma_{i,s}^2 \mathrm{d}s, \tag{4.2}$$

where h is the look-back horizon, such as 10 min, 30 min, 1 day, etc.

In this chapter, we consider the minutely logarithmic mid-price return for asset i during $(t-1,t]$ as

$$r_{i,t} := \log\left(\frac{P_{i,t}}{P_{i,t-1}}\right). \tag{4.3}$$

Here, $P_{i,t}$ is the mid-price at time t, i.e. $P_{i,t} = \frac{P_{i,t}^b + P_{i,t}^a}{2}$, and $P_{i,t}^b$ (respectively, $P_{i,t}^a$) represents the best bid (respectively, ask) price.

Andersen et al. [12] and Barndorff-Nielsen and Shephard [25] showed that the sum of squared intraday returns is a consistent estimator of the unobservable IV. Because of the availability of high-frequency intraday data, we choose to compute RV as a proxy for the unobserved IV (see Andersen et al. [12], Bollen and Inder [33], Hansen and Lunde [125], and Xiu [230]). To reduce the impact of extreme values, we

consider the logarithm, in line with Andersen et al. [13], Bucci [44], and Herskovic et al. [136]. Specifically, during a period $(t-h,t]$, the RV is defined as follows[2] :

$$RV_{i,t}^{(h)} := \log\left[\sum_{s=t-h+1}^{t} r_{i,s}^2\right]. \tag{4.4}$$

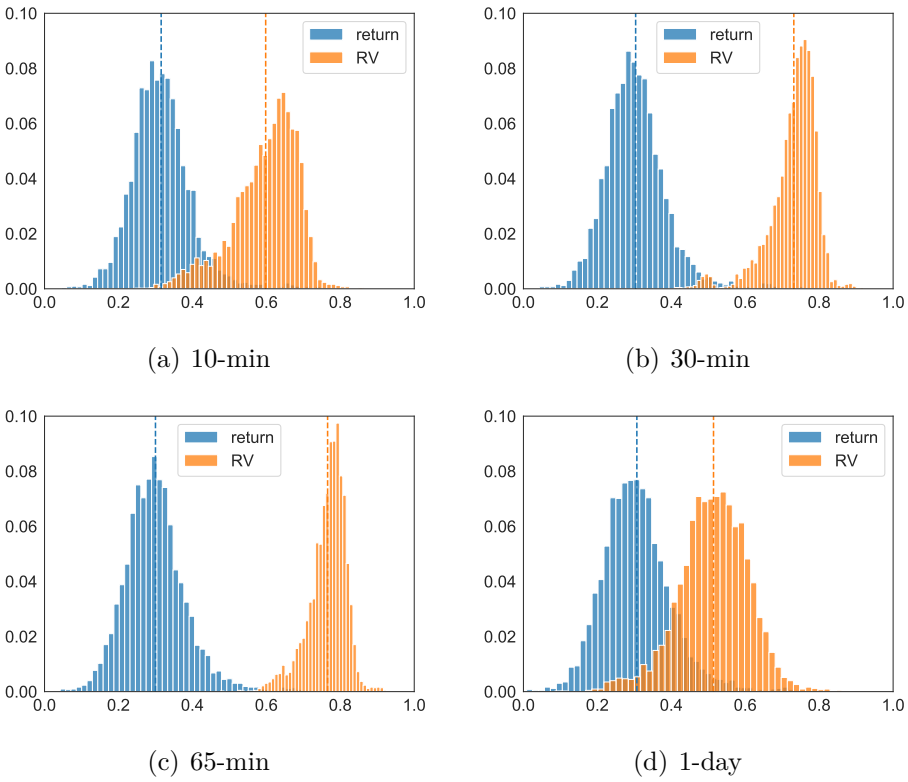
As pointed out by Pascalau and Poirier [187], there are no conclusive methods to incorporate the overnight session's information content into the daily volatility. In line with Engle and Sokalska [95], overnight information is excluded from our empirical analysis of daily volatility. For simplicity, we refer to this daily scenario (excluding the overnight) as the "1-day" scenario, throughout the rest of this chapter.

4.3.3 Summary statistics

To mitigate the effect of possibly spurious data errors, for each stock, we set the values of return/volatility below the 0.5% percentile equal to the respective 0.5 percentile, and the values above the 99.5 percentile are set equal to the 99.5% percentile, a process commonly referred to as *winsorization*. Figure 4.1 illustrates the pairwise Pearson and Spearman correlations of returns and realized volatilities. This figure depicts the empirical distribution of pairwise correlation coefficients over the entire sample period. We generally observe higher correlations in RV than the counterparts in return. Figure 4.1 also reveals that, on average, as the horizon gets longer, RV's correlations increase from 0.598 (10-min) to 0.731 (30-min) to 0.766 (65-min). However, when turning to dailyRV, correlations in RVs become weaker, with an average of 0.514. This indicates that the connections between stocks in terms of intraday volatility may be more stable and tight than the ones in daily volatility.

Figure 4.2 plots the daily RV over time. Stocks demonstrate similar time series patterns, consistent with Bollerslev et al. [35] and Herskovic et al. [136]. Additionally, the width shrinks during the periods of higher volatility, such as, stock market crashes in August 2011 (European sovereign debt crisis), between June 2015 and

[2]Liu, Patton, and Sheppard [171] demonstrate that no sub-sampling frequency significantly outperforms a 5-min interval in terms of forecasting daily RVs, making it a widely accepted time interval in the literature. In the present chapter, we use 1-min returns since our main focus is intraday RVs, such as 10-min RVs.



(a) 10-min (b) 30-min

(c) 65-min (d) 1-day

Figure 4.1: Histograms of pairwise correlations of realized volatilities and returns.

Note: (a)-(d) are based on observations in the frequency of 10-min, 30-min, 65-min, 1-day, respectively. The dashed vertical lines represent the average correlation values of RVs and returns.

June 2016 (Chinese stock market turbulence and Brexit), in March 2018 (China–U.S. trade war), in March 2020 (COVID-19). Figure 4.3 shows that the diurnal volatility forms a so-called reverse-J-shape, namely larger fluctuations near the open and close (see Engle and Sokalska [95] and Harris [128]).

4.4 Commonality estimation

Inspired by prior studies (e.g., Chordia, Roll, and Subrahmanyam [66], Dang, Moshirian, and Zhang [81], Karolyi, Lee, and Van Dijk [145], and Morck, Yeung, and Yu [178]), we follow an analogous procedure to estimate the commonality in volatility. Specifically, we use the average adjusted R^2 value from the following

Figure 4.2: Daily realized volatility (in logs).

Note: The blue curve represents cross-sectional average of daily realized volatility across stocks, with the green area covering the 25-th percentile to the 75-th percentile, and the red area covering the 5-th percentile to the 95-th percentile.

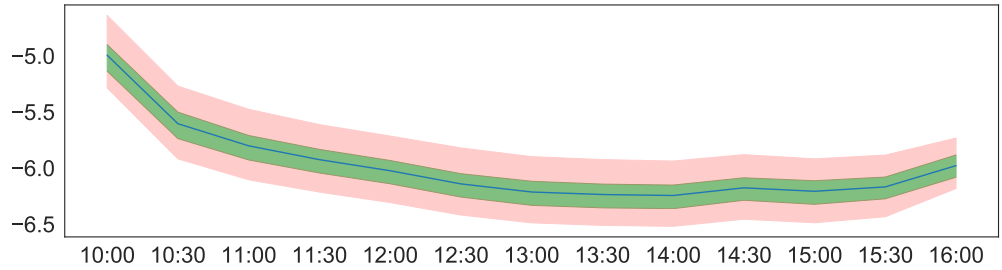


Figure 4.3: Diurnal realized volatility (in logs).

Note: The blue curve represents cross-sectional average of 30-min realized volatility across stocks and days, with the green area covering the 25-th percentile to the 75-th percentile and the red area covering the 5-th percentile to the 95-th percentile.

regressions across stocks, as a measure of commonality in volatility (denoted as $R^2_{(h)}$)[3]

$$RV^{(h)}_{i,t} = \alpha_i + \beta_i RV^{(h)}_{M,t} + \epsilon_{i,t}, \tag{4.5}$$

where $RV^{(h)}_{M,t}$ (see Bollerslev et al. [35]) is the contemporaneous market volatility during $(t-h, t]$ for stock i, which is calculated as the equally weighted average[4] of all individual stock volatilities during $(t-h, t]$, that is,

$$RV^{(h)}_{M,t} = \frac{1}{N}\sum_{i=1}^{N} RV^{(h)}_{i,t}. \tag{4.6}$$

[3]We also perform another regression, where except for contemporaneous market volatility, the lag one (thus $t-1$ in (4.5)) and lead one (thus $t+1$ in (4.5), hence not computable in real time due to the forward looking bias) in market volatility are also included, in order to explain non-contemporaneous trading, in line with Chordia, Roll, and Subrahmanyam [66], Dang, Moshirian, and Zhang [81], and Karolyi, Lee, and Van Dijk [145]. The R^2 values are similar to the ones of Eqn (4.5).

[4]We also implemented the value weighted market volatility and the results are similar to the equally weighted market volatility.

Figure 4.4 presents the commonality in RV, averaged across stocks for each month. To create this figure, we use the observations in each month, to obtain the R^2 value from Eqn (4.5). We notice that commonality effects in intraday scenarios (especially 30-min and 65-min) are substantially larger than the daily ones. For example, as reported in Table 4.2, the average commonality in 65-min data is around 74.3%, while only 35.5% in daily data. Moreover, $R^2_{(h)}$ is much more turbulent at the daily frequency. The last column in Table 4.2 also reports the results of the relation between the average commonality and the market volatility. As the horizon extends, the average commonality has a higher correlation with the market volatility.[5]

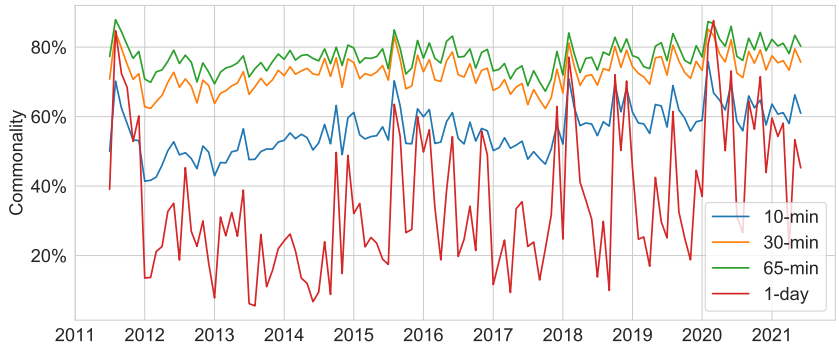


Figure 4.4: Commonality in realized volatility.

Note: The commonality is averaged across stocks for each month during the sample period of 2011-07 $\sim$ 2021-06.

Table 4.2: Statistics of the monthly average commonality in volatility.

	Mean	Std	Corr with VIX
10-min	0.560	0.068	0.536
30-min	0.725	0.048	0.574
65-min	0.743	0.041	0.609
1-day	0.355	0.198	0.690

Note: VIX represents the market volatility from the Chicago Board Options Exchange.

[5]We refer the reader to additional analysis on commonality in Appendix C.1.

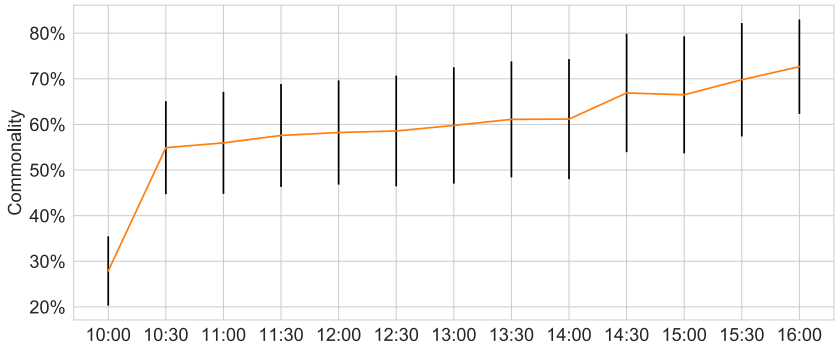


Figure 4.5: Commonality in realized volatility.

Note: The commonality is averaged across stocks for each half-hour during the sample period of 2011-07 $\sim$ 2021-06.

Figure 4.5 reports the averaged values and standard deviations (black vertical lines) of commonality for each half-hour in the trading session. To create this figure, we use the observations in a given interval, such as [09:30, 10:00], to fit Eqn (4.5). We observe a gradual increase in commonality throughout the trading session as we get closer to market close, in sharp contrast to the diurnal volatility pattern in Figure 4.3.

4.5 Methodology

In this section, we leverage the commonality for the task of predicting cross-asset volatility. We construct the prediction model as follows:

$$\begin{aligned} RV_{i,t+h}^{(h)} &= F_i\left(\mathbf{u};\theta\right) + \epsilon_{i,t+h} \\ &= F_i\left(RV_{i,t}^{(h)},\ldots,RV_{i,t-(p-1)h}^{(h)},RV_{M,t}^{(h)},\ldots,RV_{M,t-(p-1)h}^{(h)};\theta\right) + \epsilon_{i,t+h}, \end{aligned} \tag{4.7}$$

where $RV_{i,t+h}^{(h)}$ is the volatility of asset i during $(t, t+h]$. $\mathbf{u}$ represents the input features, which can be further separated into two categories: (i) a multi-dimensional vector of predictor variables for a specific stock i available up to time t, denoted as *individual features*, such as $\left(RV_{i,t}^{(h)},\ldots,RV_{i,t-(p-1)h}^{(h)}\right)'$; (ii) a vector of features for all stocks in the studied universe up to t, denoted as *market features*, such as $\left(RV_{M,t}^{(h)},\ldots,RV_{M,t-(p-1)h}^{(h)}\right)'$. θ refers to the parameters that need to be estimated. Whenever is clear from the context and no ambiguity arises, we use also use θ to denote the forecasting model. We are aiming to find a function of variables that minimizes the out-of-sample errors for future RV.

4.5.1 Models

This section summarizes the collection of ML models employed in our numerical experiments.

Seasonal ARIMA

The ARIMA model is a popular forecasting method for univariate time series data, where an initial differencing step can be applied one or more times to eliminate the non-stationarity of the trend. An ARIMA(p, d, q) is given by

$$\varphi(L)(1-L)^d RV_{i,t}^{(h)} = \rho(L)\varepsilon_{i,t}, \tag{4.8}$$

where $\varphi(L) = 1 - \sum_{k=1}^{p} \varphi_i L^i$ and $\rho(L) = 1 - \sum_{j=1}^{q} \rho_j L^j$ are the AR and MA lag polynomials, and $\epsilon_{i,t}$ is the error which is distributed as $\mathcal{N}(0, \sigma_i^2)$. Following Christensen and Prabhala [69] and Ribeiro et al. [197], we adopt ARIMA$(1, 1, 1)$ to model the daily realized volatility.

When the time series exhibits seasonality, the seasonal-differencing could be applied to eliminate the seasonal component, which is denoted as Seasonal ARIMA (SARIMA). As revealed in Figure 4.3, intraday volatility time series possess a seasonal component. For modeling intraday RV, we choose the SARIMA, where the seasonal period is the corresponding number of intraday time buckets in a day and other parameters related to the seasonal pattern are set as zero (for more details about SARIMA, see Sheppard [203]).

HAR with diurnal effects

Corsi [75] proposed a volatility model, named as HAR, which considers realized volatilities over different interval sizes. HAR has shown remarkably good forecasting performance on daily data Izzeldin et al. [143] and Patton and Sheppard [191]. For day t, the forecast of HAR is based on

$$RV_{i,t+1}^{(d)} = \alpha_i + \beta_i^{(d)} RV_{i,t}^{(d)} + \beta_i^{(w)} RV_{i,t}^{(w)} + \beta_i^{(m)} RV_{i,t}^{(m)} + \epsilon_{i,t+1}, \tag{4.9}$$

where $RV_{i,t}^{(d)}$ denotes the daily RV in the past day, and $RV_{i,t}^{(w)} = \frac{1}{5}\sum_{l=1}^{5} RV_{i,t-l}^{(d)}$, $RV_{i,t}^{(m)} = \frac{1}{21}\sum_{l=1}^{21} RV_{i,t-l}^{(d)}$ denote the weekly and monthly lagged rRV, respectively. The choice

of a daily, weekly and monthly lag is aiming to capture the long-memory dynamic dependencies observed in most RV series.

However, very little attention has been paid to forecasting intraday volatility with HAR. One closely connected model is that of Engle and Sokalska [95], who proposed an intraday volatility forecasting model, where they interpret that conditional volatility of high-frequency returns is a product of daily, diurnal, and stochastic intraday components. After the decomposition of raw returns, the authors apply a GARCH model Engle [92] to learn the stochastic intraday volatility components.

Following the spirit of Engle and Sokalska [95], we extend the daily HAR model to intraday scenarios by adding diurnal effect and previous intraday component, denoted as HAR-D, as follows[6]:

$$RV^{(h)}_{i,t+h} = \alpha_i + \beta^{(\tau)}_i D_{i,\tau_{t+h}} + \beta^{(s)}_i RV^{(h)}_{i,t} + \beta^{(d)}_i RV^{(d)}_{i,t} + \beta^{(w)}_i RV^{(w)}_{i,t} + \beta^{(m)}_i RV^{(m)}_{i,t} + \epsilon_{i,t+h}, \tag{4.10}$$

where $D_{i,\tau_{t+h}}$ denote the average diurnal RV in the bucket-of-the-day τ_{t+h} computed from the last 21 days. For example, when t = 10:30 and $h = 30$ minutes, then τ_{t+h} corresponds to the bucket 10:30–11:00. $RV^{(h)}_{i,t}$ represents the lag=1 intraday RV. $RV^{(d)}_{i,t}$ ($RV^{(w)}_{i,t}$, $RV^{(m)}_{i,t}$) denotes the aggregated daily (weekly, monthly) RV. When we consider the daily scenarios, Eqn (4.10) becomes the standard HAR model (Eqn (4.9)), by removing the diurnal term and the intraday component.

Ordinary least squares

Instead of using aggregated RV, we apply OLS to original features, as follows, with its loss function being the sum of squared errors. Recall $\mathbf{u} = (u_1, \ldots, u_p)'$ represent the vector of input features, such as past intraday RVs, and (perhaps) market RVs. Notice that the model only incorporating the past intraday RVs as features is actually an autoregressive (AR) model.

$$RV^{(h)}_{i,t+h} = \alpha_i + \sum_{l=1}^{p} \beta_l u_l + \epsilon_{i,t+h}. \tag{4.11}$$

[6]Since we use the log-version realized volatility, the multiplication of daily, diurnal, and stochastic intraday components in Engle and Sokalska [95] translates to the addition in our model (4.10).

Least absolute shrinkage and selection operator

When the number of predictors approaches the number of observations, or there are high correlations among predictor variables, the OLS model tends to overfit noise rather than signals. This is particularly burdensome for the volatility forecasting problem, where the features could be highly correlated.

LASSO is a linear regression method that can avoid overfitting via adding a penalty of parameters to the objective function. As pointed out by Hastie, Tibshirani, and Friedman [132], LASSO performs both variable selection and regularization, therefore enhances the prediction accuracy and interpretability of regression models. The objective function of LASSO is the sum of squared residuals and an additional l_1 constraint on the regression coefficients, as shown in Eqn (4.12). Here, the hyperparameter λ controls the penalty weight. In our experiments, we provide a set of hyperparameter values, and then choose the one with the best performance on the validation data, as our forecasting model.

$$L_i = \sum_t \left[RV_{i,t+h}^{(h)} - \alpha_i - \sum_{l=1}^{p} \beta_l u_l \right]^2 + \lambda \sum_{l=1}^{p} \|\beta_l\|_1 . \tag{4.12}$$

XGBoost

Linear models are unable to capture the possible non-linear relations between the dependent variable and the predictors, and the interactions among predictors. As pointed by Bucci [44], RVs are subject to structural breaks and regime-switching, hence the need to consider non-linear models. One way to add non-linearity and interactions is the decision tree, see more in Hastie, Tibshirani, and Friedman [132].

XGBoost is a decision-tree-based ensemble algorithm, implemented under a distributed gradient boosting framework by Chen and Guestrin [61]. There is abundant empirical evidence showing the success of XGBoost, such as in a large number of Kaggle competitions. In this work, we only review the essential idea behind XGBoost - tree boosting model. For more details about other important features of XGBoost, such as the scalability in various scenarios, parallelization, distributed

computing, feature importance to enhance interpretability, etc., the reader may refer to Chen and Guestrin [61]. Let $\mathbf{u}$ represent the vector of input features

$$F_i(\mathbf{u}) = \sum_{l=1}^{B} f_l(\mathbf{u}), \quad f_l \in \mathcal{F}, \tag{4.13}$$

where $\mathcal{F}$ is the space of regression trees. An example of the tree ensemble model is depicted in Figure 4.6. The tree ensemble model in Eqn (4.13) is trained sequentially. Boosting (see Friedman [107]) means that new models are added to minimize the errors made by existing models, until no further improvements are achieved.

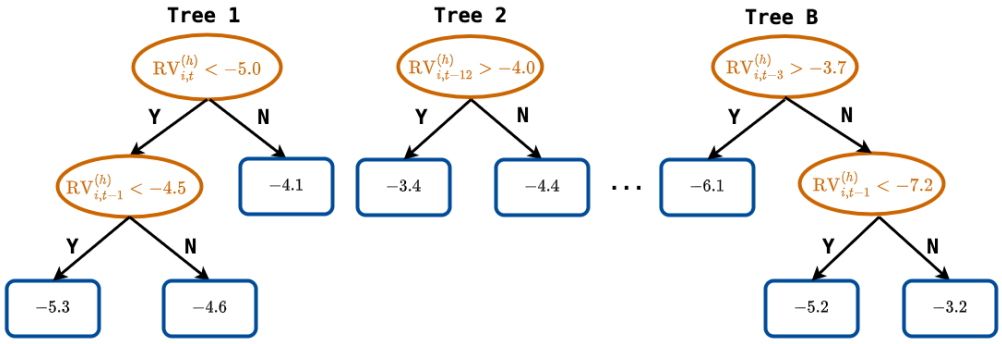


Figure 4.6: Illustration of a tree ensemble model.

Note: B represents the number of trees. The final prediction of a tree ensemble model is the sum of predictions from each tree, as shown in Eqn (4.13).

Multilayer perceptron

Another non-linear method is the NN, which has become increasingly popular for ML problems, e.g. in computer vision and natural language processing, due to its flexibility to learn complex interactions. Hill, O'Connor, and Remus [137] and Zhang [234] suggest that NN is a promising non-linear alternative to the traditional linear methods in time series forecasting. We introduce two commonly-implemented NNs in the following sections.

MLP is a class of feedforward NNs and a "universal approximator" that can learn any smooth functions (see Hornik, Stinchcombe, and White [139]). MLP has been applied to many fields, e.g. computer vision, natural language processing, etc. MLPs are composed of an input layer to receive the raw features, an output layer that makes forecasts about the input, and in-between those two, an arbitrary

number of hidden layers that are non-linear transformations. MLPs perform a static mapping between an input space and an output space Bucci [44]. The parameters in MLPs can be updated via stochastic gradient descent (SGD). In this work, we use Adam (see Kingma and Ba [150]), which is based on adaptive estimates of lower-order moments. Let $\mathbf{u} \in \mathbb{R}^p$ represent the input variables

$$F_i(\mathbf{u};\theta) = \mathbf{W}_L \cdot \sigma(\mathbf{W}_{L-1} \ldots \sigma(\mathbf{W}_1\mathbf{u} + \mathbf{b}_1) \ldots + \mathbf{b}_{L-1}) + \mathbf{b}_L, \tag{4.14}$$

where $\theta := (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_L, \mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_L)$ represents the parameters in the NN. $\mathbf{W}_l \in \mathbb{R}^{n_l \times n_{l-1}}$, $\mathbf{b}_l \in \mathbb{R}^{n_l \times 1}$ for $l = 1, 2, \ldots, L$, and $n_0 = p$. For the activation function $\sigma(\cdot)$, we choose the rectified linear unit (ReLU), i.e. $\sigma(x) = \max(x, 0)$.

Long short-term memory

A recurrent neural network (RNN) requires that historic information is retained to forecast future values. Therefore RNN is well-suited for processing, classifying, and making predictions based on time series data Bucci [44]. LSTM, proposed by Hochreiter and Schmidhuber [138], is an extension of the RNN architecture by replacing each hidden unit in RNNs with a memory block to capture the long-term effect. LSTMs have received considerable success in natural language processing, time series, generative models, etc.

For simplicity, we consider the time series for a given stock and remove the subscript for stock identity. The standard transformation in each unit of LSTM is defined as follows. For a more detailed discussion, we refer the reader to Hochreiter and Schmidhuber [138].

$$\begin{aligned}
\mathbf{f}_t &= \sigma_g(\mathbf{W}_f\mathbf{u}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{b}_f) \\
\mathbf{i}_t &= \sigma_g(\mathbf{W}_i\mathbf{u}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{b}_i) \\
\mathbf{o}_t &= \sigma_g(\mathbf{W}_o\mathbf{u}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{b}_o) \\
\tilde{\mathbf{c}}_t &= \sigma_c(\mathbf{W}_c\mathbf{u}_t + U_c\mathbf{h}_{t-1} + \mathbf{b}_c) \\
\mathbf{c}_t &= \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{c}}_t \\
\mathbf{h}_t &= \mathbf{o}_t \circ \sigma_h(\mathbf{c}_t),
\end{aligned} \tag{4.15}$$

where $\mathbf{u}_t$ is the input vector, $\mathbf{f}_t$ is the forget gate's activation vector, $\mathbf{i}_t$ is the update gate's activation vector, $\mathbf{o}_t$ is the output gate's activation vector, $\tilde{\mathbf{c}}_t$ is the cell

input activation vector, $\mathbf{c}_t$ is the cell state vector, and $\mathbf{h}_t$ is the hidden state vector, i.e. output vector of the LSTM unit. $\circ$ is the Hadamard product function. σ_g is sigmoid function, and σ_c, σ_h are hyperbolic tangent function. $\mathbf{W}_{f(i,o,c)}, \mathbf{b}_{f(i,o,c)}$ refer to weight matrices and bias vectors that need to be estimated.

To summarize, we first consider a traditional time series model ARIMA, then include three linear regression models, i.e. HAR(-D), OLS, and LASSO. To account for the nonlinear impact of individual predictors on future volatilities and the interactions among predictors, we choose an ensemble tree model XGBoost and two NNs, i.e. MLP and LSTM. The primary difference between MLP and LSTM is that LSTM has feedback connections, which allow learning the dependencies in the input sequences.

4.5.2 Training scheme

Motivated by the strong commonality in volatility across stocks, we consider the following three different schemes for the model training.

- **Single** denotes that we train customized models F_i for each stock i, as in Bucci [44] and Hansen and Lunde [124]. We use a stock's own past RVs only as predictor features, namely

  $$\mathbf{u} = \left(RV_{i,t}^{(h)}, \ldots, RV_{i,t-(p-1)h}^{(h)}\right)'$$

  and no market features, where p represents the number of lags.

- **Universal** denotes that we train models with the pooled data of all stocks in our universe. That is, F_i is same for all stocks in Eqn (4.7). As in the **Single** scheme, we use a stock's own past RVs only as predictor features and no market features. Sirignano and Cont [206] showed that the model trained on the pooled data outperforms asset-specific models trained on time series of any given stock, in the sense of forecasting the direction of price moves. Bollerslev et al. [35] and Engle and Sokalska [95] suggested that models estimated under the **Universal** setting yield superior out-of-sample risk forecasts, compared to models under the **Single** setting, when forecasting daily realized volatility.

- **Augmented** denotes that we train models with the pooled data of all stocks in our universe, but in addition, we also incorporate a predictor which takes into account the impact of the market realized volatility (e.g. Bollerslev et al. [35]) in order to leverage the commonality in volatility shown in Section 4.4. Namely, F_i is same for all stocks in Eqn (4.7). We use both individual features and market features as predictors, i.e. $\mathbf{u} = \left(RV^{(h)}_{i,t}, \ldots, RV^{(h)}_{i,t-(p-1)h}, RV^{(h)}_{M,t}, \ldots, RV^{(h)}_{M,t-(p-1)h}\right)'$. Note that for HAR-D models under the **Augmented** setting, we include aggregated market features as additional features, and use the OLS to estimate the parameters.

In summary, compared with the benchmark **Single** setting, we gradually incorporate cross-asset and market information into the training of models. The hyperparameters for each model are summarized in Appendix C.2.

4.5.3 Performance evaluation

To assess the predictive performance for RV forecasts, we compute the following metrics on the rolling out-of-sample tests (see Bollerslev et al. [35], Bucci [44], Engle and Sokalska [95], Pascalau and Poirier [187], Patton [189], Patton and Sheppard [190], and Rahimikia and Poon [194]). Both functions measure losses, so lower values are preferred. Patton and Sheppard [190] demonstrate that QLIKE has the highest power in the Diebold–Mariano (DM) test. Consequently, we focus more on the QLIKE rather than the MSE.

- Mean squared error (MSE): $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\#\mathcal{T}_{test}}\sum_{t\in\mathcal{T}_{test}}\left(RV^{(h)}_{i,t} - \widehat{RV}^{(h)}_{i,t}\right)^2$,
- Quasi-likelihood (QLIKE): $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\#\mathcal{T}_{test}}\sum_{t\in\mathcal{T}_{test}}\left[\frac{\exp(RV^{(h)}_{i,t})}{\exp(\widehat{RV}^{(h)}_{i,t})} - (RV^{(h)}_{i,t} - \widehat{RV}^{(h)}_{i,t}) - 1\right]$,

where $\widehat{RV}^{(h)}_{i,t}$ represents the predicted value of $RV^{(h)}_{i,t}$, the realized volatility for stock i during $(t-h, t]$, and $\overline{RV}^{(h)}_{i}$ is the empirical mean of $RV^{(h)}_{i,t}$ on the test data. N is the number of stocks in our universe, $\mathcal{T}_{test}$ is the testing period, and $\#\mathcal{T}_{test}$ is the length of the testing period.

Model Confidence Set. Hansen, Lunde, and Nason [126] proposed a non-parametric statistical test to formally compare the forecasting accuracy of different models, denoted as Model Confidence Set (MCS). The MCS identifies a set of models that contains the best forecasting model(s) with a given level of confidence. Specifically, given a collection of models $\mathcal{M}_0$, the test sequentially discards models with inferior predictive ability, and in the end determines a subset containing the best model(s) $\mathcal{M}^*$. The iterative elimination is based on sequentially testing the following hypothesis,

$$H_0 : \mathbb{E}\left(\Delta L_{ij,t}\right) = 0, \text{for all } i, j \in \mathcal{M}^*, \tag{4.16}$$

where $L_{ij,t}$ is the loss difference between models i and j at time t in terms of a specific loss function L, such as MSE and QLIKE. The MCS procedure makes it possible to make statements about statistical significance from multiple pairwise comparisons. For additional details, we refer to the studies of Hansen, Lunde, and Nason [126] and Hansen, Lunde, and Nason [127]. Consistent with Symitsi et al. [210], we implement the MCS procedure using a block bootstrap procedure with 10,000 replications and a block length of 2 observations.

4.5.4 Utility benefits

We have demonstrated that one can assess the out-of-sample statistical performance for each model via the above metrics and tests. However, in such an approach, the economic magnitude of the gain from complex risk models is ignored. Bollerslev et al. [35] proposed a utility-based framework, which gauges the utility benefits of an investor with mean-variance preferences investing in an asset with time-varying volatility and a constant Sharpe ratio. We implement this framework to measure the volatility forecasts. For a more detailed description, we refer the reader to Bollerslev et al. [35].

The expected utility of a mean-variance investor at t can be approximated as

$$\mathbb{E}_t\left(u\left(W_{t+1}\right)\right) = \mathbb{E}_t\left(W_{t+1}\right) - \frac{1}{2}\gamma^A \operatorname{Var}_t\left(W_{t+1}\right), \tag{4.17}$$

where W_t denotes the wealth and γ^A denotes the absolute risk aversion of the investor.

Assume that the investor allocates a fraction x_t of the current wealth to a risky asset with return r_{t+1} and the rest to a risk-free money market account earning r_t^f. Then the wealth at $t+1$ becomes $W_{t+1} = W_t\left(1 + r_t^f + x_t r_{t+1}^e\right)$, where $r_{t+1}^e \equiv r_{t+1} - r_t^f$. After dropping constant terms, the expected utility in Eqn (4.17) amounts to

$$\mathbb{E}_t\left(u\left(W_{t+1}\right)\right) := U\left(x_t\right) = W_t\left(x_t\mathbb{E}_t\left(r_{t+1}^e\right) - \frac{\gamma}{2}x_t^2\mathbb{E}_t\left(\exp(RV_{t+1})\right)\right), \tag{4.18}$$

where $\gamma \equiv \gamma^A W_t$ denotes the investor's relative risk aversion.

Suppose the conditional Sharpe ratio $SR \equiv \mathbb{E}_t\left(r_{t+1}^e\right)/\sqrt{\mathbb{E}_t\left(\exp(RV_{t+1})\right)}$ is constant, so that the expected utility is

$$U\left(x_t\right) = W_t\left(x_t SR\sqrt{\mathbb{E}_t\left(\exp(RV_{t+1})\right)} - \frac{\gamma}{2}x_t^2\mathbb{E}_t\left(\exp(RV_{t+1})\right)\right). \tag{4.19}$$

The optimal portfolio that maximizes this utility is obtained by investing the following fraction of wealth to the risky asset

$$x_t^* = \frac{SR/\gamma}{\sqrt{\mathbb{E}_t\left(\exp(RV_{t+1})\right)}}. \tag{4.20}$$

To determine the utility gains based on different risk models, the expectation based on model θ is denoted by $\mathbb{E}_t^\theta(\cdot)$. Assuming that the investor uses model θ, then the position $x_t^\theta = \frac{SR/\gamma}{\sqrt{\mathbb{E}_t^\theta(\exp(RV_{t+1}))}}$ is chosen. By plugging x_t^θ into Eqn (4.19) and replacing $\mathbb{E}_t(\exp(RV_{t+1}))$ with the RV $\exp(RV_{t+1})$, the expected utility per unit of the wealth (called realized utility, or in short RU) is given by

$$\mathrm{RU}_t = \frac{SR^2}{\gamma} \times \frac{\sqrt{\exp(RV_{t+1})}}{\sqrt{\mathbb{E}_t^\theta(\exp(RV_{t+1}))}} - \frac{SR^2}{2\gamma} \times \frac{\exp(RV_{t+1})}{\mathbb{E}_t^\theta(\exp(RV_{t+1}))}. \tag{4.21}$$

If a risk model is ideal, that is, it predicts perfectly the realized volatilities $\mathbb{E}_t^\theta(\exp(RV_{t+1})) = \exp(RV_{t+1})$, then its realized utility is $\frac{SR^2}{2\gamma}$. Alternatively, the investor is willing to give up $\frac{SR^2}{2\gamma}$ of the wealth in order to utilize the perfect risk model instead of investing only in the risk-free asset. In this chapter, the same Sharpe ratio (SR = 0.4) and the same coefficient of risk aversion ($\gamma = 2$) are applied as in Bollerslev et al. [35] and Li and Tang [167].

The previous comparisons are based on a frictionless setting, ignoring the trading cost. The case of incorporating the effect of transaction costs is also considered. Following Bollerslev et al. [35] and Li and Tang [167], we assume that transaction costs are linear in the absolute magnitude of the change in the positions, and use the full median bid-ask spread for each of the assets over the last 90 trading days. The realized utility with trading costs deducted, denoted as RU-TC, is simply the realized utility after subtracting the simulated costs. We evaluate this realized utility (with and without trading cost) empirically by averaging the corresponding realized expressions over stocks and the same rolling out-of-sample forecasts.

4.6 Experiments

4.6.1 Implementation

For each data set, we divide the observations into three non-overlapping periods and maintain their chronological order: training, validation, and testing. For a given trading day t, the training data, including the samples in the first period [July 1, $2011, t-251$], are used to estimate models subject to a given architecture. Validation data, including the recent one-year samples $[t-250, t]$, are deployed to tune the hyperparameters of the models. Finally, testing data are samples in the next year $[t+1, t+251]$; they are out-of-sample in order to provide objective assessments of the models' performance. Due to limited computational resources, models are updated annually. In other words, when we retrain the models in the next calendar year, the training data expands by one year, whereas the validation samples are rolled forward to include the samples in the most recent one-year period, following Gu, Kelly, and Xiu [119]. To examine the effect of model update frequency, we perform a robustness check for HAR-D models in Appendix C.4, and we conclude that the update frequency has insignificant effect on the model's performance. Our testing period starts from July 1, 2015 until June 30, 2016, and the corresponding training and validation samples are [July 1, 2011, June 30, 2014] and [July 1, 2014, June 30, 2015], respectively. When we predict the realized volatility in [July 1, 2016, June 30, 2017], the training and validation samples are [July 1, 2011, June

30, 2015] and [July 1, 2015, June 30, 2016], respectively. Therefore, our testing sample includes 6 years, from July 2015 to June 2021.

For HAR-D and OLS, we use both the training and validation data for training, due to no requirement of hyperparameter tuning. Given the stochastic nature of NNs, we apply an ensemble approach to MLPs and LSTMs for improving their robustness (see Gu, Kelly, and Xiu [119] and Hansen and Salamon [123]). Specifically, we train multiple NNs with different random seeds for initialization, and construct final predictions by averaging the outputs of all networks. For more information on the model settings, see Appendix C.2.

In all models, the features are based on the observations in the last 21 days. Prior to inputting variables in the models, at each rolling window estimation, we normalize them by removing the mean and scaling to unit variance.

4.6.2 Main results

Table 4.3 presents the results of each model under three training schemes.[7] Due to limited computation power, MLPs and LSTMs are only performed under the **Universal** and **Augmented** settings. We draw the following conclusions from Table 4.3.

We begin by comparing the SARIMA with HAR-D under the **Single** setting.[8] The results show that the ARIMA model achieves similar performance as HAR over the 1-day horizon, consistent with Izzeldin et al. [143]. However, HAR-D yields more accurate out-of-sample forecasts than SARIMA across intraday horizons, i.e. 10-min, 30-min, and 65-min.

Regarding linear models, we observe that for HAR-D, **Universal** shows no improvement in forecasting, compared with **Single**. HAR-D models trained under **Augmented** significantly outperform the ones trained under the other two schemes, across all horizons in our study. The average reduction in QLIKE of **Augmented**

[7]To more formally assess the statistical significance of the differences in out-of-sample volatility forecasts, Table C.4 in Appendix C.3 also reports the results of all Diebold-Mariano tests in terms of QLIKE.

[8]Note that (S)ARIMA is for single time series.

compared with **Single** is 0.031, -0.005, 0.004, and 0.010 over 10-min, 30-min, 65-min, and 1-day, respectively.

Generally speaking, there are significant improvements when moving from HAR-D models to OLS models, over 10-min, 30-min, and 65-min horizons. For example, QLIKEs are reduced from 0.453 (respectively, 0.227, 0.186) with the best HAR-D model (i.e. under **Augmented**) to 0.430 (respectively, 0.204, 0.171) with the best OLS model (i.e. under **Augmented**), across the three horizons (i.e. 10, 30, 65-min), respectively. Within the OLS models, conclusions are similar with HAR-D models, i.e. no benefits from **Universal** while significant benefits from **Augmented**. We also observe similar findings in LASSO as in OLS, suggesting that regularization does not further aid performance. On the other hand, MLPs and LSTMs achieve state-of-the-art accuracy across all measures and intraday horizons (i.e. 10, 30, 65-min), implying the complex interactions between predictors. Further analysis is provided in Section 4.6.3.

Interestingly, linear models slightly outperform MLPs and LSTMs at the 1-day horizon. This is perhaps expected, and might be due to the availability of only a small amount of data at the 1-day horizon, rendering the NNs to underperform due to lack of training data.

Echoing the findings from Panel A, OLS based on the 21-day rolling daily RVs deliver the higher utility than the HAR type models, consistent with Bollerslev et al. [35]. NNs still perform the best, with the highest realized utility achieved by LSTMs.

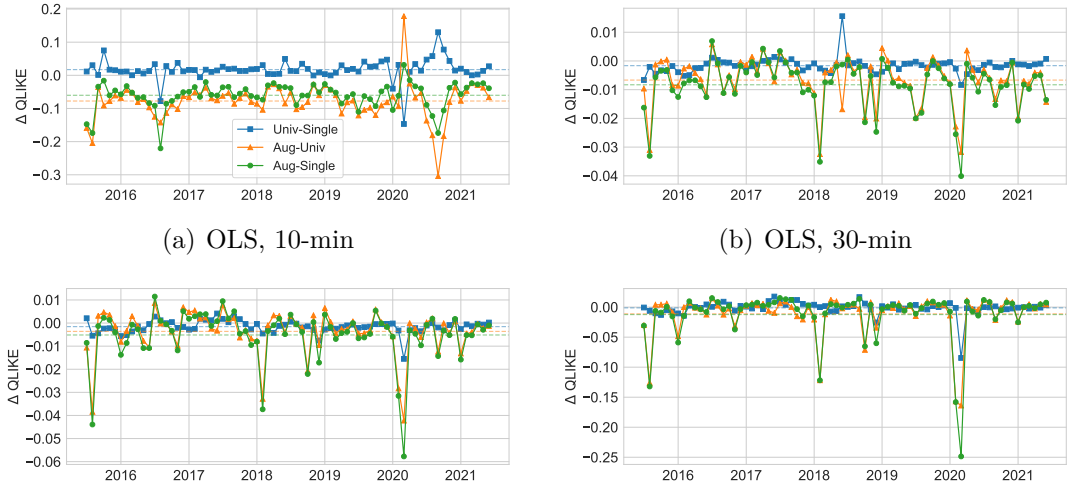
Let us now consider the OLS model as an illustrative example for understanding the relative reduction in error. We compare its QLIKEs under these three schemes, at a monthly level, as shown in Figure 4.7. For better readability, we report the reduction in error of **Universal** relative to **Single** (denoted as Univ–Single), the reduction of **Augmented** relative to **Universal** (denoted as Aug–Univ), and the reduction of **Augmented** relative to **Single** (denoted as Aug–Single). Note that Aug–Single = (Aug–Univ) + (Univ–Single). Negative values of ΔQLIKE indicate an improvement on out-of-sample data, and positive values indicate degradation. To arrive at this figure, we average the ΔQLIKE values in each month, across

Table 4.3: Out-of-sample performance.

Panel A:		10-min		30-min		65-min		1-day	
Statistical performance		MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
SARIMA	**Single**	1.319	0.617	0.524	0.338	0.362	0.231	0.268	0.190
HAR-D	**Single**	1.013	0.484	0.332	0.222	0.270	0.190	0.269	0.190
	Universal	1.021	0.518	0.333	0.230	0.270	0.191	0.269	0.190
	Augmented	0.995	0.453	0.323	0.227	0.262	0.186	0.257	0.180
OLS	**Single**	1.009	0.490	0.307	0.222	0.251	0.176	0.263	0.192
	Universal	1.008	0.507	0.307	0.221	0.250	0.175	0.260	0.191
	Augmented	0.962	0.430	0.293	0.204	0.241	0.171	0.254*	0.178*
LASSO	**Single**	1.053	0.492	0.325	0.224	0.252	0.180	0.263	0.195
	Universal	1.012	0.511	0.309	0.222	0.251	0.176	0.261	0.192
	Augmented	0.961	0.428	0.293	0.204	0.242	0.172	0.255*	0.179*
XGBoost	**Single**	1.047	0.539	0.345	0.236	0.297	0.201	0.358	0.217
	Universal	0.968	0.417	0.290	0.191	0.242	0.170	0.268	0.192
	Augmented	0.968	0.422	0.297	0.190	0.249	0.174	0.285	0.187
MLP	**Single**	-	-	-	-	-	-	-	-
	Universal	0.947	0.397	0.284	0.181	0.232	0.163	0.260	0.191
	Augmented	0.945	0.386	0.280*	0.179	0.229*	0.162	0.257	0.180
LSTM	**Single**	-	-	-	-	-	-	-	-
	Universal	0.950	0.393	0.287	0.179	0.232	0.162	0.261	0.188
	Augmented	0.934*	0.376*	0.279*	0.171*	0.229*	0.160*	0.258	0.182
Panel B:		10-min		30-min		65-min		1-day	
Realized utility		RU	RU-TC	RU	RU-TC	RU	RU-TC	RU	RU-TC
SARIMA	**Single**	2.095	1.473	3.041	2.624	3.314	2.861	3.550	3.515
HAR-D	**Single**	2.690	2.065	3.457	3.040	3.542	3.095	3.548	3.515
	Universal	2.574	1.975	3.429	3.016	3.541	3.095	3.547	3.514
	Augmented	2.790	2.280	3.428	3.022	3.552	3.107	3.571	3.536
OLS	**Single**	2.660	1.901	3.506	3.027	3.579	3.118	3.543	3.504
	Universal	2.601	2.039	3.432	2.984	3.580	3.127	3.546	3.513
	Augmented	2.845	2.271	3.485	3.036	3.587	3.130	3.576	3.536
LASSO	**Single**	2.631	1.893	3.526	3.061	3.567	3.108	3.545	3.501
	Universal	2.593	2.044	3.432	2.989	3.578	3.126	3.543	3.512
	Augmented	2.852	2.292	3.487	3.046	3.586	3.132	3.575	3.537
XGBoost	**Single**	2.492	1.552	3.408	2.888	3.520	3.039	3.508	3.449
	Universal	2.890	2.200	3.532	3.067	3.592	3.116	3.546	3.505
	Augmented	2.864	2.212	3.545	3.083	3.581	3.109	3.571	3.524
MLP	**Single**	-	-	-	-	-	-	-	-
	Universal	2.952	2.380	3.564	3.119	3.607	3.139	3.543	3.506
	Augmented	2.993	2.442	3.569	3.126	3.609	3.145	3.571	3.534
LSTM	**Single**	-	-	-	-	-	-	-	-
	Universal	2.975	2.455	3.575	3.144	3.610	3.149	3.552	3.514
	Augmented	3.028	2.532	3.595	3.170	3.614	3.166	3.567	3.533

Note: The table reports the out-of-sample results for predicting future realized volatility over multiple horizons using different models under three training schemes. For each horizon, the model with the best (second best) out-of-sample performance in QLIKE (in Panel A) / RU (in Panel B) is highlighted in red (blue), respectively. An asterisk (*) indicates models that are included in the MCS at the 5% significance level.

stocks. Figure 4.7 reveals that the improvement of **Universal** compared to **Single** is relatively small but consistent. In terms of the benefits of **Augmented**, it is typically the case that incorporating the market volatility as an additional feature helps improve the forecasting performance, especially for turmoil periods.

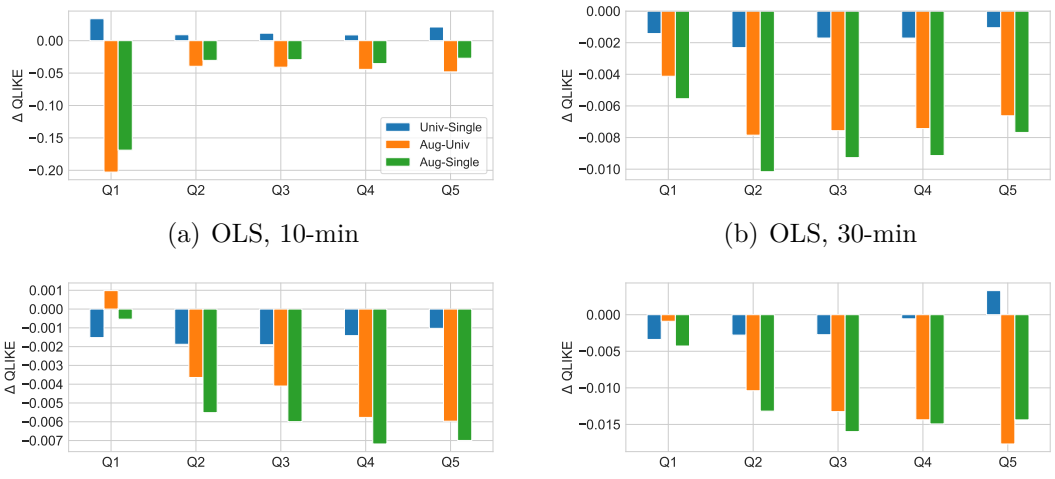


(a) OLS, 10-min (b) OLS, 30-min

(c) OLS, 65-min (d) OLS, 1-day

Figure 4.7: Pairwise ΔQLIKE of the OLS model across three training schemes.

Note: ΔQLIKE is averaged across stocks in each month during the testing period 2015-07 ∼ 2021-06. The dashed horizontal lines represent the average reductions in QLIKE.



(a) OLS, 10-min (b) OLS, 30-min

(c) OLS, 65-min (d) OLS, 1-day

Figure 4.8: Pairwise ΔQLIKE of the OLS model sorted by commonality.

Note: Q1, respectively Q5, denote the subset of stocks with the lowest, respectively highest, 20% values for the commonality.

An interesting question to investigate is whether the improvements of **Universal** or **Augmented** for individual stocks are associated with their commonality with the market volatility. To this end, we present the results in Figure 4.8 for each quintile bucket, sorted by stock commonality (computed from Eqn (4.5)). From this figure, we observe that the reduction of **Augmented** in out-of-sample QLIKE relative to **Universal** is explained by commonality to a large extent. Generally, the out-of-sample QLIKE is expected to decline steadily for stocks with higher commonality. Another interesting result arises from Figure 4.8(a), where we observe that **Universal** and **Augmented** actually reduce the out-of-sample QLIKE more for stocks (in the Q1 bucket) that are most loosely connected to the market in terms of 10-min volatility. Further research should be undertaken to investigate this finding.

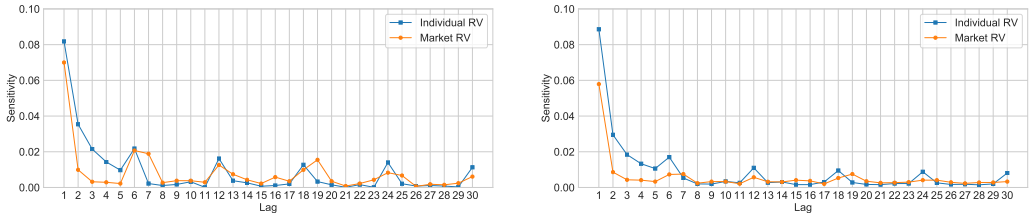
4.6.3 Variable importance and interaction effects

This section provides intuition for why NNs perform as strongly as they do, with an eye toward explainability. Due to the use of non-linear activation functions and multiple hidden layers, NNs enjoy the benefit of allowing for potentially complex interactions among predictors, albeit at the cost of considerably reducing model interpretability. To better understand such a "black-box" technique, we provide the following analysis to help illustrate the inner workings of NNs and explain their competitive performance.

Relative importance of predictors. In order to identify which variables are the most important for the prediction task at hand, we construct a metric (see Sadhwani, Giesecke, and Sirignano [199]) based on the sum of absolute partial derivatives (Sensitivity) of the predicted volatility. In particular, to quantify the importance of the k-th predictor, we compute

$$\text{Sensitivity}_k = \sum_{i=1}^{N} \sum_{t \in \mathcal{T}_{train}} \left| \frac{\partial F}{\partial u_k} \right|_{\mathbf{u}=\mathbf{u}_{i,t}} \tag{4.22}$$

Here, F is the fitted model under the **Augmented** scheme, $\mathbf{u}$ represents the vector of predictors and u_k is the k-th element in $\mathbf{u}$. $\mathbf{u}_{i,t}$ represents the input features



(a) OLS, 65-min (b) MLP, 65-min

Figure 4.9: Relative importance of lagged individual and market RVs.

Note: For ease of readability, we only report the sensitivity values for the most recent 30 lagged RVs (i.e. in the last 5 days for 65-min horizon).

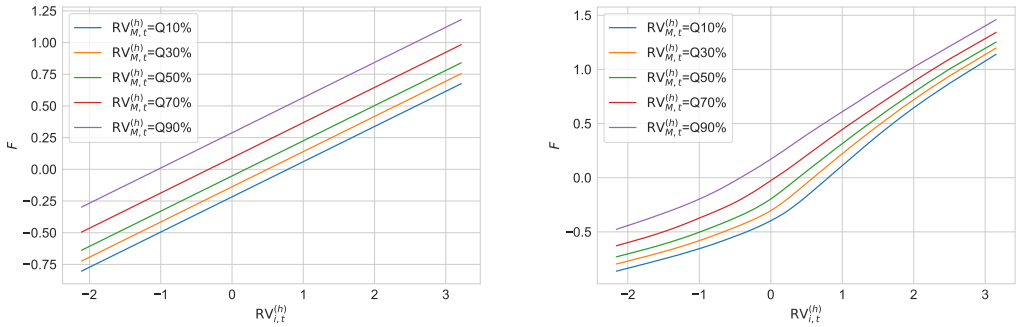
of stock i at time t. We normalize the sensitivity of all variables such that they sum up to one. In a special case of a linear regression, the sensitivity measure is the normalized absolute slope coefficient.

Considering the 65-min scenario as an example, Figure 4.9 reveals that for both OLS and MLP, there has been a tendency of the lagged features to decline in terms of sensitivity, as the lag increases. Additionally, we observe that the sensitivity values rise to a high point at every 6 lags, corresponding to 1 day. A distinct difference between the sensitivity values implied by OLS and the ones implied by MLP is that the latter places more weight on the lag=1 individual RV (Sensitivity=0.90) and less on the lag=1 market RV (Sensitivity=0.059). On the other hand, for OLS, the sensitivities of lag=1 individual (respectively, market) RV are 0.081 (respectively, 0.069).

Interaction effects. To analyze the interactions between the two most significant features implied by NNs, we adopt an approach (e.g. Choi, Jiang, and Zhang [65] and Gu, Kelly, and Xiu [119]) that focuses on the partial relations between a pair of input variables and responses, while fixing other variables at their mean values

$$F_{j|i}\left(u_j \mid u_i = q\right) = F\left(u_j, u_i = q, u_k = \bar{u}_k, k \neq i, j\right), \tag{4.23}$$

where q represent the quantile values for the i-th predictor u_i.



(a) OLS, 65-min (b) MLP, 65-min

Figure 4.10: Interactions between the lagged individual and market RV.

Note: The figure plots the pattern of predicted RV (y-axis) as a function of the lag=1 individual RV (x-axis) conditioned on various lag=1 market RV quantile values (keeping all other variables at their mean values).

Figure 4.10 illustrates how predicted volatility (i.e., the fitted values) varies as a function of the pair of predictor variables $RV_{i,t}^{(h)}$ and $RV_{M,t}^{(h)}$, over their support.[9] In particular, we analyze the interaction of the lag=1 individual RV ($RV_{i,t}^{(h)}$), with the lag=1 market RV ($RV_{M,t}^{(h)}$). As shown in Figure 4.10(a), a parallel shift between the different curves occurs if there are no interaction effects between $RV_{i,t}^{(h)}$ and $RV_{M,t}^{(h)}$. Figure 4.10(b) first reveals that the predicted volatility is non-linear in $RV_{i,t}^{(h)}$, and the slope of that relationship becomes higher after $RV_{i,t}^{(h)}$ exceeds a certain threshold (around 0.5). Furthermore, it demonstrates clear interaction effects between $RV_{i,t}^{(h)}$ and $RV_{M,t}^{(h)}$. As it can be observed from the rightmost region of Figure 4.10(b), the distances between the curves become relatively smaller, conveying the message that, when an individual stock is very volatile, the market effect on it weakens.

4.6.4 Forecasting RVs of unseen stocks

To examine the ability to generalize and address concerns regarding overfitting, we perform a stringent out-of-sample test, i.e. using the existent trained models to forecast the volatility of new stocks that have not been included in the training sample, in the spirit of Choi, Jiang, and Zhang [65] and Sirignano and Cont [206]. For better distinction, we denote the stocks used for estimating ML models as **raw**

[9] Recall that the variables are normalized by removing the mean and scaling to unit variance.

stocks, and those new stocks not in the training sample as **unseen stocks**.[10] We follow the procedure of training, validation, and testing periods described in Section 4.6.1. Specifically, to predict the RVs of unseen stocks in a particular year, we train and validate the models using the past data of raw stocks exclusively.

In this experiment, we choose OLS models trained for each unseen stock as the baseline. The results are shown in Table 4.4. Note that models trained under **Single** cannot be applied to forecast unseen stocks, since they are trained for each specific raw stock individually. From Table 4.4, we conclude that NNs trained on the pooled data of raw stocks have better forecasting performance compared to baselines, across all horizons. This presents new empirical evidence for a *universal volatility mechanism* among stocks. Furthermore, NNs significantly outperform other methods across three metrics, over 10-min, 30-min, and 65-min forecasting horizons, thus validating their robustness. Concerning the 1-day scenario, NNs obtain comparable results (QLIKE=0.252) to the best non-neural network model (MSE=0.249, attained by LASSO). The realized utility of the different risk models echoes that of the out-of-sample QLIKE.

4.7 Forecasting daily RVs with intraday RVs

Given the fact that intraday volatility exhibits a high and stable commonality (see Sections 4.3 and 4.4), we are interested in the potential benefits of using past intraday RVs to forecast daily RVs.

4.7.1 Closely related literature

Generally speaking, there are two broad families of models used to forecast daily volatility: (i) GARCH and SV models that employ daily returns; and (ii) models that use daily RVs. Previous well-established studies documented that due to the utilization of available intraday information, daily realized volatility is a superior proxy for the unobserved daily volatility, when compared to the parametric volatility

[10]The set of unseen stocks includes the following 16 tickers: AMAT, APD, BIIB, COF, DE, EQIX, EW, GPN, HUM, ICE, ILMN, ITW, NOC, NSC, PLD, SLB.

Table 4.4: Performance on unseen stocks.

Panel A:		10-min		30-min		65-min		1-day	
Statistical performance		MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
OLS	Unseen	0.664	0.372	0.329	0.219	0.287	0.205	0.348	0.254
OLS	**Universal**	0.678	0.410	0.328	0.223	0.286	0.206	0.343	0.260
	Augmented	0.639	0.359	0.317	0.222	0.278	0.208	0.327*	0.249*
LASSO	**Universal**	0.683	0.419	0.330	0.225	0.286	0.207	0.344	0.261
	Augmented	0.639	0.359	0.317	0.222	0.278	0.208	0.327*	0.249*
XGBoost	**Universal**	0.655	0.476	0.314	0.206	0.278	0.201	0.353	0.266
	Augmented	0.654	0.509	0.320	0.221	0.282	0.206	0.364	0.255
MLP	**Universal**	0.623	0.328	0.306	0.203*	0.266	0.193*	0.342	0.266
	Augmented	0.623	0.332	0.301*	0.203*	0.263*	0.194*	0.329	0.252
LSTM	**Universal**	0.637	0.348	0.311	0.211	0.267	0.195	0.339	0.265
	Augmented	0.622*	0.326*	0.303	0.205	0.263*	0.194	0.332	0.255
Panel B:		10-min		30-min		65-min		1-day	
Realized utility		RU	RU-TC	RU	RU-TC	RU	RU-TC	RU	RU-TC
OLS	Unseen	3.107	1.996	3.475	2.672	3.503	2.715	3.385	3.320
OLS	**Universal**	2.988	2.280	3.461	2.700	3.498	2.712	3.363	3.298
	Augmented	3.138	2.355	3.459	2.710	3.487	2.712	3.389	3.311
LASSO	**Universal**	2.959	2.270	3.457	2.704	3.496	2.712	3.359	3.296
	Augmented	3.137	2.376	3.458	2.720	3.485	2.716	3.389	3.315
XGBoost	**Universal**	2.688	1.640	3.510	2.711	3.511	2.701	3.349	3.269
	Augmented	2.563	1.578	3.464	2.688	3.495	2.680	3.388	3.302
MLP	**Universal**	3.233	2.396	3.515	2.736	3.529	2.730	3.340	3.266
	Augmented	3.221	2.444	3.514	2.749	3.522	2.735	3.378	3.302
LSTM	**Universal**	3.167	2.415	3.493	2.769	3.523	2.737	3.345	3.271
	Augmented	3.238	2.533	3.507	2.787	3.524	2.762	3.371	3.302

Note: The table reports the out-of-sample results for predicting future realized volatility of unseen stocks over multiple horizons using different models under three training schemes. The row OLS Unseen represents the baseline results based on OLS models estimated for each unseen stock. Other rows represent the results of models estimated on raw stocks under the **Universal** and **Augmented** settings. For each horizon, the model with the best (second best) out-of-sample performance in terms of QLIKE (in Panel A) / RU (in Panel B) is highlighted in red (blue), respectively. An asterisk (*) indicates models that are included in the MCS at the 5% significance level.

measures generated from the GARCH and SV models of daily returns (see Andersen et al. [13], Barndorff-Nielsen and Shephard [25], and Izzeldin et al. [143]). It is worth noting that in these traditional forecasting daily RV models (e.g. ARFIMA of Andersen et al. [13], HAR of Corsi [75], SHAR of Patton and Sheppard [191], HARQ of Bollerslev, Patton, and Quaedvlieg [38]), only past daily RVs (or their alternatives) are included as predictors. Even though this is a mainstream approach in the literature, it does not benefit to the full extent from the availability of intraday data. In the presidential address of SoFiE 2021, Bollerslev [34] also pointed out that "semivariation measured over shorter interday time intervals may afford additional useful information."

Intraday RV information is also studied for forecasting the one-day-ahead volatility in several previous works, e.g. the mixed data sampling (MIDAS) approach of Ghysels, Santa-Clara, and Valkanov [109, 110] and the "Rolling" approach of Pascalau and Poirier [187]. In particular, the classic MIDAS approach uses smooth-distributed lag polynomials of high-frequency predictors to forecast the low-frequency target variables, in the form $RV_{i,t+1}^{(d)} = \beta_{i,0} + \beta_{i,1}\left[a(1)^{-1}a(L)\right] RV_{i,t}^{(d)} + \epsilon_{i,t+1}$, where the $a(L)$ lag polynomial is defined by scaled beta functions. Ghysels, Santa-Clara, and Valkanov [109] find that the direct use of high-frequency data does not improve volatility predictions compared to the forecasts from a model based on daily RVs only. We reckon it is due to the restricted flexibility of MIDAS models, usually with one or two parameters determining the pattern of the weights, therefore missing the time-of-day effect of intraday RVs. Pascalau and Poirier [187] increase the training samples by rolling a fixed window of intraday returns over consecutive trading days by adding and dropping one intraday return at each end. They claim that their proposed "Rolling" approach could potentially capture the changing dynamics of serial correlation throughout the trading day, thus leading to improved volatility forecasts.

4.7.2 Proposed approach

In Section 4.5.1, we introduced a set of commonly used models, where the daily variables (lagged daily RVs) are employed as predictors when forecasting 1-day RVs. For simplicity, we refer to these models as **traditional** approaches in this section. Previous sections, such as Figure 4.9, concluded that the most recent RV plays a more important role in forecasting future volatility. Motivated by the fact that intraday volatility has a high and stable commonality, we propose a new prediction approach for forecasting daily volatility using past intraday RVs as predictors, denoted by **Intraday2Daily** approach.

In contrast to Ghysels, Santa-Clara, and Valkanov [109] and Pascalau and Poirier [187], our **Intraday2Daily** approach takes the time-of-day effect into account in an explicit way, and posits the model

$$RV^{(d)}_{i,t+1} = F_i\left(RV^{(h)}_{i,t}, \ldots, RV^{(h)}_{i,t-(p-1)h}, RV^{(d)}_{i,t-1}, \ldots, RV^{(d)}_{i,t-(p-1)}; \theta\right) + \epsilon_{i,t+1}. \quad (4.24)$$

Here $\left(RV^{(h)}_{i,t}, \ldots, RV^{(h)}_{i,t-(k-1)h}\right)$ represent the past RVs for stock i computed over shorter intraday time horizons h at day t and $\left(RV^{(d)}_{i,t-1}, \ldots, RV^{(d)}_{i,t-(p-1)}\right)$ are past daily RVs of stock i up to day $t-1$. Departing from traditional models where all the variables are computed in the daily frequency, we decompose the lag-one daily $RV^{(d)}_{i,t}$ to sub-sampled RVs, i.e. $\left(RV^{(h)}_{i,t}, \ldots, RV^{(h)}_{i,t-(k-1)h}\right)$. Under the **Augmented** training scheme, we also incorporate the market volatilities into models. Figure 4.11 illustrates the comparison between the traditional approach and our **Intraday2Daily** approach.

The advantages of the **Intraday2Daily** approach over traditional approaches can be summarized as follows. First, the **Intraday2Daily** approach significantly enriches the information content of daily volatility. Second, it contributes to the literature in the modeling of daily volatility by examining the coefficients of intraday RVs. Third, the essential idea underlying the **Intraday2Daily** approach can be possibly applied to estimate other daily risk measures, such as value-at-risk (VaR), etc. For example, one may use half-hour VaRs to forecast the one-day-ahead VaR. Finally, practitioners can better adjust their portfolios with more accurate forecasts

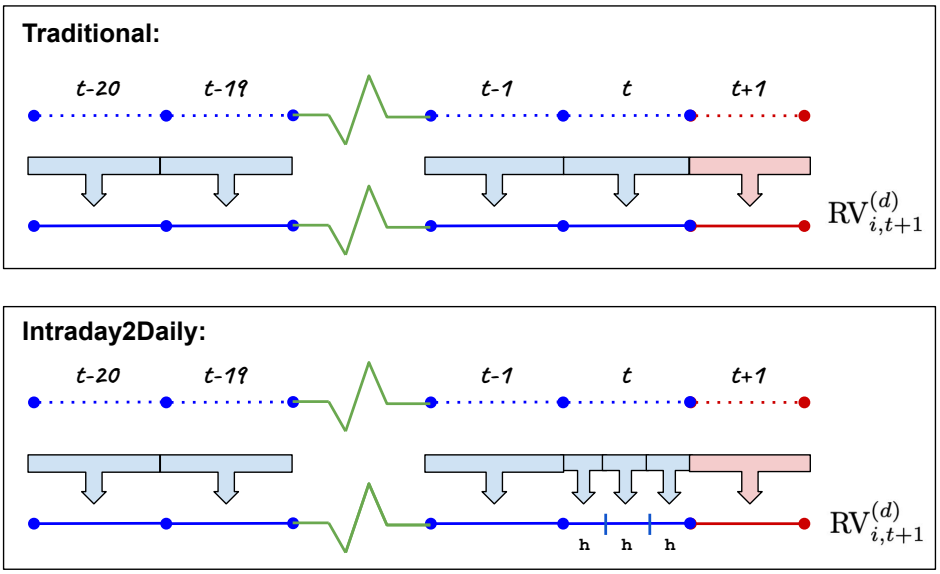


Figure 4.11: Illustration of two prediction approaches for future daily volatility (red segment).

Note: In each box, dots in the top line represent the intraday returns. The traditional approaches employ the aggregated daily (or weekly, or monthly) RVs (long blue segments) as predictors, while the **Intraday2Daily** approach employs intraday RVs (short blue segments between two adjacent vertical ticks). h represents the horizon of intraday RVs. In this example, $h = 130$ minutes.

from the **Intraday2Daily** approach rather than traditional approaches. To the best of our knowledge, this is the first study to explicitly investigate the predictive power of intraday RVs on daily volatility, and to demonstrate the additional accuracy improvements it brings to the forecasting task.

4.7.3 Experiments

The forecasting performance of traditional approaches with daily variables are already summarized in the column '1-day' of Table 4.3. Table 4.5 reports the results of models combined with the **Intraday2Daily** approach.[11] In other words, models in Table 4.5 use sub-sampled intraday RVs rather than the lag-one total RV in the column '1-day' of Table 4.3. For example, the lag-one total RV in HAR (Eqn (4.9)) is replaced by non-overlapped intraday RVs.

[11]We observe similar findings when applying the **Intraday2Daily** approach to forecast the raw volatilities (not in logs).

By comparing the column '1-day' of Table 4.3 with Table 4.5, we establish that the **Intraday2Daily** approach generally helps improve the out-of-sample performance of benchmark models. For example, under the **Single** setting, compared with OLS using daily RVs (QLIKE=0.192), 65-min RVs improve the out-of-sample performance (QLIKE=0.186).

MLPs with intraday RVs again achieve the best out-of-sample performance. For example, the QLIKEs of MLPs under **Universal** are $\{0.171, 0.171, 0.172\}$ using $\{$10-min, 30-min, 65-min$\}$ RVs as predictors, respectively. The superior performance of MLPs over linear regressions when using intraday RVs further demonstrates the advantages of NNs to learn unknown dynamics in financial markets.

In general, the improvements of the **Intraday2Daily** approach lead to higher utilities. For instance, when considering the column '1-day' of Panel B in Table 4.3, we observe that the RU (respectively, RU-TC) values of OLS under **Augmented** are 3.576% (respectively, 3.536%). OLS with 65-min RVs as predictors obtain higher RU (respectively, RU-TC) values of 3.583% (respectively, 3.547%). Overall, MLPs deliver the highest utility (RU=3.593%, RU-TC=3.550%) based on the 65-min intraday RVs, followed by LSTMs, thus hinting at potential non-linearity and complex interactions inherent in the data.

4.7.4 Robustness check

In this section, we present the empirical analysis of examining the robustness of the **Intraday2Daily** approach when incorporating new types of predictors (including semi-RV of Patton and Sheppard [191], and realized quarticity of Bollerslev, Patton, and Quaedvlieg [38]).

SHAR. Patton and Sheppard [191] proposed the Semi-variance-HAR (SHAR) model as an extension of the standard HAR model (see further details in 4.5.1), in order to exploit the well-documented leverage effect by decomposing the total RV of the first lag via signed intraday returns, as shown in Eqn (4.25) (see Barndorff-Nielsen, Kinnebrock, and Shephard [24]). In other words, the lag-one RV in SHAR

Table 4.5: Out-of-sample performance of the **Intraday2Daily** approach.

Panel A:		10-min		30-min		65-min	
Statistical performance		MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
HAR	**Single**	0.259	0.189	0.252	0.185	0.252	0.185
	Universal	0.270	0.197	0.255	0.187	0.253	0.186
	Augmented	0.256	0.179	0.249	0.174	0.249	0.175
OLS	**Single**	0.255	0.186	0.252	0.186	0.253	0.186
	Universal	0.253	0.186	0.252	0.187	0.252	0.186
	Augmented	0.249	0.173	0.248	0.173	0.249	0.173
LASSO	**Single**	0.262	0.194	0.251	0.185	0.253	0.186
	Universal	0.261	0.191	0.248	0.187	0.248	0.186
	Augmented	0.273	0.203	0.247	0.173	0.247	0.173
XGBoost	**Single**	0.323	0.204	0.330	0.201	0.332	0.200
	Universal	0.261	0.177	0.257	0.173	0.257	0.173
	Augmented	0.264	0.179	0.261	0.176	0.266	0.176
MLP	**Single**	-	-	-	-	-	-
	Universal	0.243*	0.171*	0.242*	0.171*	0.246	0.172
	Augmented	0.247	0.174	0.246	0.175	0.247	0.176
LSTM	**Single**	-	-	-	-	-	-
	Universal	0.247	0.174	0.244	0.171*	0.244	0.171
	Augmented	0.258	0.184	0.249	0.175	0.250	0.176
Panel B:		10-min		30-min		65-min	
Realized utility		RU	RU-TC	RU	RU-TC	RU	RU-TC
HAR	**Single**	3.558	3.524	3.567	3.533	3.566	3.531
	Universal	3.539	3.521	3.563	3.534	3.565	3.532
	Augmented	3.562	3.532	3.573	3.541	3.570	3.536
OLS	**Single**	3.565	3.526	3.565	3.530	3.564	3.529
	Universal	3.565	3.534	3.564	3.535	3.565	3.533
	Augmented	3.581	3.548	3.583	3.551	3.583	3.547
LASSO	**Single**	3.561	3.526	3.565	3.531	3.564	3.529
	Universal	3.561	3.524	3.564	3.535	3.573	3.536
	Augmented	3.551	3.515	3.565	3.531	3.585	3.548
XGBoost	**Single**	3.532	3.474	3.545	3.486	3.550	3.489
	Universal	3.584	3.532	3.593	3.543	3.590	3.547
	Augmented	3.579	3.527	3.586	3.535	3.589	3.538
MLP	**Single**	-	-	-	-	-	-
	Universal	3.586	3.543	3.592	3.549	3.593	3.550
	Augmented	3.562	3.521	3.583	3.541	3.581	3.540
LSTM	**Single**	-	-	-	-	-	-
	Universal	3.586	3.543	3.592	3.549	3.593	3.550
	Augmented	3.562	3.521	3.583	3.541	3.581	3.540

Note: The table reports the out-of-sample results for predicting future daily realized volatility using different models under three training schemes when combined with the **Intraday2Daily** approach. The columns ('10-min', '30-min', and '65-min') represent the frequency of predictor features and the dependent variable in this table always corresponds to future daily volatility. The model with the best (second best) out-of-sample performance in QLIKE (in Panel A) / RU (in Panel B) is highlighted in red (blue), respectively. An asterisk (*) indicates models that are included in the MCS at the 5% significance level.

(Eqn (4.26)) is split into the sum of squared positive returns and the sum of squared negative returns, as follows.

$$\begin{array}{l} RV_{i,t}^{(d)+} = \sum_{l=0}^{M-1} r_{i,t-l\cdot\Delta}^2 I_{\{r_{t-l\cdot\Delta}>0\}}, \\ RV_{i,t}^{(d)-} = \sum_{l=0}^{M-1} r_{i,t-l\cdot\Delta}^2 I_{\{r_{t-l\cdot\Delta}<0\}}, \end{array} \tag{4.25}$$

$$RV_{i,t+1}^{(d)} = \alpha_i + \beta_i^{(d)+} RV_{i,t}^{(d)+} + \beta_i^{(d)-} RV_{i,t}^{(d)-} + \beta_i^{(w)} RV_{i,w}^{(w)} + \beta_i^{(m)} RV_{i,m}^{(m)} + \epsilon_{i,t+1}. \tag{4.26}$$

In the above, Δ denotes the interval for computing the intraday returns.

HARQ. Bollerslev, Patton, and Quaedvlieg [38] pointed out that the beta coefficients in the HAR model may be affected by measurement errors in the realized volatilities. By exploiting the asymptotic theory for high-frequency realized volatility estimation, the authors propose an easy-to-implement model, termed as HARQ (Eqn (4.28)). The realized quarticity (RQ) is estimated according to Eqn (4.27), aiming to correct the measurement errors.

$$RQ_{i,t}^{(d)} = \frac{M}{3} \sum_{l=0}^{M-1} r_{t-l\cdot\Delta}^4 \tag{4.27}$$

$$RV_{i,t+1}^{(d)} = \alpha_i + \left(\beta_i^{(d)} + \beta_i^{(d)\mathrm{Q}} \sqrt{RQ_{i,t}^{(d)}} \right) RV_{i,t}^{(d)} + \beta_i^{(w)} RV_{i,t}^{(w)} + \beta_i^{(m)} RV_{i,t}^{(m)} + \epsilon_{i,t+1}. \tag{4.28}$$

We compute the corresponding intraday variables of semi-RVs and RQs and then include them as new predictors in the **Intraday2Daily** approach. From Table 4.6, we first observe that the SHAR model generally performs as well as the standard HAR model (in Table 4.3), in line with Bollerslev, Patton, and Quaedvlieg [38]. HARQ outperforms HAR and SHAR, when applied to individual stocks studied in the present chapter. Comparing the 'Traditional' column with others, we conclude that in general, replacing the daily RVs with intraday RVs as predictors helps improve the out-of-sample performance of benchmark models.

Table 4.6: Out-of-sample performance of the **Intraday2Daily** approach.

Panel A:		10-min		30-min		65-min		Traditional	
Statistical performance		MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
SHAR	**Single**	0.277	0.191	0.257	0.178	0.253	0.176	0.261	0.183
	Universal	0.285	0.198	0.263	0.183	0.255	0.178	0.261	0.182
	Augmented	0.261	0.181	0.253	0.175	0.250	0.174	0.254	0.178
HARQ	**Single**	0.264	0.204	0.254	0.178	0.253	0.176	0.256	0.179
	Universal	0.253	0.176	0.253	0.176	0.254	0.176	0.257	0.179
	Augmented	0.251	0.174	0.248*	0.172*	0.250	0.174	0.253	0.176
Panel B:		10-min		30-min		65-min		Traditional	
Realized utility		RU	RU-TC	RU	RU-TC	RU	RU-TC	RU	RU-TC
SHAR	**Single**	3.528	3.497	3.559	3.525	3.563	3.529	3.548	3.515
	Universal	3.510	3.499	3.548	3.525	3.560	3.529	3.550	3.516
	Augmented	3.563	3.533	3.576	3.545	3.578	3.545	3.571	3.537
HARQ	**Single**	3.467	3.425	3.556	3.520	3.564	3.528	3.557	3.525
	Universal	3.564	3.530	3.564	3.530	3.564	3.530	3.558	3.525
	Augmented	3.580	3.544	3.583	3.546	3.578	3.541	3.575	3.538

Note: The table reports the out-of-sample results of SHAR and HARQ for predicting future daily realized volatility under three training schemes. Columns '10-min', '30-min', and '65-min' represent the **Intraday2Daily** approach with different frequencies of predictors while the column 'Traditional' represents that lagged daily RVs are used as predictors. The dependent variable in this table always corresponds to future daily volatility. The model with the best (second best) out-of-sample performance in QLIKE (in Panel A) / RU (in Panel B) is highlighted in red (blue), respectively. An asterisk (*) indicates models that are included in the MCS at the 5% significance level.

4.7.5 Analysis of the time-of-day dependent RV

To offer a more comprehensive understanding of the performance of *time-of-day dependent* RVs, we examine the coefficients of the **Intraday2Daily** OLS model trained under **Augmented**. Recall that before we input features into the model, we rescale them to have a mean of zero and a standard deviation of one. Hence we can compare the coefficients of different lagged variables.

For better readability, we only report the first $13 = (390/30)$ coefficients of the OLS model using 30-min features in Figure 4.12, corresponding to the observations of RV in the most recent day.[12] We observe that the contributions of time-of-day dependent RVs are not even. Interestingly, *volatility near the close* (15:30-16:00) is the most important predictor, in contrast to the diurnal volatility pattern. These

[12]We attain similar results for models using intraday RVs based on other frequencies.

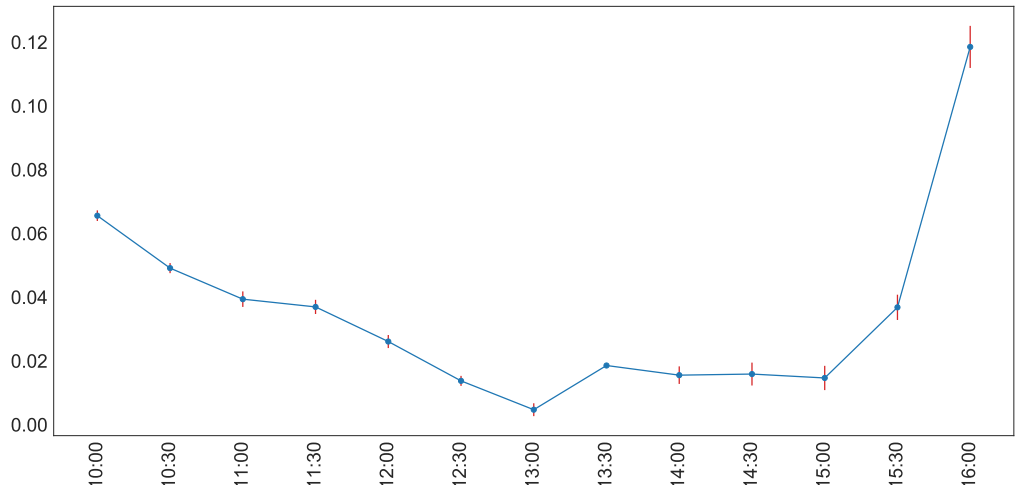


Figure 4.12: Coefficients of the **Intraday2Daily** OLS model under **Augmented**.

Note: The **Intraday2Daily** OLS model uses lagged individual 30-min RVs to forecast the next day's volatility. The x-axis represents the time of day. The y-axis represents the coefficients of lagged RVs.

results shed new light on the modeling of volatility.

To explain why the most recent half-hour RV is the most important predictor for forecasting the next day's volatility, we provide a handful of perspectives. According to Banerji [21], there is a significant fraction of the total daily trading volume in the last half-hour of the trading day. As also mentioned in Chapter 3, for the first few months of 2020 in the US equity market, about 23% of trading volume in the 3,000 largest stocks by market value has taken place after 15:30. We also conclude from Figure 4.5 that the market achieves the highest level of consensus near the close. Therefore, volatility near the close of the previous trading day might contain more useful information for predicting the next day's volatility.

4.8 Conclusion

In this chapter, the commonality in intraday volatility over multiple horizons across the U.S. equity market is studied. By leveraging the information content of commonality, we have demonstrated that for most ML models in our analysis, pooling stock data together (**Universal**) and adding the market volatility as an additional

predictor (**Augmented**) generally improves the out-of-sample performance, in comparison with asset-specific models (**Single**).

We show that NNs achieve superior performance, possibly due to their ability to uncover and model complex interactions among predictors. To alleviate concerns of overfitting, we perform a stringent out-of-sample test, applying the existent trained models to unseen stocks, and conclude that NNs still outperform traditional models.

Lastly and perhaps most importantly, motivated by the high commonality in intraday volatility, we propose a new approach (**Intraday2Daily**) to forecasting daily RVs using past intraday RVs. The empirical findings suggest that the proposed **Intraday2Daily** approach generally yields superior out-of-sample forecasts. We further examine the coefficients in **Intraday2Daily** OLS models, and the results suggest that volatility near the close (15:30-16:00) in the previous day (lag=1) is the most important predictor.

Future research directions. There are a number of interesting avenues to explore in future research. One direction pertains to the assessment of whether other characteristics, such as sector RVs, can improve the forecast of future realized volatility, since in the present work, we have only considered the individual and market RVs. Another interesting direction is to apply the underlying idea of **Intraday2Daily** approach to other risk metrics, e.g. Value-at-Risk (as considered in Chapter 2), that could potentially benefit from time-of-day dependent features.

5

Graph-based Methods for Forecasting Realized Covariances

Contents

5.1 Introduction

Covariance matrix estimation is of particular importance in areas such as risk management, asset pricing and portfolio optimization. More recently, the increased availability of reliable high-frequency intraday asset prices has motivated researchers to model and forecast the nonparametric and ex-post *Realized Covariance (RC)* measures constructed from intraday data (see Andersen et al. [13], Andersen, Bollerslev, and Meddahi [14], Barndorff-Nielsen et al. [23], Sheppard [203], and Varneskov and Voev [217]). It has been extensively documented in the literature that the high-frequency-based realized covariances are superior to those constructed from low-frequency return data, for practical investment and portfolio allocation decisions (e.g. Bollerslev, Patton, and Quaedvlieg [39], Hautsch, Kyj, and Malec [134], and Symitsi et al. [210]).

There is a large volume of studies dedicated to forecasting realized covariance matrices. Chiriac and Voev [64] extended the univariate Heterogeneous Autoregressive (HAR) model of Corsi [75], to a multivariate setting via the *Cholesky decomposition* on the covariance matrix.[1] Inspired by the dynamic conditional correlation (DCC) model of Engle [91], an alternative approach to forecasting realized covariances is based on the decomposition of the covariance matrix into the volatilities and correlations, also known as the *DRD decomposition*.[2] Bollerslev, Patton, and Quaedvlieg [39] reported that the covariance forecasts based on the DRD decomposition generally outperform the ones based on the Cholesky decomposition. Therefore, in the present chapter, we pay more attention to the methods based on DRD decomposition.

Several well-established studies have examined the interdependence of volatilities within the asset market or across various markets, a.k.a. *volatility spillover effect*. Such a type of interdependence can be leveraged to improve the volatility prediction of the target assets or markets.[3] However, previous methods (such as multivariate

[1] See Bollerslev et al. [37], Bucci [43], Fiszeder and Orzeszko [102], and Symitsi et al. [210].
[2] See Lee and Long [162], Oh and Patton [183], and Vassallo, Buccheri, and Corsi [218].
[3] See Buncic and Gisler [47], Degiannakis and Filis [83], and Wang, Pan, and Wu [223]

GARCH and Vector Autoregression) may deliver poor out-of-sample forecasts due to the curse of dimensionality, as pointed out by Callot, Kock, and Medeiros [48].

On the other hand, correlation forecasting has not yet drawn much attention in the literature, to the best of our knowledge. In addition, it remains uncertain whether and to what extent correlations influence one another, i.e. the presence of interdependence among correlation pairs.[4]

Therefore, we are particularly interested in the following research questions: (i) How do we capture the time-varying interdependence in volatilities and correlation pairs? (ii) Are the two types of interdependence predictive for covariance matrices in the U.S. equity market? (iii) Is there a parsimonious and interpretable forecasting model for realized covariances that accurately incorporates this interdependence?

To answer the above research questions, we utilize **graphs** as the mathematical model to represent two types of interdependence in the U.S. equity market. In the first type of graph, each asset is modeled as a node and an edge connecting two nodes represents the existence of any dependence between their volatilities. In the second type of graph, each node is a correlation pair of two assets and an edge connecting two nodes represents the relations between one pair and another pair.[5]

The perspective of representing the above interdependence as a graph allows us to adopt the modeling tools developed specifically for the analysis of graph-structured data. In particular, we are able to apply neighborhood[6] aggregation to generate a new input feature for every underlying asset and incorporate it into the traditional models for volatility forecasting. For correlation forecasting, we could generate a similar feature for every pair of assets. The aggregated feature summarizes the influence of the connected assets (respectively, pairs of assets) on a specific asset (respectively, a specific pair correlation).

[4]For example, the correlation between two equities APPL and IBM, denoted as APPL-IBM, might have influence on other correlations that share one of the two equities, such as IBM-MSFT and APPL-AMZN.

[5]Some related graph construction and estimation methods will be reviewed in Section 5.4.1.

[6]Andrada-Félix, Fernández-Rodrıéguez, and Fuertes [15] demonstrated the merit of the neighbor information for realized variances forecasting via the nearest neighbor approach.

Neighborhood aggregation represents that the embedding of each node is updated by aggregating the embeddings of its neighbors, which was first proposed to predict the chemical properties of molecules by Gilmer et al. [112]. It is an effective method to integrate the structural interdependence between nodes into the features while maintaining a low dimension. In fact, the aggregated features remain in the same dimension as the original features. Therefore, our model can avoid the curse of dimensionality and be applied to a large universe of assets.

To some extent, our framework may have some conceptual similarity to the work of Zhu et al. [239], who proposed a network vector autoregressive model, assuming each node's response at a given time point as a linear combination of (a) its previous value, (b) the average of its connected neighbors, (c) a set of node-specific covariates and (d) an independent noise. However, we also address the important challenges of constructing desirable graphs in the present study, not only for enhancing predictive power on covariances but also for understanding the dynamics of covariances. Our novel framework also allows for time-varying graphs, which can well capture the dynamics in graph effects of volatilities and correlations, while Zhu et al. [239] assumed a constant graph. Moreover, moving to the covariance matrix forecasting problem opens up a host of interesting and practically relevant econometric applications, as illustrated in our following analysis.

The main contributions of this research are summarized as follows. First, we put forward a methodology to model and forecast realized covariance matrices by leveraging the graph information. Our method may shed new light on the dynamic interdependence of realized volatility and correlation, through the lens of neighborhood aggregation. Second, we apply our method to forecasting realized covariance matrices over the daily, weekly, and monthly future horizons, in order to study the variation in the predictive power of graphs across time. Last, we examine the stability and robustness of our model through comprehensive experiments.

More specifically, using realized volatility data for the components of the Dow Jones Industrial Average, spanning the period from July 2007 to June 2021, we find that graph information of volatility and correlation can be used to improve the

forecasts of realized covariance matrices. Our in-sample results first show that daily volatility is the most important source for future volatility forecasting, while the correlation model attributes the most importance to the weekly component. This highlights the different dependencies in volatilities and correlations. Moreover, our in-sample analysis shows that graph-based predictors for volatility (respectively, correlation) modeling are highly significant, and including these graph components in the HAR model can substantially improve the in-sample fit accuracy of realized covariance matrices.

By performing the out-of-sample tests in a rolling window setup, our analysis provides strong evidence that the graph-based models yield superior one-day-ahead forecasts over a strong set of traditional baselines. The MCS test (Hansen, Lunde, and Nason [126]) indicates that the augmented model with a data-driven graph determined by Graphical LASSO (Friedman, Hastie, and Tibshirani [106]) is consistently included in the subset of best forecasting models irrespective of the loss functions. The graph analysis highlights the time-varying nature of graph effects. We also investigate the economic benefits of covariance forecasts from different models via the global minimum variance portfolio (GMVP). The results show that portfolios employing graph information are able to generate significantly lower out-of-sample variance compared to the traditional models without graph information or the naive equally-weighted portfolio (see DeMiguel, Garlappi, and Uppal [84]). The forecast improvements over longer horizons (such as 1-week) remain significant, but decay as the prediction horizon gets longer (e.g. 1-month). Additionally, we perform several robustness tests, which demonstrate the forecast improvements are experienced consistently over the different out-of-sample sub-periods and are insensitive to measurement errors of volatilities (Bollerslev, Patton, and Quaedvlieg [38, 39]).

The rest of the chapter is structured as follows. We briefly review the related literature on modeling and forecasting realized volatilities and correlations in Section 5.2. We then introduce the estimation of realized covariance matrices from high-frequency data and the baseline models in Section 5.3. In Section 5.4, we present the background of graphs and propose a new methodology based on graphs for modeling

and forecasting realized covariance matrices. Section 5.5 provides a description of the dataset and evaluation criteria, followed by the in-sample analysis and out-of-sample results. In Section 5.6, we perform several robustness tests. Finally, we conclude our analysis in Section 5.7 and highlight potential future research directions.

5.2 Related literature

In this section, we will briefly review the related literature on modeling and forecasting realized volatilities and correlations separately, which however should not be considered as a comprehensive survey of the subject.

On the one hand, there have been numerous contributions made to the literature on the topic of forecasting daily realized volatility (RV). Andersen et al. [13] proposed the AutoRegressive Fractionally Integrated Moving Average (ARFIMA) model for forecasting daily RVs. Corsi [75] put forward the HAR model for predicting daily RVs using various realized volatility components over different time horizons. These methods may give important information about the dynamics of volatilities, but ignore the interdependence of volatilities among assets, as pointed out by Bollerslev et al. [36] and Cubadda, Guardabascio, and Hecq [77].

Numerous studies have investigated the volatility spillover effect, which implies that a large shock of a specific asset (or market) may not only increase its own subsequent volatility, but also that of other assets (or markets). For example, Buncic and Gisler [47] revealed the cross-market volatility spillover effect from the U.S. equity market to 17 global foreign asset markets, and showed how VIX plays an important role in predicting the volatility in all of these 17 markets. Similarly, Wang, Pan, and Wu [223] showed that accounting for spillover information from the U.S. market can significantly improve the forecasting accuracy of international stock price volatility. Degiannakis and Filis [83] found evidence that different assets, such as stocks and foreign currencies, can improve the prediction of RV of crude oil price.

Various statistical models such as multivariate GARCH, stochastic volatility (SV) models, Vector Autoregression (VAR), Wishart Autoregression (WAR), and others have been employed in literature to model the volatility spillover effects.

The BEKK-GARCH introduced by Engle and Kroner [93] and the VAR-GARCH proposed by Ling and McAleer [170] are probably two popular GARCH-type models used to analyze volatility spillover with low-frequency data. The standard WAR model proposed by Gouriéroux, Jasiak, and Sufana [117] is built on latent processes and is unable to capture the long memory dependence in volatility. In terms of modeling realized volatility, Wilms, Rombouts, and Croux [227] used VAR to obtain the multivariate volatility forecasts for stock market indices. However, Callot, Kock, and Medeiros [48] highlighted that the aforementioned models may deliver poor out-of-sample forecasts due to the curse of dimensionality. This is primarily due to the fact that the dependence of each pair of financial assets or markets is modeled as an individual feature, which leads to large computational burdens. Consequently, there is a desire for a parsimonious model that can effectively incorporate the intricate interdependence.

On the other hand, to the best of our knowledge, there has been limited focus on forecasting correlations in the literature. In the classic dynamic conditional correlation (DCC) model of Engle [91], the temporal dependencies in the conditional correlations are characterized by a simple scalar GARCH$(1, 1)$ model, where the dynamics are imposed to be equal for all the correlation pairs. In terms of realized correlation modeling, Andersen et al. [12] studied the impact of past realized volatilities and past returns on forecasting realized daily correlations. Bollerslev, Patton, and Quaedvlieg [39] and Oh and Patton [183] employed a HAR-type model for forecasting realized correlations. While these models are effective to some extent, they are not flexible enough to jointly model all the elements in the correlation matrix. Generalizations of the above models, e.g. allowing each unique correlation pair to follow its specific dynamics, require a larger number of free parameters, and thus may not be feasible for high-dimensional applications.

Although there seem to be general agreements that pair correlations of asset returns vary over time, it is less clear whether and how correlations affect each other, i.e. the existence of interdependence of correlation pairs. For example, King and Wadhwani [149] noticed a significant increase in the correlations during crisis

periods. Berben and Jansen [31] documented that correlations among the U.S. stock market have increased significantly during 1980-2000. Boyer, Gibson, and Loretan [41] and Forbes and Rigobon [104] further revealed that the changes in correlations might be due to a simultaneous increase in the volatility of asset prices. Aït-Sahalia and Xiu [6] found that the increase in correlation is largely driven by the volatilities of prices. Thus, we are interested to model such co-movement or interdependence between correlation pairs, and subsequently to examine whether such effects could improve the modeling and forecasting of correlation matrices.

5.3 Traditional models

This section outlines the approach for the estimation of realized covariances, and two baseline models for forecasting realized covariances in the U.S. equity market.

5.3.1 Problem set-up

Let us consider a vector of asset returns $\boldsymbol{r}_t = (r_{1,t}, \ldots, r_{N,t})'$ as follows,

$$\boldsymbol{r}_t = \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t, \tag{5.1}$$

where N is the number of assets, and $\boldsymbol{\mu}_t$ is the conditional mean vector given the information set $\mathcal{F}_{t-1}$. $\boldsymbol{\epsilon}_t$ is a vector of innovations, which can be expressed as $\boldsymbol{\epsilon}_t = \boldsymbol{\Sigma}_t^{1/2}\boldsymbol{z}_t$, where $\boldsymbol{z}_t$ is a random vector that follows a multivariate standard normal distribution, i.e. $\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$ ($\boldsymbol{I}_N$ is an $N \times N$ identity matrix). Therefore, $\boldsymbol{\Sigma}_t$ is the conditional covariance matrix of $\boldsymbol{r}_t$.[7]

Note that $\boldsymbol{\Sigma}_t$ is unobservable. One consistent estimator of $\boldsymbol{\Sigma}_t$ is the multivariate realized covariance (RC), which is constructed from the intra-day returns, e.g. 5-min returns.[8] Specifically, we denote $\boldsymbol{r}_{t(l)}$ as the l-th vector of non-overlapping Δ-min

[7]The square root of $\boldsymbol{\Sigma}_t$ is not unique, as long as the operator satisfies the condition that $\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\Sigma}_t^{1/2} = \boldsymbol{\Sigma}_t$.

[8]Liu, Patton, and Sheppard [171] demonstrate that no sub-sampling frequency significantly outperforms a 5-min interval in terms of forecasting daily RVs, making it a widely accepted time interval in the literature.

log-returns during day t, i.e. $\boldsymbol{r}_{t(l)} = \log \boldsymbol{p}_{t(l\Delta)} - \log \boldsymbol{p}_{t((l-1)\Delta)}$, where $\boldsymbol{p}_{t(l\Delta)}$ is the price at time $l\Delta$ of day t. Then the RC matrix of day t can be computed as

$$\boldsymbol{H}_t := \sum_{l=1}^{M} \boldsymbol{r}_{t(l)} \boldsymbol{r}'_{t(l)}, \tag{5.2}$$

where $M = 390/\Delta$ is the number of non-overlapping Δ-min slots during day t.[9]

In addition, we adopt the subsampling averaging method (see Andersen, Bollerslev, and Meddahi [14], Sheppard [203], and Varneskov and Voev [217]) to improve the above RC estimation, which uses all Δ-minute returns, not just non-overlapping ones.[10] Covariances over longer horizons of k days are computed by the sum of daily realized covariances (see Symitsi et al. [210]).

The daily realized volatility for a particular asset i at day t is $RV_{i,t} = \boldsymbol{H}_t[i,i] = \sum_{l=1}^{M} r_{i,t(l)}^2$. We refer to $\boldsymbol{RV}_t = (RV_{1,t}, \ldots, RV_{N,t})'$ as the vector of cross-sectional realized volatilities.

5.3.2 HAR-Cholesky

Corsi [75] proposed a Heterogeneous Autoregressive Regression (HAR) model for modeling and forecasting the realized volatility where the lagged daily, weekly and monthly volatility components are incorporated as predictors, as also introduced in (4.9). For a given asset i, its realized volatility of day t is modeled as

$$RV_{i,t} = \alpha + \beta_d RV_{i,t-1} + \beta_w RV_{i,t-5:t-2} + \beta_m RV_{i,t-22:t-6} + u_{i,t}, \tag{5.3}$$

where $RV_{i,t-5:t-2} = \frac{1}{4}\sum_{k=2}^{5} RV_{i,t-k}, RV_{i,t-22:t-6} = \frac{1}{17}\sum_{k=6}^{22} RV_{i,t-k}$ denote the weekly and monthly lagged realized volatility, respectively.

Chiriac and Voev [64] generalized the above univariate HAR model to estimate the realized covariances as a linear combination of past daily, weekly and monthly realized covariances, in the form

[9] The general properties of the RC estimator can be found in Sheppard [203].

[10] For example, suppose prices are sampled every minute. The standard realized covariance uses 5-min returns constructed from prices sampled at 09:30, 09:35, 09:40, ..., 15:55, 16:00. The subsampled realized covariance uses returns computed from all 5-minute windows, i.e., 09:30 and 09:35, 09:31 and 09:36, 09:32 and 09:37, and so on. In the end, the subsampled realized covariance is computed by the average of all standard realized covariances.

$$\boldsymbol{y}_t = \boldsymbol{\alpha}^{(C)} + \beta_d^{(C)}\boldsymbol{y}_{t-1} + \beta_w^{(C)}\boldsymbol{y}_{t-5:t-2} + \beta_m^{(C)}\boldsymbol{y}_{t-22:t-6} + \boldsymbol{u}_t^{(C)}, \tag{5.4}$$

where $\boldsymbol{y}_t = \text{vech}_{Chol}(\boldsymbol{H}_t)$ denotes the $N^* = N(N+1)/2$ dimensional vectorized version of the Cholesky decomposition of $\boldsymbol{H}_t$ and $\boldsymbol{y}_{t-5:t-2}$ (respectively, $\boldsymbol{y}_{t-22:t-6}$) is computed as $\frac{1}{4}\sum_{k=2}^{5}\boldsymbol{y}_{t-k}$ (respectively, $\frac{1}{17}\sum_{k=6}^{22}\boldsymbol{y}_{t-k}$), following Bollerslev, Patton, and Quaedvlieg [39] and Symitsi et al. [210]. The intercept $\boldsymbol{\alpha}^{(C)}$ is a N^*-dimensional vector, while the $\beta_d^{(C)}$, $\beta_w^{(C)}$ and $\beta_m^{(C)}$ parameters are all assumed to be scalar. $\boldsymbol{u}_t^{(C)}$ is the error term with conditional mean of zeros.

5.3.3 HAR-DRD

Andersen et al. [11] demonstrated that the dynamic dependencies in the correlations are different from the volatilities, also known as *correlation breakdown*. Oh and Patton [183] thus proposed the HAR-DRD model, which is based on the decomposition of the covariance matrix into the diagonal matrix of realized volatilities and the correlation matrix

$$\boldsymbol{H}_t = \boldsymbol{D}_t\boldsymbol{R}_t\boldsymbol{D}_t, \tag{5.5}$$

where $\boldsymbol{D}_t$ is the diagonal matrix with the elements of the square roots of $\boldsymbol{RV}_t$ on the main diagonal, i.e. $\boldsymbol{D}_t[i,i] = \sqrt{RV_{i,t}}, \forall i$, and $\boldsymbol{D}_t[i,j] = 0, \forall i \neq j$. $\boldsymbol{R}_t$ is the correlation matrix.

In the HAR-DRD model, the realized variance vector is modeled by vectorized HAR model as in Eqn (5.6). The realized correlation matrix is modeled using another vectorized HAR model as in Eqn (5.7), in consistent with Bollerslev, Patton, and Quaedvlieg [39].

$$\boldsymbol{RV}_t = \boldsymbol{\alpha}_0^{(D)} + \beta_d^{(D)}\boldsymbol{RV}_{t-1} + \beta_w^{(D)}\boldsymbol{RV}_{t-5:t-2} + \beta_m^{(D)}\boldsymbol{RV}_{t-22:t-6} + \boldsymbol{u}_t^{(D)}, \tag{5.6}$$

$$\boldsymbol{x}_t = \boldsymbol{\alpha}_0^{(R)} + \beta_d^{(R)}\boldsymbol{x}_{t-1} + \beta_w^{(R)}\boldsymbol{x}_{t-5:t-2} + \beta_m^{(R)}\boldsymbol{x}_{t-22:t-6} + \boldsymbol{u}_t^{(R)}, \tag{5.7}$$

where $\boldsymbol{x}_t = \text{vech}_{Tri}(\boldsymbol{R}_t)$ is the $N^{\#} = N(N-1)/2$ dimensional vectorized version of the lower triangular of $\boldsymbol{R}_t$ and $\boldsymbol{x}_{t-5:t-2}$ (respectively, $\boldsymbol{x}_{t-22:t-6}$) is computed as $\frac{1}{4}\sum_{k=2}^{5}\boldsymbol{x}_{t-k}$ (respectively, $\frac{1}{17}\sum_{k=6}^{22}\boldsymbol{x}_{t-k}$).

5.4 Proposed models

Before introducing our new forecasting models based on graphs, we first provide some graph-related definitions for better understanding.

5.4.1 Background on graph construction and estimation

Definition 5.4.1 (Graph). *A graph $\mathcal{G}$ is defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_1, \ldots, v_N\}$ is a set of N nodes and $\mathcal{E}$ is a set of edges, where $e_{ij} = (v_i, v_j) \in \mathcal{E}$ denotes an edge connecting node v_i and node v_j.*

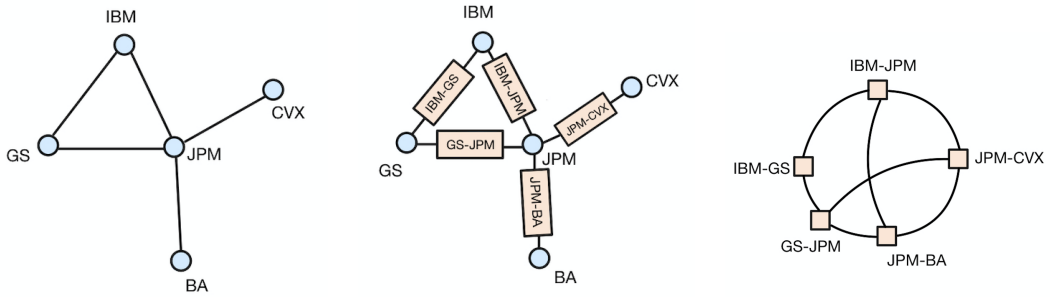
Definition 5.4.2 (Adjacency Matrix). *An adjacency matrix $\mathbf{A}$ is a square matrix whose dimension is $N \times N$, where $\mathbf{A}[i,j] = 1$ represents the connection between v_i and v_j in the graph $\mathcal{G}$, and $\mathbf{A}[i,j] = 0$ otherwise.*

Example 5.4.3 (Two Trivial Graphs). *The adjacency matrix of a **complete** graph (often referred to as K in the literature) contains all ones except along the diagonal where there are only zeros. The adjacency matrix of an **empty** graph is a zero matrix.*

Example 5.4.4 (The GICS Sector Graph). *The Global Industry Classification Standard (GICS) is an industry taxonomy developed for classifying major public companies. A GICS sector graph considers an edge between two asset nodes if they belong to the same GICS sector. Table 5.2 will introduce the corresponding sector of each stock under consideration.*

Definition 5.4.5 (Line Graph). *Given a graph $\mathcal{G}$, its line graph $L(\mathcal{G})$ is a graph such that*

- *each node of $L(\mathcal{G})$ represents an edge of $\mathcal{G}$;*
- *two nodes of $L(\mathcal{G})$ are adjacent if and only if their corresponding edges share a common endpoint in $\mathcal{G}$.*



(a) Graph $\mathcal{G}$ for volatility (b) Labeled edges in $\mathcal{G}$ (c) Line graph $L(\mathcal{G})$ for correlation

Figure 5.1: An illustration of the process of building the line graph $L(\mathcal{G})$ for $N = 5$ assets.

Figure 5.1 shows an illustration of graphs defined in our framework. Figure 5.1(a) models the interdependence between the volatility of assets, where each node represents an asset. Their edges, as shown in Figure 5.1(b), constitute a node in the line graph shown in Figure 5.1(c). The edges in the line graph capture the interdependence between two correlation pairs that have an asset in common.

In addition to the aforementioned domain-knowledge-based graphs, we also take into account a data-driven approach for the construction of graphs for volatilities.

Graphical LASSO (GLASSO) is a sparsity-penalized maximum likelihood estimator for the precision matrix (i.e. inverse of the covariance matrix) proposed by Friedman, Hastie, and Tibshirani [106]. Assume observations of $\boldsymbol{\epsilon}_t$ in Eqn (5.1) are drawn from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The GLASSO algorithm aims to obtain an estimator $\hat{\boldsymbol{\Theta}}$ that minimizes the following objective function

$$\hat{\boldsymbol{\Theta}} = \text{argmin}_{\boldsymbol{\Theta} \succeq 0} \left(\text{Tr}(\boldsymbol{S}\boldsymbol{\Theta}) - \log\det(\boldsymbol{\Theta}) + \lambda \sum_{j \neq k} |\boldsymbol{\Theta}_{jk}| \right),$$

where $\boldsymbol{S}$ is the sample covariance of $\boldsymbol{\epsilon}_t$, and λ is the penalizing trade-off parameter that controls the sparsity of $\hat{\boldsymbol{\Theta}}$. In this work, we choose λ on the training data using the standard cross-validation procedure.

An important and elegant feature of graphical models is the conditional independence properties of variables under the multivariate Gaussian distribution. If

the ij-th component of the precision matrix is zero, then the i-th and j-th variables are conditionally independent, given the other variables (for further details, see Friedman, Hastie, and Tibshirani [106] and Meinshausen and Bühlmann [175]). Alternatively, zeros in the precision matrix indicate conditional independence. We thus remove these zero elements in the estimated precision matrix and consider the non-zero elements as edges in the GLASSO graph, as below.

Example 5.4.6 (The GLASSO Graph)**.** *The adjacency matrix* $\boldsymbol{A}$ *from GLASSO is defined as* $\boldsymbol{A}[i,j] = 1$ *if* $\hat{\boldsymbol{\Theta}}[i,j] \neq 0$*; otherwise* $\boldsymbol{A}[i,j] = 0$*.*

5.4.2 Forecasting the realized covariance matrix with graphs

Based on the HAR-DRD model in Eqn (5.5), we propose to forecast the realized covariance matrix by modeling the realized volatilites and correlation matrix separately. For each subtask, we incorporate new features that represent the structural information contained in graphs via neighborhood aggregation.

Forecasting realized volatilities with graphs

The HAR model (Eqn (5.3)) assumes that the future volatility of a target asset only relies on its own past volatilities. Considering that volatilities of connected assets may have predictive power on the volatility of the target asset, we propose the following model to aggregate the spillover effects from its neighbors (i.e. connected assets), which we denote as the Graph-HAR model (or in short GHAR)

$$\begin{aligned}\textbf{GHAR}(\boldsymbol{A}):\ \boldsymbol{RV}_t = \boldsymbol{\alpha}^{(D)} &+ \underbrace{\beta_d^{(D)}\boldsymbol{RV}_{t-1} + \beta_w^{(D)}\boldsymbol{RV}_{t-5:t-2} + \beta_m^{(D)}\boldsymbol{RV}_{t-22:t-6}}_{Self} \\ &+ \underbrace{\gamma_d^{(D)}\boldsymbol{W}\cdot\boldsymbol{RV}_{t-1} + \gamma_w^{(D)}\boldsymbol{W}\cdot\boldsymbol{RV}_{t-5:t-2} + \gamma_m^{(D)}\boldsymbol{W}\cdot\boldsymbol{RV}_{t-22:t-6}}_{Graph} + \boldsymbol{u}_t^{(D)}\end{aligned} \tag{5.8}$$

where $\boldsymbol{W} = \boldsymbol{O}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{O}^{-\frac{1}{2}}$ is the normalized adjacency matrix, following Kipf and Welling [151] and Zhu et al. [239].[11] Specifically, $\boldsymbol{A}$ is a $N \times N$ adjacency matrix indicating the connections between assets with diagonal elements as 0,

[11] Normalizing the adjacency matrix makes it easier to compare the effects of *Self* and *Graph*.

and $\boldsymbol{O} = \text{diag}\{n_1, \ldots, n_N\}$, where $n_i = \sum_j \boldsymbol{A}[i,j]$, $\forall i$. Therefore $\boldsymbol{W} \cdot \boldsymbol{RV}_{t-1}, \boldsymbol{W} \cdot \boldsymbol{RV}_{t-5:t-2}, \boldsymbol{W} \cdot \boldsymbol{RV}_{t-22:t-6}$ represent the **neighborhood aggregation** over daily, weekly and monthly horizons. γ_d, γ_w, γ_m represent the spillover effects of connected neighbors over different horizons. For simplicity, we denote the effect from a stock's own historical events as *Self*, and that from its neighbors as *Graph*, as shown in Eqn (5.8). We use the Ordinary Least Squares (OLS) to obtain the estimates of Eqn (5.8).

In terms of the choice of adjacency matrices for volatility modeling, we first consider an empty graph, i.e. the elements of $\boldsymbol{A}$ are all 0s. In this case, Eqn (5.8) reduces to the standard vectorized HAR model (5.6) for modeling volatilities. When the off-diagonal elements of $\boldsymbol{A}$ are all 1s, i.e. a complete graph, $\boldsymbol{W} \cdot \boldsymbol{RV}_{t-1}$ represents the global volatility as studied in Bollerslev et al. [35]. Given that companies from the same sector tend to behave similarly and also influence each other, the GICS sector is another frequently-used graph for the construction of $\boldsymbol{A}$, e.g. Fan, Furger, and Xiu [98]. In addition, we employ a data-driven approach, i.e. graphical LASSO, to compute the adjacency matrix (see Hallac et al. [121]).

Forecasting realized correlations with graphs

Moreover, we apply the idea of the graph effect to modeling correlations according to the model

$$\begin{aligned}\textbf{GHAR}(\widetilde{\boldsymbol{A}}): \ \boldsymbol{x}_t = \boldsymbol{\alpha}^{(R)} &+ \underbrace{\beta_d^{(R)}\boldsymbol{x}_{t-1} + \beta_w^{(R)}\boldsymbol{x}_{t-5:t-2} + \beta_m^{(R)}\boldsymbol{x}_{t-22:t-6}}_{Self} \\ &+ \underbrace{\gamma_d^{(R)}\widetilde{\boldsymbol{W}}\boldsymbol{x}_{t-1} + \gamma_w^{(R)}\widetilde{\boldsymbol{W}}\boldsymbol{x}_{t-5:t-2} + \gamma_m^{(R)}\widetilde{\boldsymbol{W}}\boldsymbol{x}_{t-22:t-6}}_{Graph} + \boldsymbol{u}_t^{(R)},\end{aligned} \tag{5.9}$$

where $\widetilde{\boldsymbol{W}} = \widetilde{\boldsymbol{O}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{O}}^{-\frac{1}{2}}$ is the normalized adjacency matrix. Specifically, $\widetilde{\boldsymbol{A}}$ is a $N^{\#} \times N^{\#}$ (recall $N^{\#} = N(N-1)/2$) adjacency matrix indicating the connections between pairwise correlations with diagonal elements as 0, and $\widetilde{\boldsymbol{O}} = \text{diag}\{\tilde{n}_1, \ldots, \tilde{n}_{N^{\#}}\}$, where $\tilde{n}_i = \sum_j \widetilde{\boldsymbol{A}}[i,j]$, $\forall i$.

In terms of adjacency matrices for pair correlations, in addition to the empty graph (all 0's) and complete graph (all 1's), we also adopt the line graph, where the correlation $\boldsymbol{R}[i,j]$ between asset i and j $(i \neq j)$ is connected to another correlation

Table 5.1: Choices of adjacency matrices for modeling volatilities and correlations, and the joint models for forecasting realized covariances.

$\boldsymbol{A}$ \ $\widetilde{\boldsymbol{A}}$	Empty	Complete	Line
Empty	HAR-DRD	GHAR(-, $\widetilde{K}$)	GHAR(-, $\widetilde{L}$)
Complete	GHAR(K, -)	GHAR(K, $\widetilde{K}$)	GHAR(K, $\widetilde{L}$)
Sector	GHAR(S, -)	GHAR(S, $\widetilde{K}$)	GHAR(S, $\widetilde{L}$)
GLASSO	GHAR(GL, -)	GHAR(GL, $\widetilde{K}$)	GHAR(GL, $\widetilde{L}$)

$\boldsymbol{R}[k,l]$ between k and l ($k \neq l$), iff $\{i,j\} \cap \{k,l\} \neq \varnothing$ and $\{i,j\} \neq \{k,l\}$. It is not difficult to obtain that the ratio between the degree [12] sum of a line graph and the degree sum of a complete graph is $\frac{N^{\#}(N-2)}{N^{\#}(N^{\#}-1)/2} = \frac{4(N-2)}{N(N-1)-2}$. To the best of our knowledge, this is the first study to explore the predictive information of line graphs for forecasting realized correlation matrices.

As a notational convention, we use $\boldsymbol{A}$ (respectively, $\widetilde{\boldsymbol{A}}$) to denote the adjacency matrix for volatility (respectively, correlation). Additionally, we refer to GHAR($\boldsymbol{A}$) (respectively, GHAR($\widetilde{\boldsymbol{A}}$)) as the augmented HAR model with graph information $\boldsymbol{A}$ (respectively, $\widetilde{\boldsymbol{A}}$) for forecasting volatilities (respectively, correlations). In the end, GHAR($\boldsymbol{A}, \widetilde{\boldsymbol{A}}$) is referred to as the model for forecasting realized covariances, which combines the forecasts of volatilities based on $\boldsymbol{A}$ and correlations based on $\widetilde{\boldsymbol{A}}$. Table 5.1 lists the adjacency matrices for realized volatilities and correlations, and the corresponding joint models.

5.5 Empirical analysis

5.5.1 Data

We obtain intraday data on the components of Dow Jones Industrial Average (DJIA) from LOBSTER for the period from July 1, 2007 to July 1, 2021. Following Bollerslev, Patton, and Quaedvlieg [38], only stocks that traded continuously from the start to the end of our sample are maintained. As a result, 27 Dow Jones constituents are in the final sample and their ticker symbols are listed in Table

[12] In graph theory, the degree of a node of a graph is the number of edges that are adjacent to the node.

Table 5.2: Summary statistics of realized volatilities.

	Mean	Std	Min	25%	50%	75%	Max	Sector
AAPL	2.30	3.39	0.07	0.70	1.25	2.46	38.30	Information Technology
MSFT	1.82	2.51	0.11	0.67	1.09	1.92	30.64	Information Technology
INTC	2.29	3.12	0.14	0.86	1.39	2.44	42.90	Information Technology
CSCO	1.98	2.92	0.14	0.70	1.13	2.09	43.74	Information Technology
CRM	4.00	4.93	0.22	1.44	2.41	4.64	61.67	Information Technology
IBM	1.38	2.33	0.11	0.47	0.75	1.34	30.22	Information Technology
DIS	1.89	3.04	0.12	0.60	1.01	1.88	40.56	Communication Services
VZ	1.40	2.36	0.10	0.50	0.77	1.33	34.19	Communication Services
HD	2.11	3.59	0.15	0.62	1.02	2.01	48.22	Consumer Discretionary
MCD	1.17	2.15	0.08	0.39	0.61	1.13	37.57	Consumer Discretionary
NKE	2.03	3.00	0.14	0.74	1.15	2.02	47.87	Consumer Discretionary
PG	1.00	1.76	0.09	0.38	0.58	0.98	31.60	Consumer Staples
KO	0.99	1.68	0.07	0.37	0.58	1.00	25.00	Consumer Staples
WMT	1.18	1.76	0.11	0.45	0.67	1.18	27.18	Consumer Staples
JNJ	0.92	1.56	0.06	0.35	0.54	0.90	24.74	Health Care
UNH	2.70	4.34	0.16	0.78	1.35	2.57	52.54	Health Care
MRK	1.65	2.45	0.12	0.58	0.92	1.74	30.99	Health Care
AMGN	1.91	2.34	0.16	0.82	1.27	2.14	33.44	Health Care
JPM	3.46	7.04	0.15	0.74	1.36	2.82	108.17	Financials
AXP	3.19	6.32	0.12	0.64	1.15	2.67	91.45	Financials
GS	3.24	6.27	0.19	0.92	1.49	2.81	112.41	Financials
TRV	2.04	4.09	0.11	0.49	0.81	1.76	57.95	Financials
BA	2.69	5.00	0.13	0.78	1.35	2.60	90.65	Industrials
HON	1.85	3.25	0.10	0.52	0.97	1.84	49.64	Industrials
MMM	1.43	2.25	0.08	0.46	0.81	1.49	31.11	Industrials
CAT	2.79	4.00	0.15	0.94	1.58	2.89	45.26	Industrials
CVX	2.03	3.51	0.13	0.61	1.07	2.04	48.07	Energy

Note: The table reports summary statistics for the daily realized volatility and the corresponding sector of 27 stocks in DJIA. The statistics are averaged across each trading day.

5.2, where we also provide summary statistics for the volatility estimates and the corresponding sector of each stock.

Figure 5.2 reports the average daily realized correlations across the 27 stocks. We calculate the realized correlation matrix of stock returns for each day and then compute an average across all trading days. All cross-correlations are positive and range between 0.17 and 0.63. The correlation matrix displays a block-diagonal structure that corresponds to GICS sector classification, which is in line with previous studies (e.g. Benzaquen et al. [30]).

In addition to the usual summary measures, we also report the autocorrelations for realized volatilities and correlations. Figure 5.3 reveals that the autocorrelation

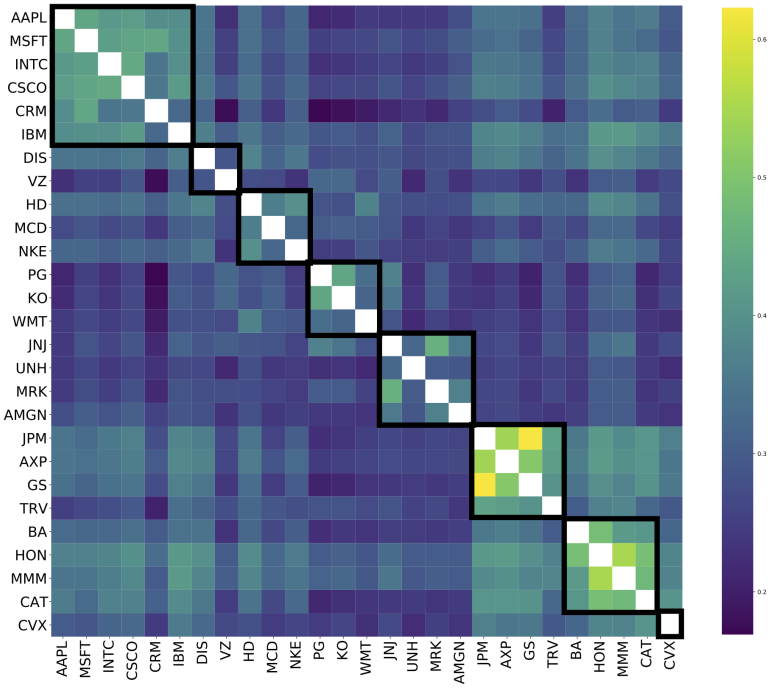


Figure 5.2: Realized correlation matrix.

Note: The figure shows the daily realized cross-correlations of 27 stocks in DJIA, averaged across each trading day. Stocks are sorted by their industrial sectors.

coefficients of volatilities are consistently larger than those of correlations, directly highlighting the importance of individual models for volatilities and correlations (see Andersen et al. [11]).

5.5.2 Forecast evaluation metrics

Following the convention in the literature, we use hats to denote forecasts. For example, we refer to the forecast of the conditional covariance matrix on day t as $\widehat{\mathbf{\Sigma}}_t$. Consistent with previous studies (e.g. Bollerslev, Patton, and Quaedvlieg [39] and Symitsi et al. [210]), we consider the following loss functions to measure the average distance between predicted covariances and realized covariances for the model comparisons,

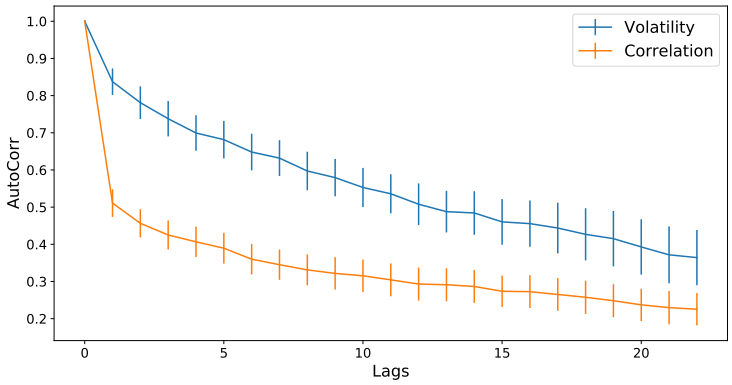


Figure 5.3: Autocorrelation of realized volatilities and correlations.

Note: The figure provides the standard autocorrelation coefficients and their standard deviations (vertical error bars) of the daily realized volatilities and correlations. The autocorrelation statistics for volatilities (respectively, correlations) are averaged across trading days and stocks (respectively, stock pairs).

$$\begin{aligned}
\mathcal{L}_t^E &= \sqrt{\mathrm{vech}_{Tri}\left(\boldsymbol{H}_t - \widehat{\boldsymbol{\Sigma}}_t\right)' \mathrm{vech}_{Tri}\left(\boldsymbol{H}_t - \widehat{\boldsymbol{\Sigma}}_t\right)}, \\
\mathcal{L}_t^F &= \sqrt{\mathrm{Tr}\left[\left(\boldsymbol{H}_t - \widehat{\boldsymbol{\Sigma}}_t\right)' \left(\boldsymbol{H}_t - \widehat{\boldsymbol{\Sigma}}_t\right)\right]}, \\
\mathcal{L}_t^Q &= \log\det(\widehat{\boldsymbol{\Sigma}}_t) + \mathrm{Tr}\left[\widehat{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{H}_t\right].
\end{aligned} \tag{5.10}$$

Here $\mathcal{L}_t^E$ represents the Euclidean distance between the vectorization of forecast covariances and ex-post realized covariances. $\mathcal{L}_t^F$ is the Frobenius distance between the two matrices. $\mathcal{L}_t^Q$ is the quasi-likelihood (QLIKE) loss based on the negative log-likelihood of a multivariate normal. All of these functions measure losses, therefore lower values are preferred.

To assess the statistical significance of our results, we also employ the MCS test (see more in 4.5.3) to compare the models under consideration.

5.5.3 In-sample results

Following the workflow in Bollerslev, Patton, and Quaedvlieg [38, 39], we start the empirical analysis with an in-sample test of each model under consideration, followed by a graph structure study.

Model evaluation

Table 5.3 reports the estimated coefficients for modeling the one-day-ahead realized covariances under the Cholesky decomposition, and the estimated coefficients for modeling one-day-ahead realized volatilities and correlations under the DRD decomposition. The HAR-Cholesky model attributes the most importance to mid-term RV, followed by the monthly and daily components.[13] Additionally, our results support a stronger effect of the first lag on volatilities ($\beta_d = 0.609$) than the respective one on correlations ($\beta_d = 0.289$), consistent with the existing literature (e.g. Bollerslev, Patton, and Quaedvlieg [39]).

When examining the graph effect, we find the GHAR model for forecasting volatilities redistributes some weights from a stock's own daily lag to the daily lag of its connected neighbors. We observe a similar shift for correlation forecasting. Interestingly, the estimated coefficients of the weekly and monthly lags from neighbors are negative and statistically significant in the in-sample test. Future studies on the negative graph effects of longer lags are therefore recommended.

Table 5.4 reports the average in-sample predictive performance of each model under consideration. To conserve space, we present (i) the ratio of loss functions when comparing each model with HAR-DRD; (ii) the rank of loss values in 13 model candidates; (iii) p-values of the MCS test. In this manner, a value of ratio less than 1 indicates the predictive performance is better than the baseline model HAR-DRD; a rank of 1 indicates the model has the best in-sample fit; p-val greater than 5% indicates the model(s) are superior to all other models identified by the MCS test.

From Table 5.4, we summarize the following findings. First, we conclude that HAR-Cholesky produces worse results than HAR-DRD, which is consistent with the conclusion of Bollerslev, Patton, and Quaedvlieg [39]. The results further show that the GHAR(GL, $\widetilde{L}$) model achieves the best in-sample fit across all loss functions. For example, looking at the value of the $\mathcal{L}^E$ (respectively, $\mathcal{L}^F$, $\mathcal{L}^Q$) loss function, we observe that the average loss of GHAR(GL, $\widetilde{L}$) is about 98.3% (respectively,

[13]The HAR-Cholesky in Bollerslev, Patton, and Quaedvlieg [39] model allocates roughly 0.247, 0.410, and 0.244 to the daily, weekly, and monthly components, respectively.

Table 5.3: In-sample estimation.

		β_d	β_w	β_m	γ_d	γ_w	γ_m
Covariance	HAR-Cholesky	0.311	0.413	0.242			
		(0.001)	(0.001)	(0.001)			
Volatility	HAR	0.609	0.271	0.057			
		(0.003)	(0.004)	(0.003)			
	GHAR(K)	0.483	0.279	0.169	0.239	-0.075	-0.150
		(0.004)	(0.006)	(0.005)	(0.006)	(0.008)	(0.006)
	GHAR(S)	0.492	0.277	0.155	0.187	-0.048	-0.120
		(0.004)	(0.006)	(0.006)	(0.005)	(0.007)	(0.006)
	GHAR(GL)	0.471	0.285	0.144	0.224	-0.075	-0.103
		(0.004)	(0.006)	(0.006)	(0.005)	(0.007)	(0.006)
Correlation	HAR	0.289	0.386	0.213			
		(0.001)	(0.001)	(0.001)			
	GHAR($\widetilde{K}$)	0.144	0.268	0.484	0.348	0.068	-0.413
		(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
	GHAR($\widetilde{L}$)	0.076	0.222	0.595	0.407	0.121	-0.521
		(0.001)	(0.002)	(0.003)	(0.002)	(0.003)	(0.004)

Note: The table reports the in-sample estimates for coefficients and their standard errors (in parentheses) of every model, calculated from the entire sample period. The Covariance row reports the HAR-Cholesky model, for forecasting the Cholesky decomposition vector of the realized covariance matrix. The Volatility rows include different models for forecasting multivariate realized volatilities, while the Correlation rows include different models for forecasting the lower triangular part of the realized correlation matrix.

98.4%, 98.2%) of the baseline, HAR-DRD. In general, all graph models outperform the baseline, providing strong evidence for the importance of leveraging graph information and the existence of certain spillover effects.

Graph analysis

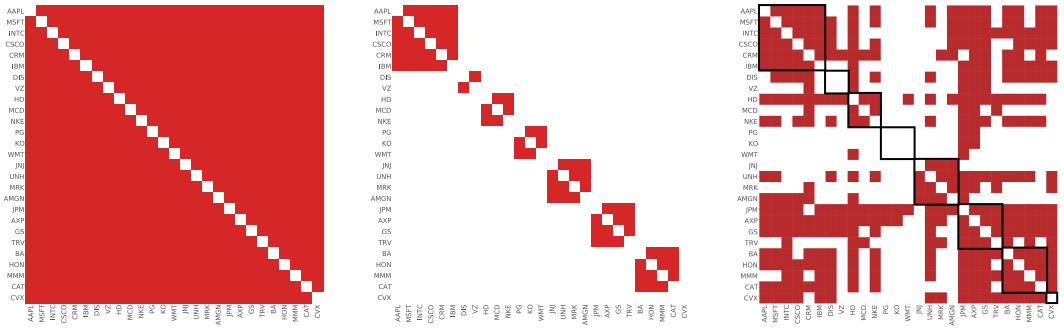
Figure 5.4 displays the adjacency matrices of various graphs for volatilities. Recall that the neighborhood aggregation via a complete graph (shown in Figure 5.4(a)) is calculating the global (or market) volatility, and the neighborhood aggregation via the GICS sector graph (shown in Figure 5.4(b)) is equivalent to calculating the average volatility within each sector, while ignoring the inter-sector links.

The adjacency matrix chosen by GLASSO (shown in Figure 5.4(c)) maintains a GICS sector structure to some extent, such as information technology, consumer discretionary, health care, financials, and industrials. Meanwhile, we observe substantial inter-sector connections, especially in financials and information technology,

Table 5.4: In-sample losses.

	$\mathcal{L}^E$			$\mathcal{L}^F$			$\mathcal{L}^Q$		
	Ratio	Rank	p-val	Ratio	Rank	p-val	Ratio	Rank	p-val
HAR-Cholesky	1.023	13	<0.001	1.032	13	<0.001	1.051	13	<0.001
HAR-DRD	1.000	12	<0.001	1.000	12	<0.001	1.000	10	<0.001
GHAR(-, $\widetilde{K}$)	0.997	11	<0.001	0.997	11	<0.001	0.987	6	<0.001
GHAR(-, $\widetilde{L}$)	0.995	10	<0.001	0.994	10	<0.001	0.985	4	<0.001
GHAR(S, -)	0.990	9	<0.001	0.992	9	<0.001	1.001	11	<0.001
GHAR(S, $\widetilde{K}$)	0.988	8	<0.001	0.989	8	<0.001	0.987	5	<0.001
GHAR(S, $\widetilde{L}$)	0.986	7	<0.001	0.987	5	<0.001	0.985	3	<0.001
GHAR(K, -)	0.986	5	<0.001	0.987	6	<0.001	1.006	12	<0.001
GHAR(K, $\widetilde{K}$)	0.985	4	<0.001	0.986	4	<0.001	0.991	8	<0.001
GHAR(K, $\widetilde{L}$)	0.983	2	0.557	0.984	2	0.745	0.989	7	<0.001
GHAR(GL, -)	0.986	6	<0.001	0.988	7	<0.001	0.998	9	<0.001
GHAR(GL, $\widetilde{K}$)	0.984	3	<0.001	0.986	3	<0.001	0.985	2	<0.001
GHAR(GL, $\widetilde{L}$)	0.983	1	1.000	0.984	1	1.000	0.982	1	1.000

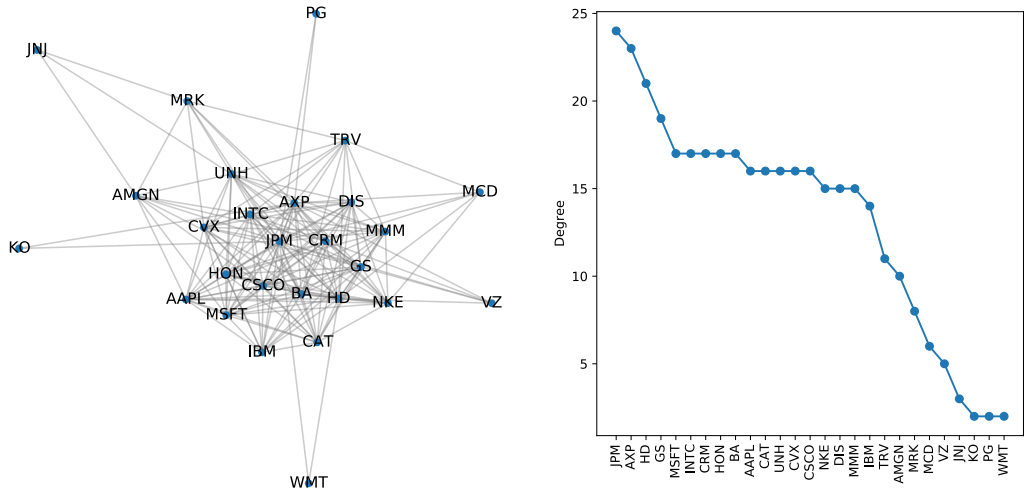
Note: The table reports the in-sample losses of various models over the 1-day forecast horizons, averaged over the entire sample. $\mathcal{L}^E$ is the Euclidean distance, $\mathcal{L}^F$ is the Frobenius distance, and $\mathcal{L}^Q$ is the Quasi-Likelihood loss function.



(a) Complete (K) (b) Sector (S) (c) GLASSO (GL)

Figure 5.4: Adjacency matrices for realized volatility.

Note: The figure shows the adjacency matrices from various methods, for forecasting realized volatility. The stocks are sorted by their industrial sectors, as listed in Table 5.2. The red unit represents 1, i.e. a connection between two stocks, while the white unit represents 0, i.e. no connection.



(a) Connected components (b) Degree-rank plot

Figure 5.5: Illustration of the graph corresponding to the adjacency matrix in 5.4(c) and degree-rank plot.

for capturing market information.

To better illustrate the interactions between stocks modeled by GLASSO, we construct a network based on the adjacency matrix of 5.4(c), shown in Figure 5.5(a). We also calculate the node degrees and sort stocks in descending order of degree. Figure 5.5(b) reveals that financial stocks, such as JPM, AXP, and GS, appear to be the center of the volatility network, followed by technology companies, including MSFT, INTC, etc.

5.5.4 Out-of-sample results

We now focus our discussion on the out-of-sample predictive performance of the competing models in a rolling window setup.

Model evaluation

Following Bollerslev, Patton, and Quaedvlieg [38, 39], Pascalau and Poirier [187], and Symitsi et al. [210], we estimate the parameters of each model using a fixed length rolling window of the most recent 1000 observations and obtain the rolling 1-day forecasts for each trading day in the next month. To this end, the out-

of-sample period lasts from July 2011 to July 2021. All of the models are re-estimated every month.[14]

To ensure that the predicted correlations lie in $[-1, 1]$, we apply an "insanity filter", meaning that we set the values below -1 to -1, and those above 1 to 1 in the following experiments.[15] Furthermore, to ensure the positiveness and definiteness of covariances, we replace the negative definite covariance matrices with the simple average realized covariance matrix over the relevant estimation sample period (same as Bollerslev, Patton, and Quaedvlieg [39]).

Table 5.5 illustrates the average losses over the entire out-of-sample period, i.e. July 2011-July 2021. We first focus on the comparison between HAR-Cholesky and HAR-DRD. Consistent with the in-sample results, HAR-DRD significantly outperforms HAR-Cholesky across all losses. Furthermore, the results show that the GHAR(GL, $\widetilde{L}$) model yields the most accurate out-of-sample forecasts across all loss functions. Specifically, in terms of $\mathcal{L}_E$ (respectively, $\mathcal{L}_F$, $\mathcal{L}_Q$), the GHAR(GL, $\widetilde{L}$) model has about 2.5% (respectively, 2.5%, 1.8%) lower average forecast error compared to the baseline HAR-DRD. To disentangle the prediction improvements in covariances, we compare the performance of HAR-DRD, GHAR(GL, -), GHAR(-, $\widetilde{L}$), and GHAR(GL, $\widetilde{L}$). We observe that the graph information helps improve the predictive performance in both volatilities and correlations forecasting, which subsequently leads to an improved prediction of the covariance matrix. Indeed, GHAR(GL, $\widetilde{L}$), which utilizes both graphs in volatility and correlation simultaneously, clearly yields superior results compared to the models that only utilize one separate graph either in volatility or correlation. The MCS test indicates that in addition to GHAR(GL, $\widetilde{L}$), the GHAR(K, $\widetilde{L}$) model is also included in the subset of best models, across all loss functions. One possible reason why GHAR(K, $\widetilde{L}$) also performs well is that the studied universe consists of highly liquid stocks that may

[14] In Appendix D.3, we also report the performance of the models updated more frequently, such as daily and weekly, which leads to similar conclusions.

[15] Another common method is to use some transformation functions to convert volatilities and correlations, e.g. log transformation for volatilities (see Andersen et al. [13], Bucci [44], and Zhang et al. [233]) and Fisher transformation for correlations (see Andersen et al. [11]). We report the results in Appendix D.4.

Table 5.5: Out-of-sample losses over 1-day forecast horizon.

	$\mathcal{L}^E$			$\mathcal{L}^F$			$\mathcal{L}^Q$		
	Ratio	Rank	p-val	Ratio	Rank	p-val	Ratio	Rank	p-val
HAR-Cholesky	1.031	13	<0.001	1.032	13	<0.001	1.071	13	<0.001
HAR-DRD	1.000	12	0.002	1.000	12	0.003	1.000	9	<0.001
GHAR(-, $\widetilde{K}$)	0.993	11	0.002	0.991	11	0.003	0.986	7	<0.001
GHAR(-, $\widetilde{L}$)	0.991	10	0.004	0.989	9	0.006	0.984	3	0.048
GHAR(S, -)	0.990	9	0.002	0.990	10	0.003	1.002	12	<0.001
GHAR(S, $\widetilde{K}$)	0.983	8	0.003	0.983	7	0.005	0.987	8	<0.001
GHAR(S, $\widetilde{L}$)	0.982	7	0.004	0.981	5	0.006	0.985	6	<0.001
GHAR(K, -)	0.982	5	0.004	0.983	6	0.006	1.001	11	<0.001
GHAR(K, $\widetilde{K}$)	0.976	4	0.004	0.976	3	0.006	0.985	5	<0.001
GHAR(K, $\widetilde{L}$)	0.975	2	0.869	0.975	1	1.000	0.982	2	0.277
GHAR(GL, -)	0.982	6	0.004	0.983	8	0.006	1.001	10	<0.001
GHAR(GL, $\widetilde{K}$)	0.976	3	0.005	0.976	4	0.006	0.984	4	<0.001
GHAR(GL, $\widetilde{L}$)	0.975	1	1.000	0.975	2	0.768	0.982	1	1.000

Note: The table reports the out-of-sample losses of various models over the 1–day forecast horizon, averaged over the entire testing sample. $\mathcal{L}^E$ is the Euclidean distance, $\mathcal{L}^F$ is the Frobenius distance, and $\mathcal{L}^Q$ is the Quasi-Likelihood loss function.

have strong correlations and thus are primarily influenced by the market. Further research on a larger universe is recommended.

In addition, GHAR(S, $\cdot$) has an inferior predictive performance, compared to GHAR(GL, $\cdot$). Except for the fact that there are no inter-sector links in sector graphs, another possible reason might be the lack of temporal dynamics in sector graphs.[16] This also motivates us to examine the dynamics of graphs captured by GLASSO in the following subsection.

We further investigate whether the coefficients in each model vary across time. Figure 5.6 displays the rolling betas (and gammas) and their respective 95% confidence intervals for modeling volatilities over the testing period. The plots of HAR first show that there are three substantial shocks in β_d, which occur at the end of 2012, the end of 2015, and March 2020. In all GHAR models, β_d rapidly drops at the end of 2012 and 2015 as well. However, β_d appears to be more consistent during the pandemic period (March 2020), while the graph coefficient γ_d displays a high

[16]We use the up-to-date GICS sector to construct a static graph employed during the entire period.

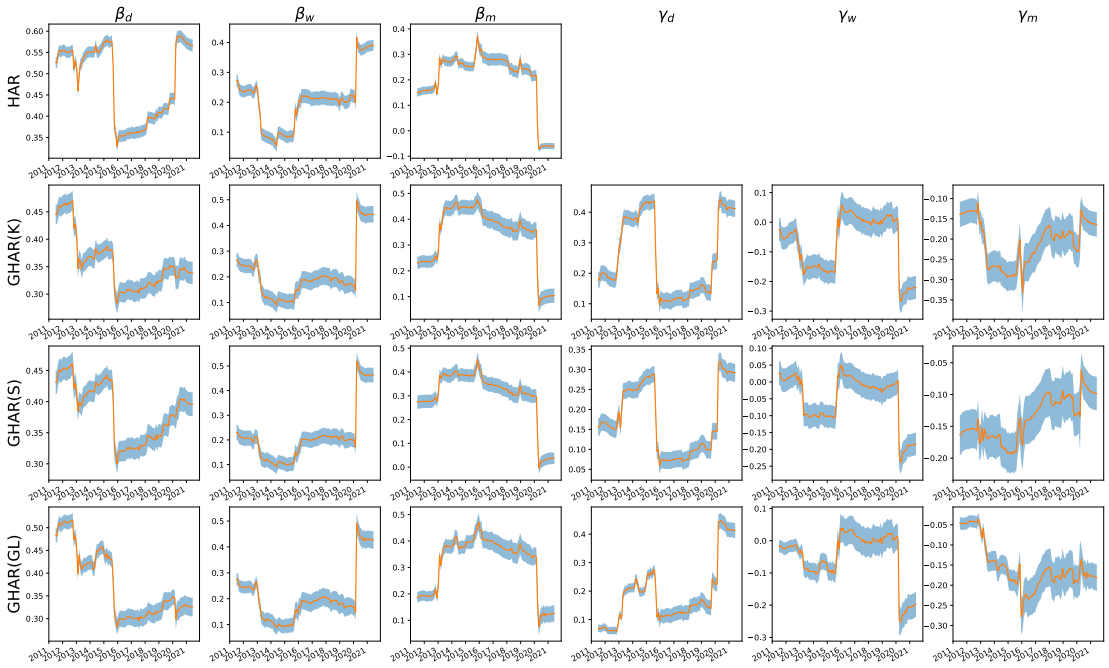


Figure 5.6: Rolling coefficients of various models for forecasting volatilities.

Note: The orange curve represents the estimated coefficients, with the blue area covering the 95% confidence interval.

jump during the COVID-19 pandemic. This may indicate that the spillover effect is a major driver of market volatility during the pandemic. Another surprising aspect of the plots is that β_w has a similar trend across various models, and similarly for β_m. We also observe that γ_w is insignificant during 2016-2020 at the 5% confidence level. These findings indicate that mid- and long-term graph effects may have a minor influence on volatility dynamics.[17]

We provide a similar analysis of the models for correlations, illustrated in Figure 5.7. First, the rolling coefficients appear to be less volatile than their counterparts when modeling volatilities. Second, the rolling coefficients look similar across graph-based methods. In addition, we observe that the coefficients β_d, β_w, and β_m are predominantly smoother for graph-based models compared to their counterparts in the standard HAR. Interestingly, the short-term graph effect γ_d in graph-based models exhibits a similar trend as β_d in HAR.

[17]We study the graph-based model with only short-term graph effect in Appendix D.2.

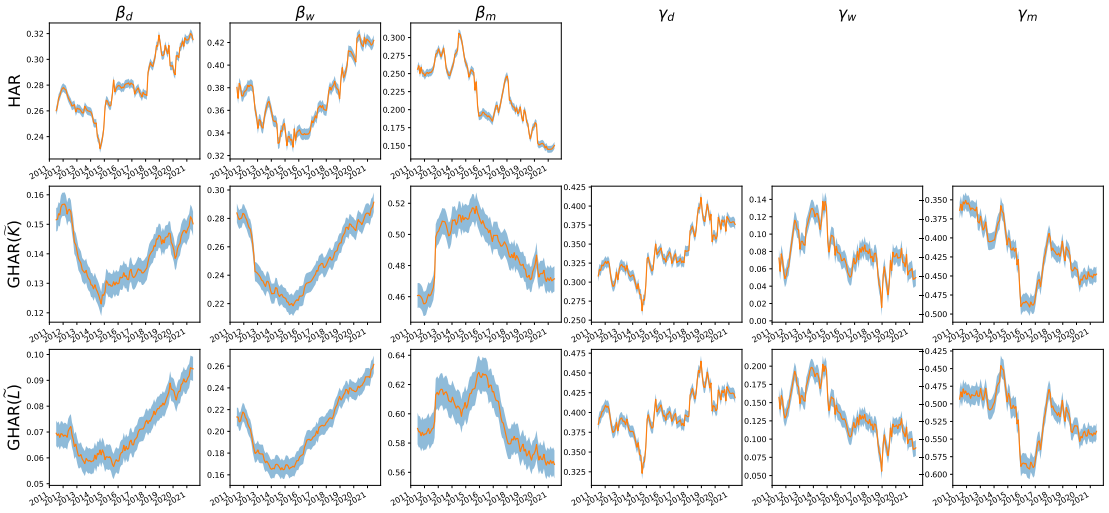


Figure 5.7: Rolling coefficients of various models for forecasting correlations.

Note: The orange curve represents the estimated coefficients, with the blue area covering the 95% confidence interval.

Graph analysis

Since GHAR(GL, $\widetilde{L}$) delivers the best forecasting results, we further dive into the dynamics of the volatility graphs uncovered by GLASSO. For each adjacency matrix at a timestamp, we calculate the fraction of non-sector stocks that are connected to sector members, a.k.a. *group degree centrality* (see more in Everett and Borgatti [97]). For ease of visualization, in Figure 5.8, we only plot the group degree centrality of four GICS sectors with most constituents. We first observe that the centrality for each sector maintains at a higher level since 2015, implying that sector connectedness becomes increasingly important in the U.S. equity market,[18] which also helps explain why GHAR(K, $\widetilde{L}$) performs well. In addition, we observe that at the beginning of the COVID-19 pandemic, healthcare and information technology companies exhibit a slightly increasing influence on other stocks, which is not surprising as they are two critical sectors whose prices have been largely affected during the pandemic.

To study the stability and temporal dynamics of graphs learned by GLASSO, we also calculate the similarity of two consecutive graphs at monthly basis in terms of their edge sets. The Jaccard index of the edge set is defined as the ratio between

[18] In line with the findings of Baruník, Kočenda, and Vácha [26] and Costa, Matos, and Silva [76].

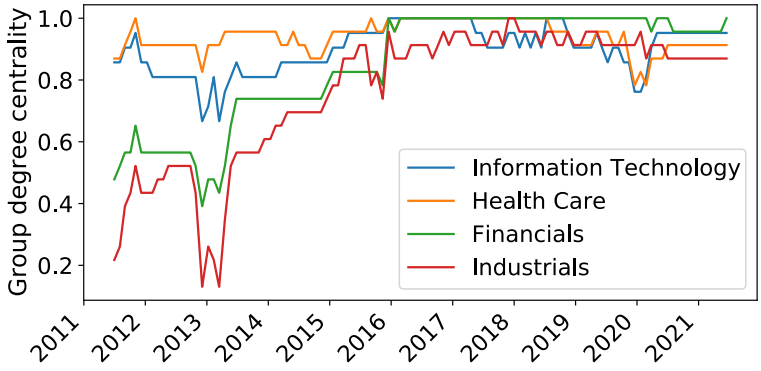


Figure 5.8: Group degree centrality of 4 sectors in the graphs from GLASSO.

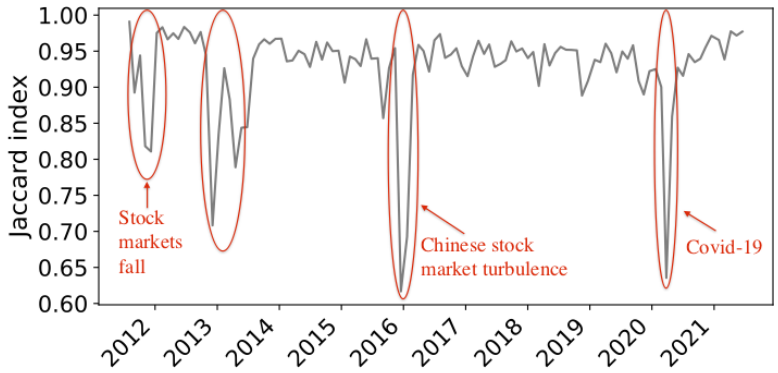


Figure 5.9: Jaccard index of similarity between consecutive graphs.

the number of intersection edges and the number of union edges of two consecutive graphs. Specifically, at time t, the edge set of the GLASSO estimate is denoted as $\mathcal{E}^{(t)}$, with the Jaccard index computed as $\frac{|\mathcal{E}^{(t)} \cap \mathcal{E}^{(t-1)}|}{|\mathcal{E}^{(t)} \cup \mathcal{E}^{(t-1)}|}$.

We plot the Jaccard index between consecutive GLASSO estimates against the timespan in Figure 5.9. The graphs are stable in general, with only a few exceptions: Aug 2011 - Dec 2011, Nov 2012 - June 2013, Dec 2015 - Feb 2016, and March 2020 - April 2020. This suggests that the equity market might be experiencing significant changes during these periods. In fact, some periods correspond to stock market turmoils, as highlighted in Figure 5.9. To the best of our knowledge, there was no major financial event from the end of 2012 until mid 2013 documented in the literature, yet very recent methods in the change-point detection machine learning literature have also picked up this regime (see Sulem et al. [209]).

5.5.5 Portfolio performance

To further quantify the benefit of our proposed GHAR models, we examine the portfolio performance based on the covariance forecasts (see Bollerslev, Patton, and Quaedvlieg [39] and Symitsi et al. [210]). We construct the global minimum variance portfolio (GMVP) to bypass the problem of forecasting future returns, as the scope of this chapter is on covariance forecasting. We calculate the weights of GMVP as

$$\min_{\boldsymbol{w}_t} \boldsymbol{w}_t' \widehat{\boldsymbol{\Sigma}}_t \boldsymbol{w}_t \quad \text{s.t.} \quad \boldsymbol{w}_t' \mathbf{1} = 1 \tag{5.11}$$

where $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^N$. The optimal weights of the GMVP are given by $\hat{\boldsymbol{w}}_t = \frac{\widehat{\boldsymbol{\Sigma}}_t^{-1}\mathbf{1}}{\mathbf{1}'\widehat{\boldsymbol{\Sigma}}_t^{-1}\mathbf{1}}$. Finally, we compute the portfolio returns as $r_t^{(p)} = \hat{\boldsymbol{w}}_t' \boldsymbol{r}_t$.

In addition, we consider another portfolio imposed with the **long-only** constraint, denoted as GMVP$^+$. In this case, we are not allowed to short-sell stocks, i.e. $w_{i,t} \geq 0, \ \forall i, t$. Since there is no closed-form solution for GMVP$^+$, we use the quadratic programming solver in the built-in Python package CVXPY[19] to compute the optimized portfolio weights.

Based on the covariance forecasts from the competing models, we obtain multiple corresponding portfolios. We then use the following metrics to evaluate the out-of-sample performance of each portfolio.

- (1) Annualized portfolio standard deviation:

$$\sigma^{(p)} = \sqrt{252 \times \frac{1}{T} \sum_{t=1}^{T} \left(r_t^{(p)} - \bar{r}_t^{(p)} \right)^2}, \tag{5.12}$$

where T is the length of the out-of-sample period, and $\bar{r}_t^{(p)}$ is the average return of a particular portfolio over the out-of-sample period.

- (2) Average portfolio turnover:

$$\begin{aligned} TO_t &= \sum_{i=1}^{N} \left| w_{i,t+1} - w_{i,t} \frac{1 + r_{t,i}}{1 + \boldsymbol{w}_t' \boldsymbol{r}_t} \right|, \\ \tau^{(p)} &= \frac{1}{T-1} \sum_{t=1}^{T-1} TO_t, \end{aligned} \tag{5.13}$$

[19] https://www.cvxpy.org

where $w_{i,t}$ (respectively, $r_{i,t}$) is the i-th element in $\boldsymbol{w}_t$ (respectively, $\boldsymbol{r}_t$) and TO_t represents the portfolio turnover from day t to day $t+1$.

To further gauge the economic value of covariance forecasts, we also compare the aforementioned portfolios with the naive equally-weighted portfolio (denoted as 1/N), a common benchmark in the literature.[20] Table 5.6 illustrates the annualized standard deviation and average turnover for GMVP and GMVP$^+$ over the out-of-sample period, assuming no transaction cost. The results show that using high-frequency data for forecasting realized covariance leads to superior portfolios compared to the naive 1/N portfolio in terms of standard deviation. Portfolios with graph information, especially GHAR(GL, $\widetilde{L}$), are able to generate lower out-of-sample variance compared to the one without any graphs, i.e. HAR-DRD, further proving evidence for the existence of spillover effects.

With regards to portfolio turnover, except for the 1/N portfolio, the portfolios constructed from models that employ complete graphs yield the lowest average turnover. This is not surprising as the market volatility with equal weights on individual stocks is incorporated into such models. Moreover, portfolios based on GLASSO graphs generally exhibit lower turnover relative to the portfolios based on sector graphs, or without graphs.

5.5.6 Longer future horizons

One-day-ahead forecasting is not the only time horizon of interest to practitioners. In this section, we consider various prediction horizons, which allows us to examine how the impact of graph information varies on future covariances. To obtain longer horizon forecasts, we adopt the direct approach (in line with Andersen et al. [13] and Bollerslev, Patton, and Quaedvlieg [39]), rather than the iterated approach.[21] Specifically, for the GHAR formulation of volatility in Eqn (5.8), we implement

[20] DeMiguel, Garlappi, and Uppal [84] demonstrated that of the 14 models they evaluated across seven empirical datasets, none is consistently better than the 1/N portfolio in terms of Sharpe ratio, certainty-equivalent return, or turnover.

[21] Bollerslev, Patton, and Quaedvlieg [39] demonstrated that even minor model mis-specifications will be amplified in the iterated one-day-ahead forecasts, and result in worse estimations than the direct forecast procedure in practice.

Table 5.6: Out-of-sample portfolio performance over 1-day forecast horizon.

	GMVP		GMVP$^+$	
	$\sigma^{(p)}$	$\tau^{(p)}$	$\sigma^{(p)}$	$\tau^{(p)}$
1/N	11.890	0.007	11.890	0.007
HAR-Cholesky	9.948	0.664	9.917	0.411
HAR-DRD	9.918	0.771	9.936	0.511
GHAR(-, $\widetilde{K}$)	9.967	0.730	9.935	0.482
GHAR(-, $\widetilde{L}$)	9.970	0.761	9.951	0.510
GHAR(S, -)	9.906	0.738	10.005	0.503
GHAR(S, $\widetilde{K}$)	9.933	0.675	9.983	0.468
GHAR(S, $\widetilde{L}$)	9.930	0.717	9.993	0.499
GHAR(K, -)	9.857	0.653	9.947	0.432
GHAR(K, $\widetilde{K}$)	9.896	0.590	9.930	0.389
GHAR(K, $\widetilde{L}$)	9.893	0.634	9.945	0.426
GHAR(GL, -)	9.831	0.695	9.911	0.475
GHAR(GL, $\widetilde{K}$)	9.825	0.630	**9.883**	0.435
GHAR(GL, $\widetilde{L}$)	**9.799**	0.675	9.901	0.470

Note: The table reports the out-of-sample performance of GMVP and GMVP$^+$ constructed using the one-day-ahead covariance forecasts of various models. $\sigma^{(p)}$ is the annualized portfolio standard deviation, and $\tau^{(p)}$ is the average portfolio turnover. The lowest $\sigma^{(p)}$ in each column is indicated in bold.

the following model over longer forecast horizons. We apply the same idea to the modeling of correlations, or the vector of Cholesky decomposition of covariances.

$$\begin{aligned}\textbf{GHAR}(\boldsymbol{A}) : \boldsymbol{RV}_{t:t+h} = \boldsymbol{\alpha}^{(D)} + \beta_d^{(D)}\boldsymbol{RV}_{t-1} + \beta_w^{(D)}\boldsymbol{RV}_{t-5:t-2} + \beta_m^{(D)}\boldsymbol{RV}_{t-22:t-6} \\ + \gamma_d^{(D)}\boldsymbol{W}\cdot\boldsymbol{RV}_{t-1} + \gamma_w^{(D)}\boldsymbol{W}\cdot\boldsymbol{RV}_{t-5:t-2} + \gamma_m^{(D)}\boldsymbol{W}\cdot\boldsymbol{RV}_{t-22:t-6},\end{aligned} \tag{5.14}$$

where $\boldsymbol{RV}_{t:t+h} = \sum_{k=0}^{h}\boldsymbol{RV}_{t+k}$, and $h = 4$ and 21 for the weekly and monthly forecasts, respectively.

Table 5.7 illustrates the statistical performance of GHAR models over longer prediction horizons.[22] We observe that GHAR(GL, $\widetilde{L}$) continuously outperforms the baseline HAR-DRD, with only one exception: the $\mathcal{L}^Q$ loss over the monthly forecasting horizon. Furthermore, the ratios converge to one as the prediction horizon gets longer, as expected. Longer time horizon forecasting is less sensitive to graph information. A possible explanation for this might be that unexpected

[22]We report the portfolio performance results in the Appendix D.1.

Table 5.7: Out-of-sample losses over longer forecast horizons.

	$\mathcal{L}^E$			$\mathcal{L}^F$			$\mathcal{L}^Q$		
	Ratio	Rank	p-val	Ratio	Rank	p-val	Ratio	Rank	p-val
Panel A: 1-Week									
HAR-Cholesky	1.051	13	<0.001	1.058	13	<0.001	1.028	13	<0.001
HAR-DRD	1.000	12	0.006	1.000	12	0.013	1.000	7	0.002
GHAR(-, $\widetilde{K}$)	0.993	11	0.008	0.992	11	0.016	0.997	4	0.492
GHAR(-, $\widetilde{L}$)	0.992	10	0.013	0.991	9	0.025	0.997	3	0.492
GHAR(S, -)	0.991	9	0.008	0.992	10	0.011	1.005	12	<0.001
GHAR(S, $\widetilde{K}$)	0.985	8	0.008	0.985	8	0.016	1.003	10	0.002
GHAR(S, $\widetilde{L}$)	0.984	7	0.015	0.983	7	0.025	1.003	11	0.054
GHAR(K, -)	0.981	6	0.028	0.982	6	0.027	1.002	9	<0.001
GHAR(K, $\widetilde{K}$)	0.976	4	0.030	0.977	4	0.027	0.998	5	0.062
GHAR(K, $\widetilde{L}$)	0.975	3	0.064	0.976	2	0.449	0.998	6	0.104
GHAR(GL, -)	0.980	5	0.030	0.982	5	0.027	1.000	8	<0.001
GHAR(GL, $\widetilde{K}$)	0.975	2	0.064	0.976	3	0.055	0.997	2	0.492
GHAR(GL, $\widetilde{L}$)	0.974	1	1.000	0.975	1	1.000	0.997	1	1.000
Panel B: 1-Month									
HAR-Cholesky	1.108	13	<0.001	1.102	13	<0.001	1.023	13	<0.001
HAR-DRD	1.000	11	0.105	1.000	11	0.114	1.000	3	0.006
GHAR(-, $\widetilde{K}$)	0.994	4	0.480	0.993	4	0.581	0.999	1	1.000
GHAR(-, $\widetilde{L}$)	0.993	3	0.697	0.992	3	0.837	0.999	2	0.006
GHAR(S, -)	0.999	9	0.053	0.999	9	0.055	1.010	11	0.006
GHAR(S, $\widetilde{K}$)	0.995	6	0.314	0.994	6	0.501	1.010	10	0.006
GHAR(S, $\widetilde{L}$)	0.994	5	0.374	0.993	5	0.542	1.011	12	0.006
GHAR(K, -)	1.000	12	0.036	1.000	12	0.036	1.010	9	0.005
GHAR(K, $\widetilde{K}$)	1.000	10	0.065	1.000	10	0.064	1.008	7	0.006
GHAR(K, $\widetilde{L}$)	0.999	8	0.146	0.999	8	0.160	1.008	8	0.006
GHAR(GL, -)	0.996	7	0.322	0.996	7	0.393	1.003	6	0.006
GHAR(GL, $\widetilde{K}$)	0.992	2	0.697	0.991	2	0.837	1.001	4	0.006
GHAR(GL, $\widetilde{L}$)	0.991	1	1.000	0.991	1	1.000	1.001	5	0.006

Note: The table reports the out-of-sample losses of forecasting 1-week-ahead (Panel A) and 1-month-ahead (Panel B) realized covariances.

information (e.g. shocks) of a specific stock is diffused to the broad equity market in the short term (such as several days), rendering less necessary the need to explicitly incorporate graph effects into the model, at longer horizons.

5.6 Robustness analysis

In this section, we address potential concerns regarding the robustness of our results. In particular, we aim to answer two important questions related to (i) the stability across different market regimes, and (ii) the measurement errors of volatilities. Due to space considerations and to improve the readability, discussions about other relevant concerns, such as the short-term graph effect, alternative model update frequencies, and transformations for volatilities and correlations, are provided in Appendices D.2, D.3, and D.4.

5.6.1 Stability across market regimes

The first essential question is whether the model performance varies across different market regimes. To this end, we perform a stratified out-of-sample analysis over two sub-samples: a relatively calm period when the realized volatility of S&P500 ETF index is below the 90% quantile of its entire sample distribution, and a turbulent period when this realized volatility is above its 90% quantile (Pascalau and Poirier [187]).

Table 5.8 indicates the relative losses, ranks and p-values of the MCS test for the bottom 90% (low and moderate volatility regimes) and top 10% (high volatility regimes) of market RVs. In general, it appears that GHAR(GL, $\widetilde{L}$) is superior across market regimes in terms of all loss functions. Interestingly, the larger percentage improvements in terms of $\mathcal{L}^E$ and $\mathcal{L}^F$ stem from the turbulent period, while the opposite holds true for $\mathcal{L}^Q$. A potential explanation could be that the $\mathcal{L}^Q$ loss metric is typically less influenced by extreme observations when compared to the other two measures, as stated in the work of Patton [189] on volatility.

Additionally, it is worth noting that HAR-Cholesky achieves worse predictive performance relative to HAR-DRD during the turbulent period. This finding provides further robust evidence in favor of the DRD decomposition for modeling covariance matrices, especially during periods of high volatility.[23]

[23]This is in line with findings reported in Andersen et al. [11], where the authors proposed to characterize the dependencies in volatilities and correlations by regime-switching models.

Table 5.8: Stratified out-of-sample losses.

	$\mathcal{L}^E$			$\mathcal{L}^F$			$\mathcal{L}^Q$		
	Ratio	Rank	p-val	Ratio	Rank	p-val	Ratio	Rank	p-val
Panel A: Bottom 90%									
HAR-Cholesky	0.987	7	<0.001	0.988	7	<0.001	1.067	13	<0.001
HAR-DRD	1.000	13	<0.001	1.000	13	<0.001	1.000	9	<0.001
GHAR(-, $\widetilde{K}$)	0.999	12	<0.001	0.998	12	<0.001	0.983	7	<0.001
GHAR(-, $\widetilde{L}$)	0.997	11	<0.001	0.996	11	<0.001	0.980	3	0.729
GHAR(S, -)	0.992	10	<0.001	0.993	10	<0.001	1.003	12	<0.001
GHAR(S, $\widetilde{K}$)	0.992	9	<0.001	0.993	9	<0.001	0.984	8	<0.001
GHAR(S, $\widetilde{L}$)	0.990	8	<0.001	0.990	8	<0.001	0.981	4	0.001
GHAR(K, -)	0.986	3	0.676	0.987	3	0.503	1.002	11	<0.001
GHAR(K, $\widetilde{K}$)	0.986	6	<0.001	0.988	5	<0.001	0.982	5	<0.001
GHAR(K, $\widetilde{L}$)	0.985	2	0.780	0.986	2	0.664	0.979	2	0.965
GHAR(GL, -)	0.986	4	0.651	0.988	6	0.367	1.002	10	<0.001
GHAR(GL, $\widetilde{K}$)	0.986	5	<0.001	0.988	4	<0.001	0.982	6	<0.001
GHAR(GL, $\widetilde{L}$)	0.984	1	1.000	0.986	1	1.000	0.979	1	1.000
Panel B: Top 10%									
HAR-Cholesky	1.119	13	<0.001	1.115	13	<0.001	1.091	13	<0.001
HAR-DRD	1.000	12	0.034	1.000	12	0.035	1.000	12	0.011
GHAR(-, $\widetilde{K}$)	0.983	10	0.071	0.981	10	0.117	0.999	10	0.015
GHAR(-, $\widetilde{L}$)	0.982	9	0.147	0.981	9	0.136	0.998	9	0.015
GHAR(S, -)	0.986	11	0.028	0.986	11	0.029	1.000	11	0.006
GHAR(S, $\widetilde{K}$)	0.970	6	0.116	0.968	5	0.210	0.996	7	0.015
GHAR(S, $\widetilde{L}$)	0.970	5	0.292	0.968	6	0.343	0.996	6	0.015
GHAR(K, -)	0.976	7	0.055	0.976	7	0.057	0.998	8	0.015
GHAR(K, $\widetilde{K}$)	0.961	2	0.991	0.959	1	1.000	0.993	4	0.024
GHAR(K, $\widetilde{L}$)	0.961	4	0.991	0.960	2	0.985	0.992	3	0.050
GHAR(GL, -)	0.976	8	0.071	0.976	8	0.076	0.996	5	0.015
GHAR(GL, $\widetilde{K}$)	0.961	3	0.991	0.960	3	0.985	0.991	2	0.188
GHAR(GL, $\widetilde{L}$)	0.961	1	1.000	0.960	4	0.980	0.991	1	1.000

Note: Stratified losses during trading days with the bottom 90% (Panel A) and top 10% (Panel B) realized volatility of S&P500 ETF index.

5.6.2 Measurement errors of volatilities

Bollerslev, Patton, and Quaedvlieg [38] revealed that the beta coefficients in the standard HAR model may be affected by measurement errors in the realized volatilities. By exploiting the asymptotic theory for high-frequency realized volatility estimation, the authors proposed an easy-to-implement model, termed as HARQ, shown in Eqn (5.16). The realized quarticity (RQ) is estimated according to Eqn

(5.15), aiming to correct the measurement errors.

$$RQ_{i,t} = \frac{M}{3}\sum_{l=1}^{M} r_{i,t(l)}^4, \quad \boldsymbol{RQ}_{t-1} = (RQ_{1,t}, \ldots, RQ_{N,t})', \tag{5.15}$$

$$\boldsymbol{RV}_t = \boldsymbol{\alpha}_0^{(Q)} + \beta_d^{(Q)}\boldsymbol{RV}_{t-1} + \phi_d^{(Q)}\boldsymbol{RQ}_{t-1} \circ \boldsymbol{RV}_{t-1} + \beta_w^{(Q)}\boldsymbol{RV}_{t-5:t-2} + \beta_m^{(Q)}\boldsymbol{RV}_{t-22:t-6} + \boldsymbol{u}_t^{(Q)}, \tag{5.16}$$

where $\circ$ denotes the Hadamard (element-wise) product. Bollerslev, Patton, and Quaedvlieg [38] showed that the HARQ model outperforms HAR, when applied to the S&P500 ETF index and individual stocks.

To examine if our results are sensitive to measurement errors of volatilities, we also implement the following models with graph information, denoted as **GHARQ** (Eqn (5.17)). As demonstrated by Bollerslev, Patton, and Quaedvlieg [39], the measurement errors for the correlations are fairly small; therefore, we follow their setting and ignore the attenuation for the correlations. Finally, as summarized in Table 5.1, we also consider various graphs for modeling volatilities and correlations.

$$\begin{aligned}\boldsymbol{RV}_t = \boldsymbol{\alpha}^{(Q)} &+ \beta_d^{(Q)}\boldsymbol{RV}_{t-1} + \phi_d^{(Q)}\boldsymbol{RQ}_{t-1} \circ \boldsymbol{RV}_{t-1} + \beta_w^{(Q)}\boldsymbol{RV}_{t-5:t-2} + \beta_m^{(Q)}\boldsymbol{RV}_{t-22:t-6} \\ &+ \gamma_d^{(Q)}\boldsymbol{W}\cdot\boldsymbol{RV}_{t-1} + \gamma_w^{(Q)}\boldsymbol{W}\cdot\boldsymbol{RV}_{t-5:t-2} + \gamma_m^{(Q)}\boldsymbol{W}\cdot\boldsymbol{RV}_{t-22:t-6} + \boldsymbol{u}_t^{(Q)}.\end{aligned} \tag{5.17}$$

Table 5.9 presents the results obtained from **GHARQ**. We conclude from this table that GHARQ(GL, $\widetilde{L}$) remains the best-performing model. This indicates that volatility and correlation graphs indeed provide additional information for covariance forecasting.

5.7 Conclusion

This study investigates whether the graph information can provide additional predictive power when forecasting realized variances and covariances in the U.S. equity market. We propose a new approach that augments the HAR-DRD model by considering graph effects from connected neighbors. We evaluate the in-sample and out-of-sample contributions of such graph effects in realized covariance matrices forecasting, and observe that the proposed model GHAR(GL, $\widetilde{L}$) achieves the best predictive accuracy consistently across various evaluation metrics. Furthermore,

Table 5.9: Out-of-sample losses of HARQ and GHARQ.

	$\mathcal{L}^E$			$\mathcal{L}^F$			$\mathcal{L}^Q$		
	Ratio	Rank	p-val	Ratio	Rank	p-val	Ratio	Rank	p-val
HARQ-DRD	1.000	12	<0.001	1.000	12	0.003	1.000	8	<0.001
GHARQ(-, $\widetilde{K}$)	0.995	10	<0.001	0.994	10	0.003	0.992	6	<0.001
GHARQ(-, $\widetilde{L}$)	0.993	9	0.006	0.992	8	0.016	0.989	4	<0.001
GHARQ(S, -)	0.996	11	0.003	0.997	11	0.003	1.010	11	<0.001
GHARQ(S, $\widetilde{K}$)	0.992	8	0.003	0.993	9	0.003	0.991	5	<0.001
GHARQ(S, $\widetilde{L}$)	0.991	7	0.008	0.991	7	0.016	0.989	3	<0.001
GHARQ(K, -)	0.989	6	0.008	0.990	6	0.016	1.023	12	<0.001
GHARQ(K, $\widetilde{K}$)	0.988	5	0.008	0.989	5	0.016	1.002	10	<0.001
GHARQ(K, $\widetilde{L}$)	0.986	3	0.008	0.987	3	0.016	0.999	7	<0.001
GHARQ(GL, -)	0.987	4	0.008	0.988	4	0.016	1.001	9	<0.001
GHARQ(GL, $\widetilde{K}$)	0.985	2	0.008	0.986	2	0.016	0.984	2	0.006
GHARQ(GL, $\widetilde{L}$)	0.983	1	1.000	0.985	1	1.000	0.982	1	1.000

Note: The table reports the out-of-sample losses of forecasting one-day-ahead realized covariances of HARQ and GHARQ with different graphs. The baseline here is the HARQ-DRD model.

based on the covariance predictions, we construct a global minimum variance portfolio. The portfolio analysis shows that portfolios considering graph effects are able to attain significantly lower out-of-sample variance compared to the traditional models without any graph information or the naive equally-weighted portfolio. An assessment of the forecast performance over time shows that the forecast improvements over longer horizons (1-week) remain significant, but start to decay as the prediction horizon increases (such as 1-month). Furthermore, the robustness tests demonstrate the forecast improvements are observed consistently over the different out-of-sample sub-periods and are insensitive to measurement errors of volatilities. Overall, our results show that graph structures in volatilities and correlations are informative for forecasting realized covariance matrices.

Future research directions. There are a number of interesting avenues to explore in future research. One direction pertains to other types of graphs, e.g. supply/demand chains (see Herskovic et al. [135] and Tokman et al. [214]), shared analyst coverage (Ali and Hirshleifer [7]), and news co-mentions (Sidorov et al. [205]) that may contain rich structural information about equities. Our framework can be

directly applied to new sources of graphs. It would be very interesting to compare (and potentially ensemble) the predictive power of different types of graphs.

Another interesting direction pertains to performing a similar analysis as in the present chapter, but across different markets. For example, according to Buncic and Gisler [47], Choi, Jiang, and Zhang [65], and Rapach, Strauss, and Zhou [195], U.S.-based equity market information can be used to improve returns/volatility forecasts in a large cross-section of international equity markets. However, there seems to be less information about the impact of international equity markets on the U.S. market (Wilms, Rombouts, and Croux [227]). It would be an interesting study to explore the two-way interplay between U.S. equity market and international equity markets, in a bipartite (and potentially multipartite) graph setting.

Appendices

A
Appendix of Chapter 2

[noframenumbering,plain]Outline

Contents

[currentsection,hideothersubsections,currentsubsection]

A.1 Proofs

Proof of Proposition 2.2.2. First we check the elicitability condition for $H_1(v)$ and $H_2(e)$ on region $\mathcal{B}$. When $H_2(e) = \frac{\alpha}{2}e^2$, we have $H_2'(e) = \alpha\, e$ and $H_2''(e) = \alpha$. For any $(v, e) \in \mathcal{B}$, this amounts to

$$\frac{\partial R_\alpha(v,e)}{\partial v} = e - W_\alpha v \geq 0,$$

where $R_\alpha(v, e)$ is defined in (2.1).

Recall the score function $S_\alpha(v, e, x)$ defined in (2.5), and $s_\alpha(v, e)$ defined in (2.4). Then

$$\begin{aligned} s_\alpha(v,e) &= -(\mu(X \leq v) - \alpha)\frac{W_\alpha}{2}v^2 + \frac{W_\alpha}{2}\int_{-\infty}^{v} x^2\mu(\mathrm{d}x) + \mu(X \leq v)v\, e \\ &\quad -e\int_{-\infty}^{v} x\mu(\mathrm{d}x) + \alpha e\left(\frac{e}{2} - v\right) + \text{const.} \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\partial s_\alpha}{\partial v}(v,e) &= \Big(\mu(X \leq v) - \alpha\Big)\left(-W_\alpha v + e\right), \\ \frac{\partial s_\alpha}{\partial e}(v,e) &= \mu(X \leq v)v - \int_{-\infty}^{v} x\mu(\mathrm{d}x) + \alpha(e - v). \end{aligned}$$

And hence

$$\begin{aligned} \frac{\partial^2 s_\alpha}{\partial v^2}(v,e) &= \frac{\mu(\mathrm{d}v)}{\mathrm{d}v}\left(-W_\alpha v + e\right) - W_\alpha(\mu(X \leq v) - \alpha), \\ \frac{\partial^2 s_\alpha}{\partial e^2}(v,e) &= \alpha, \\ \frac{\partial^2 s_\alpha}{\partial e \partial v}(v,e) &= \mu(X \leq v) - \alpha. \end{aligned}$$

Since $\frac{\mu(X\in \mathrm{d}v)}{\mathrm{d}v} \geq 0$ and $-W_\alpha v + e > 0$ hold on region $\mathcal{B}$, we have

$$\frac{\partial^2 s_\alpha}{\partial v^2}(v,e) \geq -W_\alpha(\mu(X\leq v)-\alpha), \ \textit{on } \mathcal{B}.$$

Therefore $\frac{\partial^2 s_\alpha}{\partial v^2}(v,e) \geq 0$ holds since $v \leq \mathrm{VaR}_\alpha(\mu)$ on region $\mathcal{B}$. Next when $(v,e)\in\mathcal{B}$,

$$\begin{aligned}
&\frac{\partial^2 s_\alpha}{\partial v^2}\frac{\partial^2 s_\alpha}{\partial e^2} - \left(\frac{\partial^2 s_\alpha}{\partial v \partial e}\right)^2 = \alpha\,\frac{\mu(X\in \mathrm{d}v)}{\mathrm{d}v}\,(-W_\alpha v + e) - \alpha W_\alpha(\mu(X\leq v)-\alpha) - (\mu(X\leq v)-\alpha)^2 \\
\geq\ & (\alpha - \mu(X\leq v))(\alpha W_\alpha - \alpha + \mu(X\leq v)) \qquad (A.1)\\
\geq\ & (\alpha - \mu(X\leq v))\mu(X\leq v) \geq 0. \qquad (A.2)
\end{aligned}$$

Note that (A.1) holds since $-W_\alpha v + e \geq 0$, and (A.2) holds since $W_\alpha \geq 1$ and $\mu(X\leq v)\leq \alpha$ on $\mathcal{B}$. Therefore $\nabla^2 s_\alpha$ is positive semi-definite on the region $\mathcal{B}$.

In addition, when condition (2.6) holds, we show that $s_\alpha(v,e)$ is positive semi-definite on $\widetilde{\mathcal{B}}$.

Denote $\widetilde{\mathcal{B}}^1 = \widetilde{\mathcal{B}} \cap \left\{(v,e)\in\mathbb{R}^2 \,|\, v\leq \mathrm{VaR}_\alpha(\mu)\right\}$ and $\widetilde{\mathcal{B}}^2 = \widetilde{\mathcal{B}} \cap \{(v,e)\in\mathbb{R}^2 | v > \mathrm{VaR}_\alpha(\mu)\}$. Then $\widetilde{\mathcal{B}}^1 \cup \widetilde{\mathcal{B}}^2 = \widetilde{\mathcal{B}}$ and $\widetilde{\mathcal{B}}^1 \cap \widetilde{\mathcal{B}}^2 = \emptyset$. The positive semi-definiteness property of s_α on $\widetilde{\mathcal{B}}^1$ follows a similar proof as above.

We only need to show that s_α is positive semi-definite on $\widetilde{\mathcal{B}}^2$. In this case, we have

$$\begin{aligned}
\frac{\partial^2 s_\alpha}{\partial v^2}(v,e) &= \frac{\mu(\mathrm{d}v)}{\mathrm{d}v}\Big(-W_\alpha v + e\Big) - W_\alpha(\mu(X\leq v)-\alpha) \\
&\geq \delta_\alpha\, z_\alpha - W_\alpha(\mu(X\leq v)-\alpha) \geq 0, \qquad (A.3)
\end{aligned}$$

which holds since $\frac{\delta_\alpha z_\alpha}{W_\alpha} + \alpha \geq \beta_\alpha + \alpha \geq \mu(X\leq v)$ on $\widetilde{\mathcal{B}}$. In addition,

$$\begin{aligned}
&\frac{\partial^2 s_\alpha}{\partial v^2}\frac{\partial^2 s_\alpha}{\partial e^2} - \left(\frac{\partial^2 s_\alpha}{\partial v \partial e}\right)^2 = \alpha\,\frac{\mu(\mathrm{d}v)}{\mathrm{d}v}\,(-W_\alpha v + e) - \alpha W_\alpha(\mu(X\leq v)-\alpha) - (\mu(X\leq v)-\alpha)^2 \\
\geq\ & \alpha\delta_\alpha\, z_\alpha + (\mu(X\leq v)-\alpha)(-\alpha W_\alpha + \alpha - \mu(X\leq v)) \qquad (A.4)\\
\geq\ & \alpha\delta_\alpha\, z_\alpha - \beta_\alpha(\alpha W_\alpha + \beta_\alpha) \ \geq 0. \qquad (A.5)
\end{aligned}$$

Here (A.4) holds since $\frac{\mu(\mathrm{d}v)}{\mathrm{d}v} \geq \delta_\alpha$ and $z_\alpha \geq (-W_\alpha v + e)$. (A.5) holds since $\mu(X\leq v)\in(\alpha, \alpha+\beta_\alpha]$ on $\widetilde{\mathcal{B}}^2$. To show (A.5), it suffices to show

$$\alpha\delta_\alpha\, z_\alpha - \frac{\delta_\alpha\, z_\alpha}{2W_\alpha}\left(\alpha W_\alpha + \frac{\delta_\alpha\, z_\alpha}{2W_\alpha}\right) \geq 0, \qquad (A.6)$$

since $\beta_\alpha \leq \frac{\delta_\alpha z_\alpha}{2W_\alpha}$. Finally, (A.6) holds since $W_\alpha > \frac{1}{\sqrt{\alpha}}$, $\delta_\alpha \in (0,1)$, and $z_\alpha \in \left(0, \frac{1}{2} - \alpha\right)$. This completes the proof. □

Proof of Theorem 2.3.3. Step 1. Consider the optimal transport problem in the semi-discrete setting: the source measure $\mathbb{P}_z$ is continuous and the target measure P_n is discrete. Under Assumption 2.3.2, we can write $\mathbb{P}_z(\mathrm{d}x) = m(x)\mathrm{d}x$ for some probability density m. P_n is discrete and we can write $P_n = \sum_{i=1}^n \nu_i \delta_{y_i}$ for some $\{y_i\}_{i=1}^n \subset \mathbb{R}^{M\times T}$, $\nu_j \geq 0$ and $\sum_{j=1}^n \nu_j = 1$. In this semi-discrete setting, the Monge's problem is defined as

$$\inf_{\Phi} \int \frac{1}{2}\|x - \Phi(x)\|^2 m(x)\mathrm{d}x \quad \text{s.t.} \int_{\Phi^{-1}(y_j)} \mathrm{d}\mathbb{P}_z = \nu_j,\ j = 1, 2, \cdots, n. \tag{A.7}$$

In this case, the transport map assigns each point $x \in \mathbb{R}^{M\times T}$ to one of these y_j. Moreover, by taking advantage of the discreteness of the measure ν, one sees that the dual Kantorovich problem in the semi-discrete case is maximizing the following functional:

$$\mathcal{F}(\psi) = \mathcal{F}(\psi_1, \cdots, \psi_n) = \int \inf_j \left(\frac{1}{2}\|x - y_j\|^2 - \psi_j\right) m(x)\mathrm{d}x + \sum_{j=1}^n \psi_j \nu_j. \tag{A.8}$$

The optimal transport map of the Monge's problem (A.7) can be characterized by the maximizer of $\mathcal{F}$. To see this, let us introduce the concept of power diagram. Given a finite set of points $\{y_j\}_{j=1}^n \subset \mathbb{R}^{M\times T}$ and the scalars $\psi = \{\psi_j\}_{j=1}^n$, the power diagrams associated to the scalars ψ and the points $\{y_j\}_{j=1}^n$ are the sets:

$$S_j = \left\{ x \in \mathbb{R}^{M\times T} \;\middle|\; \frac{1}{2}\|x - y_j\|^2 - \psi_j \leq \frac{1}{2}\|x - y_k\|^2 - \psi_k,\ \forall k \neq j \right\}.$$

By grouping the points according to the power diagrams S_j, we have from (A.8) that

$$\mathcal{F}(\psi) = \sum_{j=1}^n \left[\int_{S_j} \left(\frac{1}{2}\|x - y_j\|^2 - \psi_j\right) m(x)\mathrm{d}x + \psi_j \nu_j \right]. \tag{A.9}$$

According to Theorem 4.2 in Lu and Lu [172], the optimal transport plan Φ to solve the semi-discrete Monge's problem is given by

$$\Phi(x) = \nabla \bar{\psi}(x),$$

where $\bar{\psi}(x) = \max_j\{x \cdot y_j + m_j\}$ for some $m_j \in \mathbb{R}$. Specifically, $\Phi(x) = y_j$ if $x \in S_j(x)$. Here $\psi = (\psi_1, \cdots, \psi_n)$ is an maximizer of $\mathcal{F}$ defined in (A.8) and $\{S_j\}_{j=1}^n$ denotes the power diagrams associated to $\{y_j\}_{j=1}^n$ and ψ.

Proposition 4.1 in Lu and Lu [172] guarantees that there exists a feed-forward neural network $G(\cdot;\gamma)$ with $L = \lceil \log n \rceil$ fully connected layers of equal width $N = 2^L$ and ReLU activation such that $\bar{\psi}(\cdot) = G(\cdot;\gamma)$.

Step 2. Denote $\mathbb{P}_r^{(n)}(\cdot) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\cdot = \mathbf{p}_i\}$ as an empirical measure to approximate $\mathbb{P}_r \in \mathcal{P}(\mathbb{R}^{M\times T})$ using n i.i.d. samples $\{\mathbf{p}_i\}_{i=1}^n$. Let $\{\Pi^k(\mathbf{p}_{[i]})\}_{i=1}^n$ be the order statistics of $\{\Pi^k(\mathbf{p}_i)\}_{i=1}^n$, i.e., $\Pi^k(\mathbf{p}_{[1]}) \le \cdots \le \Pi^k(\mathbf{p}_{[n]})$.

Let $\hat{v}_{n,\alpha}^k$ and $\hat{e}_{n,\alpha}^k$ denote the estimates of VaR and ES at level α using the n samples above. These quantities are defined as follows (Serfling [201]):

$$\begin{aligned}\hat{v}_{n,\alpha}^k &= \Pi^k(\mathbf{p}_{[\lceil \alpha n \rceil]}), \text{ and}\\ \hat{e}_{n,\alpha}^k &= \frac{1}{n(1-\alpha)}\sum_{i=1}^n \Pi^k(\mathbf{p}_i)\mathbf{1}\{\Pi^k(\mathbf{p}_i) \le \hat{v}_{n,\alpha}^k\}.\end{aligned}$$

We first prove the result under the VaR criteria. According to [153, Proposition 2], with probability at least $\frac{1}{2}$ it holds that

$$\left|\hat{v}_{n,\alpha}^k - \mathrm{VaR}_\alpha(\Pi^k \# \mathbb{P}_r)\right| \le \sqrt{\frac{\log(4)}{2nc}}, \tag{A.10}$$

where $c = c(\delta_k, \eta_k)$ is a constant that depends on δ_k and η_k, which are specified in Assumption **A3**. Setting the RHS of (A.10) as ε, we have $n = \mathcal{O}(\varepsilon^{-2})$. Under this choice of n, we have $\left|\hat{v}_{n,\alpha}^k - \mathrm{VaR}(\Pi^k \# \mathbb{P}_r)\right| < \varepsilon$ holds with probability at least $\frac{1}{2}$. This implies that there must exist an empirical measure $\mathbb{P}_r^{(n)*}$ such that the corresponding $\hat{v}_{n,\alpha}^k$ satisfies $\left|\hat{v}_{n,\alpha}^k - \mathrm{VaR}(\Pi^k \# \mathbb{P}_r)\right| < \varepsilon$. $\mathbb{P}_r^{(n)*}$ will be the target (empirical) measure we input in Step 1. This concludes the main result for the universal approximation under the VaR criteria.

We next prove the result under the ES criteria. Under Assumptions 2.3.1 and 2.3.2, we have

$$\mathbb{E}_{\mathbb{P}_r}\left[|\Pi^k(\mathbf{p})|^\beta\right] \le (\ell_k)^\beta \mathbb{E}_{\mathbb{P}_r}\left[\|\mathbf{p}\|^\beta\right] < \infty. \tag{A.11}$$

Take $n > \frac{16\log(8)}{(\eta_k\delta_k(1-\alpha))^2}$. Under (A.11) and Assumption **A3**, with probability $\frac{1}{2}$ it holds that

$$\begin{aligned}\left|\hat{e}^k_{n,\alpha} - \mathrm{ES}(\Pi^k\#\mathbb{P}_r)\right| \;\leq\; & \frac{\left(5\left(\mathbb{E}_{\mathbb{P}_r}[\|\Pi^k(\mathbf{p})\|^\beta]\right)^{1/\beta} - \mathrm{VaR}(\Pi^k\#\mathbb{P}_r)\right)}{(1-\alpha)}\left(\frac{1}{n}\right)^{1-\frac{1}{\beta}}\sqrt{\log(6)} \\ & + \frac{4}{\eta_k(1-\alpha)}\sqrt{\frac{\log(8)}{n}}, \end{aligned} \tag{A.12}$$

where η_k and δ_k are as defined in **A3**. The result in (A.12) is a slight modification of [192, Theorem 4.1]. Setting the RHS of (A.12) as ε, we have $n = \mathcal{O}(\varepsilon^{-\frac{\beta}{\beta-1}})$. Under this choice of n, we have $\left|\hat{e}^k_{n,\alpha} - \mathrm{ES}(\Pi^k\#\mathbb{P}_r)\right| < \varepsilon$ holds with probability at least $\frac{1}{2}$. This implies that there must exist an empirical measure $\mathbb{P}_r^{(n)*}$ such that $\left|\hat{e}^k_{n,\alpha} - \mathrm{ES}(\Pi^k\#\mathbb{P}_r)\right| < \varepsilon$ holds. $\mathbb{P}_r^{(n)*}$ will be the target (empirical) measure we input in Step 1. This concludes the main result for the universal approximation under the ES criterion. □

Proof of Theorem 2.3.4. Step 1 is the same as Theorem 2.3.3. It is sufficient to prove the corresponding Step 2.

Step 2. Denote $\mathbb{P}_r^{(n)}(\cdot) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\cdot = \mathbf{p}_i\}$ as an empirical measure to approximate $\mathbb{P}_r \in \mathcal{P}(\mathbb{R}^{M\times T})$ using n i.i.d. samples $\{\mathbf{p}_i\}_{i=1}^n$. Denote $M_\beta := \mathbb{E}_{\mathbb{P}_r}[\|\mathbf{p}\|^\beta] < \infty$. From [164, Theorem 3.1] we have

$$\mathbb{E}\mathcal{W}_1(\mu_n, \mu) \leq c_\beta M_\beta n^{-\frac{1}{(2\beta)\vee(M\times T)}\wedge(1-\frac{1}{\beta})}(\log n)^{\zeta_{\beta,M\times T}}, \tag{A.13}$$

where c_β is a constant depending only on β (not $M \times T$)

$$\zeta_{\beta,M\times T} = \begin{cases} 2 & \text{if } M\times T = \beta = 2, \\ 1 & \text{if ``}M\times T \neq 2 \text{ and } \beta = \frac{M\times T}{M\times T - 1}\wedge 2\text{'' or ``}\beta > M\times T = 2\text{''}, \\ 0 & \text{otherwise.} \end{cases}$$

By the Kantorovich-Rubenstein duality, we have

$$\begin{aligned}\mathcal{W}_1\left(\Pi^k\#\mathbb{P}_r, \Pi^k\#\mathbb{P}_r^{(n)}\right) &= \frac{1}{\ell}\sup_{\|f\|_L\leq\ell}\mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[f(\Pi^k(\mathbf{p}))\right] - \mathbb{E}_{\mathbf{q}\sim\mathbb{P}_r^{(n)}}\left[f(\Pi^k(\mathbf{q}))\right] && \text{(A.14)} \\ &\leq \frac{1}{\ell}\sup_{\|g\|_L\leq\ell\ell_k}\mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[g(\mathbf{p})\right] - \mathbb{E}_{\mathbf{q}\sim\mathbb{P}_r^{(n)}}\left[g(\mathbf{q})\right] && \text{(A.15)} \\ &\leq \ell_k\mathcal{W}_1\left(\mathbb{P}_r, \mathbb{P}_r^{(n)}\right) && \text{(A.16)}\end{aligned}$$

where $\|\cdot\|_L$ is the Lipschitz norm. (A.15) holds since $f(\Pi^k(\cdot))$ is $\ell\ell_k$-Lipschitz when f is ℓ-Lipschitz and Π^k is ℓ_k-Lipschitz. (A.16) holds by the Kantorovich-Rubenstein duality.

Taking expectation on (2.15) and applying (A.13) and (A.16), we have

$$\begin{aligned}
\mathbb{E}\left|\rho(\Pi^k\#\mathbb{P}_r) - \rho(\Pi^k\#\mathbb{P}_r^{(n)})\right| &\leq L\mathbb{E}\left(\mathcal{W}_1\left(\Pi^k\#\mathbb{P}_r,\, \Pi^k\#\mathbb{P}_r^{(n)}\right)\right)^{\kappa} && \text{(A.17)}\\
&\leq L\left(\mathbb{E}\left[\mathcal{W}_1\left(\Pi^k\#\mathbb{P}_r,\, \Pi^k\#\mathbb{P}_r^{(n)}\right)\right]\right)^{\kappa} && \text{(A.18)}\\
&\leq L\left(\ell_k c_\beta M_\beta n^{-\frac{1}{2\vee d}\wedge(1-\frac{1}{\beta})}(\log n)^{\zeta_{\beta,M\times T}}\right)^{\kappa}, && \text{(A.19)}
\end{aligned}$$

where (A.18) holds by Jensen's inequality since $\kappa \in (0, 1]$.

(A.17) implies that there must exist an empirical measure $\mathbb{P}_r^{(n)*}$ such that $\left|\rho(\Pi^k\#\mathbb{P}_r) - \rho(\Pi^k\#\mathbb{P}_r^{(n)})\right| < \varepsilon$ holds. This $\mathbb{P}_r^{(n)*}$ will be the target (empirical) measure we input in Step 1.

It is easy to check that

- $\frac{1}{2\vee(M\times T)} \wedge (1 - \frac{1}{\beta}) = 1 - \frac{1}{\beta}$ when $M = T = 1$ and $1 < \beta \leq 2$;
- $\frac{1}{2\vee(M\times T)} \wedge (1 - \frac{1}{\beta}) = \frac{1}{2}$ when $M = T = 1$ and $\beta \geq 2$;
- $\frac{1}{2\vee(M\times T)} \wedge (1 - \frac{1}{\beta}) = \frac{1}{M\times T}$ when $M \times T \geq 2$ and $\frac{1}{M\times T} + \frac{1}{\beta} < 1$;
- $\frac{1}{2\vee(M\times T)} \wedge (1 - \frac{1}{\beta}) = 1 - \frac{1}{\beta}$ when $M \times T \geq 2$ and $\frac{1}{M\times T} + \frac{1}{\beta} \geq 1$.

This concludes the main result for the universal approximation under risk measures that are Hölder continuous. □

Proof of Theorem 2.3.6. For any $\overline{D} \in \overline{\mathcal{D}}_0$, by definition there exists $\mu := \mu(\overline{D}) \in \mathcal{P}(\mathbb{R}^{M\times T})$ with a finite first moment such that

$$\overline{D}(\Pi^k\#\mu) = \Big(\mathrm{VaR}_\alpha(\Pi^k\#\mathbb{P}_r), \mathrm{ES}_\alpha(\Pi^k\#\mathbb{P}_r)\Big), \quad \forall k = 1, 2, \cdots, K. \tag{A.20}$$

Denote $\Sigma(\overline{D})$ as the set of all such $\mu \in \mathcal{P}(\mathbb{R}^{M\times T})$ with finite first moment that satisfies (A.20). Then given that both $\mu \in \Sigma(\overline{D})$ and $\mathbb{P}_z$ have finite first moments and that $\mathbb{P}_z$ is absolutely continuous with respect to the Lebesgue measure, we could find

a mapping $\overline{G} \in \overline{\mathcal{G}}$ such that $\overline{G}\#\mathbb{P}_z \in \Sigma(\overline{D})$ so that $\mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[\ S_\alpha\Big(\overline{D}\Big(\Pi^k\#\mathbb{P}_{\overline{G}}\Big), \Pi^k(\mathbf{p})\Big)\right]$ is minimized (Theorem 7.1 in Ambrosio et al. [8]). That is,

$$\min_{\overline{G}\in\overline{\mathcal{G}}} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[\ S_\alpha\Big(\overline{D}(\Pi^k\#\mathbb{P}_{\overline{G}}), \Pi^k(\mathbf{p})\Big)\right] = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[\ S_\alpha\Big(\Big(\mathrm{VaR}_\alpha(\Pi^k\#\mathbb{P}_r), \mathrm{ES}_\alpha(\Pi^k\#\mathbb{P}_r)\Big), \Pi^k(\mathbf{p})\Big)\right].$$

In this case, for the maximization problem of $\overline{D}$ over $\overline{\mathcal{D}}_0$,

$$\begin{aligned}(2.18) \quad &= \max_{\overline{D}\in\overline{\mathcal{D}}_0} \frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[\ S_\alpha\Big(\Big(\mathrm{VaR}_\alpha(\Pi^k\#\mathbb{P}_r), \mathrm{ES}_\alpha(\Pi^k\#\mathbb{P}_r)\Big), \Pi^k(\mathbf{p})\Big)\right] - \lambda\ \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[S_\alpha\Big(\overline{D}(\Pi^k\#\mathbb{P}_r),\ \Pi^k(\mathbf{p})\Big)\right]\ \right] \\ &= -\lambda \min_{\overline{D}\in\overline{\mathcal{D}}_0} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[S_\alpha\Big(\overline{D}(\Pi^k\#\mathbb{P}_r),\ \Pi^k(\mathbf{p})\Big)\right].\end{aligned}$$

By the definition of $\overline{\mathcal{D}}_0$, we have

$$\begin{aligned}&\min_{\overline{D}\in\overline{\mathcal{D}}_0} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[S_\alpha\Big(\overline{D}(\Pi^k\#\mathbb{P}_r),\ \Pi^k(\mathbf{p})\Big)\right] \\ = \ &\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[S_\alpha\Big(\Big(\mathrm{VaR}_\alpha(\Pi^k\#\mathbb{P}_r), \mathrm{ES}_\alpha(\Pi^k\#\mathbb{P}_r)\Big),\ \Pi^k(\mathbf{p})\Big)\right] \\ = \ &\min_{\overline{D}\in\overline{\mathcal{D}}} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{p}\sim\mathbb{P}_r}\left[S_\alpha\Big(\overline{D}(\Pi^k\#\mathbb{P}_r),\ \Pi^k(\mathbf{p})\Big)\right],\end{aligned}$$

which is equivalent to (2.17). Denote this minimizer as $\overline{D}^*$, plugging this into the optimization problem for $\overline{G}$ in the max-min game leads to the upper-level optimization problem (2.16). □

A.2 Implementation details

A.2.1 Setup of parameters in the synthetic data set

Mathematically, for any given time $t \in [0, T]$, we first sample $\mathbf{u}_t = (u_{1,t}, \ldots, u_{5,t})^\top \sim \mathcal{N}(0, \Sigma)$ with covariance matrix $\Sigma \in \mathbb{R}^{5\times5}$, $v_{1,t} \sim \chi^2(\nu_1)$ and $v_{2,t} \sim \chi^2(\nu_2)$. Here $v_{1,t}$, $v_{2,t}$ are independent of $\mathbf{u}_t$. We then calculate the price increments according to the following equations

$$\begin{aligned}\Delta p_{1,t} &= u_{1,t}, \\ \Delta p_{2,t} &= \phi_1 \Delta p_{2,t-1} + u_{2,t}, \\ \Delta p_{3,t} &= \phi_2 \Delta p_{3,t-1} + u_{3,t}, \\ \Delta p_{4,t} &= \varepsilon_{4,t} = \sigma_{4,t}\eta_{1,t}, \\ \Delta p_{5,t} &= \varepsilon_{5,t} = \sigma_{5,t}\eta_{2,t},\end{aligned}$$

where $\sigma^2_{4,t} = \gamma_4 + \kappa_4\varepsilon^2_{4,t-1} + \beta_4\sigma^2_{4,t-1}, \eta_{1,t} = \frac{u_{4,t}}{\sqrt{v_{1,t}/\nu_1}}$, and $\sigma^2_{5,t} = \gamma_5 + \kappa_5\varepsilon^2_{5,t-1} + \beta_5\sigma^2_{5,t-1}, \eta_{2,t} = \frac{u_{5,t}}{\sqrt{v_{2,t}/\nu_2}}$.

We set $T = 100$ as the number of observations over one trading day. We first generate a correlation matrix ρ with elements uniformly sampled from $[0, 1]$. We then sample the annualized standard deviations s with values between 0.3 and 0.5, and set $\Sigma_{ij} = \frac{s_i}{255\times T}\frac{s_j}{255\times T}\rho_{ij}$ $(i, j = 1, 2, \ldots, 5)$; $\phi_1 = 0.5$ and $\phi_2 = -0.15$; $\nu_1 = 5$ and $\nu_2 = 10$; κ_4 and κ_5 are sampled uniformly from [0.08, 0.12]; β_4 and β_5 are sampled uniformly from [0.825, 0.875]; and finally γ_4 and γ_5 are sampled uniformly from [0.03, 0.07]. We choose one quantile $\alpha = 0.05$ for this experiment.

Table A.1 reports the 5%-VaR and 5%-ES values of several strategies calculated with the synthetic financial scenarios designed above.

	Static buy-and-hold		Mean-reversion		Trend-following	
	VaR	ES	VaR	ES	VaR	ES
Gaussian	-0.489	-0.615	-0.432	-0.553	-0.409	-0.515
AR(1) with $\phi_1 = 0.5$	-0.876	-1.100	-0.850	-1.066	-0.671	-0.829
AR(1) with $\phi_2 = -0.12$	-0.461	-0.581	-0.399	-0.513	-0.387	-0.488
GARCH(1,1) with $t(5)$	-0.480	-0.603	-0.420	-0.535	-0.400	-0.501
GARCH(1,1) with $t(10)$	-0.403	-0.507	-0.354	-0.453	-0.328	-0.410

Table A.1: Empirical VaR and ES values for trading strategies evaluated on the training data.

A.2.2 Setup of the configuration

	Configuration	Values
Discriminator	Architecture	Fully-connected layers
	Activation	Leaky ReLU
	Number of neurons in each layer	(1000, 256, 128, 2)
	Learning rate	10^{-7}
	Dual parameter (λ)	1
	Batch normalization	No
Generator	Architecture	Fully-connected layers
	Activation	Leaky ReLU
	Number of neurons in each layer	(1000, 128, 256, 512, 1024, 5×100)
	Learning rate	10^{-6}
	Batch normalization	Yes
Strategies	Static portfolio with single asset	5
	Static portfolio with multiple assets	50
	Mean-reversion strategies	5
	Trend-following strategies	5
Additional parameters	Size of training data (N)	50,000
	Number of PnL samples (N_B)	1,000
	Noise dimension (N_z)	1,000
	Noise distribution	$t(5)$
	H_1, H_2	$H_1(v) = -5v^2$, $H_2(e) = \frac{\alpha}{2}e^2$

Table A.2: Network architecture configuration.

Discussion on the configuration.

- **Choice of λ:** Theorem 2.3.6 suggests that Tail-GAN is effective as long as $\lambda > 0$. In our experiments, we set $\lambda = 1$ and also tested values of 2, 10, and 100 to address the issue of hyper-parameter selection. We observed that $\lambda = 2$ and $\lambda = 10$ resulted in a similar performance to $\lambda = 1$, while larger values such as $\lambda = 100$ led to a worse performance similar to that of the supervised learning method. This may be due to the fact that larger λ values could potentially harm the model's generalization power in practical settings.

- **Choice of S_α (H_1 and H_2):** Proposition 2.2.2 demonstrates that choosing H_1 and H_2 as quadratic functions (as proposed in Acerbi and Szekely [3]) results in a positive semi-definite score function in a neighborhood region around the global minimum. This is the first theoretical evidence that highlights the optimization landscape advantages of selecting quadratic functions for H_1 and H_2.

- **Neural network architecture:** Theorem 2.3.3 implies that a feed-forward neural network with fully connected layers of equal width and ReLU activation is capable of generating financial scenarios that are arbitrarily close to the scenarios sampled from the true distribution $\mathbb{P}_r$ under VaR and ES criteria. This sheds light on using a simple network architecture such as multi-layer perceptron (MLP) in the training of Tail-GAN.

  While a more sophisticated neural network architecture may improve practical performance, our focus is not to compare different architectures, but rather to demonstrate the benefits of incorporating the essential component of tail risks of trading strategies into our Tail-GAN framework. Therefore, we choose to use a simple MLP, the same architecture used in Wasserstein GAN (Arjovsky, Chintala, and Bottou [16]). On the other hand, it has been reported in recent literature that simple and shallow neural network architectures attain better performance on financial applications compared to more advanced architectures (Chen, Pelger, and Zhu [60] and Gu, Kelly, and Xiu [119]).

A.2.3 Differentiable neural sorting

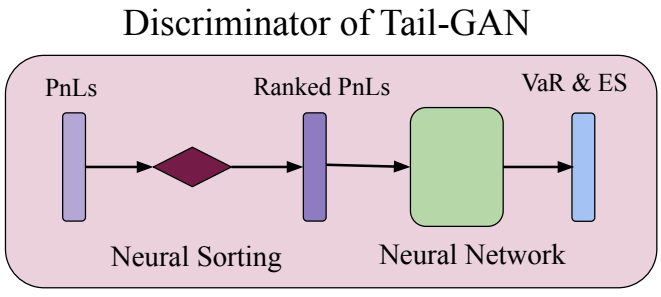


Figure A.1: Architecture of the Tail-GAN discriminator.

The architecture of the Tail-GAN Discriminator has two key ingredients, as depicted in Figure A.1. For the first ingredient, a differentiable sorting algorithm proposed by Grover et al. [118] is employed to rank the PnLs. The second part adopts a standard neural network architecture, taking the ranked PnLs as the input, and providing the estimated α-VaR and α-ES values as the output.

We follow the design in Grover et al. [118] to include the differentiable sorting architecture, so that the input of the discriminator will be the ranked PnL's (sorted in decreasing order). This design, based on the idea of using the SOFT-MAX operator to approximate the ARG-MAX operator, enables back-propagation of the gradient of the sorting function during the network training process.

Denote $\mathbf{x}^k = (x_1^k, x_2^k, \ldots, x_n^k)^\top$ as a real-valued vector of length n, representing the PnL samples of strategy k. Let $B(\mathbf{x}^k)$ denote the matrix of absolute pairwise differences of the elements of $\mathbf{x}^k$, such that $B_{i,j}(\mathbf{x}^k) = |x_i^k - x_j^k|$. We then define the following permutation matrix $\Gamma(\mathbf{x}^k)$ following Grover et al. [118] and Ogryczak and Tamir [182]

$$\Gamma_{i,j}(\mathbf{x}^k) = \begin{cases} 1, & \text{if } j = \arg\max((n+1-2i) - B(\mathbf{x}^k)\mathbf{1}), \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbf{1}$ is the all-ones vector. Then, $\Gamma(\mathbf{x}^k)\mathbf{x}^k$ provides a ranked vector of $\mathbf{x}^k$ ([182, Lemma 1] and [118, Corollary 3]). However, the ARG-MAX operator is *non-differentiable* which prohibits the direct usage of the permutation matrix for gradient computation. Instead, Grover et al. [118] propose to replace the ARG-MAX operator with SOFT-MAX, in order to obtain a continuous relaxation $\widehat{\Gamma}^\tau$ with a temperature parameter $\tau > 0$. In particular, the (i,j)-th element of $\widehat{\Gamma}^\tau(\mathbf{x}^k)$ is given by

$$\widehat{\Gamma}^\tau_{i,j}(\mathbf{x}^k) = \frac{\exp\Big(\Big((n+1-2i) - B(\mathbf{x}^k)_j\mathbf{1}\Big)/\tau\Big)}{\sum_{l=1}^n \exp\Big(\Big((n+1-2i) - B(\mathbf{x}^k)_l\mathbf{1}\Big)/\tau\Big)},$$

in which $B(\mathbf{x}^k)_l$ is the l-th row of matrix $B(\mathbf{x}^k)$. This relaxation is continuous everywhere and differentiable almost everywhere with respect to the elements of $\mathbf{x}^k$. In addition, [118, Theorem 4] shows that $\widehat{\Gamma}^\tau_{i,j}(\mathbf{x}^k)$ converges to $\Gamma_{i,j}(\mathbf{x}^k)$ almost surely when $x_1^k, \ldots, x_n^k$ are sampled IID from a distribution which is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}$.

Finally we could set in (2.11):

$$\widetilde{\Gamma}(\mathbf{x}) = \widehat{\Gamma}^\tau(\mathbf{x})\mathbf{x}.$$

A.2.4 Divergence functions and GOM

Divergence functions. Here, we provide two choices for the divergence function. First, the quantile divergence between two distributions P and Q is defined as Ostrovski, Dabney, and Munos [186]

$$q(P,Q) := \int_0^1 \left[\int_{F_P^{-1}(\tau)}^{F_Q^{-1}(\tau)} \left(F_P(x) - \tau\right) \mathrm{d}x \right] \mathrm{d}\tau,$$

where F_P (resp. F_Q) is the CDF of P (resp. Q). Motivated by this definition of quantile divergence, we propose the following local-"divergence", which focuses on the tail distribution of the strategy PnLs

$$d_q(\mathbb{P}_r, \mathbb{P}_G) := \frac{1}{K} \sum_{k=1}^{K} \int_0^{\alpha} \left[\int_{F^{-1}_{\Pi^k \# \mathbb{P}_r}(\tau)}^{F^{-1}_{\Pi^k \# \mathbb{P}_G}(\tau)} \left(F_{\Pi^k \# \mathbb{P}_r}(x) - \tau\right) \mathrm{d}x \right] \mathrm{d}\tau, \tag{A.21}$$

where $F_{\Pi^k \# \mathbb{P}_r}$ (resp. $F_{\Pi^k \# \mathbb{P}_G}$) is the CDF of $\Pi^k(\mathbf{p})$ with $\mathbf{p} \sim \mathbb{P}_r$ (resp. $\Pi^k(\mathbf{q})$ with $\mathbf{q} \sim \mathbb{P}_G$). Recall that the score function used in the loss function (2.18) can also be constructed as a "divergence" to measure the difference between two distributions in terms of their respective VaR and ES values

$$\begin{aligned} d_s\left(\mathbb{P}_r, \mathbb{P}_G\right) := \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\mathbf{p} \sim \mathbb{P}_r} \Bigg[& S_\alpha\Big(\mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_G\right), \mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_G\right), \Pi^k(\mathbf{p})\Big) \\ & - S_\alpha\Big(\mathrm{VaR}_\alpha\left(\Pi^k \# \mathbb{P}_r\right), \mathrm{ES}_\alpha\left(\Pi^k \# \mathbb{P}_r\right), \Pi^k(\mathbf{p})\Big) \Bigg]. \end{aligned} \tag{A.22}$$

Setup for GOM. We follow the procedure in Section A.2.3 and use $\left(x^k_{(\lfloor \alpha n \rfloor)}, \frac{1}{\lfloor \alpha n \rfloor} \sum_{i=1}^{\lfloor \alpha n \rfloor} x^k_{(i)}\right)$ to estimate the α-VaR and α-ES values, where $x^k_{(n)} \geq \ldots \geq x^k_{(2)} \geq x^k_{(1)}$ are the sorted PnLs of $\mathbf{x}^k$ via the differentiable neural sorting architecture. We train the GOM on synthetic price scenarios[1], with both multi-asset portfolio and dynamic strategies. The setting of GOM is the same as that of Tail-GAN (described in Table A.2), except that there is no discriminator.

[1]We have also trained GOM on real price scenarios, and observed that the performance and conclusion are similar.

A.2.5 Construction of eigenportfolios

We construct eigenportfolios from the principal components of the empirical correlation matrix $\hat{\rho}$ of returns, ranked in decreasing order of eigenvalues: $\hat{\rho} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ where $\mathbf{Q}$ is the orthogonal matrix with the i-th column being the eigenvector $\mathbf{q}_i \in \mathbb{R}^M$ of $\hat{\rho}$, and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, such that $\mathbf{\Lambda}_{1,1} \geq \mathbf{\Lambda}_{2,2} \geq \cdots \geq \mathbf{\Lambda}_{M,M} \geq 0$.

Eigenportfolios are constructed from the principal components as follows. Denote $\mathbf{h} = \mathrm{diag}(\sigma_1, \ldots, \sigma_M)$, where σ_i is the empirical standard deviation of asset i. For the i-th eigenvector $\mathbf{q}_i$, we consider its corresponding eigenportfolio

$$\frac{(\mathbf{h}^{-1}\mathbf{q}_i)^T\mathbf{p}}{\|\mathbf{h}^{-1}\mathbf{q}_i\|_1},$$

where $\mathbf{p} \in \mathbb{R}^{M\times T}$ is the price scenario, and $\|\mathbf{h}^{-1}\mathbf{q}_i\|_1$ is used to normalize the portfolio weights so that the absolute weights sum to unity.

A.3 Additional numerical experiments

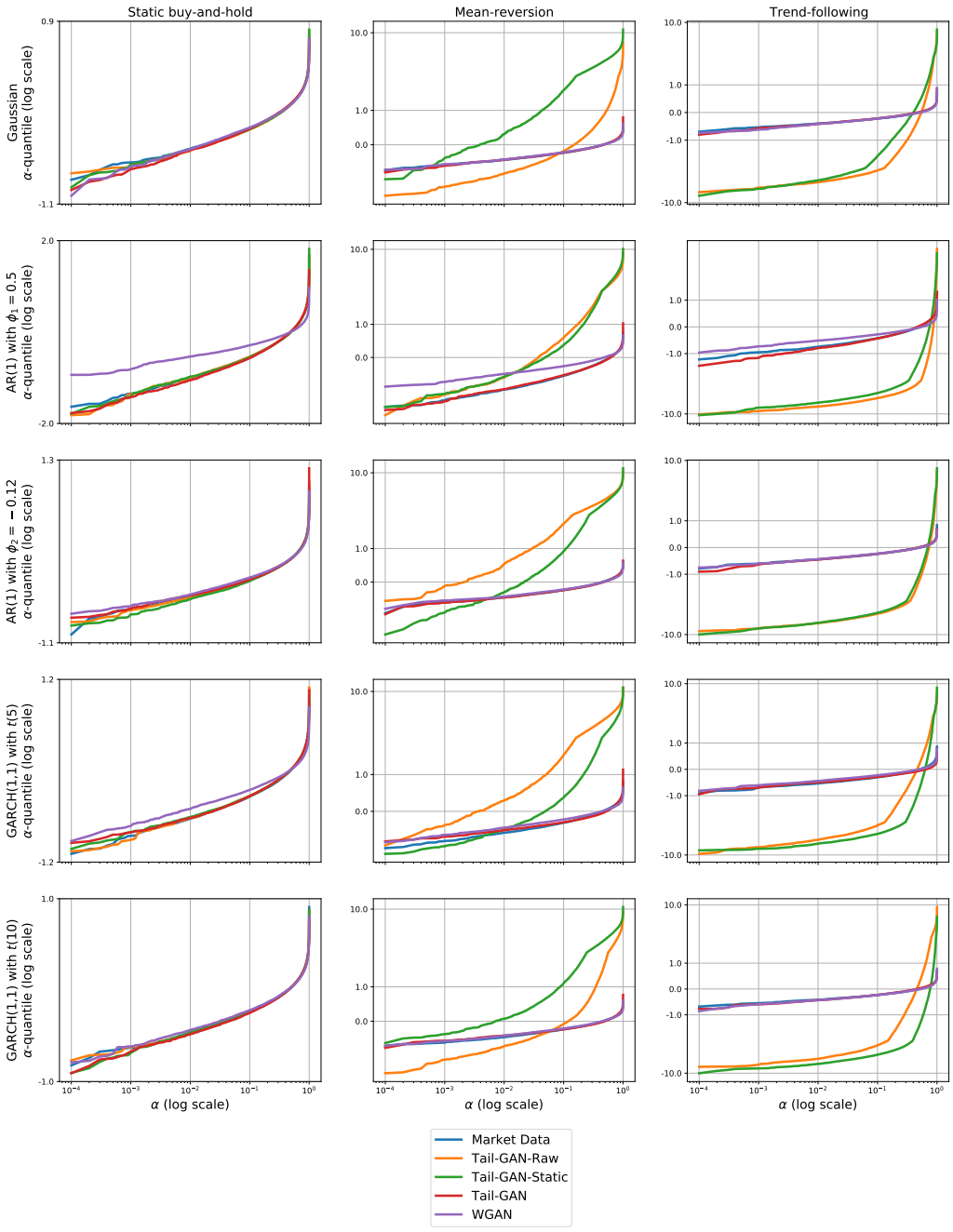


Figure A.2: Tail behavior.

Note: This figure presents the empirical rank-frequency distribution of the strategy PnL. The rows index the various models used for generating the ground truth synthetic data, while the columns represent the strategy types.

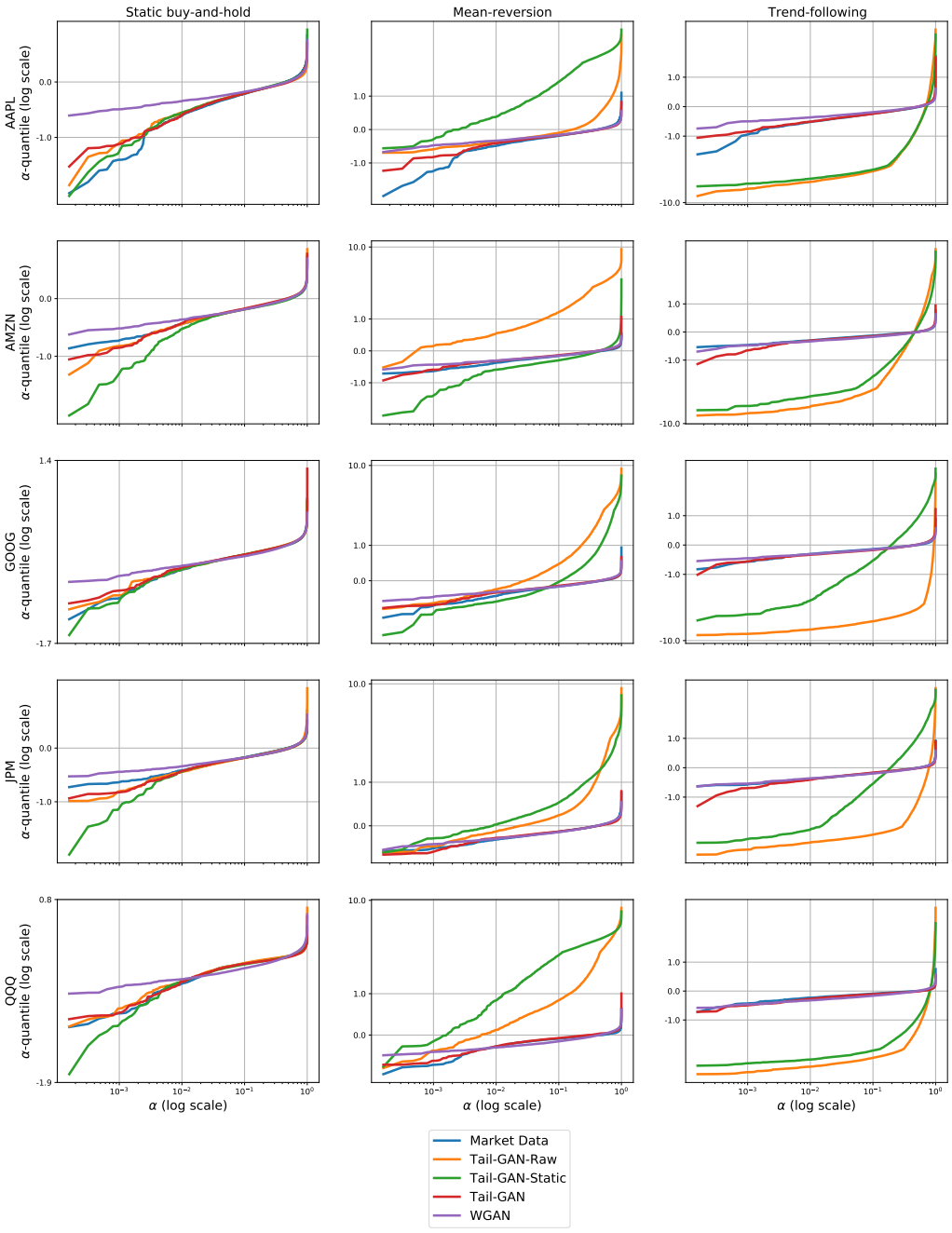


Figure A.3: Tail behavior.

Note: This figure presents the empirical rank-frequency distribution of the strategy PnL. The rows index various stocks, while the columns represent the strategy types.

B

Appendix of Chapter 3

Contents

B.1 Contemporaneous price impact of multi-level OFIs

To explicitly identify the impact of deeper-level OFIs, we also consider an extended version of $\text{PI}^{[1]}$ by incorporating multi-level OFIs as features in the model

$$\text{PI}^{[m]}: \quad r_{i,t}^{(h)} = \alpha_i^{[m]} + \sum_{k=1}^{m} \beta_i^{[m],k} \text{ofi}_{i,t}^{k,(h)} + \epsilon_{i,t}^{[m]}. \tag{B.1}$$

Recall that $\text{ofi}_{i,t}^{k,(h)}$ is the OFI at level k. We refer to this model as $\text{PI}^{[m]}$, and use OLS to estimate it.

The top panel of Table B.1 shows that the in-sample R^2 values increase as more multi-level OFIs are included as features, which is not surprising given that $\text{PI}^{[m]}$ is a nested model of $\text{PI}^{[m+1]}$. However the increments of the in-sample R^2 are descending, indicating that much deeper LOB data might be unable to provide

additional information. This argument is confirmed by the models' performance on out-of-sample data, as shown at the bottom panel of Table B.1. Out-of-sample R^2 reaches a peak at $\text{PI}^{[8]}$.

Table B.1: Performance of price-impact models with multi-level OFIs.

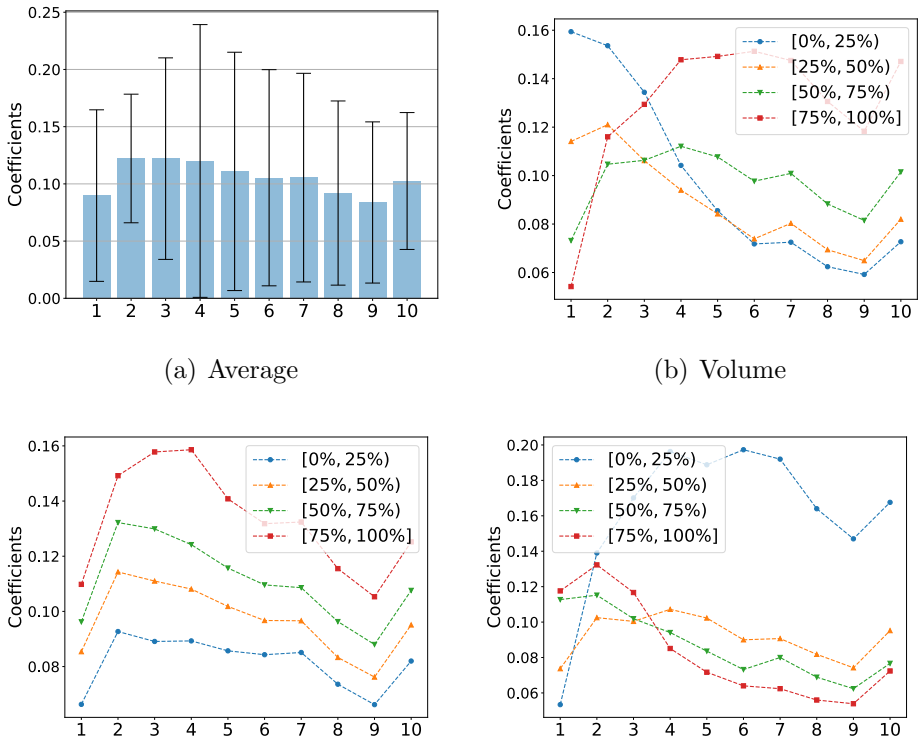
	$\text{PI}^{[1]}$	$\text{PI}^{[2]}$	$\text{PI}^{[3]}$	$\text{PI}^{[4]}$	$\text{PI}^{[5]}$	$\text{PI}^{[6]}$	$\text{PI}^{[7]}$	$\text{PI}^{[8]}$	$\text{PI}^{[9]}$	$\text{PI}^{[10]}$
IS R^2	71.16	81.61	85.07	86.69	87.66	88.30	88.74	89.04	89.24	**89.38**
	(13.80)	(11.80)	(10.76)	(10.30)	(10.05)	(9.86)	(9.71)	(9.57)	(9.45)	(9.34)
OS R^2	64.64	75.81	79.47	81.13	82.05	82.65	83.01	**83.16**	83.15	83.11
	(21.82)	(19.83)	(18.87)	(18.61)	(18.58)	(18.65)	(18.78)	(18.93)	(19.49)	(20.93)

Note: The table reports the mean values and standard deviations (in parentheses) of both in-sample and out-of-sample R^2 (in percentage points) of $\text{PI}^{[m]}$ ($m = 1, \ldots, 10$) when modeling contemporaneous returns. These statistics are averaged across each stock and each regression window.

Impact comparison between multi-level OFIs. An interesting question is whether the OFIs at different price levels contribute evenly in terms of price impact. Based on Figure B.1(a), we conclude that multi-level OFIs have different contributions to price movements. Generally, OFIs at the second-best level manifest greater influence than OFIs at the best level in model $\text{PI}^{[10]}$, which is perhaps counter-intuitive, at first sight.

We further investigate how the coefficients vary across stocks with different characteristics, such as volume, volatility, and bid-ask spread. Figure B.1 (b)-(d) reveals that for stocks with *high-volume* and *small-spread*, order flow posted deeper in the LOB has more influence on price movements. The results regarding spread are in line with Xu, Gould, and Howison [231], where it is observed that for large-spread stocks (AMZN, TSLA, and NFLX), the coefficients of ofi^m (OFIs at the m-th level) tend to get smaller as the LOB level m increases, while for small-spread stocks (ORCL, CSCO, and MU), the coefficients of ofi^m may become larger as m increases.

Cont, Kukanov, and Stoikov [74] conclude that the effect of $\text{ofi}^m (m \geq 2)$ on price changes is only second-order or null. There are two likely causes for the differences between their findings and ours. First, the data used in Cont, Kukanov, and Stoikov [74] includes 50 stocks (randomly picked from S&P500 constituents)



(a) Average (b) Volume

(c) Volatility (d) Spread

Figure B.1: Coefficients of the model $\text{PI}^{[10]}$.

Note: Plot (a) reports average coefficients and one standard deviation (error bars); Plots (b)-(d) show coefficients sorted by stock characteristics. Volume: trading volume on the previous trading day. Volatility: volatility of one-minute returns during the previous trading day. Spread: average bid-ask spread during the previous trading day. $[0\%, 25\%)$, respectively $[75\%, 100\%]$, denote the subset of stocks with the lowest, respectively highest, 25% values for a given stock characteristic. The x-axis represents different levels of OFIs and the y-axis represents the coefficients.

for a single month in 2010, while we use the top 100 large-cap stocks for 36 months during 2017-2019. Second, Cont, Kukanov, and Stoikov [74] consider the average of the coefficients across 50 stocks. In our work, we first group 100 stocks by firm characteristics, and then study the average coefficients of each subset. Therefore, our results are based on a more granular analysis, across a significantly longer period of time.

Time-series variation. When taking an aggregate view on the results in Tables 3.3, 3.5, and B.1, we observe that PI^I is superior to all $\text{PI}^{[m]}$ $(m = 1, \ldots, 10)$. Furthermore, we compare PI^I with $\text{PI}^{[m]}$ by month to study the robustness of PI^I,

(a) In-Sample R^2 (b) Out-Of-Sample R^2

Figure B.2: Time-series variation of R^2 by month.

Note: Plots (a) and (b) are based on the average results in the in-sample tests and out-of-sample tests, respectively.

as shown in Figure B.2. For better readability, we only report the results when $m = \{1, 8\}$, that correspond to the most parsimonious model and the model with highest out-of-sample R^2, respectively. To arrive at this figure, we average R^2 values in each month. Figure B.2 reveals that the improvement of PI^I compared to $\text{PI}^{[1]}$ in the in-sample tests is consistent. In terms of out-of-sample tests, it is typically the case that the univariate regression PI^I can explain slightly more about price movements compared to the best multivariate regression, $\text{PI}^{[8]}$.

Cross-sectional variation. Next, we ask the question whether the integrated OFIs affect stocks differently, depending on their characteristics. To this end, we report the results in Table B.2, both for in-sample and out-of-sample, for the best-level OFI and the integrated OFI, for each quartile bucket of the distribution obtained from the volume, volatility, and spread of the stocks in our universe. We conclude from Table B.2 that the integrated OFIs can improve R^2 in price impact consistently across different subsets grouped by firm characteristics. Table B.2 also shows that price-impact models can better explain *high-volume, low-volatility, and small-spread stocks*. These results shed new light on the modeling of price dynamics, viewed in light of stylized properties of the assets.

Table B.2: Model performance across different stock characteristics.

			[0%, 25%)	[25%, 50%)	[50%, 75%)	[75%, 100%]
Best-level	Volume	IS	57.02	70.87	77.11	**80.43**
		OS	48.87	64.63	71.72	**73.86**
	Volatility	IS	**73.25**	72.24	70.81	67.67
		OS	**66.98**	66.28	64.61	59.71
	Spread	IS	**83.16**	77.43	69.61	54.37
		OS	**77.39**	71.84	63.64	45.37
Integrated	Volume	IS	70.24	86.27	92.97	**96.31**
		OS	65.94	83.37	90.09	**93.09**
	Volatility	IS	**88.73**	87.42	85.82	83.31
		OS	**85.99**	84.48	82.52	79.00
	Spread	IS	**97.77**	93.84	84.16	70.12
		OS	**94.74**	91.15	81.42	65.29

Note: The table presents the R^2 values (in percentage) of the price-impact model, sorted by stock characteristics. R^2 is measured for both in-sample and out-of-sample tests. The top panel is based on experiments using OFIs at the best level. The bottom panel is based on experiments using integrated OFIs. Volume: trading volume on the previous trading day. Volatility: volatility of one-minute returns during the previous trading day. Spread: average bid-ask spread during the previous trading day. $[0\%, 25\%)$, respectively $[75\%, 100\%]$, denote the subset of stocks with the lowest, respectively highest, 25% values for a given stock characteristic.

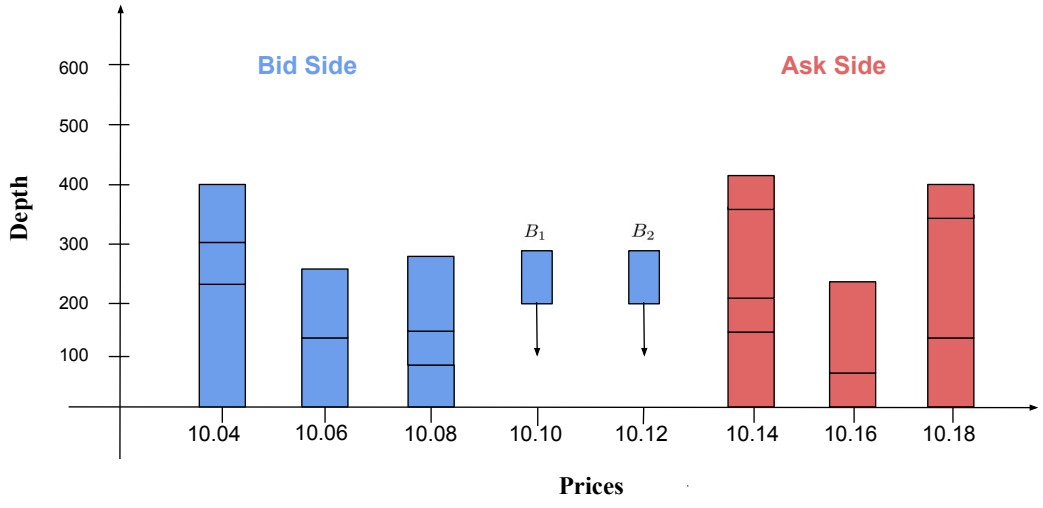


Figure B.3: Illustration of the multi-level price-impact model.

Note: Two potential scenarios when a limit buy order, B_1 (B_2), arrives within the bid-ask spread with a specific target price. Assuming B_1 and B_2 have the same size but different prices, then they will lead to the same multi-level OFI vectors but different mid-prices.

Take spread as an example; for large-spread stocks, new orders are very likely to

arrive inside the bid-ask spread as demonstrated by Xu, Gould, and Howison [231]. Now assume two limit buy orders, B_1 and B_2 (as depicted in Figure B.3), have the same size but different target prices; they will lead to the same multi-level OFI vectors but different mid-prices. Hence, it is more difficult to explain the impact of OFIs on prices for large-spread stocks. In terms of volatility and volume, a possible explanation may be due to their correlations with the bid-ask spread. Previous studies (such as Abhyankar et al. [1] and Wyart et al. [228]) found that there is a strong *positive* correlation between spread and volatility, while trading volume is *negatively* correlated with spread. For further studies on the price impact model, it is recommended to take these findings into account.

B.2 Comparison with a previous model

One closely related work is Capponi and Cont [53] (CC hereafter), where the authors propose a two-step procedure to justify the significance of cross-impact terms and render a different conclusion about cross-impact.

In the first step, the authors use OLS to decompose each stock's OFIs ($\text{ofi}_{i,t}^{1,(h)}$) into the common factor of OFIs ($F_{\text{ofi},t}^{(h)}$), that is the first principal component of the multi-asset order flow imbalances, and obtain the idiosyncratic components ($\tau_{i,t}^{(h)}$) of the OFIs, for each individual stock.

$$\text{ofi}_{i,t}^{1,(h)} = \mu_i + \gamma_i F_{\text{ofi},t}^{(h)} + \tau_{i,t}^{(h)}. \tag{B.2}$$

In the second step, they regress returns ($r_{i,t}^{(h)}$) of stock i against (i) the common factor of OFIs ($F_{\text{ofi},t}^{(h)}$), (ii) the idiosyncratic components of its own OFIs ($\tau_{i,t}^{(h)}$), and (iii) the idiosyncratic components of the OFIs of other stocks ($\tau_{j,t}^{(h)}, j \neq i$). Finally, we arrive at the cross-impact model proposed by Capponi and Cont [53] in Eqn (B.3), denoted as CI^{CC}.

$$\text{CI}^{CC}: \quad r_{i,t}^{(h)} = \alpha_i^{CC} + \beta_{i0}^{CC} F_{\text{ofi},t}^{(h)} + \sum_{j=1}^{N} \beta_{ij}^{CC} \tau_{j,t}^{(h)} + \eta_{i,t}^{CC}. \tag{B.3}$$

Table B.3: Performance of CC's models.

	PI^{CC}	CI^{CC}
IS R^2	72.58	73.95
	(13.22)	(12.56)
OS R^2	64.78	65.36
	(19.95)	(18.68)

Note: The table reports the mean values and standard deviations (in parentheses) of both in-sample and out-of-sample R^2 (in percentage points) of PI^{CC} and CI^{CC} when modeling contemporaneous returns. These statistics are averaged across each stock and each regression window.

CI^{CC} is compared with a parsimonious model PI^{CC} (Eqn (B.4)), in which only the common order flow factor and a stock's own idiosyncratic OFI are utilized.

$$\text{PI}^{CC}: \quad r_{i,t}^{(h)} = \alpha_i^{CC} + \beta_{i0}^{CC} F_{\text{ofi},t}^{(h)} + \beta_{ii}^{CC} \tau_{i,t}^{(h)} + \epsilon_{i,t}^{CC}. \tag{B.4}$$

We estimate the PI^{CC} and CI^{CC} models on historical data, under the same setting as in Section 3.3.2. Given that there are more features than observations, we employ LASSO in the second step to testify the intraday cross-impact of the idiosyncratic OFIs.

Similarly, we present the both in-sample and out-of-sample R^2 values of PI^{CC} and CI^{CC} in Table B.3. We observe small improvements (1.37% in in-sample tests, 0.58% in out-of-sample tests) from PI^{CC} to CI^{CC}. From considering Tables 3.3, 3.5, and B.3, we also observe that introducing the common factor leads to quite small changes in the model's explanatory power of price dynamics in the in-sample and out-of-sample tests. Moreover, the models employing integrated OFIs continually outperform others.

In summary, our present study differs from Capponi and Cont [53] in the following several aspects. First, Capponi and Cont [53] only focus on the in-sample performance. Second, our model takes into account the potential sparsity of the cross-impact terms, while Capponi and Cont [53] ignore this aspect. In addition to examining the cross-impact of best-level OFIs, we also consider the cross-impact from multi-level OFIs, in order to gauge a comprehensive understanding of the

Table B.4: Performance of various models under *one-minute update frequency*.

	Best-level OFIs		Integrated OFIs	
	$\text{PI}^{[1]}$	$\text{CI}^{[1]}$	PI^{I}	CI^{I}
IS R^2	70.80 (13.10)	73.55 (12.73)	86.10 (9.64)	86.84 (8.79)
OS R^2	59.67 (23.15)	61.46 (18.96)	78.88 (16.78)	78.91 (15.02)

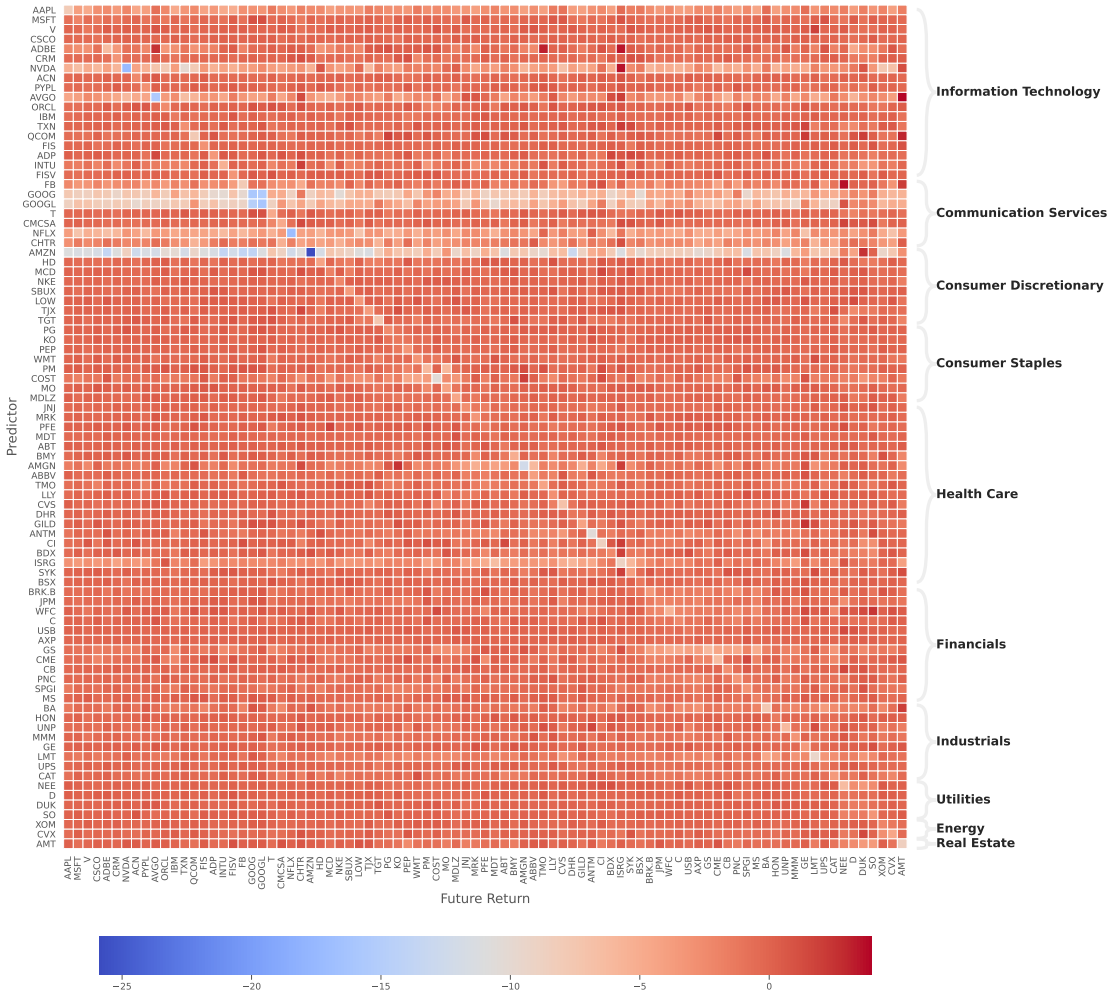
Note: The table reports the mean values and standard deviations (in parentheses) of both in-sample and out-of-sample R^2 (in percentage points) of various models when modeling contemporaneous returns in one-minute update frequency. The models include $\text{PI}^{[1]}$ (Eqn (3.6)), $\text{CI}^{[1]}$ (Eqn (3.8)), PI^{I} (Eqn (3.7)), and CI^{I} (Eqn (3.9)). These statistics are averaged across each stock and each regression window.

relations between multi-level OFIs of different assets and individual returns. To the best of our knowledge, this is the first study to investigate such relations. In the end, Capponi and Cont [53] claim that the main determinants of impact is from idiosyncratic order flow imbalance as well as a market order flow factor common across stocks, while we conclude that as long as the multi-level idiosyncratic OFIs are included, additional cross-impact terms are not necessary. The results also reveal that the sparse price impact model with integrated (or multi-level) OFIs can explain the price dynamics much better than the models proposed by Capponi and Cont [53].

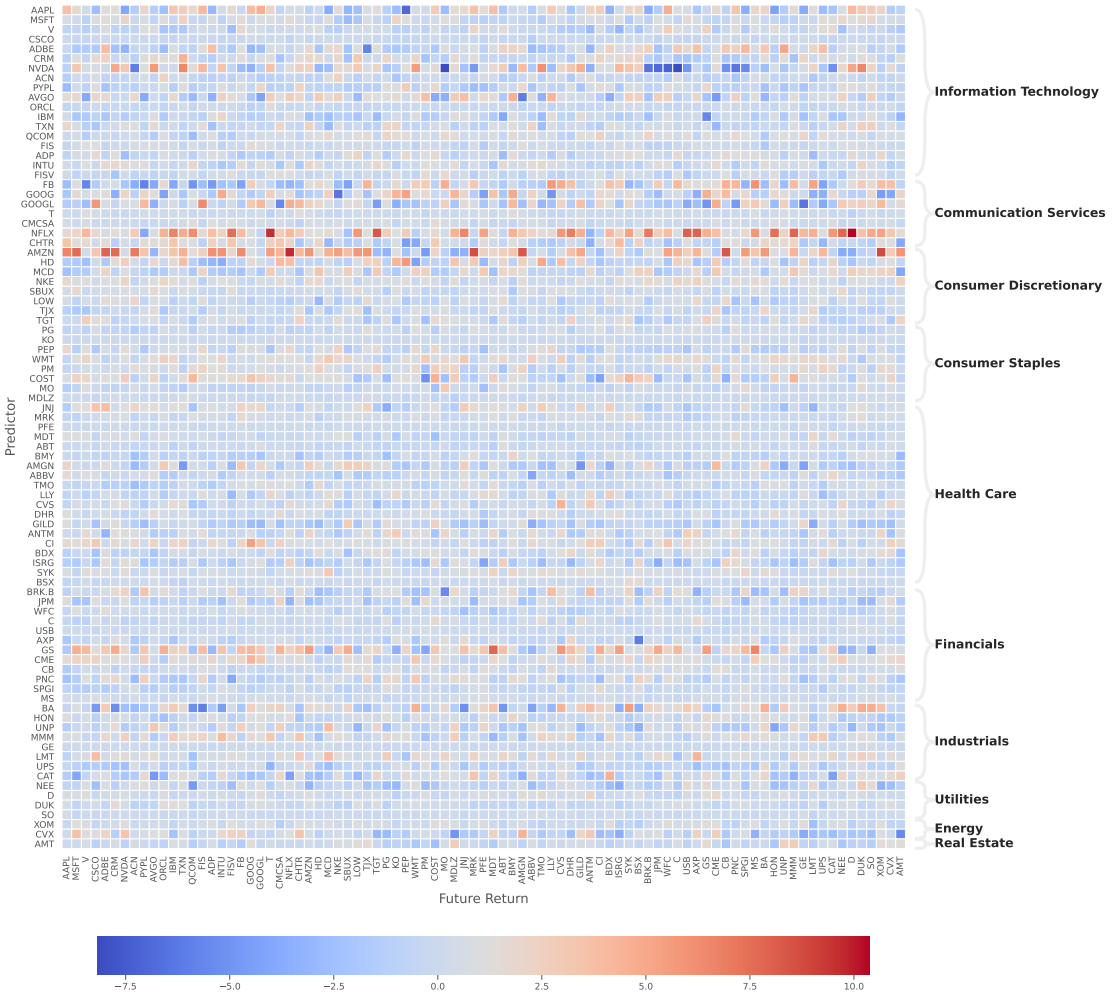
B.3 High-frequency updates of contemporaneous models

In this experiment, we use a 30-minute window to estimate contemporaneous models. We then apply the estimated coefficients to fit data in *the next one minute*, and repeat this procedure every minute. The results summarized in Table B.4 reveal similar conclusions as in Section 3.3, illustrating the robustness of our findings.

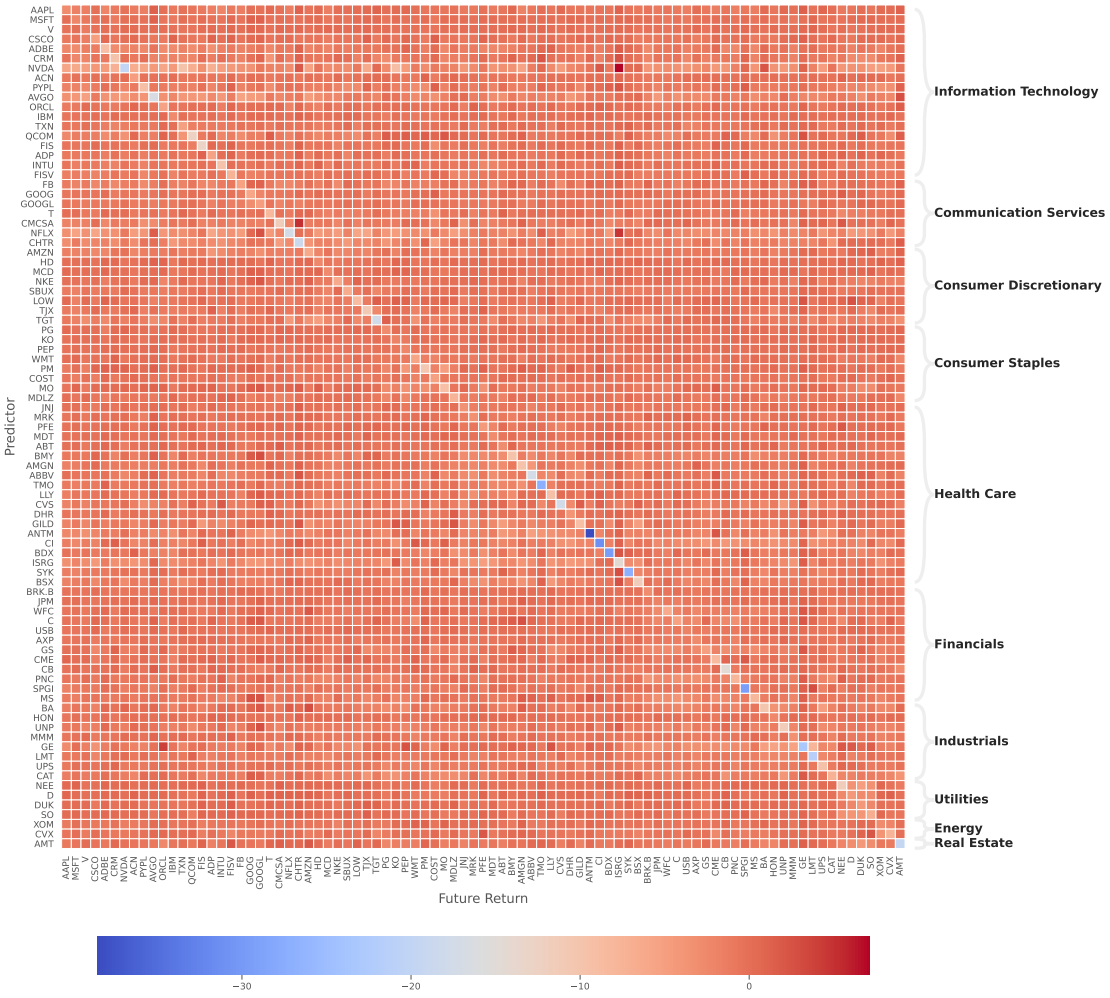
B.4 Additional results of Section 3.4



(a) Coefficient matrix of **FCI**[1]



(b) Coefficient matrix of $\mathbf{FCI}^I$



(c) Coefficient matrix of **CAR**

Figure B.4: Average coefficient matrices constructed from forward-looking cross-impact models.

C

Appendix of Chapter 4

Contents

C.1 What may drive commonality in volatility?

Previous studies, especially in the behavioral finance field, have shown that investor sentiments could affect stock prices (e.g. Baker and Wurgler [19], Bollerslev et al. [35], Da, Engelberg, and Gao [79, 80], Hameed, Kang, and Viswanathan [122], Karolyi, Lee, and Van Dijk [145], and Kogan et al. [152]). Keynes [148] argued that animal spirits affect consumer confidence, thereby moving prices in times of high levels of uncertainty. De Long et al. [82], Kogan et al. [152], and Shleifer and Summers [204] found that investor sentiments induce excess volatility. Karolyi, Lee, and Van Dijk [145] considered the investor sentiment index as an important source of commonality in liquidity. Bollerslev et al. [35] found a monotonic relationship between volatility and sentiment, possibly driven by correlated trading. In this

section, we are interested in the relationship between investor sentiments and commonality in volatility.

Traditionally, there are two approaches to measuring investor sentiments (see Da, Engelberg, and Gao [80]), i.e. market-based measures and survey-based indices. Following Baker and Wurgler [19], we consider the daily market volatility index (VIX) from Chicago Board Options Exchange to be the market sentiment measure. We use the Consumer Sentiment Index (CSI)[1] by the University of Michigan's Survey Research Center as a proxy for survey-based indices (see Carroll, Fuhrer, and Wilcox [54] and Lemmon and Portniaguina [165]). Generally speaking, CSI is a consumer confidence index, calculated by subtracting the percentage of unfavorable consumer replies from the percentage of favorable ones. Following Da, Engelberg, and Gao [80], we also include a news-based index EPU[2] proposed by Baker, Bloom, and Davis [20] to measure policy-related economic uncertainty (Amengual and Xiu [9]).

As suggested by Morck, Yeung, and Yu [178], the raw monthly commonality measures $R^2_{(h),m}$ (computed based on Eqn (4.5)) are inappropriate to use as the dependent variable in regressions, because they are bounded by 0 and 1. Consistent with Karolyi, Lee, and Van Dijk [145] and Morck, Yeung, and Yu [178] and Dang, Moshirian, and Zhang [81], we take the logistic transformation of $R^2_{(h),m}$, i.e. $\log\left[R^2_{(h),m}/(1-R^2_{(h),m})\right]$, denoted by $(R^2_{(h),m})_L$, in our following empirical analysis. To explain the commonality in volatility, we regress $(R^2_{(h),m})_L$ against the aforementioned three indices, as shown in Eqn (C.1)

$$(R^2_{(h),m})_L = \alpha + \beta_1 \mathrm{CSI}_m + \beta_2 \mathrm{VIX}_m + \beta_3 \mathrm{EPU}_m + \epsilon_{i,t}. \tag{C.1}$$

Table C.1 reports the estimation results. First, we notice that a large proportion of the variance for the commonality is explained by these three sentiment factors. For example, the commonality for the 1-day scenario is 51.6%. In terms of intraday scenarios, the R-squared values for 30-min and 65-min horizons are slightly small, 48.6% and 48.1%, respectively. The results on 10-min data are somewhat surprising, where the R-squared reaches to 55.6%. One possible reason is that economic policy

[1] http://www.sca.isr.umich.edu
[2] https://www.policyuncertainty.com

Table C.1: Time series regression of commonality.

	10-min	30-min	65-min	1-day
VIX	0.233*	0.196*	0.192*	0.714*
	(0.030)	(0.024)	(0.023)	(0.084)
CSI	0.214*	0.097*	0.066*	0.237*
	(0.025)	(0.020)	(0.019)	(0.070)
EPU	0.079*	0.025	0.022	0.114
	(0.029)	(0.023)	(0.022)	(0.080)
Constant	0.161*	0.982*	1.267*	−0.689*
	(0.023)	(0.018)	(0.018)	(0.063)
Adjusted R^2 (%)	55.6	48.6	48.1	51.6

Note: The table reports the results of time series regressions of average commonality in volatility $(R^2_{(h),m})_L$ over different horizons against three sentiment measures, VIX, CSI, and EPU. Superscript * denotes the significance levels of 5%. To compare the effects of various investor sentiments, we normalize each explanatory variable by removing its mean and scaling to the unit variance.

uncertainty is significant in the 10-min scenario. In another unreported robustness test, we estimate the regressions without the EPU factor. The adjusted R^2 value in the regression of 10-min data declines 2.5% while for other regressions, the changes in adjusted R^2 are subtle.

Besides the market volatility (VIX), we also find a significant effect of consumer sentiment (CSI) on the commonality of volatility over every studied horizon. The level of commonality is higher in times of higher market volatility and consumer sentiments. In addition, we observe that the coefficients of VIX and CSI for commonality in intraday volatility (especially for 30-min, 65-min) are substantially smaller than those in the daily case.

C.2 Hyperparameter tuning

There is no hyperparameter to tune in HAR-D and OLS. For LASSO, we use the standard 5-fold cross-validation method to determine λ_1. Hyperparameters for other models in the main analysis are summarized as follows.

To assess the robustness of neural networks to different architectures, we repeat the main analysis using 1, 2, and 3 hidden layers.[3] The results reported in Table

[3]The number of neurons is chosen based on the geometric pyramid rule, following Gu, Kelly,

Table C.2: Hyperparameters in XGBoost, MLP, LSTM.

	XGBoost	MLP	LSTM
Learning rate	0.1	0.001	0.001
Early stopping rounds	10	10	10
Ensemble	2000	10	10
Max depth	10	-	-
Batch size	-	1024	1024
Epoches	-	100	100
No. of hidden layers	-	3	2
Batch normalization	-	✓	✗

Table C.3: Out-of-sample performance of alternative hyperparameters in NNs.

		10-min		30-min		65-min		1-day	
		MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
MLP1	**Universal**	0.949	0.398	0.286	0.182	0.234	0.164	0.261	0.191
	Augmented	0.947	0.388	0.281	0.181	0.229	0.162	0.257	0.181
MLP2	**Universal**	0.948	0.398	0.284	0.182	0.232	0.164	0.260	0.190
	Augmented	0.947	0.387	0.281	0.180	0.229	0.163	0.256	0.180
MLP3	**Universal**	0.947	0.397	0.284	0.181	0.232	0.163	0.260	0.191
	Augmented	0.945	0.386	0.280	0.179	0.229	0.162	0.257	0.180
LSTM1	**Universal**	0.956	0.398	0.293	0.190	0.232	0.163	0.262	0.189
	Augmented	0.938	0.383	0.286	0.181	0.230	0.161	0.259	0.182
LSTM2	**Universal**	0.950	0.393	0.287	0.179	0.232	0.162	0.261	0.188
	Augmented	0.934	0.376	0.279	0.171	0.229	0.160	0.258	0.182
LSTM3	**Universal**	0.949	0.392	0.286	0.178	0.232	0.163	0.260	0.187
	Augmented	0.933	0.376	0.280	0.171	0.229	0.161	0.256	0.181

Note: MLP1 has single hidden layer with 128 neurons. MLP2 has two hidden layers of 128 and 64 neurons, respectively. MLP3 has three hidden layers of 128, 64 and 32 neurons, respectively. LSTM variants have similar meanings.

C.3 are generally consistent with those reported in Table 4.3.

C.3 Diebold-Mariano test

Diebold-Mariano (DM) test is used to discriminate the significant differences of forecasting accuracy between different time series models (for example Diebold [85] and Diebold and Mariano [87]). Denote the loss associated with forecast error e_t by $L(e_t)$, e.g. $L(e_t) = e_t^2$. Then the loss difference between the forecasts of models a

and Xiu [119].

and b is given by $d_t^{(a-b)} = L(e_t^{(a)}) - L(e_t^{(b)})$, where $e_t^{(a)}$ ($e_t^{(b)}$) represents the forecast error from model a (b), respectively. The DM test makes one assumption that $d_t^{(a-b)}$ is covariance stationary. The null hypothesis is that $\mathbb{E}(d_t^{(a-b)}) = 0$. Under the covariance stationary assumption, we have the test statistic

$$DM_{12} = \frac{\bar{d}^{(a-b)}}{\hat{\sigma}^{(a-b)}} \to N(0,1), \tag{C.2}$$

where $\bar{d}^{(a-b)} = \frac{1}{T}\sum_{t=1}^{T} d_t^{(a-b)}$ is the sample mean of $d_t^{(a-b)}$, and $\hat{\sigma}^{(a-b)}$ is a consistent estimate of the standard deviation of $\bar{d}^{(a-b)}$.

Following Gu, Kelly, and Xiu [119], we apply a modified DM test, to make pairwise comparisons of models' performance when forecasting multi-asset volatility. Specifically, the modified DM test compares the cross-sectional average of prediction errors from each model, rather than comparing errors for individual asset, i.e.

$$d_t^{(a-b)} = \frac{1}{N}\sum_{i=1}^{N}\left(L(e_{i,t}^{(a)}) - L(e_{i,t}^{(b)})\right), \tag{C.3}$$

where $e_{i,t}^{(a)}$ ($e_{i,t}^{(b)}$) refers to the forecast error for stock i at time t from model a (b), respectively.

To assess the statistical significance of the differences in out-of-sample volatility forecasts as shown in Table 4.3, we report the results of all DM tests in terms of QLIKE for each horizon.

Table C.4: Statistics of Diebold-Mariano tests.

Panel A: 10-min.

Univ \ Univ	LASSO	OLS	LASSO	XGBoost	MLP	LSTM
LASSO		42.33*	36.17*	56.55*	83.26*	72.43*
OLS			−32.84*	33.30*	62.29*	52.90*
Lasso				35.00*	62.15*	54.17*
XGBoost					25.86*	20.31*
MLP						−3.39*
LSTM						
Single vs	−30.07*	−1.02	31.27*	59.38*		

Aug \ Aug	LASSO	OLS	LASSO	XGBoost	MLP	LSTM
LASSO		56.60*	58.95*	33.75*	68.81*	66.94*
OLS			5.69*	−6.02*	31.82*	31.71*
Lasso				−6.52*	30.99*	31.43*
XGBoost					21.54*	28.72*
MLP						14.51*
LSTM						
Univ vs	46.23*	51.28*	53.53*	−0.32	6.24*	29.63*

Panel B: 30-min.

Univ \ Univ	HAR-D	OLS	LASSO	XGBoost	MLP	LSTM
HAR-D		44.74*	44.00*	42.96*	52.99*	46.17*
OLS			−23.57*	22.63*	35.25*	26.19*
LASSO				24.55*	37.07*	28.04*
XGBoost					16.46*	8.35*
MLP						−8.72*
LSTM						
Single vs	−7.47*	−0.48	23.14*	48.57*		

Aug \ Aug	HAR-D	OLS	LASSO	XGBoost	MLP	LSTM
HAR-D		36.36*	38.00*	24.16*	45.12*	44.43*
OLS			−2.11*	−4.50*	20.55*	18.26*
LASSO				−4.34*	21.21*	19.05*
XGBoost					21.04*	23.02*
MLP						2.24*
LSTM						
Univ vs	17.70*	22.51*	25.24*	−9.59*	9.56*	23.23*

Panel C: 65-min.

Univ \ Univ	HAR-D	OLS	LASSO	XGBoost	MLP	LSTM
HAR-D		28.27*	27.75*	21.56*	30.06*	29.00*
OLS			−11.73*	7.78*	20.22*	18.67*
LASSO				8.81*	20.91*	19.35*
XGBoost					19.83*	18.17*
MLP						0.68*
LSTM						
Single vs	−1.87	8.17*	8.53*	41.26*		

Aug \ Aug	HAR-D	OLS	LASSO	XGBoost	MLP	LSTM
HAR-D		22.11*	22.67*	9.87*	25.92*	26.01*
OLS			−4.89*	−5.56*	12.84*	11.79*
LASSO				−5.11*	13.72*	12.67*
XGBoost					17.71*	18.54*
MLP						0.94*
LSTM						
Univ vs	10.92*	12.47*	13.35*	−7.52*	7.15*	7.94*

Panel D: 1-day.

Univ \ Univ	HAR-D	OLS	LASSO	XGBoost	MLP	LSTM
HAR-D		0.50	−0.12	−4.29*	2.69*	3.86*
OLS			−4.94*	−5.50*	1.91	3.36*
LASSO				−4.29*	2.76*	3.88*
XGBoost					9.49*	10.90*
MLP						3.03*
LSTM						
Single vs	−1.32	3.41*	5.42*	20.53*		

Aug \ Aug	HAR-D	OLS	LASSO	XGBoost	MLP	LSTM
HAR-D		−0.12	5.39*	−8.97*	3.20*	1.87
OLS			−1.10	−12.63*	−3.77*	−3.99*
LASSO				−12.68*	−3.34*	−3.91*
XGBoost					12.30*	11.61*
MLP						−2.20*
LSTM						
Univ vs	4.50*	5.81*	6.30*	−5.01*	4.63*	2.78*

Note: In each panel, the left sub-table represents the pairwise comparison of forecasting performance of six models trained under **Universal** and the right one represents the pairwise comparison of forecasting performance of six models trained under **Augmented**. The bottom row in each sub-table represents the comparison of forecasting performance of the same model under two different training schemes. Positive numbers indicate the column model outperforms the row model. Superscript * deNote the significance levels of 5%.

C.4 Model update frequency

In the present paper, we choose to update each risk model annually due to the limited computation resources. To understand whether the model's performance might change with respect to the update frequency, we update the HAR-D model with different frequencies, i.e. weekly, monthly, and yearly, and results are summarized as follows. The conclusions are generally consistent with those from our main analysis.

Table C.5: Frequency of updating HAR-D for predicting intraday RVs.

Panel A:		10-min		30-min		65-min		1-day	
Statistical performance		MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
Weekly	**Single**	1.013	0.483	0.332	0.221	0.270	0.190	0.267	0.188
	Universal	1.021	0.517	0.333	0.230	0.270	0.190	0.268	0.189
	Augmented	0.995	0.453	0.323	0.228	0.262	0.185	0.256	0.180
Monthly	**Single**	1.013	0.483	0.332	0.222	0.270	0.190	0.267	0.189
	Universal	1.021	0.517	0.333	0.230	0.270	0.191	0.268	0.190
	Augmented	0.995	0.453	0.323	0.227	0.262	0.185	0.256	0.180
Yearly	**Single**	1.013	0.484	0.332	0.222	0.270	0.190	0.269	0.190
	Universal	1.021	0.518	0.333	0.230	0.270	0.191	0.269	0.190
	Augmented	0.995	0.453	0.323	0.227	0.262	0.186	0.257	0.180
Panel B:		10-min		30-min		65-min		1-day	
Realized utility		RU	RU-TC	RU	RU-TC	RU	RU-TC	RU	RU-TC
Weekly	**Single**	2.694	2.069	3.459	3.042	3.543	3.096	3.551	3.518
	Universal	2.575	1.972	3.427	3.014	3.541	3.095	3.548	3.516
	Augmented	2.790	2.280	3.427	3.020	3.553	3.108	3.571	3.536
Monthly	**Single**	2.693	2.068	3.458	3.042	3.542	3.096	3.549	3.516
	Universal	2.574	1.972	3.427	3.014	3.541	3.095	3.547	3.514
	Augmented	2.789	2.279	3.426	3.020	3.553	3.107	3.571	3.536
Yearly	**Single**	2.690	2.065	3.457	3.040	3.542	3.095	3.548	3.515
	Universal	2.574	1.975	3.429	3.016	3.541	3.095	3.547	3.514
	Augmented	2.790	2.280	3.428	3.022	3.552	3.107	3.571	3.536

D

Appendix of Chapter 5

Contents

D.1 Additional results of Section 5.5.6

Following Bollerslev, Patton, and Quaedvlieg [39], the results for GMVP and GMVP$^+$ based on longer horizon forecasting models are averaged across all possible weekly (i.e., Monday-to-Monday, Tuesday-to-Tuesday, etc.) and monthly (i.e., 1st June-to-1st July, 2nd June-to-2nd July, etc.) horizons. Table D.1 details the findings of portfolios re-balanced weekly (Panel A) and monthly (Panel B). To begin, note that modeling volatilities and correlations separately (HAR-DRD) manifests larger economic benefits than the one based on Cholesky decomposition. Compared to the HAR-DRD portfolio, the GHAR portfolios (especially GHAR(GL, $\widetilde{K}$)) achieve better ex-post standard deviations and reduce the turnover over the 1-week horizon. In contrast, the performances of most GHAR models are worse over 1-month horizon.

Table D.1: Out-of-sample portfolio performance over longer forecast horizons.

	GMVP		GMVP$^+$	
	$\sigma^{(p)}$	$\tau^{(p)}$	$\sigma^{(p)}$	$\tau^{(p)}$
Panel A: 1-Week				
1/N	10.679	0.016	10.679	0.016
HAR-Cholesky	10.562	0.808	9.641	0.455
HAR-DRD	10.015	0.787	9.536	0.530
GHAR(-, $\widetilde{K}$)	10.000	0.740	9.483	0.488
GHAR(-, $\widetilde{L}$)	10.014	0.768	9.520	0.517
GHAR(S, -)	10.200	0.804	9.703	0.558
GHAR(S, $\widetilde{K}$)	10.177	0.731	9.624	0.510
GHAR(S, $\widetilde{L}$)	10.183	0.770	9.658	0.540
GHAR(K, -)	10.040	0.702	9.700	0.473
GHAR(K, $\widetilde{K}$)	10.033	0.620	9.624	0.414
GHAR(K, $\widetilde{L}$)	10.040	0.664	9.660	0.450
GHAR(GL, -)	9.924	0.754	9.516	0.519
GHAR(GL, $\widetilde{K}$)	**9.884**	0.671	**9.434**	0.463
GHAR(GL, $\widetilde{L}$)	9.892	0.717	9.479	0.498
Panel B: 1-Month				
1/N	10.286	0.035	10.286	0.035
HAR-Cholesky	10.721	1.107	10.130	0.630
HAR-DRD	9.990	0.778	**9.598**	0.558
GHAR(-, $\widetilde{K}$)	9.903	0.814	9.633	0.566
GHAR(-, $\widetilde{L}$)	**9.869**	0.812	9.601	0.578
GHAR(S, -)	10.085	0.840	9.684	0.605
GHAR(S, $\widetilde{K}$)	10.019	0.864	9.714	0.608
GHAR(S, $\widetilde{L}$)	9.989	0.870	9.704	0.622
GHAR(K, -)	10.085	0.795	9.717	0.575
GHAR(K, $\widetilde{K}$)	10.000	0.813	9.727	0.572
GHAR(K, $\widetilde{L}$)	9.969	0.821	9.705	0.588
GHAR(GL, -)	10.258	0.844	9.814	0.615
GHAR(GL, $\widetilde{K}$)	10.194	0.865	9.832	0.615
GHAR(GL, $\widetilde{L}$)	10.178	0.875	9.828	0.633

Note: The table reports the out-of-sample performance of GMVP and GMVP$^+$ constructed using the 1-week-ahead (Panel A) or 1-month-ahead (Panel B) covariance forecasts of various models. $\sigma^{(p)}$ is the annualized portfolio standard deviation, and $\tau^{(p)}$ is the average portfolio turnover. The lowest $\sigma^{(p)}$ attained in each column is indicated in bold.

This again provides evidence that the graph effect in the equity market mainly manifests itself at short-term horizons and decays rapidly in time.

D.2 Short-term graph effect

The main aim of this section is to assess the performance of models with only the short-term graph effect. In other words, we consider models similar to **GHAR**($\boldsymbol{A}$) (Eqn (5.8)) and **GHAR**($\widetilde{\boldsymbol{A}}$) (Eqn (5.9)), but without the weekly and monthly graph effects. We refer to these models as **GHAR-S**. The results in Table D.2 show that forecast improvements relative to the benchmark HAR-DRD model remain significant even when we only incorporate the short-term graph effect.

Table D.2: Out-of-sample losses with the short-term graph effect.

	$\mathcal{L}^E$			$\mathcal{L}^F$			$\mathcal{L}^Q$		
	Ratio	Rank	p-val	Ratio	Rank	p-val	Ratio	Rank	p-val
HAR-DRD	1.000	12	0.010	1.000	12	0.019	1.000	10	<0.001
GHAR-S(-, $\widetilde{K}$)	0.997	11	0.008	0.997	11	0.013	0.992	8	<0.001
GHAR-S(-, $\widetilde{L}$)	0.996	10	0.010	0.995	9	0.019	0.990	4	0.005
GHAR-S(S, -)	0.994	9	0.010	0.995	10	0.019	1.000	12	<0.001
GHAR-S(S, $\widetilde{K}$)	0.992	8	0.010	0.993	8	0.019	0.992	7	<0.001
GHAR-S(S, $\widetilde{L}$)	0.991	7	0.010	0.991	6	0.019	0.990	3	0.001
GHAR-S(K, -)	0.990	6	0.010	0.991	7	0.019	1.000	9	<0.001
GHAR-S(K, $\widetilde{K}$)	0.989	5	0.010	0.990	5	0.019	0.991	5	<0.001
GHAR-S(K, $\widetilde{L}$)	0.988	3	0.010	0.989	3	0.019	0.989	2	0.387
GHAR-S(GL, -)	0.988	4	0.028	0.990	4	0.026	1.000	11	<0.001
GHAR-S(GL, $\widetilde{K}$)	0.987	2	0.028	0.988	2	0.026	0.991	6	<0.001
GHAR-S(GL, $\widetilde{L}$)	0.986	1	1.000	0.987	1	1.000	0.988	1	1.000

Note: The table reports the out-of-sample losses of the models with the short-term graph effect over the 1–day forecast horizon, averaged over the entire testing sample. $\mathcal{L}^E$ is the Euclidean distance, $\mathcal{L}^F$ is the Frobenius distance, and $\mathcal{L}^Q$ is the Quasi-Likelihood loss function.

D.3 Alternative model update frequencies

A potential concern one could raise is that the superior forecasting ability of our GHAR models could be driven to some extent by the low update frequency (every month in our previous analysis) of models. To investigate this possibility, we re-run our analysis and update all models at higher frequencies, specifically daily and weekly. The results in Table D.1 are generally consistent with those from our main analysis.

Table D.1: Out-of-sample losses under alternative model update frequencies.

	$\mathcal{L}^E$			$\mathcal{L}^F$			$\mathcal{L}^Q$		
	Ratio	Rank	p-val	Ratio	Rank	p-val	Ratio	Rank	p-val
Panel A: Daily									
HAR-Cholesky	1.026	13	<0.001	1.027	13	<0.001	1.115	13	<0.001
HAR-DRD	1.000	12	0.015	1.000	12	0.023	1.000	9	<0.001
GHAR(-, $\widetilde{K}$)	0.993	10	0.022	0.992	8	0.035	0.986	3	<0.001
GHAR(-, $\widetilde{L}$)	0.991	7	0.035	0.990	7	0.067	0.984	1	1.000
GHAR(S, -)	0.995	11	0.019	0.996	11	0.020	1.007	11	<0.001
GHAR(S, $\widetilde{K}$)	0.989	6	0.030	0.989	6	0.035	0.991	6	<0.001
GHAR(S, $\widetilde{L}$)	0.988	5	0.062	0.988	4	0.110	0.988	5	<0.001
GHAR(K, -)	0.992	9	0.022	0.994	10	0.032	1.015	12	<0.001
GHAR(K, $\widetilde{K}$)	0.987	4	0.035	0.988	5	0.067	0.997	8	<0.001
GHAR(K, $\widetilde{L}$)	0.986	3	0.062	0.987	2	0.140	0.995	7	<0.001
GHAR(GL, -)	0.992	8	0.030	0.993	9	0.033	1.003	10	<0.001
GHAR(GL, $\widetilde{K}$)	0.986	2	0.062	0.987	3	0.110	0.987	4	<0.001
GHAR(GL, $\widetilde{L}$)	0.986	1	1.000	0.986	1	1.000	0.984	2	0.713
Panel B: Weekly									
HAR-Cholesky	1.029	13	<0.001	1.031	13	<0.001	1.104	13	<0.001
HAR-DRD	1.000	12	0.003	1.000	12	0.006	1.000	9	<0.001
GHAR(-, $\widetilde{K}$)	0.991	10	0.008	0.990	8	0.021	0.986	4	<0.001
GHAR(-, $\widetilde{L}$)	0.990	8	0.014	0.988	7	0.037	0.984	2	0.542
GHAR(S, -)	0.994	11	0.006	0.994	11	0.008	1.006	11	<0.001
GHAR(S, $\widetilde{K}$)	0.986	6	0.014	0.986	6	0.033	0.991	8	<0.001
GHAR(S, $\widetilde{L}$)	0.985	5	0.039	0.985	4	0.094	0.988	5	<0.001
GHAR(K, -)	0.990	9	0.010	0.992	10	0.012	1.008	12	<0.001
GHAR(K, $\widetilde{K}$)	0.984	4	0.014	0.985	5	0.037	0.991	7	<0.001
GHAR(K, $\widetilde{L}$)	0.983	3	0.039	0.983	3	0.094	0.988	6	<0.001
GHAR(GL, -)	0.989	7	0.012	0.991	9	0.016	1.002	10	<0.001
GHAR(GL, $\widetilde{K}$)	0.983	2	0.039	0.983	2	0.094	0.986	3	<0.001
GHAR(GL, $\widetilde{L}$)	0.982	1	1.000	0.982	1	1.000	0.983	1	1.000

Note: The table reports the out-of-sample losses of the competitive models when updated every day (Panel A) and every week (Panel B).

D.4 Transformations for volatilities and correlations

In line with Andersen et al. [11], we now assess the ability of our proposed models under the transformation for volatilities and correlations. Specifically, we apply the log transformation to volatilities (e.g. Andersen et al. [13], Bucci [44], Herskovic et al. [136], and Zhang et al. [233]) and the Fisher transformation to correlations

(see Andersen et al. [11]). Table D.1 presents the results for modeling transformed volatilities and correlations. As shown, GHAR(K, $\widetilde{L}$) is the best performing model, followed by GHAR(GL, $\widetilde{L}$). The baseline model HAR-DRD is never included in the MCS, demonstrating the importance of graph structural information.

Table D.1: Out-of-sample losses for forecasting transformed volatilities and correlations.

	$\mathcal{L}^E$			$\mathcal{L}^F$			$\mathcal{L}^Q$		
	Ratio	Rank	p-val	Ratio	Rank	p-val	Ratio	Rank	p-val
HAR-DRD	1.000	12	0.003	1.000	12	0.002	1.000	12	<0.001
GHAR(-, $\widetilde{K}$)	0.992	11	0.004	0.991	11	0.003	0.979	8	<0.001
GHAR(-, $\widetilde{L}$)	0.991	10	0.004	0.990	10	0.004	0.975	6	<0.001
GHAR(S, -)	0.987	9	0.004	0.986	9	0.002	0.998	11	<0.001
GHAR(S, $\widetilde{K}$)	0.980	8	0.004	0.978	8	0.004	0.976	7	<0.001
GHAR(S, $\widetilde{L}$)	0.979	7	0.004	0.977	7	0.004	0.972	3	<0.001
GHAR(K, -)	0.974	5	0.004	0.973	5	0.004	0.996	9	<0.001
GHAR(K, $\widetilde{K}$)	0.969	2	0.004	0.967	2	0.004	0.973	4	<0.001
GHAR(K, $\widetilde{L}$)	0.968	1	1.000	0.966	1	1.000	0.968	1	1.000
GHAR(GL, -)	0.978	6	0.004	0.977	6	0.004	0.998	10	<0.001
GHAR(GL, $\widetilde{K}$)	0.972	4	0.004	0.970	4	0.004	0.975	5	<0.001
GHAR(GL, $\widetilde{L}$)	0.972	3	0.004	0.970	3	0.004	0.969	2	0.523

Note: The table reports the out-of-sample losses of the models for forecasting transformed volatilities and correlations over the 1-day forecast horizon, averaged over the entire testing sample. $\mathcal{L}^E$ is the Euclidean distance, $\mathcal{L}^F$ is the Frobenius distance, and $\mathcal{L}^Q$ is the Quasi-Likelihood loss function.

References

[1] Abhay Abhyankar, Dipak Ghosh, Eric Levin, and RJ Limmack. "Bid-Ask Spreads, Trading Volume and Volatility: Intra-Day Evidence from the London Stock Exchange". In: *Journal of Business Finance & Accounting* 24.3 (1997), pp. 343–362.

[2] Carlo Acerbi. "Spectral Measures of Risk: A Coherent Representation of Subjective Risk Aversion". In: *Journal of Banking & Finance* 26.7 (2002), pp. 1505–1518.

[3] Carlo Acerbi and Balazs Szekely. "Back-Testing Expected Shortfall". In: *Risk* 27.11 (2014), pp. 76–81.

[4] Hee-Joon Ahn, Kee-Hong Bae, and Kalok Chan. "Limit Orders, Depth, and Volatility: Evidence from the Stock Exchange of Hong Kong". In: *Journal of Finance* 56.2 (2001), pp. 767–788.

[5] Yacine Aït-Sahalia, Jianqing Fan, Lirong Xue, and Yifeng Zhou. "How and When Are High-Frequency Stock Returns Predictable?" In: *Available at SSRN 4095405* (2022).

[6] Yacine Aït-Sahalia and Dacheng Xiu. "Increased Correlation among Asset Classes: Are Volatility or Jumps to Blame, or Both?" In: *Journal of Econometrics* 194.2 (2016), pp. 205–219.

[7] Usman Ali and David Hirshleifer. "Shared Analyst Coverage: Unifying Momentum Spillover Effects". In: *Journal of Financial Economics* 136.3 (2020), pp. 649–675.

[8] Luigi Ambrosio, Luis A Caffarelli, Yann Brenier, Giuseppe Buttazzo, Cedric Villani, Sandro Salsa, Luigi Ambrosio, and Aldo Pratelli. "Existence and Stability Results in the L 1 Theory of Optimal Transportation". In: *Optimal Transportation and Applications: Lectures given at the CIME Summer School, held in Martina Franca, Italy, September 2-8, 2001* (2003), pp. 123–160.

[9] Dante Amengual and Dacheng Xiu. "Resolution of Policy Uncertainty and Sudden Declines in Volatility". In: *Journal of Econometrics* 203.2 (2018), pp. 297–315.

[10] Torben G Andersen and Tim Bollerslev. "Intraday Periodicity and Volatility Persistence in Financial Markets". In: *Journal of Empirical Finance* 4.2-3 (1997), pp. 115–158.

[11] Torben G Andersen, Tim Bollerslev, Peter F Christoffersen, and Francis X Diebold. "Volatility and Correlation Forecasting". In: *Handbook of Economic Forecasting* 1 (2006), pp. 777–878.

[12] Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Heiko Ebens. "The Distribution of Realized Stock Return Volatility". In: *Journal of Financial Economics* 61.1 (2001), pp. 43–76.

[13] Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Paul Labys. "Modeling and Forecasting Realized Volatility". In: *Econometrica* 71.2 (2003), pp. 579–625.

[14] Torben G Andersen, Tim Bollerslev, and Nour Meddahi. "Realized Volatility Forecasting and Market Microstructure Noise". In: *Journal of Econometrics* 160.1 (2011), pp. 220–234.

[15] Julián Andrada-Félix, Fernando Fernández-Rodrıéguez, and Ana-Maria Fuertes. "Combining Nearest Neighbor Predictions and Model-Based Predictions of Realized Variance: Does It Pay?" In: *International Journal of Forecasting* 32.3 (2016), pp. 695–715.

[16] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein Generative Adversarial Networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 214–223.

[17] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. "Generalization and Equilibrium in Generative Adversarial Nets (GANs)". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 224–232.

[18] Marco Avellaneda and Jeong-Hyun Lee. "Statistical Arbitrage in the US Equities Market". In: *Quantitative Finance* 10.7 (2010), pp. 761–782.

[19] Malcolm Baker and Jeffrey Wurgler. "Investor Sentiment in the Stock Market". In: *Journal of Economic Perspectives* 21.2 (2007), pp. 129–152.

[20] Scott R Baker, Nicholas Bloom, and Steven J Davis. "Measuring Economic Policy Uncertainty". In: *Quarterly Journal of Economics* 131.4 (2016), pp. 1593–1636.

[21] Gunjan Banerji. *The 30 Minutes that Can Make or Break the Trading Day*. 2020. URL: `https://www.wsj.com/articles/the-30-minutes-that-can-make-or-break-the-trading-day-11583886131?reflink=desktopwebshare_permalink` (visited on 03/11/2020).

[22] Bank for International Settlements. "Minimum Capital Requirements for Market Risk". In: *https://www.bis.org/bcbs/publ/d457.pdf* (2019).

[23] Ole E Barndorff-Nielsen, Peter Reinhard Hansen, Asger Lunde, and Neil Shephard. "Multivariate Realised Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading". In: *Journal of Econometrics* 162.2 (2011), pp. 149–169.

[24] Ole E Barndorff-Nielsen, Silja Kinnebrock, and Neil Shephard. "Measuring Downside Risk-Realised Semivariance". In: *CREATES Research Paper* 2008-42 (2008).

[25] Ole E Barndorff-Nielsen and Neil Shephard. "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.2 (2002), pp. 253–280.

[26] Jozef Barunıék, Evžen Kočenda, and Lukáš Vácha. "Asymmetric Connectedness on the US Stock Market: Bad and Good Volatility Spillovers". In: *Journal of Financial Markets* 27 (2016), pp. 55–78.

[27] David S Bates. "How Crashes Develop: Intradaily Volatility and Crash Evolution". In: *Journal of Finance* 74.1 (2019), pp. 193–238.

[28] Aharon Ben-Tal and Marc Teboulle. "An Old-New Concept of Convex Risk Measures: The Optimized Certainty Equivalent". In: *Mathematical Finance* 17.3 (2007), pp. 449–476.

[29] Aharon Ben-Tal and Marc Teboulle. "Expected Utility, Penalty Functions, and Duality in Stochastic Nonlinear Programming". In: *Management Science* 32.11 (1986), pp. 1445–1466.

[30] Michael Benzaquen, Iacopo Mastromatteo, Zoltan Eisler, and Jean-Philippe Bouchaud. "Dissecting Cross-Impact on Stock Markets: An Empirical Analysis". In: *Journal of Statistical Mechanics: Theory and Experiment* 2017.2 (2017), p. 023406.

[31] Robert-Paul Berben and W Jos Jansen. "Comovement in International Equity Markets: A Sectoral View". In: *Journal of International Money and Finance* 24.5 (2005), pp. 832–857.

[32] Siddharth Bhatia, Arjit Jain, and Bryan Hooi. "ExGAN: Adversarial Generation of Extreme Samples". In: *arXiv preprint arXiv:2009.08454* (2020).

[33] Bernard Bollen and Brett Inder. "Estimating Daily Volatility in Financial Markets Utilizing Intraday Data". In: *Journal of Empirical Finance* 9.5 (2002), pp. 551–562.

[34] Tim Bollerslev. "Realized Semi (Co) Variation: Signs that All Volatilities Are not Created Equal". In: *Journal of Financial Econometrics* 20.2 (2022), pp. 219–252.

[35] Tim Bollerslev, Benjamin Hood, John Huss, and Lasse Heje Pedersen. "Risk Everywhere: Modeling and Managing Volatility". In: *Review of Financial Studies* 31.7 (2018), pp. 2729–2773.

[36] Tim Bollerslev, James Marrone, Lai Xu, and Hao Zhou. "Stock Return Predictability and Variance Risk Premia: Statistical Inference and International Evidence". In: *Journal of Financial and Quantitative Analysis* 49.3 (2014), pp. 633–661.

[37] Tim Bollerslev, Marcelo C Medeiros, Andrew J Patton, and Rogier Quaedvlieg. "From Zero to Hero: Realized Partial (Co) Variances". In: *Journal of Econometrics* (2021).

[38] Tim Bollerslev, Andrew J Patton, and Rogier Quaedvlieg. "Exploiting the Errors: A Simple Approach for Improved Volatility Forecasting". In: *Journal of Econometrics* 192.1 (2016), pp. 1–18.

[39] Tim Bollerslev, Andrew J Patton, and Rogier Quaedvlieg. "Modeling and Forecasting (Un)Reliable Realized Covariances for More Reliable Financial Decisions". In: *Journal of Econometrics* 207.1 (2018), pp. 71–91.

[40] Jean-Philippe Bouchaud. "Price Impact". In: *Encyclopedia of Quantitative Finance* (2010).

[41] Brian H Boyer, Michael S Gibson, and Mico Loretan. "Pitfalls in Tests for Changes in Correlations". In: (1999).

[42] Giuseppe Buccheri, Fulvio Corsi, and Stefano Peluso. "High-Frequency Lead-Lag Effects and Cross-Asset Linkages: A Multi-Asset Lagged Adjustment Model". In: *Journal of Business & Economic Statistics* 39.3 (2021), pp. 605–621.

[43] Andrea Bucci. "Cholesky–ANN Models for Predicting Multivariate Realized Volatility". In: *Journal of Forecasting* 39.6 (2020), pp. 865–876.

[44] Andrea Bucci. "Realized Volatility Forecasting with Neural Networks". In: *Journal of Financial Econometrics* 18.3 (2020), pp. 502–531.

[45] Hans Buehler, Blanka Horvath, Terry Lyons, Imanol Perez Arribas, and Ben Wood. "A Data-Driven Market Simulator for Small Data Environments". In: *Available at SSRN 3632431* (2020).

[46] Hans Buehler, Blanka Horvath, Terry Lyons, Imanol Perez Arribas, and Ben Wood. "Generating Financial Markets with Signatures". In: *Available at SSRN 3657366* (2020).

[47] Daniel Buncic and Katja IM Gisler. "Global Equity Market Volatility Spillovers: A Broader Role for the United States". In: *International Journal of Forecasting* 32.4 (2016), pp. 1317–1339.

[48] Laurent AF Callot, Anders B Kock, and Marcelo C Medeiros. "Modeling and Forecasting Large Realized Covariance Matrices and Portfolio Choice". In: *Journal of Applied Econometrics* 32.1 (2017), pp. 140–158.

[49] Laurent E Calvet, Adlai J Fisher, and Samuel B Thompson. "Volatility Comovement: A Multifrequency Approach". In: *Journal of Econometrics* 131.1-2 (2006), pp. 179–215.

[50] Sean D Campbell. "A Review of Backtesting and Backtesting Procedures". In: *The Journal of Risk* 9.2 (2006), p. 1.

[51] Charles Cao, Oliver Hansch, and Xiaoxin Wang. "The Information Content of an Open Limit-Order Book". In: *Journal of Futures Markets: Futures, Options, and Other Derivative Products* 29.1 (2009), pp. 16–41.

[52] Haoyang Cao, Xin Guo, and Guan Wang. "Meta-Learning with GANs for Anomaly Detection, with Deployment in High-Speed Rail Inspection System". In: *arXiv preprint arXiv:2202.05795* (2022).

[53] Francesco Capponi and Rama Cont. "Multi-Asset Market Impact and Order Flow Commonality". In: *Available at SSRN* (2020).

[54] Christopher D Carroll, Jeffrey C Fuhrer, and David W Wilcox. "Does Consumer Sentiment Forecast Household Spending? If So, Why?" In: *American Economic Review* 84.5 (1994), pp. 1397–1408.

[55] Álvaro Cartea, Ryan Donnelly, and Sebastian Jaimungal. "Enhancing Trading Strategies with Order Book Signals". In: *Applied Mathematical Finance* 25.1 (2018), pp. 1–35.

[56] Álvaro Cartea, Luhui Gan, and Sebastian Jaimungal. "Trading Co-Integrated Assets with Price Impact". In: *Mathematical Finance* 29.2 (2019), pp. 542–567.

[57] Álvaro Cartea and Sebastian Jaimungal. "Incorporating Order-Flow into Optimal Execution". In: *Mathematics and Financial Economics* 10 (2016), pp. 339–364.

[58] Tolga Cenesizoglu, Georges Dionne, and Xiaozhou Zhou. "Asymmetric Effects of the Limit Order Book on Price Dynamics". In: *Journal of Empirical Finance* 65 (2022), pp. 77–98.

[59] Bidisha Chakrabarty, Terrence Hendershott, Samarpan Nawn, and Roberto Pascual. "Order Exposure in High Frequency Markets". In: *Available at SSRN 3074049* (2022).

[60] Luyang Chen, Markus Pelger, and Jason Zhu. "Deep Learning in Asset Pricing". In: *Management Science* (2023).

[61] Tianqi Chen and Carlos Guestrin. "Xgboost: A Scalable Tree Boosting System". In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.

[62] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems* 29 (2016).

[63] Alex Chinco, Adam D Clark-Joseph, and Mao Ye. "Sparse Signals in the Cross-Section of Returns". In: *Journal of Finance* 74.1 (2019), pp. 449–492.

[64] Roxana Chiriac and Valeri Voev. "Modelling and Forecasting Multivariate Realized Volatility". In: *Journal of Applied Econometrics* 26.6 (2011), pp. 922–947.

[65] Darwin Choi, Wenxi Jiang, and Chao Zhang. "Alpha Go Everywhere: Machine Learning and International Stock Returns". In: *Available at SSRN 3489679* (2022).

[66] Tarun Chordia, Richard Roll, and Avanidhar Subrahmanyam. "Commonality in Liquidity". In: *Journal of Financial Economics* 56.1 (2000), pp. 3–28.

[67] Tarun Chordia, Richard Roll, and Avanidhar Subrahmanyam. "Order Imbalance, Liquidity, and Market Returns". In: *Journal of Financial Economics* 65.1 (2002), pp. 111–130.

[68] Tarun Chordia and Avanidhar Subrahmanyam. "Order Imbalance and Individual Stock Returns: Theory and Evidence". In: *Journal of Financial Economics* 72.3 (2004), pp. 485–518.

[69] Bent J Christensen and Nagpurnanand R Prabhala. "The Relation Between Implied and Realized Volatility". In: *Journal of Financial Economics* 50.2 (1998), pp. 125–150.

[70] Kim Christensen, Mathias Siggaard, Bezirgen Veliyev, et al. *A Machine Learning Approach to Volatility Forecasting*. Vol. 3. Department of Economics and Business Economics, Aarhus University, 2021.

[71] Lauren Cohen and Andrea Frazzini. "Economic Links and Predictable Returns". In: *Journal of Finance* 63.4 (2008), pp. 1977–2011.

[72] Rama Cont. "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues". In: *Quantitative finance* 1.2 (2001), p. 223.

[73] Rama Cont, Romain Deguest, and Xue Dong He. "Loss-Based Risk Measures". In: *Statistics and Risk Modeling* 30.2 (2013), pp. 133–167. URL: https://doi.org/10.1524/strm.2013.1132.

[74] Rama Cont, Arseniy Kukanov, and Sasha Stoikov. "The Price Impact of Order Book Events". In: *Journal of Financial Econometrics* 12.1 (2014), pp. 47–88.

[75] Fulvio Corsi. "A Simple Approximate Long-Memory Model of Realized Volatility". In: *Journal of Financial Econometrics* 7.2 (2009), pp. 174–196.

[76] Antonio Costa, Paulo Matos, and Cristiano da Silva. "Sectoral Connectedness: New Evidence from US Stock Market during Covid-19 Pandemics". In: *Finance Research Letters* 45 (2022), p. 102124.

[77] Gianluca Cubadda, Barbara Guardabascio, and Alain Hecq. "A Vector Heterogeneous Autoregressive Index Model for Realized Volatility Measures". In: *International Journal of Forecasting* 33.2 (2017), pp. 337–344.

[78] Chester Curme, Michele Tumminello, Rosario N Mantegna, H Eugene Stanley, and Dror Y Kenett. "Emergence of Statistically Validated Financial Intraday Lead-Lag Relationships". In: *Quantitative Finance* 15.8 (2015), pp. 1375–1386.

[79] Zhi Da, Joseph Engelberg, and Pengjie Gao. "In Search of Attention". In: *Journal of Finance* 66.5 (2011), pp. 1461–1499.

[80] Zhi Da, Joseph Engelberg, and Pengjie Gao. "The Sum of All Fears Investor Sentiment and Asset Prices". In: *Review of Financial Studies* 28.1 (2015), pp. 1–32.

[81] Tung Lam Dang, Fariborz Moshirian, and Bohui Zhang. "Commonality in News Around the World". In: *Journal of Financial Economics* 116.1 (2015), pp. 82–110.

[82] J Bradford De Long, Andrei Shleifer, Lawrence H Summers, and Robert J Waldmann. "Noise Trader Risk in Financial Markets". In: *Journal of Political Economy* 98.4 (1990), pp. 703–738.

[83] Stavros Degiannakis and George Filis. "Forecasting Oil Price Realized Volatility Using Information Channels from Other Asset Classes". In: *Journal of International Money and Finance* 76 (2017), pp. 28–49.

[84] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. "Optimal Versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?" In: *Review of Financial Studies* 22.5 (2009), pp. 1915–1953.

[85] Francis X Diebold. "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests". In: *Journal of Business & Economic Statistics* 33.1 (2015), pp. 1–1.

[86] Francis X Diebold and Robert S Mariano. "Comparing Predictive Accuracy". In: *Journal of Business & Economic statistics* 20.1 (2002), pp. 134–144.

[87] Francis X Diebold and Roberto S Mariano. "Comparing Predictive Accuracy". In: *Journal of Business & Economic Statistics* 13.3 (1995), pp. 253–263.

[88] Chris Donahue, Julian McAuley, and Miller Puckette. "Adversarial Audio Synthesis". In: *arXiv preprint arXiv:1802.04208* (2018).

[89] J Doyne Farmer, Laszlo Gillemot, Fabrizio Lillo, Szabolcs Mike, and Anindya Sen. "What Really Causes Large Price Changes?" In: *Quantitative Finance* 4.4 (2004), pp. 383–397.

[90] Zaichao Du and Juan Carlos Escanciano. "Backtesting Expected Shortfall: Accounting for Tail Risk". In: *Management Science* 63.4 (2017), pp. 940–958.

[91] Robert Engle. "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models". In: *Journal of Business & Economic Statistics* 20.3 (2002), pp. 339–350.

[92] Robert F Engle. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation". In: *Econometrica: Journal of the Econometric Society* (1982), pp. 987–1007.

[93] Robert F Engle and Kenneth F Kroner. "Multivariate Simultaneous Generalized ARCH". In: *Econometric Theory* 11.1 (1995), pp. 122–150.

[94] Robert F Engle and Andrew J Patton. "What Good Is a Volatility Model?" In: *Forecasting Volatility in the Financial Markets*. Elsevier, 2007, pp. 47–63.

[95] Robert F Engle and Magdalena E Sokalska. "Forecasting Intraday Volatility in the US Equity Market. Multiplicative Component GARCH". In: *Journal of Financial Econometrics* 10.1 (2012), pp. 54–83.

[96] Thomas W Epps. "Comovements in Stock Prices in the Very Short Run". In: *Journal of the American Statistical Association* 74.366a (1979), pp. 291–298.

[97] Martin G Everett and Stephen P Borgatti. "The Centrality of Groups and Classes". In: *The Journal of Mathematical Sociology* 23.3 (1999), pp. 181–201.

[98] Jianqing Fan, Alex Furger, and Dacheng Xiu. "Incorporating Global Industrial Classification Standard into Portfolio Allocation: A Simple Factor-Based Large Covariance Matrix Estimator with High-Frequency Data". In: *Journal of Business & Economic Statistics* 34.4 (2016), pp. 489–503.

[99] William Fedus, Ian Goodfellow, and Andrew M Dai. "MaskGAN: Better Text Generation via Filling in the _". In: *arXiv preprint arXiv:1801.07736* (2018).

[100] Tobias Fissler, Johanna F Ziegel, et al. "Higher Order Elicitability and Osband's Principle". In: *Annals of Statistics* 44.4 (2016), pp. 1680–1707.

[101] Tobias Fissler, Johanna F Ziegel, and Tilmann Gneiting. "Expected Shortfall Is Jointly Elicitable with Value at Risk-Implications for Backtesting". In: *arXiv preprint arXiv:1507.00244* (2015).

[102] Piotr Fiszeder and Witold Orzeszko. "Covariance Matrix Forecasting Using Support Vector Regression". In: *Applied Intelligence* 51.10 (2021), pp. 7029–7042.

[103] Hans Föllmer and Alexander Schied. "Convex Measures of Risk and Trading Constraints". In: *Finance and Stochastics* 6.4 (2002), pp. 429–447.

[104] Kristin J Forbes and Roberto Rigobon. "No Contagion, Only Interdependence: Measuring Stock Market Comovements". In: *Journal of Finance* 57.5 (2002), pp. 2223–2261.

[105] Andrea Frazzini, Ronen Israel, and Tobias J Moskowitz. "Trading Costs of Asset Pricing Anomalies". In: *Fama-Miller Working Paper, Chicago Booth Research Paper* 14-05 (2012).

[106] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse Inverse Covariance Estimation with the Graphical LASSO". In: *Biostatistics* 9.3 (2008), pp. 432–441.

[107] Jerome H Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". In: *Annals of Statistics* (2001), pp. 1189–1232.

[108] Rao Fu, Jie Chen, Shutian Zeng, Yiping Zhuang, and Agus Sudjianto. "Time Series Simulation by Conditional Generative Adversarial Net". In: *arXiv preprint arXiv:1904.11419* (2019).

[109] Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov. "Predicting Volatility: Getting the Most out of Return Data Sampled at Different Frequencies". In: *Journal of Econometrics* 131.1-2 (2006), pp. 59–95.

[110] Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov. "There Is a Risk-Return Trade-Off After All". In: *Journal of Financial Economics* 76.3 (2005), pp. 509–548.

[111] Raffaella Giacomini and Halbert White. "Tests of Conditional Predictive Ability". In: *Econometrica* 74.6 (2006), pp. 1545–1578.

[112] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. "Neural Message Passing for Quantum Chemistry". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1263–1272.

[113] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2003.

[114] Tilmann Gneiting. "Making and Evaluating Point Forecasts". In: *Journal of the American Statistical Association* 106.494 (2011), pp. 746–762.

[115] Itay Goldstein, Chester S Spatt, and Mao Ye. "Big Data in Finance". In: *Review of Financial Studies* 34.7 (2021), pp. 3213–3225.

[116] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems* 27 (2014), pp. 2672–2680.

[117] Christian Gouriéroux, Joann Jasiak, and Razvan Sufana. "The Wishart Autoregressive Process of Multivariate Stochastic Volatility". In: *Journal of Econometrics* 150.2 (2009), pp. 167–181.

[118] Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. "Stochastic Optimization of Sorting Networks via Continuous Relaxations". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=H1eSS3CcKX.

[119] Shihao Gu, Bryan Kelly, and Dacheng Xiu. "Empirical Asset Pricing via Machine Learning". In: *Review of Financial Studies* 33.5 (2020), pp. 2223–2273.

[120] Xin Guo and Mihail Zervos. "Optimal Execution with Multiplicative Price Impact". In: *SIAM Journal on Financial Mathematics* 6.1 (2015), pp. 281–306.

[121] David Hallac, Youngsuk Park, Stephen Boyd, and Jure Leskovec. "Network Inference via the Time-Varying Graphical LASSO". In: *Proceedings of the 23Rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 205–213.

[122] Allaudeen Hameed, Wenjin Kang, and Shivesh Viswanathan. "Stock Market Declines and Liquidity". In: *Journal of Finance* 65.1 (2010), pp. 257–293.

[123] Lars Kai Hansen and Peter Salamon. "Neural Network Ensembles". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.10 (1990), pp. 993–1001.

[124] Peter R Hansen and Asger Lunde. "A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH (1, 1)?" In: *Journal of Applied Econometrics* 20.7 (2005), pp. 873–889.

[125] Peter R Hansen and Asger Lunde. "Realized Variance and Market Microstructure Noise". In: *Journal of Business & Economic Statistics* 24.2 (2006), pp. 127–161.

[126] Peter R Hansen, Asger Lunde, and James M Nason. "The Model Confidence Set". In: *Econometrica* 79.2 (2011), pp. 453–497.

[127] Peter Reinhard Hansen, Asger Lunde, and James M Nason. "Choosing the Best Volatility Models: The Model Confidence Set Approach". In: *Oxford Bulletin of Economics and Statistics* 65 (2003), pp. 839–861.

[128] Lawrence Harris. "A Transaction Data Study of Weekly and Intradaily Patterns in Stock Returns". In: *Journal of Financial Economics* 16.1 (1986), pp. 99–117.

[129] Lawrence E Harris and Venkatesh Panchapagesan. "The Information Content of the Limit Order Book: Evidence from Nyse Specialist Trading Decisions". In: *Journal of Financial Markets* 8.1 (2005), pp. 25–67.

[130] Joel Hasbrouck and Gideon Saar. "Limit Orders and Volatility in a Hybrid Market: The Island Ecn". In: *Stern School of Business Dept. of Finance Working Paper FIN-01-025* (2002).

[131] Joel Hasbrouck and Duane J Seppi. "Common Factors in Prices, Order Flows, and Liquidity". In: *Journal of Financial Economics* 59.3 (2001), pp. 383–411.

[132] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[133] Nikolaus Hautsch and Ruihong Huang. "The Market Impact of a Limit Order". In: *Journal of Economic Dynamics and Control* 36.4 (2012), pp. 501–522.

[134] Nikolaus Hautsch, Lada M Kyj, and Peter Malec. "Do High-Frequency Data Improve High-Dimensional Portfolio Allocations?" In: *Journal of Applied Econometrics* 30.2 (2015), pp. 263–290.

[135] Bernard Herskovic, Bryan Kelly, Hanno Lustig, and Stijn Van Nieuwerburgh. "Firm Volatility in Granular Networks". In: *Journal of Political Economy* 128.11 (2020), pp. 4097–4162.

[136] Bernard Herskovic, Bryan Kelly, Hanno Lustig, and Stijn Van Nieuwerburgh. "The Common Factor in Idiosyncratic Volatility: Quantitative Asset Pricing Implications". In: *Journal of Financial Economics* 119.2 (2016), pp. 249–283.

[137] Tim Hill, Marcus O'Connor, and William Remus. "Neural Network Models for Time Series Forecasts". In: *Management Science* 42.7 (1996), pp. 1082–1092.

[138] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[139] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer Feedforward Networks Are Universal Approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366.

[140] Kewei Hou. "Industry Information Diffusion and the Lead-Lag Effect in Stock Returns". In: *Review of Financial Studies* 20.4 (2007), pp. 1113–1138.

[141] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. "Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2439–2448.

[142] Nicolas Huck. "Large Data Sets and Machine Learning: Applications to Statistical Arbitrage". In: *European Journal of Operational Research* 278.1 (2019), pp. 330–342.

[143] Marwan Izzeldin, M Kabir Hassan, Vasileios Pappas, and Mike Tsionas. "Forecasting Realised Volatility Using ARFIMA and HAR Models". In: *Quantitative Finance* 19.10 (2019), pp. 1627–1638.

[144] Ravindran Kannan and Santosh Vempala. "Spectral Algorithms". In: *Foundations and Trends® in Theoretical Computer Science* 4.3–4 (2009), pp. 157–288.

[145] G Andrew Karolyi, Kuan-Hui Lee, and Mathijs A Van Dijk. "Understanding Commonality in Liquidity Around the World". In: *Journal of Financial Economics* 105.1 (2012), pp. 82–112.

[146] Bryan T Kelly, Semyon Malamud, and Kangying Zhou. "The Virtue of Complexity in Return Prediction". In: *Journal of Finance, forthcoming* (2023).

[147] Dror Y Kenett, Michele Tumminello, Asaf Madi, Gitit Gur-Gershgoren, Rosario N Mantegna, and Eshel Ben-Jacob. "Dominating Clasp of the Financial Sector Revealed by Partial Correlation Analysis of the Stock Market". In: *PloS One* 5.12 (2010), e15032.

[148] John Maynard Keynes. *The General Theory of Employment, Interest, and Money*. Springer, 2018.

[149] Mervyn A King and Sushil Wadhwani. "Transmission of Volatility Between Stock Markets". In: *Review of Financial Studies* 3.1 (1990), pp. 5–33.

[150] Diederik P Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[151] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *International Conference on Learning Representations* (ICLR). 2017.

[152] Leonid Kogan, Stephen A Ross, Jiang Wang, and Mark M Westerfield. "The Price Impact and Survival of Irrational Traders". In: *Journal of Finance* 61.1 (2006), pp. 195–229.

[153] Ravi Kumar Kolla, LA Prashanth, Sanjay P Bhat, and Krishna Jagannathan. "Concentration Bounds for Empirical Conditional Value-At-Risk: The Unbounded Case". In: *Operations Research Letters* 47.1 (2019), pp. 16–20.

[154] Petter N Kolm, Jeremy Turiel, and Nicholas Westray. "Deep Order Flow Imbalance: Extracting Alpha at Multiple Horizons from the Limit Order Book". In: *Available at SSRN 3900141* (2021).

[155] Adriano Koshiyama, Nick Firoozye, and Philip Treleaven. "Generative Adversarial Networks for Financial Trading Strategies Fine-Tuning and Combination". In: *Quantitative Finance* (2020), pp. 1–17.

[156] Christopher Krauss, Xuan Anh Do, and Nicolas Huck. "Deep Neural Networks, Gradient-Boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500". In: *European Journal of Operational Research* 259.2 (2017), pp. 689–702.

[157] Paul Kupiec. "Techniques for Verifying the Accuracy of Risk Measurement Models". In: *The Journal of Derivatives* 3.2 (1995).

[158] Shigeo Kusuoka. "On Law Invariant Coherent Risk Measures". In: *Advances in Mathematical Economics*. Springer, 2001, pp. 83–95.

[159] Albert S Kyle. "Continuous Auctions and Insider Trading". In: *Econometrica: Journal of the Econometric Society* (1985), pp. 1315–1335.

[160] Laurent Laloux, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud. "Random Matrix Theory and Financial Correlations". In: *International Journal of Theoretical and Applied Finance* 3.03 (2000), pp. 391–397.

[161] Nicolas S Lambert, David M Pennock, and Yoav Shoham. "Eliciting Properties of Probability Distributions". In: *Proceedings of the 9Th ACM Conference on Electronic Commerce*. 2008, pp. 129–138.

[162] Tae-Hwy Lee and Xiangdong Long. "Copula-Based Multivariate GARCH Model with Uncorrelated Dependent Errors". In: *Journal of Econometrics* 150.2 (2009), pp. 207–218.

[163] Erich L Lehmann and Joseph P Romano. *Testing Statistical Hypotheses*. Springer Science & Business Media, 2006.

[164] Jing Lei. "Convergence and Concentration of Empirical Measures under Wasserstein Distance in Unbounded Functional Spaces". In: *Bernoulli* 26.1 (2020), pp. 767–798.

[165] Michael Lemmon and Evgenia Portniaguina. "Consumer Confidence and Asset Prices: Some Empirical Evidence". In: *Review of Financial Studies* 19.4 (2006), pp. 1499–1529.

[166] Junyi Li, Xintong Wang, Yaoyang Lin, Arunesh Sinha, and Michael Wellman. "Generating Realistic Stock Market Order Streams". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.01 (2020), pp. 727–734.

[167] Sophia Zhengzi Li and Yushan Tang. "Forecasting Realized Volatility: An Automatic System Using Many Features and Many Machine Learning Algorithms". In: *Working paper* (2020).

[168] Fabrizio Lillo and J Doyne Farmer. "The Long Memory of the Efficient Market". In: *Studies in Nonlinear Dynamics & Econometrics* 8.3 (2004).

[169] Fabrizio Lillo, J Doyne Farmer, and Rosario N Mantegna. "Master Curve for Price-Impact Function". In: *Nature* 421.6919 (2003), pp. 129–130.

[170] Shiqing Ling and Michael McAleer. "Asymptotic theory for a Vector ARMA-GARCH Model". In: *Econometric Theory* 19.2 (2003), pp. 280–310.

[171] Lily Y Liu, Andrew J Patton, and Kevin Sheppard. "Does Anything Beat 5-minute RV? A Comparison of Realized Measures Across Multiple Asset Classes". In: *Journal of Econometrics* 187.1 (2015), pp. 293–311.

[172] Yulong Lu and Jianfeng Lu. "A Universal Approximation Theorem of Deep Neural Networks for Expressing Probability Distributions". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3094–3105.

[173] Ananth Madhavan, Matthew Richardson, and Mark Roomans. "Why Do Security Prices Change? A Transaction-Level Analysis of Nyse Stocks". In: *Review of Financial Studies* 10.4 (1997), pp. 1035–1064.

[174] Gautier Marti. "CorrGAN: Sampling Realistic Financial Correlation Matrices Using Generative Adversarial Networks". In: *Icassp 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (Icassp)*. IEEE. 2020, pp. 8459–8463.

[175] Nicolai Meinshausen and Peter Bühlmann. "High-Dimensional Graphs and Variable Selection with the LASSO". In: *Annals of Statistics* 34.3 (2006), pp. 1436–1462.

[176] Lior Menzly and Oguzhan Ozbas. "Market Segmentation and Cross-Predictability of Returns". In: *Journal of Finance* 65.4 (2010), pp. 1555–1580.

[177] Mehdi Mirza and Simon Osindero. "Conditional Generative Adversarial Nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[178] Randall Morck, Bernard Yeung, and Wayne Yu. "The Information Content of Stock Markets: Why Do Emerging Markets Have Synchronous Stock Price Movements?" In: *Journal of Financial Economics* 58.1-2 (2000), pp. 215–260.

[179] Mark Newman. "Networks of Information". In: *Networks*. Oxford University Press.

[180] Hao Ni, Lukasz Szpruch, Magnus Wiese, Shujian Liao, and Baoren Xiao. "Conditional Sig-Wasserstein GANs for Time Series Generation". In: *arXiv preprint arXiv:2006.05421* (2020).

[181] Sophie X Ni, Jun Pan, and Allen M Poteshman. "Volatility Information Trading in the Option Market". In: *Journal of Finance* 63.3 (2008), pp. 1059–1091.

[182] Wlodzimierz Ogryczak and Arie Tamir. "Minimizing the Sum of the K Largest Functions in Linear Time". In: *Information Processing Letters* 85.3 (2003), pp. 117–122.

[183] Dong Hwan Oh and Andrew J Patton. "High-Dimensional Copula-Based Distributions with Mixed Frequency Data". In: *Journal of Econometrics* 193.2 (2016), pp. 349–366.

[184] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "Wavenet: A Generative Model for Raw Audio". In: *arXiv preprint arXiv:1609.03499* (2016).

[185] Kent Osband. "Providing Incentives for Better Cost Forecasting". PhD thesis. University of California, Berkeley, 1985.

[186] Georg Ostrovski, Will Dabney, and Rémi Munos. "Autoregressive Quantile Networks for Generative Modeling". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 3936–3945.

[187] Razvan Pascalau and Ryan Poirier. "Increasing the Information Content of Realized Volatility Forecasts". In: *Journal of Financial Econometrics* (2021).

[188] Paolo Pasquariello and Clara Vega. "Strategic Cross-Trading in the US Stock Market". In: *Review of Finance* 19.1 (2015), pp. 229–282.

[189] Andrew J Patton. "Volatility Forecast Comparison Using Imperfect Volatility Proxies". In: *Journal of Econometrics* 160.1 (2011), pp. 246–256.

[190] Andrew J Patton and Kevin Sheppard. "Evaluating Volatility and Correlation Forecasts". In: *Handbook of Financial Time Series*. Springer, 2009, pp. 801–838.

[191] Andrew J Patton and Kevin Sheppard. "Good Volatility, Bad Volatility: Signed Jumps and the Persistence of Volatility". In: *Review of Economics and Statistics* 97.3 (2015), pp. 683–697.

[192] LA Prashanth, Krishna Jagannathan, and Ravi Kumar Kolla. "Concentration Bounds for CVAR Estimation: The Cases of Light-Tailed and Heavy-Tailed Distributions". In: *International Conference on Machine Learning*. 2020, pp. 5577–5586.

[193] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *arXiv preprint arXiv:1511.06434* (2015).

[194] Eghbal Rahimikia and Ser-Huang Poon. "Machine Learning for Realised Volatility Forecasting". In: *Working paper* (2020).

[195] David E Rapach, Jack K Strauss, and Guofu Zhou. "International Stock Return Predictability: What Is the Role of the United States?" In: *Journal of Finance* 68.4 (2013), pp. 1633–1662.

[196] Roberto Renò. "A Closer Look at the Epps Effect". In: *International Journal of Theoretical and Applied Finance* 6.01 (2003), pp. 87–102.

[197] Gabriel Trierweiler Ribeiro, André Alves Portela Santos, Viviana Cocco Mariani, and Leandro dos Santos Coelho. "Novel Hybrid Model Based on Echo State Neural Network Applied to the Prediction of Stock Price Return Volatility". In: *Expert Systems with Applications* 184 (2021), p. 115490.

[198] Mathieu Rosenbaum and Mehdi Tomas. "A Characterisation of Cross-Impact Kernels". In: *arXiv preprint arXiv:2107.08684* (2021).

[199] Apaar Sadhwani, Kay Giesecke, and Justin Sirignano. "Deep Learning for Mortgage Risk". In: *Journal of Financial Econometrics* 19.2 (2021), pp. 313–368.

[200] Michael Schneider and Fabrizio Lillo. "Cross-Impact and No-Dynamic-Arbitrage". In: *Quantitative Finance* 19.1 (2019), pp. 137–154.

[201] Robert J Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.

[202] Jun Shao. "Linear Model Selection by Cross-Validation". In: *Journal of the American statistical Association* 88.422 (1993), pp. 486–494.

[203] Kevin Sheppard. "Financial Econometrics Notes". In: *University of Oxford* (2010), pp. 333–426.

[204] Andrei Shleifer and Lawrence H Summers. "The Noise Trader Approach to Finance". In: *Journal of Economic Perspectives* 4.2 (1990), pp. 19–33.

[205] SP Sidorov, AR Faizliev, VA Balash, AA Gudkov, AZ Chekmareva, and PK Anikin. "Company Co-Mention Network Analysis". In: *International Conference on Network Analysis*. Springer. 2016, pp. 341–354.

[206] Justin Sirignano and Rama Cont. "Universal Features of Price Formation in Financial Markets: Perspectives from Deep Learning". In: *Quantitative Finance* 19.9 (2019), pp. 1449–1459.

[207] Justin A Sirignano. "Deep Learning for Limit Order Books". In: *Quantitative Finance* 19.4 (2019), pp. 549–570.

[208] Jonathan R Stroud and Michael S Johannes. "Bayesian Modeling and Forecasting of 24-hour High-Frequency Volatility". In: *Journal of the American Statistical Association* 109.508 (2014), pp. 1368–1384.

[209] Deborah Sulem, Henry Kenlay, Mihai Cucuringu, and Xiaowen Dong. "Graph Similarity Learning for Change-Point Detection in Dynamic Networks". In: *arXiv preprint arXiv:2203.15470* (2022).

[210] Efthymia Symitsi, Lazaros Symeonidis, Apostolos Kourtis, and Raphael Markellos. "Covariance Forecasting in Equity Markets". In: *Journal of Banking & Finance* 96 (2018), pp. 153–168.

[211] Shuntaro Takahashi, Yu Chen, and Kumiko Tanaka-Ishii. "Modeling Financial Time-Series with Generative Adversarial Networks". In: *Physica A: Statistical Mechanics and its Applications* 527 (2019), p. 121261.

[212] Daigo Tashiro, Hiroyasu Matsushima, Kiyoshi Izumi, and Hiroki Sakaji. "Encoding of High-Frequency Order Information and Prediction of Short-Term Stock Price by Deep Learning". In: *Quantitative Finance* 19.9 (2019), pp. 1499–1506.

[213] Stephen J Taylor and Xinzhong Xu. "The Incremental Volatility Information in One Million Foreign Exchange Quotations". In: *Journal of Empirical Finance* 4.4 (1997), pp. 317–340.

[214] Mert Tokman, R Glenn Richey, Louis D Marino, and K Mark Weaver. "Exploration, Exploitation and Satisfaction in Supply Chain Portfolio Strategy". In: *Journal of Business Logistics* 28.1 (2007), pp. 25–56.

[215] Mehdi Tomas, Iacopo Mastromatteo, and Michael Benzaquen. "How to Build a Cross-Impact Model from First Principles: Theoretical Requirements and Empirical Results". In: *Quantitative Finance* 22.6 (2022), pp. 1017–1036.

[216] Bence Tóth and János Kertész. "The Epps Effect Revisited". In: *Quantitative Finance* 9.7 (2009), pp. 793–802.

[217] Rasmus Varneskov and Valeri Voev. "The Role of Realized Ex-Post Covariance Measures and Dynamic Model Choice on the Quality of Covariance Forecasts". In: *Journal of Empirical Finance* 20 (2013), pp. 83–95.

[218] Danilo Vassallo, Giuseppe Buccheri, and Fulvio Corsi. "A DCC-Type Approach for Realized Covariance Modeling with Score-Driven Dynamics". In: *International Journal of Forecasting* 37.2 (2021), pp. 569–586.

[219] Milena Vuletić, Felix Prenzel, and Mihai Cucuringu. "Fin-GAN: Forecasting and Classifying Financial Time Series via Generative Adversarial Networks". In: *Available at SSRN 4328302* (2023).

[220] Shanshan Wang, Sebastian Neusüß, and Thomas Guhr. "Statistical Properties of Market Collective Responses". In: *European Physical Journal B* 91 (2018), pp. 1–11.

[221] Shanshan Wang, Rudi Schäfer, and Thomas Guhr. "Average Cross-Responses in Correlated Financial Markets". In: *The European Physical Journal B* 89.9 (2016), p. 207.

[222] Shanshan Wang, Rudi Schäfer, and Thomas Guhr. "Cross-Response in Correlated Financial Markets: Individual Stocks". In: *The European Physical Journal B* 89.4 (2016), p. 105.

[223] Yudong Wang, Zhiyuan Pan, and Chongfeng Wu. "Volatility Spillover from the US to International Stock Markets: A Heterogeneous Volatility Spillover GARCH Model". In: *Journal of Forecasting* 37.3 (2018), pp. 385–400.

[224] Michael D Ward and John S Ahlquist. *Maximum Likelihood for Social Science: Strategies for Analysis*. Cambridge University Press, 2018.

[225] Stefan Weber. "Distribution-Invariant Risk Measures, Information, and Dynamic Consistency". In: *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* 16.2 (2006), pp. 419–441.

[226] Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. "Quant GANs: Deep Generation of Financial Time Series". In: *Quantitative Finance* (2020), pp. 1–22.

[227] Ines Wilms, Jeroen Rombouts, and Christophe Croux. "Multivariate Volatility Forecasts for Stock Market Indices". In: *International Journal of Forecasting* 37.2 (2021), pp. 484–499.

[228] Matthieu Wyart, Jean-Philippe Bouchaud, Julien Kockelkoren, Marc Potters, and Michele Vettorazzo. "Relation Between Bid–Ask Spread, Impact and Volatility in Order-Driven Markets". In: *Quantitative Finance* 8.1 (2008), pp. 41–57.

[229] Ruoxuan Xiong, Eric P Nichols, and Yuan Shen. "Deep Learning Stock Volatility with Google Domestic Trends". In: *Working paper* (2015).

[230] Dacheng Xiu. "Quasi-Maximum Likelihood Estimation of Volatility with High Frequency Data". In: *Journal of Econometrics* 159.1 (2010), pp. 235–250.

[231] Ke Xu, Martin D Gould, and Sam D Howison. "Multi-Level Order-Flow Imbalance in a Limit Order Book". In: *Market Microstructure and Liquidity* 4.3-4 (2018), p. 1950011.

[232] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. "Time-Series Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 5508–5518.

[233] Chao Zhang, Yihuang Zhang, Mihai Cucuringu, and Zhongmin Qian. "Volatility Forecasting with Machine Learning and Intraday Commonality". In: *Journal of Financial Econometrics, forthcoming* (2023).

[234] G Peter Zhang. "Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model". In: *Neurocomputing* 50 (2003), pp. 159–175.

[235] Lan Zhang. "Estimating Covariation: Epps Effect, Microstructure Noise". In: *Journal of Econometrics* 160.1 (2011), pp. 33–47.

[236] Yaojie Zhang, Feng Ma, and Yudong Wang. "Forecasting Crude Oil Prices with a Large Set of Predictors: Can LASSO Select Powerful Predictors?" In: *Journal of Empirical Finance* 54 (2019), pp. 97–117.

[237] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. "Adversarial Feature Matching for Text Generation". In: *arXiv preprint arXiv:1706.03850* (2017).

[238] Peng Zhou, Lingxi Xie, Bingbing Ni, Cong Geng, and Qi Tian. "Omni-GAN: on the Secrets of CGANs and Beyond". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2021, pp. 14061–14071.

[239] Xuening Zhu, Rui Pan, Guodong Li, Yuewen Liu, and Hansheng Wang. "Network Vector Autoregression". In: *Annals of Statistics* 45.3 (2017), pp. 1096–1123.