Journal of Empirical Finance 78 (2024) 101533

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Big portfolio selection by graph-based conditional moments method☆

Zhoufan Zhu [a], Ningning Zhang [b], Ke Zhu [b,*]

[a] Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, China
[b] Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong

ARTICLE INFO

Keywords:
Asset pricing knowledge
Big data
Big portfolio selection
Domain knowledge
High-dimensional time series
Machine learning
Quantiled conditional moments

ABSTRACT

This paper proposes a new graph-based conditional moments (GRACE) method to do portfolio selection based on thousands of stocks or even more. The GRACE method first learns the conditional quantiles and mean of stock returns via a factor-augmented temporal graph convolutional network, which is guided by the set of stock-to-stock relations as well as the set of factor-to-stock relations. Next, the GRACE method learns the conditional variance, skewness, and kurtosis of stock returns from the learned conditional quantiles via the quantiled conditional moment method. Finally, the GRACE method uses the learned conditional mean, variance, skewness, and kurtosis to construct several performance measures, which are criteria to sort the stocks to proceed the portfolio selection in the well-known 10-decile framework. An application to NASDAQ and NYSE stock markets shows that the GRACE method performs much better than its competitors, particularly when the performance measures are comprised of conditional variance, skewness, and kurtosis.

1. Introduction

In conjunction with the huge growth of stock market capitalization, the number of existing stocks in the financial market is increasing rapidly nowadays, raising a big challenge to researchers and practitioners on how to do portfolio selection based on thousands of stocks or even more. Suppose there are N different stocks with their prices over T timepoints. Let $r_{i,t}$ denote the return of individual stock i at time t with the conditional mean $\mu_{i,t} \equiv \mathbb{E}\left(r_{i,t} \mid \mathcal{F}_{t-1}\right)$, where $i = 1, \ldots, N$, $t = 1, \ldots, T$, and $\mathcal{F}_t \equiv \sigma\left(r_{i,s}; i = 1, \ldots, N, s \le t\right)$ is the available information set up to time t. Conventionally, all considered N stocks are sorted into 10 deciles according to their predicted values of $\mu_{i,t}$ (from the smallest to the largest), and then the long-short portfolio is constructed at time $t-1$ by buying the 10% highest ranking stocks (decile 10) and selling the 10% lowest (decile 1); see, for example, Gu et al. (2020, 2021). However, the resulting mean (M) portfolio that uses $\mu_{i,t}$ as the performance measure to sort the stocks has two major shortcomings: First, it ignores the impact of conditional variance $h_{i,t} \equiv \mathrm{Var}\left(r_{i,t} \mid \mathcal{F}_{t-1}\right)$, which is the risk of uncertainty for guiding portfolio selection under the mean–variance criterion (Markowitz, 1952) or Sharpe ratio criterion (Sharpe, 1994); Second, it does not accommodate the observation that rational investors prefer assets with higher skewness and lower kurtosis in the market (Scott and Horvath, 1980; Dittmar, 2002), implying the necessity of considering the conditional skewness $s_{i,t} \equiv \mathrm{Skew}\left(r_{i,t} \mid \mathcal{F}_{t-1}\right)$ for the asymmetry risk and conditional kurtosis $k_{i,t} \equiv \mathrm{Kurt}\left(r_{i,t} \mid \mathcal{F}_{t-1}\right)$ for the tail risk to proceed the portfolio selection.

☆ We thank the anonymous referee for valuable comments. K. Zhu's work is supported by the GRF, RGC of Hong Kong (Nos. 17302424, 17304723, 17312622, and 17304421).

* Correspondence to: Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong.
E-mail addresses: tylerzzf1103@gmail.com (Z. Zhu), xcxks1@connect.hku.hk (N. Zhang), mazhuke@hku.hk (K. Zhu).

https://doi.org/10.1016/j.jempfin.2024.101533
Received 4 January 2024; Received in revised form 26 July 2024; Accepted 10 August 2024
Available online 13 August 2024

To overcome the two aforementioned shortcomings, we aim to sort the stocks for portfolio selection by using the performance measures below:

$$\text{Mean–variance (MV): } \mu_{i,t} - \lambda_1 h_{i,t}; \tag{1}$$

$$\text{Mean–variance with skewness and kurtosis (MVSK): } \mu_{i,t} - \lambda_1 h_{i,t} + \lambda_2 s_{i,t} - \lambda_3 k_{i,t}; \tag{2}$$

$$\text{Sharpe ratio (SR): } \mu_{i,t}/\sqrt{h_{i,t}}; \tag{3}$$

$$\text{Sharpe ratio with skewness and kurtosis (SRSK): } \mu_{i,t}/\sqrt{h_{i,t}} + \lambda_2 s_{i,t} - \lambda_3 k_{i,t}, \tag{4}$$

where λ_l, $l = 1, 2, 3$, are positive hyperparameters, and they determine how much penalty one needs to pay for the large values of $h_{i,t}$ and $k_{i,t}$, or how much reward one can gain for the large values of $s_{i,t}$. To implement these four performance measures, we need to learn $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ dynamically for N stocks. When $N = 1$, these three higher-order conditional moments are studied via some variants of univariate generalized autoregressive conditional heteroskedasticity (GARCH) model (Engle, 1982; Bollerslev, 1986); see, for example, Jondeau and Rockinger (2003), León et al. (2005), León and Ñíguez (2020), and references therein. However, those univariate GARCH-type methods have the risk of model mis-specification and the instability of model estimation particularly when the dynamics of $s_{i,t}$ and $k_{i,t}$ are considered. When N is large (say, e.g., $N = 1000$), no clear feasible manner so far has been offered in the literature to estimate $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ by using high-dimensional GARCH-type models, which are formed to study the dynamics of $\boldsymbol{r}_t \equiv (r_{1,t}, \ldots, r_{N,t})'$.

This paper contributes to the literature by proposing a new graph-based conditional moments (GRACE) method for portfolio selection under four performance measures in (1)–(4). The GRACE method has two core engines. Its first engine is to study the conditional quantiles of $r_{i,t}$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$ by a graph-based quantile model, which can be directly estimated via the quantile loss function (Koenker and Bassett, 1978). Our graph-based quantile model is based on a new factor-augmented temporal graph convolutional network (FTGCN), and thus it is called the FTGCN-based quantile model. This FTGCN-based quantile model uses the stock and factor features to extract both temporal and spatial information for all stocks, and then takes the extracted information to learn the conditional quantiles under the guidance of a factor-augmented hypergraph. The factor-augmented hypergraph is neither random nor time-variant, and it combines the domain knowledge of the multiple types of relation between any two stocks and the asset pricing knowledge of the impact of common factors on all stocks. Our factor-augmented hypergraph has a linkage with the hypergraph in TGCN (Feng et al., 2019) that also exploits the domain knowledge to build the graph structure among stocks, where the domain knowledge comes from the public information of stocks (e.g., the industrial background, financial statement, and shareholder information).[1] However, the hypergraph in TGCN overlooks an important fact from the asset pricing literature that some common factors can globally affect all stocks in the market (Fama and French, 1993, 2015, 2018; Hou et al., 2011). This asset pricing knowledge is obviously as informative as the domain knowledge, and it motivates the design of our factor-augmented hypergraph. Using the similar idea above, our GRACE method further proposes an FTGCN-based mean model to estimate $\mu_{i,t}$.

Based on the estimated conditional quantiles of $r_{i,t}$ at K different quantile levels from our FTGCN-based quantile model, the second engine of our GRACE method is to estimate $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ via their corresponding quantiled conditional moments (QCMs) in Zhang and Zhu (2023). In principle, the QCM method transforms the estimation of $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ to that of conditional quantiles, and this brings us two substantial advantages over the GARCH-type method. First, the QCM method is easy-to-implement as long as the estimated conditional quantiles of $r_{i,t}$ are provided. Note that our FTGCN-based quantile model can estimate conditional quantiles of $r_{i,t}$ for large N and T cases by a supervised learning through the use of quantile loss function. Therefore, unlike the estimation of high-dimensional GARCH-type models (Engle et al., 2019; Pakel et al., 2021), no assumption on the distribution of $\boldsymbol{r}_t$ is needed to estimate our FTGCN-based quantile model. This is the reason why the QCM method can make the estimation of higher-order moments feasible for large N cases, although it needs to estimate the quantile model K different times. Second, the QCM method largely alleviates the risk of model mis-specification, since the QCMs of $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ are proposed without any estimator of $\mu_{i,t}$, and more importantly, they are consistent even when the conditional quantile estimators of $r_{i,t}$ are biased to some extent. In this sense, the QCMs are consistent, as long as the specification of our FTGCN-based quantile model does not largely deviate from the true specification of conditional quantile of $r_{i,t}$.

We apply our GRACE method to construct long-short portfolios based on two benchmark stock data sets in Feng et al. (2019), which contain 1026 and 1737 stocks in NASDAQ and NYSE, respectively. To build the factor-augmented hypergraph, we use the Wiki company-based relations (Feng et al., 2019) as the domain knowledge to specify the multiple types of relation between any two stocks, and meanwhile, we take the Fama–French five factors (Fama and French, 2015) as the asset pricing knowledge to capture the impact of common factors on all stocks. From an economic viewpoint, our empirical results are encouraging in four aspects. First, all of the MV, MVSK, SR, and SRSK portfolios have larger values of out-of-sample annualized SR than the M portfolio in the GRACE method. Second, the SRSK portfolio from the GRACE method performs the best, and its values of out-of-sample annualized SR are 4.81 and 3.48 in NASDAQ and NYSE, respectively, which are 236% and 21% higher than those of the M portfolio from the benchmark method in Feng et al. (2019). Third, the GRACE method always dominates the simple GRACE method in portfolio selection by a wide margin, where the simple GRACE method adopts the linear structure (Zhu et al., 2017, 2019) to extract the information from stock and factor features to learn the conditional quantiles and mean of stock returns. Fourth, regardless of performance measure,

[1] The usefulness of the public information of stocks has been well documented by Livingston (1977), Cohen and Frazzini (2008), Lee et al. (2019), Burt and Hrdlicka (2021), and many others.

the portfolios from the GRACE method have a more robust performance than those from its competitors over the set of stock-to-stock relations, the choice of hyperparameters, and the level of transection cost. All of these findings indicate the importance of using the higher-order conditional moments to form the performance measure, the asset pricing knowledge to build the hypergraph, and the network structure to extract the feature information. Finally, we also apply our GRACE method to study the expanded dataset that contains daily stock data from January 2, 2013 to December 30, 2022, and we find that its advantage over competitors remains.

The remaining paper is organized as follows. Section 2 presents our entire methodology, including the network architecture of FTGCN, the training procedure of FTGCN-based quantile and mean models, the formal estimation procedure of the QCMs, and the implementation details of the GRACE method. Section 3 presents our empirical studies of big portfolio selection in NASDAQ and NYSE stock markets. Concluding remarks are offered in Section 4. Some additional empirical results are deferred into the supplementary materials.

2. Methodology

2.1. Graph-based learning for conditional quantiles

Let $\boldsymbol{Q}_t(\tau) = (Q_{1,t}(\tau), \ldots, Q_{N,t}(\tau))'$ be the high-dimensional vector of τ-th conditional quantiles, where $Q_{i,t}(\tau)$ is the τ-th conditional quantile of $r_{i,t}$ given $\mathcal{F}_{t-1}$. We study $\boldsymbol{Q}_t(\tau)$ by a new FTGCN-based quantile model defined as

$$\boldsymbol{Q}_t(\tau) = f(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\tau), \tag{5}$$

where $\boldsymbol{X}_{t-1} \in \mathcal{R}^{(N+B)\times P\times S}$ is a feature tensor built on $\mathcal{F}_{t-1}$ including the information of N stocks and B factors, and $f(\cdot; \mathcal{G}, \boldsymbol{\theta}_\tau) : \mathcal{R}^{(N+B)\times P\times S} \to \mathcal{R}^{N\times 1}$ is the FTGCN depending on a factor-augmented hypergraph $\mathcal{G}$ and a vector of unknown parameters $\boldsymbol{\theta}_\tau$. Here, $\boldsymbol{X}_{t-1} = [\boldsymbol{X}_{1,t-1}, \ldots, \boldsymbol{X}_{N,t-1}, \boldsymbol{X}_{N+1,t-1}, \ldots, \boldsymbol{X}_{N+B,t-1}]$ with $\boldsymbol{X}_{i,t-1} \in \mathcal{R}^{P\times S}$ having its sth column $\boldsymbol{x}_{i,t-1-S+s} \in \mathcal{R}^P$, where $\boldsymbol{X}_{i,t-1}$ for $i = 1, \ldots, N$ is the feature matrix for stock i, $\boldsymbol{X}_{N+b,t-1}$ for $b = 1, \ldots, B$ is the feature matrix for factor b, P with a potential high dimension is the number of stock or factor features, S is the number of lagged values of each feature, and $\boldsymbol{x}_{i,t}$ (or $\boldsymbol{x}_{N+b,t}$) is the vector of P different features of stock i (or factor b) at time t. Below, we show the four modules to construct the FTGCN $f(\cdot; \mathcal{G}, \boldsymbol{\theta}_\tau)$.

2.1.1. Module I: Feature extraction

In the first module, we employ a one-layer long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997) to extract the temporal embedding $\boldsymbol{x}_{i,t}^L \in \mathcal{R}^d$ from the feature matrix $\boldsymbol{X}_{i,t-1}$ at time $t-1$. Using LSTM networks to extract temporal information from financial time series is a widely employed methodology; see, for example, Kim and Won (2018), Feng et al. (2019), Ghosh et al. (2022), and references therein. Specifically, we let $\underline{\boldsymbol{x}}_{i,t-1}^s = \boldsymbol{x}_{i,t-1-S+s}$ and compute the hidden state vectors $\boldsymbol{h}_{i,t}^s$, $s = 1, \ldots, S$, recursively from the LSTM network:

$$\begin{aligned}
\boldsymbol{z}_{i,t}^s &= \tanh(\boldsymbol{W}_{1x}\underline{\boldsymbol{x}}_{i,t-1}^s + \boldsymbol{W}_{1h}\boldsymbol{h}_{i,t}^{s-1} + \boldsymbol{b}_1),\\
\boldsymbol{i}_{i,t}^s &= \text{sigmoid}(\boldsymbol{W}_{2x}\underline{\boldsymbol{x}}_{i,t-1}^s + \boldsymbol{W}_{2h}\boldsymbol{h}_{i,t}^{s-1} + \boldsymbol{b}_2),\\
\boldsymbol{f}_{i,t}^s &= \text{sigmoid}(\boldsymbol{W}_{3x}\underline{\boldsymbol{x}}_{i,t-1}^s + \boldsymbol{W}_{3h}\boldsymbol{h}_{i,t}^{s-1} + \boldsymbol{b}_3),\\
\boldsymbol{c}_{i,t}^s &= \boldsymbol{f}_{i,t}^s \odot \boldsymbol{c}_{i,t}^{s-1} + \boldsymbol{i}_{i,t}^s \odot \boldsymbol{z}_{i,t}^s,\\
\boldsymbol{o}_{i,t}^s &= \text{sigmoid}(\boldsymbol{W}_{4x}\underline{\boldsymbol{x}}_{i,t-1}^s + \boldsymbol{W}_{4h}\boldsymbol{h}_{i,t}^{s-1} + \boldsymbol{b}_4),\\
\boldsymbol{h}_{i,t}^s &= \boldsymbol{o}_{i,t}^s \odot \tanh(\boldsymbol{c}_{i,t}^s),
\end{aligned} \tag{6}$$

where $\boldsymbol{W}_{1x}, \boldsymbol{W}_{2x}, \boldsymbol{W}_{3x}, \boldsymbol{W}_{4x} \in \mathcal{R}^{d\times P}$ and $\boldsymbol{W}_{1h}, \boldsymbol{W}_{2h}, \boldsymbol{W}_{3h}, \boldsymbol{W}_{4h} \in \mathcal{R}^{d\times d}$ are matrices of weight parameters, $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3, \boldsymbol{b}_4 \in \mathcal{R}^d$ are vectors of bias parameters, d is the dimension of hidden vectors (e.g., $\boldsymbol{h}_{i,t}^s$) controlling the network complexity, $\tanh(\cdot)$ and $\text{sigmoid}(\cdot)$ are two entry-wise vector-valued functions, $\odot$ is element-wise production operation, and the initial values $\boldsymbol{c}_{i,t}^0$ and $\boldsymbol{h}_{i,t}^0$ are conventionally set to the d-dimensional vector of zeros. Then, our temporal embedding $\boldsymbol{x}_{i,t}^L$ is taken as the last hidden state vector $\boldsymbol{h}_{i,t}^S$ (the output of LSTM network), that is,

$$\boldsymbol{x}_{i,t}^L = \boldsymbol{h}_{i,t}^S = h(\boldsymbol{X}_{i,t-1}; \boldsymbol{\theta}_L) \in \mathcal{R}^d, \tag{7}$$

where $h(\cdot; \boldsymbol{\theta}_L)$ is the LSTM network in (6) indexed by $\boldsymbol{\theta}_L$, and $\boldsymbol{\theta}_L$ contains all the parameters in $\{\boldsymbol{W}_{jx}, \boldsymbol{W}_{jh}, \boldsymbol{b}_j : j = 1, 2, 3, 4\}$. Clearly, the purpose of this module is to extract an expressive vector $\boldsymbol{x}_{i,t}^L$, which stores the long-term temporal information of P features up to time $t-1$. It is expected that all the temporal information carried by $\{\boldsymbol{x}_{1,t}^L, \ldots, \boldsymbol{x}_{N+B,t}^L\}$ can help us to predict the behavior of future returns $\{r_{1,t}, \ldots, r_{N,t}\}$ at time t.

2.1.2. Module II: Hypergraph construction

Besides the temporal information from the stock features, the spatial information (i.e., interdependence relations) among all stocks is also important for predictions. For example, (i) MSFT LLC and Google LLC could have an industry-specific relation, since both of them belong to "Computer Software: Programming" industry; and (ii) Boeing Inc. and United Airlines Inc. could have a corporate relation, in view of the fact that Boeing Inc. produces Boeing airplanes for United Airlines Inc. Needless to say, these stock-to-stock (S2S) relations are informative and should not be ignored. In general, we can have M different types of S2S relation (denoted by $\boldsymbol{E}_{stock} = \{e_1, \ldots, e_M\}$) between any two stocks based on the domain knowledge.

Along with the S2S relations, the factor-to-stock (F2S) relations also exist in the market, since the stock returns can move together driven by the common factors; see the vast evidence in the asset pricing literature (Fama and French, 2018; Lettau and Pelger, 2020; Gu et al., 2021). The F2S relations convey the spatial information from factors to stocks, and they are highly possible to be factor-specific. Therefore, based on the asset pricing knowledge, we consider B different F2S relations (denoted by $\boldsymbol{E}_{factor} = \{e_{M+1}, \ldots, e_{M+B}\}$), where the F2S relation e_{M+b} is induced by factor b.

To describe all of S2S and F2S relations above, we build a factor-augmented hypergraph

$$\mathcal{G} = (\boldsymbol{V}, \boldsymbol{A}), \tag{8}$$

where $\boldsymbol{V} = \{\boldsymbol{V}_{stock}, \boldsymbol{V}_{factor}\}$ is the set of vertices, and $\boldsymbol{A} = \{\boldsymbol{A}_{stock}, \boldsymbol{A}_{factor}\}$ is the set of adjacency matrices. Here, $\boldsymbol{V}_{stock} = \{1, \ldots, N\}$ is the set of stock vertices with the vertex $i \in \boldsymbol{V}_{stock}$ representing the stock i, $\boldsymbol{V}_{factor} = \{N+1, \ldots, N+B\}$ is the set of factor vertices with the vertex $N+b \in \boldsymbol{V}_{factor}$ representing the factor b, $\boldsymbol{A}_{stock} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_M\}$ is the set of adjacency matrices with the matrix $\boldsymbol{A}_m \in \boldsymbol{A}_{stock}$ representing the S2S relation e_m, and $\boldsymbol{A}_{factor} = \{\boldsymbol{A}_{M+1}, \ldots, \boldsymbol{A}_{M+B}\}$ is the set of adjacency matrices with the matrix $\boldsymbol{A}_{M+b} \in \boldsymbol{A}_{factor}$ representing the F2S relation e_{M+b}, where $\boldsymbol{A}_m$ has its (i,j)-th entry

$$a_{i,j,m} = \begin{cases} 1, & \text{if there is an S2S relation } e_m \text{ between vertices } i \in \boldsymbol{V}_{stock} \text{ and } j \in \boldsymbol{V}_{stock}, \\ 0, & \text{otherwise,} \end{cases}$$

and $\boldsymbol{A}_{M+b}$ has its (i,j)-th entry

$$a_{i,j,M+b} = \begin{cases} 1, & \text{if } i \in \boldsymbol{V}_{stock} \text{ and } j = N+b \in \boldsymbol{V}_{factor} \text{ or } j \in \boldsymbol{V}_{stock} \text{ and } i = N+b \in \boldsymbol{V}_{factor}, \\ 0, & \text{otherwise.} \end{cases}$$

According to the definitions of $\boldsymbol{A}_m$ and $\boldsymbol{A}_{M+b}$, the factor-augmented hypergraph $\mathcal{G}$ ensures that (i) two stock vertices in $\boldsymbol{V}_{stock}$ are linked when they have up to M different S2S relations; and (ii) each factor vertex in $\boldsymbol{V}_{factor}$ is linked to all of stock vertices in $\boldsymbol{V}_{stock}$ indicating the corresponding F2S relation. Since our target is to study the dynamics of stocks rather than factors, we do not need to specify the connections of factors, which are irrelevant to our analysis results below.

In sum, the overall relation between any two vertices i and j in $\mathcal{G}$ can be represented by the vector

$$\boldsymbol{a}_{i,j} = (a_{i,j,1}, \ldots, a_{i,j,M}, a_{i,j,M+1}, \ldots, a_{i,j,M+B})' \in \mathcal{R}^{M+B}, \tag{9}$$

where the first M entries and the remaining B entries carry the information of S2S relations and F2S relations, respectively.

2.1.3. Module III: Hypergraph learning

Having known the relations among all stocks and factors in $\mathcal{G}$, it is natural to capture how much temporal information the stock i can receive from its linked stocks and factors. To fulfill this goal, we define the aggregated temporal embedding for stock i as

$$\boldsymbol{x}_{i,t}^{P} = \sum_{j \in \boldsymbol{V}_{stock}, j \neq i} \frac{g\left(\boldsymbol{a}_{i,j}, \boldsymbol{x}_{i,t}^{L}, \boldsymbol{x}_{j,t}^{L}; \mathcal{G}, \boldsymbol{\theta}_P\right)}{d_j} \boldsymbol{x}_{j,t}^{L} + \sum_{j \in \boldsymbol{V}_{factor}} \frac{g\left(\boldsymbol{a}_{i,j}, \boldsymbol{x}_{i,t}^{L}, \boldsymbol{x}_{j,t}^{L}; \mathcal{G}, \boldsymbol{\theta}_P\right)}{N} \boldsymbol{x}_{j,t}^{L} \in \mathcal{R}^{d}, \tag{10}$$

where $i = 1, \ldots, N$, $\boldsymbol{a}_{i,j}$ in (9) represents the overall relation between vertices i and j, $\boldsymbol{x}_{i,t}^{L}$ and $\boldsymbol{x}_{j,t}^{L}$ are the temporal embeddings extracted by (7), $g(\cdot) : \mathcal{R}^{M+B} \times \mathcal{R}^{d} \times \mathcal{R}^{d} \to \mathcal{R}$ is a nonlinear function indexed by $\boldsymbol{\theta}_P$ to measure the intensity of interplay between vertices i and j, and $d_j = \sum_{i \neq j} I(\sum_{m=1}^{M} a_{i,j,m} > 0)$ is the number of stocks that are connected to stock j. Here, $I(\cdot)$ is the indicator function. Following Feng et al. (2019), we take

$$g\left(\boldsymbol{a}_{i,j}, \boldsymbol{x}_{i,t}^{L}, \boldsymbol{x}_{j,t}^{L}; \mathcal{G}, \boldsymbol{\theta}_P\right) = \mathrm{softmax}\left(\boldsymbol{W}_5 (\boldsymbol{x}_{i,t}^{L\prime}, \boldsymbol{x}_{j,t}^{L\prime}, \boldsymbol{a}_{i,j}')' + b_5\right) \in \mathcal{R}, \tag{11}$$

where $\boldsymbol{W}_5 \in \mathcal{R}^{1 \times (M+B+2d)}$ is a vector of weight parameters, $b_5 \in \mathcal{R}$ is a bias parameter, $\boldsymbol{\theta}_P$ contains all the parameters in $\boldsymbol{W}_5$ and b_5, and $\mathrm{softmax}(\cdot)$ is used to normalize the value of $g(\cdot)$ into $(0,1)$. The specification of $g(\cdot)$ in (11) has two merits: First, it allows the intensity of interplay between any two vertices to be stock-, factor-, and relation-specific; Second, it aims to capture some missing relations that are not described in $\boldsymbol{a}_{i,j}$ (i.e., $\boldsymbol{a}_{i,j} \equiv 0$) but presented by the similarity of $\boldsymbol{x}_{i,t}^{L}$ and $\boldsymbol{x}_{j,t}^{L}$, since the term $\boldsymbol{W}_5 (\boldsymbol{x}_{i,t}^{L\prime}, \boldsymbol{x}_{j,t}^{L\prime}, \boldsymbol{a}_{i,j}')'$ is still informative even when $\boldsymbol{a}_{i,j} \equiv 0$.

As a temporal graph convolution (TGC), the third module combines the temporal embedding $\boldsymbol{x}_{i,t}^{L}$ in (7) and the aggregated temporal embedding $\boldsymbol{x}_{i,t}^{P}$ in (10) to form

$$\boldsymbol{x}_{i,t}^{TGC} = \left(\boldsymbol{x}_{i,t}^{L\prime}, \boldsymbol{x}_{i,t}^{P\prime}\right)' \in \mathcal{R}^{2d} \text{ for } i = 1, \ldots, N. \tag{12}$$

The advantage of using $\boldsymbol{x}_{i,t}^{TGC}$ is apparent, since $\boldsymbol{x}_{i,t}^{TGC}$ captures the spatial and temporal information of stock features simultaneously.

2.1.4. Module IV: Quantile output

Our last module applies a fully connected (FC) network to revise the spatial–temporal information $\boldsymbol{x}_{i,t}^{TGC}$ in (12) to $\boldsymbol{Q}_t(\tau)$ in (5). Let $\boldsymbol{\theta}_C$ contain all the parameters in $\boldsymbol{W}_6$ and b_6, where $\boldsymbol{W}_6 \in \mathcal{R}^{1 \times 2d}$ is a vector of weight parameters, and $b_6 \in \mathcal{R}$ is a bias parameter. Then, we set the form of FTGCN as

$$\begin{aligned} &f(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}) \equiv (f_1(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}), \ldots, f_N(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}))' \\ &\text{with } f_i(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}) = \boldsymbol{W}_6 \boldsymbol{x}_{i,t}^{TGC} + b_6 \text{ for } i = 1, \ldots, N, \end{aligned} \tag{13}$$

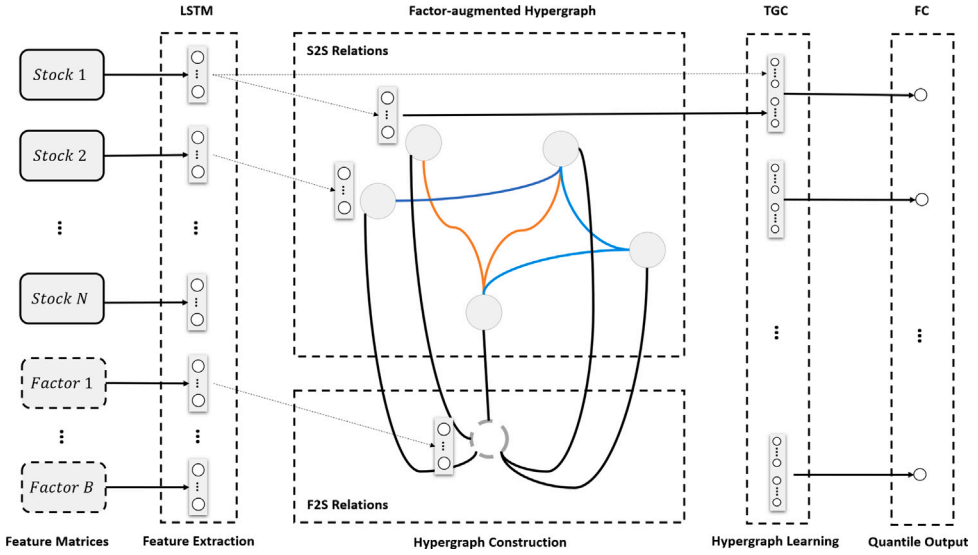

Fig. 1. The architecture of FTGCN-based quantile model.

where $\mathcal{G}$ is the factor-augmented hypergraph in (8), and $\boldsymbol{\theta}$ contains all the parameters in $\boldsymbol{\theta}_L$, $\boldsymbol{\theta}_P$, and $\boldsymbol{\theta}_C$. Consequently, our FTGCN-based quantile model has the specification:

$$\boldsymbol{Q}_t(\tau) = f(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\tau) \text{ with } Q_{i,t}(\tau) = f_i(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\tau) \text{ for } i = 1, \ldots, N \tag{14}$$

(see its network architecture in Fig. 1), where $f(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\tau)$ and $f_i(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\tau)$ are defined as in (13).

2.1.5. Estimation of the FTGCN-based quantile model

As $Q_{i,t}(\tau)$ is the τ-th conditional quantile of $r_{i,t}$ given $\mathcal{F}_{t-1}$, we estimate $\boldsymbol{\theta}_\tau$ in (14) by the following quantile estimator:

$$\widehat{\boldsymbol{\theta}}_\tau = \underset{\boldsymbol{\theta}_\tau}{\operatorname{argmin}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho_\tau \left(r_{i,t} - f_i(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\tau) \right) \equiv \underset{\boldsymbol{\theta}_\tau}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} \ell(\boldsymbol{r}_t, \boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\tau, \tau), \tag{15}$$

where $\rho_\tau(x) = x[\tau - I(x < 0)]$ is the quantile loss function (Koenker and Bassett, 1978). Due to the massive data volume, we adopt the adaptive moment estimation (Adam) algorithm in Kingma and Ba (2015) to compute $\widehat{\boldsymbol{\theta}}_\tau$ in (15); see Algorithm 1 for the details. Using $\widehat{\boldsymbol{\theta}}_\tau$, we obtain

$$\widehat{\boldsymbol{Q}}_t(\tau) = f(\boldsymbol{X}_{t-1}; \mathcal{G}, \widehat{\boldsymbol{\theta}}_\tau), \tag{16}$$

which is the estimator of $\boldsymbol{Q}_t(\tau)$ at the quantile level τ.

Algorithm 1 The training procedure of $\widehat{\boldsymbol{\theta}}_\tau$ by the Adam algorithm.

Input:
The sample: $\{(\boldsymbol{r}_t, \boldsymbol{X}_{t-1})\}$;
The initial value of network parameters in the FTGCN-based quantile model: $\boldsymbol{\theta}_\tau^{(0)}$;
The factor-augmented hypergraph: $\mathcal{G}$;
Hyperparameter: learning rate γ;

$l = 0$;
repeat
  $(\boldsymbol{r}_l, \boldsymbol{X}_{l-1}) \leftarrow$ draw a random data point from $\{(\boldsymbol{r}_t, \boldsymbol{X}_{t-1})\}$ (A cross-sectional minibatch);
  $\boldsymbol{g}_\tau^{(l)} \leftarrow \nabla_{\boldsymbol{\theta}_\tau} \left[\ell(\boldsymbol{r}_l, \boldsymbol{X}_{l-1}; \mathcal{G}, \boldsymbol{\theta}_\tau^{(l)}, \tau) \right]$ (Gradients of minibatch estimator);
  $\boldsymbol{\theta}_\tau^{(l+1)} \leftarrow$ update parameters using learning rate γ and gradients $\boldsymbol{g}_\tau^{(l)}$ (Adam);
  $l \leftarrow l + 1$;
until convergence of parameters $\boldsymbol{\theta}_\tau^{(l+1)}$;

Output:
The value of $\boldsymbol{\theta}_\tau^{(l+1)}$, which is taken as the quantile estimator $\widehat{\boldsymbol{\theta}}_\tau$.

2.1.6. Comparison with the existing models

Our FTGCN-based quantile model in (14) has a linkage with the TGCN-based model in Feng et al. (2019) with regard to the network structure. As the pioneering work, the TGCN-based model applies the domain knowledge to construct a hypergraph for taking multiple types of S2S relation into account. The main difference between the FTGCN-based quantile model and the TGCN-based model is two-fold. First, the FTGCN-based quantile model aims to learn the conditional quantiles of $r_{i,t}$, whereas the TGCN-based model focuses on the conditional mean of $r_{i,t}$. Second, the FTGCN-based quantile model incorporates the F2S relations to build the factor-augmented hypergraph, but the TGCN-based model does not consider this kind of important information in its hypergraph.

Besides our FTGCN-based quantile model, many other models are existing in the literature to study the conditional quantile of high-dimensional data; see, for example, Koenker (2004), Kato et al. (2012), and Galvao and Kato (2016) for the quantile individual fixed effects models, Ando and Bai (2020), Chen et al. (2021), Ma et al. (2021), and Yang et al. (2024) for the quantile factor models, and Härdle et al. (2016), Zhu et al. (2019), and Xu et al. (2024) for the quantile network models. However, except for the factor-augmented dynamic network quantile regression (FDNQR) model in Xu et al. (2024), none of the aforementioned models takes the domain knowledge and asset pricing knowledge simultaneously into account to guide the estimation of conditional quantile. Specifically, the FDNQR model uses the domain knowledge to propose a weighted adjacency matrix $\boldsymbol{W} \in \mathcal{R}^{N\times N}$ with the (i,j)-th entry $w_{i,j}$, where $w_{i,j} = a_{i,j}/n_i$, $n_i = \sum_{j=1}^{N} a_{i,j}$, $a_{i,j} = 1$ if the stock i has the connection with another stock j, and $a_{i,j} = 0$ otherwise. Based on $\boldsymbol{W}$, the FDNQR model assumes

$$Q_{i,t}(\tau) = \alpha_\tau + \boldsymbol{\beta}_\tau' \boldsymbol{z}_i + \gamma_\tau \sum_{j=1}^{N} w_{i,j} r_{j,t-1} + \zeta_\tau r_{i,t-1} + \sum_{s=1}^{S} \boldsymbol{\varsigma}_{s,\tau}' \boldsymbol{F}_{t-s}, \tag{17}$$

where $\alpha_\tau \in \mathcal{R}$, $\boldsymbol{\beta}_\tau \in \mathcal{R}^{Q\times 1}$, $\gamma_\tau \in \mathcal{R}$, and $\boldsymbol{\varsigma}_{s,\tau} \in \mathcal{R}^{B\times 1}$ are quantile regression coefficients, $\boldsymbol{z}_i \in \mathcal{R}^{Q\times 1}$ is a Q-dimensional vector of time-invariant stock features, and $\boldsymbol{F}_t = (f_{1,t}, \ldots, f_{B,t})' \in \mathcal{R}^{B\times 1}$ is a B-dimensional vector of time-variant factors. In model (17), α_τ is the constant intercept term for all stocks, $\boldsymbol{\beta}_\tau$ is the constant intensity of the impact from stock features on stock i, γ_τ is the constant intensity of spatial impact on stock i caused by its connected stocks, ζ_τ is the constant intensity of temporal impact on stock i caused by its lagged term, and $\boldsymbol{\varsigma}_{s,\tau}$ is the constant intensity of factor impact on all stocks caused by the lagged factors. Clearly, our FTGCN-based quantile model is much more general than model (17), since it captures multiple types of relation separately, extracts the information of time-variant stock and factor features in a non-linear way, and allows for the time-variant heterogeneous intensity of spatial and temporal impacts on each stock caused by either its connected stocks or factors.

Note that model (17) does not include the contemporaneous variables in the original FDNQR model for the purpose of prediction, and it nests the network quantile autoregressive model in Zhu et al. (2019). As the contemporaneous variables are absent, model (17) now can be consistently estimated by using the quantile loss function as in Zhu et al. (2019).

2.2. Graph-based learning for conditional mean

So far, we have introduced the FTGCN to learn the conditional quantile of $r_{i,t}$. Following the similar idea, we can learn the conditional mean of $r_{i,t}$ by an FTGCN-based mean model:

$$r_{i,t} = f_i(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\mu) + \varepsilon_{i,t}, \tag{18}$$

where $f_i(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\mu)$ is defined as in (13), and $\varepsilon_{i,t}$ is the error term with zero mean. Note that model (18) reduces to the TGCN model in Feng et al. (2019), when the factors and their features are absent. To estimate model (18), we consider the penalized least squares (PLS) estimator of $\boldsymbol{\theta}_\mu$ given by

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_\mu &= \underset{\boldsymbol{\theta}_\mu}{\operatorname{argmin}} \frac{1}{T}\sum_{t=1}^{T}\Bigg(\frac{1}{N}\sum_{i=1}^{N}[r_{i,t} - f_i(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\mu)]^2 \\ &\quad + \frac{\lambda^*}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\max\left\{0, -[f_i(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\mu) - f_j(\boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\mu)](r_{i,t} - r_{j,t})\right\}\Bigg) \\ &\equiv \underset{\boldsymbol{\theta}_\mu}{\operatorname{argmin}} \frac{1}{T}\sum_{t=1}^{T} \ell_\mu(\boldsymbol{r}_t, \boldsymbol{X}_{t-1}; \mathcal{G}, \boldsymbol{\theta}_\mu, \lambda^*),\end{aligned} \tag{19}$$

where the penalty term is utilized to ensure that the orders of $r_{i,t}$ and $r_{j,t}$ do not deviate largely from those of their predicted values, and λ^* is a positive hyperparameter. As for $\widehat{\boldsymbol{\theta}}_\tau$, we adopt the Adam algorithm to compute $\widehat{\boldsymbol{\theta}}_\mu$; see Algorithm 2 for the details. Using $\widehat{\boldsymbol{\theta}}_\mu$, we then estimate the conditional mean $\mu_{i,t}$ by

$$\widehat{\mu}_{i,t} = f_i(\boldsymbol{X}_{t-1}; \mathcal{G}, \widehat{\boldsymbol{\theta}}_\mu) \quad \text{for } i = 1, \ldots, N. \tag{20}$$

It is worthwhile mentioning that the order-preserving penalty in (19) has been widely adopted in the literature to improve the learning efficiency for conditional mean prediction (Zheng et al., 2007; Socher et al., 2013; Feng et al., 2019). However, this penalized method is inappropriate for the conditional quantile estimation, since $Q_{i,t}(\tau)$ does not tend to be larger than $Q_{j,t}(\tau)$ when $r_{i,t}$ is larger than $r_{j,t}$.

Algorithm 2 The training procedure of $\widehat{\theta}_{\mu}$ by the Adam algorithm.

Input:
The sample: $\{(\boldsymbol{r}_t, \boldsymbol{X}_{t-1})\}$;
The initial value of network parameters in the FTGCN-based mean model: $\theta_{\mu}^{(0)}$;
The factor-augmented hypergraph: $\mathcal{G}$;
Hyperparameters: λ^*, learning rate γ;
$l = 0$;
repeat
  $(\boldsymbol{r}_l, \boldsymbol{X}_{l-1}) \leftarrow$ draw a random data point from $\{(\boldsymbol{r}_t, \boldsymbol{X}_{t-1})\}$ (A cross-sectional minibatch);
  $\boldsymbol{g}^{(l)} \leftarrow \nabla_{\theta_{\mu}} \left[\ell_{\mu}(\boldsymbol{r}_l, \boldsymbol{X}_{l-1}; \mathcal{G}, \theta_{\mu}^{(l)}, \lambda^*)\right]$ (Gradients of minibatch estimator);
  $\theta_{\mu}^{(l+1)} \leftarrow$ update parameters using learning rate γ and gradients $\boldsymbol{g}^{(l)}$ (Adam);
  $l \leftarrow l + 1$;
until convergence of parameters $\theta_{\mu}^{(l+1)}$;
Output:
The value of $\theta_{\mu}^{(l+1)}$, which is taken as the PLS estimator $\widehat{\theta}_{\mu}$.

2.3. The QCM learning for higher-order conditional moments

Let $\widehat{\boldsymbol{Q}}_t(\tau_1), \ldots, \widehat{\boldsymbol{Q}}_t(\tau_K)$ be the vectors of estimated conditional quantiles at K different quantile levels $\tau_1, \ldots, \tau_K$, where $\widehat{\boldsymbol{Q}}_t(\tau_k) \equiv (\widehat{Q}_{1,t}(\tau_k), \ldots, \widehat{Q}_{N,t}(\tau_k))'$ for $k = 1, \ldots, K$ are computed as in (16). Below, we elaborate on how to estimate $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ by the QCM method in Zhang and Zhu (2023) for the fixed values of i and t, based on $\widehat{Q}_{i,t}(\tau_1), \ldots, \widehat{Q}_{i,t}(\tau_K)$.

The QCM method is motivated by the Cornish-Fisher expansion (Cornish and Fisher, 1938), which shows the fundamental linkage between conditional quantiles and conditional moments:

$$\begin{aligned} Q_{i,t}(\tau_k) = \mu_{i,t} + z(\tau_k)\sqrt{h_{i,t}} + \left[z(\tau_k)^2 - 1\right]\frac{\sqrt{h_{i,t}}s_{i,t}}{6} \\ + \left[z(\tau_k)^3 - 3z(\tau_k)\right]\frac{\sqrt{h_{i,t}}(k_{i,t} - 3)}{24} + \sqrt{h_{i,t}}\omega_{i,t}(\tau_k) \end{aligned} \tag{21}$$

for $k = 1, \ldots, K$, where $z(\tau_k)$ is the τ_k-th quantile of standard normal distribution, and $\omega_{i,t}(\tau_k)$ is the remainder of this expansion. Define

$$\begin{aligned} &\varepsilon_{i,t,k}^{\bullet} = \varepsilon_{i,t,k}^{*} + \varepsilon_{i,t,k}^{\circ} \text{ with } \varepsilon_{i,t,k}^{*} = \sqrt{h_{i,t}}\omega_{i,t}(\tau_k) \text{ and } \varepsilon_{i,t,k}^{\circ} = \widehat{Q}_{i,t}(\tau_k) - Q_{i,t}(\tau_k), \\ &\boldsymbol{Z}_k = \left(z(\tau_k), z(\tau_k)^2 - 1, z(\tau_k)^3 - 3z(\tau_k)\right)', \\ &\boldsymbol{\beta}_{i,t} \equiv (\beta_{i,t,1}, \beta_{i,t,2}, \beta_{i,t,3})' = \left(\sqrt{h_{i,t}}, \frac{\sqrt{h_{i,t}}s_{i,t}}{6}, \frac{\sqrt{h_{i,t}}(k_{i,t} - 3)}{24}\right)', \end{aligned}$$

where $\widehat{Q}_{i,t}(\tau_k)$ is the estimator of $Q_{i,t}(\tau_k)$. Then, we can rewrite (21) as follows:

$$\widehat{Q}_{i,t}(\tau_k) = \mu_{i,t} + \boldsymbol{Z}_k'\boldsymbol{\beta}_{i,t} + \varepsilon_{i,t,k}^{\bullet} \quad \text{for } k = 1, \ldots, K, \tag{22}$$

where $\varepsilon_{i,t,k}^{\bullet}$ is the gross error containing the expansion error $\varepsilon_{i,t,k}^{*}$ and the quantile estimation error $\varepsilon_{i,t,k}^{\circ}$. Clearly, Eq. (22) is a linear regression model with the response variable $\widehat{Q}_{i,t}(\tau_k)$, explanatory variables $\boldsymbol{Z}_k$, parameter vector $(\mu_{i,t}, \boldsymbol{\beta}_{i,t}')'$, and error term $\varepsilon_{i,t,k}^{\bullet}$. Since $\varepsilon_{i,t,k}^{\bullet}$ may not have zero mean for model identification, we add an additional deterministic intercept term $\gamma_{i,t}$ into Eq. (22) to form the following linear regression model:

$$\widehat{Q}_{i,t}(\tau_k) = (\mu_{i,t} + \gamma_{i,t}) + \boldsymbol{Z}_k'\boldsymbol{\beta}_{i,t} + \varepsilon_{i,t,k} \equiv \bar{\boldsymbol{Z}}_k'\boldsymbol{\theta}_{i,t} + \varepsilon_{i,t,k} \quad \text{for } k = 1, \ldots, K, \tag{23}$$

where $\varepsilon_{i,t,k} = \varepsilon_{i,t,k}^{\bullet} - \gamma_{i,t}$, $\bar{\boldsymbol{Z}}_k = (1, \boldsymbol{Z}_k')'$, and $\boldsymbol{\theta}_{i,t} = (\beta_{i,t,0}, \boldsymbol{\beta}_{i,t}')'$ with $\beta_{i,t,0} = \mu_{i,t} + \gamma_{i,t}$.

Let $\boldsymbol{Y}_{i,t}$ be a $K \times 1$ vector with entries $\widehat{Q}_{i,t}(\tau_k)$, $\bar{\boldsymbol{Z}}$ be a $K \times 4$ matrix with rows $\bar{\boldsymbol{Z}}_k'$, and $\boldsymbol{\varepsilon}_{i,t}$ be a $K \times 1$ vector with entries $\varepsilon_{i,t,k}$. Then, the ordinary least squares (OLS) estimator of $\boldsymbol{\theta}_{i,t}$ in (23) is

$$\widehat{\boldsymbol{\theta}}_{i,t} \equiv (\widehat{\beta}_{i,t,0}, \widehat{\boldsymbol{\beta}}_{i,t}')' = (\bar{\boldsymbol{Z}}'\bar{\boldsymbol{Z}})^{-1}\bar{\boldsymbol{Z}}'\boldsymbol{Y}_{i,t}, \tag{24}$$

where $\widehat{\boldsymbol{\beta}}_{i,t} = (\widehat{\beta}_{i,t,1}, \widehat{\beta}_{i,t,2}, \widehat{\beta}_{i,t,3})'$. Zhang and Zhu (2023) show that $\widehat{\boldsymbol{\theta}}_{i,t} - \boldsymbol{\theta}_{i,t} \to \mathbf{0}$ in probability as $K \to \infty$ under the following two classical conditions in the regression literature:

Condition 1. *$\bar{\boldsymbol{Z}}'\bar{\boldsymbol{Z}}$ is positive definite.*

Condition 2. *$\bar{\boldsymbol{Z}}'\boldsymbol{\varepsilon}_{i,t}/K \to \mathbf{0}$ in probability as $K \to \infty$.*

Consequently, by the continuous mapping theorem, we have

$$\widehat{h}_{i,t} \equiv \widehat{\beta}^2_{i,t,1} \to h_{i,t}, \quad \widehat{s}_{i,t} \equiv \frac{6\widehat{\beta}_{i,t,2}}{\widehat{\beta}_{i,t,1}} \to s_{i,t}, \quad \text{and} \quad \widehat{k}_{i,t} \equiv \frac{24\widehat{\beta}_{i,t,3}}{\widehat{\beta}_{i,t,1}} + 3 \to k_{i,t} \tag{25}$$

in probability as $K \to \infty$, where $\widehat{h}_{i,t}$, $\widehat{s}_{i,t}$, and $\widehat{k}_{i,t}$ are the QCMs of $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$, respectively. In order to make sure that $\widehat{h}_{i,t}$, $\widehat{s}_{i,t}$, and $\widehat{k}_{i,t}$ are moments under certain distribution of $r_{i,t}$, two necessary constraints are required:

$$\widehat{h}_{i,t} \geq 0 \text{ and } \widehat{k}_{i,t} \geq \widehat{s}^2_{i,t} + 1.$$

Clearly, the first constraint holds automatically, and the second constraint can be checked directly based on the values of $\widehat{k}_{i,t}$ and $\widehat{s}_{i,t}$. If the second constraint does not hold, we can easily replace $\widehat{\theta}_{i,t}$ in (24) by a constrained least squares estimator, so that the resulting $\widehat{k}_{i,t}$ and $\widehat{s}_{i,t}$ satisfy this constraint; see more detailed discussions in Zhang and Zhu (2023). Moreover, it should be noted that we cannot estimate $\mu_{i,t}$ by the QCM method. The reason is that $\mu_{i,t}$ cannot be estimated by $\widehat{\beta}_{i,t,0}$ in (24) due to the presence of $\gamma_{i,t}$. Therefore, we have to estimate $\mu_{i,t}$ separately by other methods (see, e.g., the graph-based method in Section 2.2 above).

As we observed, the core idea of QCM method is to transform the estimation of conditional moments to that of conditional quantiles, giving us two remarkable advantages particularly in the realm of high-dimensional data analysis.

First, the QCM method is easy-to-implement, since it only requires the estimated conditional quantiles as the input to compute the OLS estimator $\widehat{\theta}_{i,t}$. When N is large, a direct estimation for the higher-order conditional moments $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ via high-dimensional GARCH-type models is computationally infeasible. The reason is that the high-dimensional GARCH-type models are fitted by the QML estimation method, which relies on a certain distribution of $\boldsymbol{r}_t$ to write down the log-likelihood function. However, the log-likelihood function is too complex to be optimized for large N cases. For example, the often used Gaussian log-likelihood function depends on the inverse of many $N \times N$ covariance matrices, making its optimization infeasible. Transforming the estimation of higher-order moments to that of quantiles circumvents this annoying difficulty, since the estimation of quantiles is a classical supervised learning but that of higher-order moments is not. To be more specific, the supervised learning is a machine learning paradigm, and it aims to learn a function f_0 that maps features (say, x) to labels (say, y) supervised by a certain loss function without assuming the distribution of x or y. For example, the τ-th quantile of y can be learned by $f_0(x)$ using the quantile loss function $\rho_\tau(y - f_0(x))$ as the supervisor. However, it is unclear how to design an appropriate loss function as the supervisor for learning the variance, skewness, or kurtosis of y by $f_0(x)$, unless certain distributional assumption is made for x or y. This indicates that the direct estimation of $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ has to rely on a certain distribution of $r_{i,t}$, as done by the QML estimation in the high-dimensional GARCH-type models. Owing to the supervised learning feature of quantiles, our indirect estimation of $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ from the QCM method does not need any distributional assumption of $r_{i,t}$, so it bypasses the computational difficulty raised in the direct estimation method to deal with large N cases.

Second, the QCM method can largely reduce the risk of model mis-specification on conditional mean and quantiles. The behind reason is two-fold: (i) $\widehat{h}_{i,t}$, $\widehat{s}_{i,t}$, and $\widehat{k}_{i,t}$ are simultaneously computed without any prior estimation of $\mu_{i,t}$; and (ii) their consistency could hold when the specification of $Q_{i,t}(\tau)$ is mis-specified to some extent. This advantage is far beyond our expectations, since normally we have to first estimate $\mu_{i,t}$ and then $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ using some parametric models that are needed to be correctly specified to generate consistent estimators. The reason leading to this advantage is that the QCM method is regression-based, so that the impact of $\mu_{i,t}$ is eliminated by absorbing it as one part of the intercept term, and the bias from the use of wrongly specified $Q_{i,t}(\tau)$ is aggregately offset by the other part of the intercept term $\gamma_{i,t}$; see Zhang and Zhu (2023) for more discussions on this aspect.

2.4. Implementation details of the GRACE method

Due to the use of FTGCN, the GRACE method first needs to alleviate the risk of overfitting, a prevalent deficiency of the neural network. Following the standard approach to circumvent overfitting, we chronologically divide the full data into three disjoint parts: training sample, validation sample, and testing sample. The training and validation samples are taken to do parameter estimation, and the testing sample is used to evaluate the truly out-of-sample performance of the GRACE method. To be more specific, we compute $\theta^{(l)}_\tau$ at lth iteration in Algorithm 1 based on the training sample, and then calculate its corresponding validation sample error. Here, the validation sample error is the value of the objective function in (15) based on the validation sample and $\theta_\tau = \theta^{(l)}_\tau$. To regularize against overfitting, we utilize the early stopping method, which terminates the iteration process early in Algorithm 1 when the validation sample error increases for several iterations. Then, we select the estimator $\theta^{(l)}_\tau$ that has the smallest validation sample error as the quantile estimator $\widehat{\theta}_\tau$. Similarly, based on the training and validation samples, we obtain the PLS estimator $\widehat{\theta}_\mu$ under Algorithm 2.

Next, the GRACE method uses $\widehat{\theta}_{\tau_k}$ and $\widehat{\theta}_\mu$ to predict the values of $\widehat{Q}_{i,t}(\tau_k)$ and $\mu_{i,t}$ on the testing sample, respectively, where $\tau_k = k/(K+1)$, $k = 1, \ldots, K$, for simplicity. In the large pool of $\widehat{Q}_{i,t}(\tau_k)$, some values of $\widehat{Q}_{i,t}(\tau_k)$ might be invalid. Intuitively, it is reasonable to exclude those invalid $\widehat{Q}_{i,t}(\tau_k)$ for the computation of QCMs. To achieve this goal, we make use of the unconditional coverage test LR_{uc} in Kupiec (1995) and conditional coverage test LR_{cc} in Christoffersen (1998). Specifically, we compute $\widehat{Q}_{i,t}(\tau_k)$ on the training and validation samples, and apply LR_{uc} and LR_{cc} to detect whether the sequence of estimated conditional quantiles $\mathcal{Q}_i(\tau_k) \equiv \{\widehat{Q}_{i,t}(\tau_k) : t \in \text{training and validation samples}\}$ is valid at the significance level α for each stock i and quantile level τ_k. Then, we build a valid quantile level set for stock i:

$$\Omega_i = \{\tau_k : \text{ the validity of } \mathcal{Q}_i(\tau_k) \text{ is accepted by both } LR_{uc} \text{ and } LR_{cc} \text{ at the level } \alpha\}. \tag{26}$$

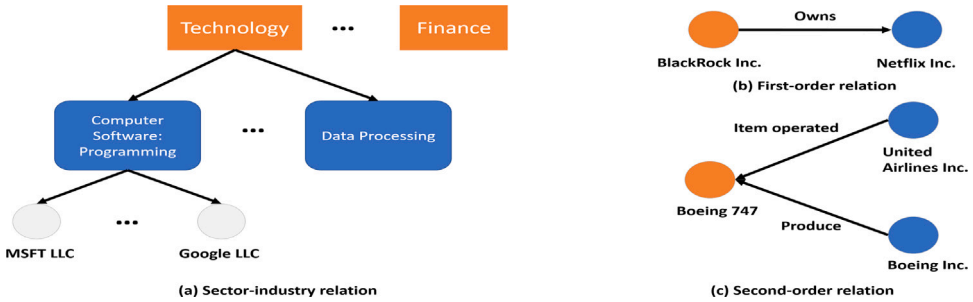

Fig. 2. Examples of the sector-industry, first-order, and second-order relations.

Table 1
Summary of S2S relations.

Market	Sector-industry relation		Wiki company-based relation	
	Relation types	Relation ratio (pairwise)	Relation types	Relation ratio (pairwise)
NASDAQ	112	5.00%	42	0.21%
NYSE	130	9.37%	32	0.30%

That is, Ω_i groups all of those quantile levels τ_k, for which the sequence $Q_i(\tau_k)$ is valid. Clearly, Ω_i depends on α and K jointly, where its size (denoted by $|\Omega_i|$) is decreasing with the value of α while increasing with the value of K. In particular, we know that $|\Omega_i| = K$ when $\alpha = 0$. After having Ω_i, we use the predicted values $\{\hat{Q}_{i,t}(\tau_k) : \tau_k \in \Omega_i\}$ to predict the values of $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ for stock i via the related QCMs on the testing sample. As K is essentially replaced by $|\Omega_i|$ from the above manipulation, we need a large value of $|\Omega_i|$ to ensure the consistency of the QCMs. This motivates us to discard those stocks having the value of $|\Omega_i|$ less than a predetermined tolerance K_0 (say, e.g., $K_0 = 30$).

Finally, the GRACE method employs different performance measures from the predicted values of $\mu_{i,t}$, $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ to construct portfolios, based on all of remaining stocks.

3. Empirical analysis

3.1. Benchmark data

We apply our GRACE method to construct portfolios based on the stocks in two major exchanges: NASDAQ and NYSE. The stock data we consider are the same as those in Feng et al. (2019), and they contain daily prices from January 2, 2013 to December 8, 2017 for 1026 and 1737 stocks in NASDAQ and NYSE, respectively. Alone with the stock price data, we also take the S2S relation data in Feng et al. (2019) to describe the multiple types of S2S relation. Based on the domain knowledge, the S2S relations can be divided into two groups: Sector-industry relations $\boldsymbol{E}^{si}_{stock}$ and Wiki company-based relations $\boldsymbol{E}^{wiki}_{stock}$ (see the Appendix A of Feng et al. (2019) for their detailed definitions). Specifically, two stocks (say, stock i and stock j) have a sector-industry relation if they belong to the same industry, where the industries are classified by the GICS standard. For example, all 1026 stocks in NASDAQ are divided into 13 different sectors, where each sector contains several industries; see Fig. 2(a) for the sector-industry hierarchy of all 1026 stocks in this market. From Fig. 2(a), we know that MSFT LLC and Google LLC have an S2S relation since they belong to the same industry "Computer Software: Programming".

Meanwhile, two stocks can also have a Wiki company-based relation if they have an either first-order or second-order relation. The first-order and second-order relations have the format of "company $i \xrightarrow{R}$ company j" and "company $i \xrightarrow{R_1}$ entity $k \xleftarrow{R_2}$ company j", respectively, where the companies i and j bridged by an entity k are corresponding to the stocks i and j, respectively, and the relations R, R_1, and R_2 are defined in Wikidata (https://www.wikidata.org/wiki/Wikidata:List_of_properties/all). It turns out that there are 5 and 53 different types of first-order and second-order relations, respectively. Fig. 2(b) and 2(c) give some illustrating examples on the first-order and second-order relations. We see from this figure that BlackRock Inc. has a first-order relation with Netflix Inc. since BlackRock Inc. owns Netflix Inc., and United Airlines Inc. and Boeing Inc. have a second-order relation since Boeing Inc. produces Boeing 747 that is sold to United Airlines Inc. In sum, Table 1 lists the number of S2S relation types and the ratio of S2S relations to all possible stock pairs in NASDAQ and NYSE. Since the ratio of S2S relations is always less than 10%, it indicates that the S2S relations in both markets are sparse.

In addition, the F2S relation data are also needed to facilitate our GRACE method. Among a great variety of factors, we use the prevalent daily Fama–French five factors (Fama and French, 2015) to specify the related F2S relations $\boldsymbol{E}^{ff5}_{factor}$, based on the asset-pricing knowledge. These five factors are excess market return, RMW, HML, SMB, and CMA, and their daily data are available at https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. The reason to choose these five factors is because they are updated frequently with incremental information for doing predictions.

Table 2
List of hyperparameters.

Hyperparameter	Description	Value
λ^*	tuning parameter in the computation of PLS estimator	0.1
γ	learning rate in the Adam algorithm	10^{-3}
S	length of lagged features in the LSTM	16
d	dimension of hidden vectors in the LSTM	64
α	significance level of LR_{uc} and LR_{cc} tests	0.01
K	number of predetermined quantile levels	199

For stock i, we let $r_{i,t}$ be its 1-day return and $r_{i,t}^{(k_*)} = \frac{1}{k_*}\sum_{s=1}^{k_*} r_{i,t+1-s}$ be its k_*-day moving average of returns at day t, and use the OLS method to calculate its 5-dimensional vector of factor exposures (denoted by $\lambda_{i,t} = (\lambda_{i,t}^{(1)}, \ldots, \lambda_{i,t}^{(5)})'$) based on the sample of 1-day returns in a half-year rolling window up to day t. For factor b, we similarly let $f_{b,t}$ be its value, $f_{b,t}^{(k_*)}$ be its k_*-day moving average, and $\bar{\lambda}_{b,t}^{(b_*)}$ be its exposure on factor b_* at day t, where we assume that factor b has one exposure on itself and zero exposure on other factors (w.r.t., $b_* \neq b$). Now, based on ten stock features $r_{i,t}^{(k_*)}$ and $\lambda_{i,t}^{(b_*)}$ together with ten factor features $f_{b,t}^{(k_*)}$ and $\bar{\lambda}_{b,t}^{(b_*)}$ for $k_* = 1, 5, 10, 20, 30$ and $b_* = 1, 2, 3, 4, 5$, our feature tensor $\boldsymbol{X}_{t-1} = [\boldsymbol{X}_{1,t-1}, \ldots, \boldsymbol{X}_{N,t-1}, \boldsymbol{X}_{N+1,t-1}, \ldots, \boldsymbol{X}_{N+B,t-1}]$ in (5) is taken as

$$\boldsymbol{X}_{i,t-1} = \begin{pmatrix} r_{i,t-S}^{(1)} & \cdots & r_{i,t-1}^{(1)} \\ \cdots & \cdots & \cdots \\ r_{i,t-S}^{(30)} & \cdots & r_{i,t-1}^{(30)} \\ \lambda_{i,t-S}^{(1)} & \cdots & \lambda_{i,t-1}^{(1)} \\ \cdots & \cdots & \cdots \\ \lambda_{i,t-S}^{(5)} & \cdots & \lambda_{i,t-1}^{(5)} \end{pmatrix} \text{ and } \boldsymbol{X}_{N+b,t-1} = \begin{pmatrix} f_{b,t-S}^{(1)} & \cdots & f_{b,t-1}^{(1)} \\ \cdots & \cdots & \cdots \\ f_{b,t-S}^{(30)} & \cdots & f_{b,t-1}^{(30)} \\ \bar{\lambda}_{b,t-S}^{(1)} & \cdots & \bar{\lambda}_{b,t-1}^{(1)} \\ \cdots & \cdots & \cdots \\ \bar{\lambda}_{b,t-S}^{(5)} & \cdots & \bar{\lambda}_{b,t-1}^{(5)} \end{pmatrix}, \tag{27}$$

for $i = 1, \ldots, N$ and $b = 1, \ldots, B$, where $N = 1026$ (or 1737) for the NASDAQ (or the NYSE) market, $B = 5$, $\boldsymbol{X}_{i,t-1} \in \mathcal{R}^{P\times S}$, and $\boldsymbol{X}_{N+b,t-1} \in \mathcal{R}^{P\times S}$ with $P = 10$. It is worthy noting that each entry in $\boldsymbol{X}_{i,t-1}$ and $\boldsymbol{X}_{N+b,t-1}$ is normalized by its range in the training sample to reduce its skewness and leptokurtosis; see the similar implementation in Feng et al. (2019).

With the full data sample $\{\boldsymbol{r}_t, \boldsymbol{X}_{t-1}\}$ in hand, we divide it into three disjoint parts in the same way as Feng et al. (2019): The training sample from January 2, 2013 to December 31, 2015 has 756 trading days, the validation sample follows and ends on December 30, 2016 with 252 trading days, and the testing sample covers the remaining 237 trading days from January 3, 2017 to December 8, 2017 (i.e., the out-of-sample period). Now, based on the values of hyperparameters listed in Table 2, our GRACE method is implemented as the details specified in Section 2.4 above. Here, we follow the tuning results in Feng et al. (2019) to select λ^*, γ, S, and d, and we examine the effects of K and α in the supplementary materials.

3.2. Comparison methods

Besides our GRACE method, other graph-based methods can also be adopted to select portfolios using the same idea, except for different models to predict the conditional quantiles and mean of $r_{i,t}$. Below, we introduce two alternative graph-based methods for the purpose of comparison.

The first competitor is labeled as GRACE_1, which replaces the factor-augmented hypergraph in the GRACE method with the hypergraph in Feng et al. (2019), and leaves other mechanisms (including the input features and the selection of tuning hyperparameters) unchanged. The comparison between the GRACE and GRACE_1 methods is to verify whether incorporating the asset pricing knowledge into the hypergraph is informative for portfolio selection.

The second competitor is the simple GRACE (denoted by GRACE_2) method, which is motivated by the network autoregression model in Zhu et al. (2017) and the FDNQR model in Xu et al. (2024). Specifically, the GRACE_2 method predicts the conditional quantiles of $r_{i,t}$ based on the following specification:

$$Q_{i,t}(\tau) = \alpha_\tau + \gamma_\tau \sum_{j=1}^{N} w_{i,j} r_{j,t-1} + \boldsymbol{\zeta}_\tau' \boldsymbol{x}_{i,t-1} + \boldsymbol{\varsigma}_{1,\tau}' \boldsymbol{F}_{t-1}, \tag{28}$$

where $\boldsymbol{x}_{i,t-1}$ is the last column of $\boldsymbol{X}_{i,t-1}$ in (27), $\boldsymbol{F}_{t-1}$ is the 5-dimensional vector containing the values of Fama–French five factors at $t-1$, and other notations are inherited from (17). By construction, model (28) uses the term $\boldsymbol{\zeta}_\tau' \boldsymbol{x}_{i,t-1}$ (replacing the term $\boldsymbol{\beta}_\tau \boldsymbol{z}_i + \zeta r_{i,t-1}$ in (17)) to account for the lag-1 stock features, and it takes the term $\boldsymbol{\varsigma}_{1,\tau}' \boldsymbol{F}_{t-1}$ to include the lag-1 factor features. Similarly, the GRACE_2 method predicts the conditional mean of $r_{i,t}$ by the following factor-augmented network autoregressive specification:

$$r_{i,t} = \alpha_\mu + \gamma_\mu \sum_{j=1}^{N} w_{i,j} r_{j,t-1} + \boldsymbol{\zeta}_\mu' \boldsymbol{x}_{i,t-1} + \boldsymbol{\varsigma}_{1,\mu}' \boldsymbol{F}_{t-1} + \varepsilon_{i,t}^*, \tag{29}$$

where $\varepsilon_{i,t}^*$ is the error term with zero mean, and α_μ, γ_μ, $\boldsymbol{\zeta}_\mu$, and $\boldsymbol{\varsigma}_{1,\mu}$ are unknown regression coefficients. Here, $w_{i,j} = a_{i,j}/n_i$ in models (28) and (29) is determined by $\boldsymbol{E}_{stock}$ in the GRACE method, such that $a_{i,j} = 1$ if there is any S2S relation between stock i

and stock j, and $a_{i,j} = 0$ otherwise. We estimate models (28) and (29) respectively via the quantile loss and L_2 loss functions using the data from the combination of training and validation samples, and then proceed the portfolio selection on the testing sample. One may extend both models to contain the stock and factor features up to lag-S. However, our unreported analysis shows that this extension makes model estimation less stable, leading to the worse performance in portfolio selection. Clearly, the comparison between the GRACE and GRACE$_2$ methods aims to check whether those simple settings in models (28) and (29) are adequate enough for portfolio selection.

As usual, we are also of interest in comparing our GRACE method with benchmark GARCH-type methods. To ensure computational feasibility for $N > 1000$, we consider two GARCH-type methods, namely GARCH$_1$ and GARCH$_2$, to predict the first four conditional moments of each $r_{i,t}$ for proceeding the portfolios. Specifically, the GARCH$_1$ and GARCH$_2$ methods do the prediction for each $r_{i,t}$ by using the first-order univariate AR-GARCH model with skewed t distribution in Hansen (1994) and semi-nonparametric AR-GJR model in León and Ñíguez (2020), respectively. These two GARCH-type methods acting as the third and forth competitors of our GRACE method, ignore the inter-dependencies among stocks and learn conditional moments stock by stock.

We shall mention that another standard way to do the sorting for portfolio selection is based on some asset characteristics, such as size, value, and momentum. In the supplementary materials, we find that our GRACE method outperforms those characteristics-based methods by a wide margin. For saving space, the related results are not reported below.

3.3. Economic performance evaluation

This subsection evaluates the out-of-sample performance of the long-short portfolios selected by the GRACE method. After sorting all stocks via a certain performance measure into 10 deciles, the long-short portfolio is re-balanced on every trading day via buying the 10% highest ranking stocks (decile 10) and selling the 10% lowest ranking stocks (decile 1) with equal weights. The performance measures to sort all stocks include the M, MV, MVSK, SR, and SRSK, where the performance measure M is only related to $\mu_{i,t}$, and the definitions of the last four performance measures are given in (1)–(4). The values of all five performance measures are computed based on the predicted values of $\mu_{i,t}$, $h_{i,t}$, $s_{i,t}$, and $k_{i,t}$ from the GRACE method with $\boldsymbol{E}_{stock} = \boldsymbol{E}^{wiki}_{stock}$, $\boldsymbol{E}_{factor} = \boldsymbol{E}^{ff5}_{factor}$, and hyperparameters taken as in Table 2. For the performance measures MV, MVSK, and SRSK, the hyperparameters λ_1, λ_2, and λ_3 are tuned by the grid search within the sets A_1, A_2, and A_3, respectively, to maximize the value of SR of each long-short portfolio re-balanced on training and validation samples. Here, due to different ranges of $h_{i,t}$ and $s_{i,t}$ (or $k_{i,t}$), we take

$$A_1 = \{a \times 10^{-b} : a = 1, 2, \dots, 9 \text{ and } b = -1, 0, \dots, 3\},$$
$$A_2 = A_3 = \{a \times 10^{-b} : a = 1, 2, \dots, 9 \text{ and } b = 2, 3, \dots, 6\}.$$

As a comparison, the GRACE$_1$ and GRACE$_2$ methods are also used to select the long-short portfolios under the similar procedure as above, where the GRACE$_1$ method with the performance measure M is the benchmark method proposed by Feng et al. (2019). Since the QCM method relies on the choices of α and K, the stock pool for portfolio selection under the performance measures MV, MVSK, SR, and SRSK varies with the choice of α or K. For the sake of consistency, the portfolio selection under the performance measure M will use the same stock pool as for other performance measures in the sequel, although the implementation of performance measure M is independent of higher-order conditional moments and not affected by the choice of α or K technically.

From an economic viewpoint, we compare all of the selected portfolios in terms of their annualized SR, which is the ratio of annualized excess return to annualized risk. To compute the annualized excess return, we use the Treasury bill rate as a proxy for the risk-free return, and consider the transaction cost of 3‰ (i.e., 30 basis points) for buying and selling as done in Engle et al. (2012). Note that when there is no ambiguity, we omit the wording "annualized" below for ease of presentation.

3.3.1. Impacts of method and performance measure

We first assess how the graph-based methods and performance measures affect the out-of-sample performance of portfolios. Table 3 reports the values of (excess) return, risk, and SR of the out-of-sample portfolios selected by five different methods with five different performance measures in NASDAQ and NYSE, as well as market portfolios being comprised of only index returns.[2] From this table, we can have the following findings:

(i) For the GRACE method, the SRSK and M portfolios have the largest and smallest values of SR, respectively, in both markets, implying the necessity of using three higher-order conditional moments for portfolio selection. Particularly, the advantage of using higher-order conditional moments is more evident in NASDAQ than NYSE by observing that the value of SR for the SRSK portfolio is 61% (or 12%) higher than that for the M portfolio in NASDAQ (or NYSE). Moreover, the values of return and risk indicate that the SRSK portfolio has a larger value of SR mainly because it can generate a much larger (or smaller) value of return (or risk) than the M portfolio in NASDAQ (or NYSE). Another difference between NASDAQ and NYSE is the influence of conditional skewness and kurtosis. Specifically, using the conditional skewness and kurtosis can well decrease the portfolio risk in NASDAQ while only marginally increase the portfolio return in NYSE, according to the comparison between MV and

[2] Due to the extensive parameters in the FTGCN and TGCN, the GRACE and GRACE$_1$ methods require approximately 370 and 360 hours to reproduce the empirical results in Table 3, using a *single NVIDIA 4090 GPU*. In contrast, the GRACE$_2$, GARCH$_1$, and GARCH$_2$ methods require less than 6 hours to reproduce the results in Table 3, using *two Intel(R) Xeon(R) Gold 6240 CPUs*. The different computational complexities above are rational, since the deep learning methods usually impose significantly heavier computational burdens as a tradeoff for higher prediction accuracy.

Table 3
Out-of-sample performances of long-short portfolios across different methods and performance measures.

Method	Measure	NASDAQ			NYSE		
		Return (%)	Risk (%)	SR	Return (%)	Risk (%)	SR
GRACE	M	25.40	8.53	2.98	29.04	8.74	3.10
	MV	38.81	10.43	3.72	29.36	8.62	3.17
	MVSK	37.04	8.43	4.39	30.01	8.62	3.25
	SR	37.48	9.03	4.15	13.90	3.43	3.47
	SRSK	39.21	8.15	4.81	13.95	3.43	3.48
$GRACE_1$	M	12.08	8.45	1.43	25.74	8.97	2.87
	MV	13.55	7.83	1.73	25.82	8.97	2.88
	MVSK	33.45	8.24	4.06	25.82	8.97	2.88
	SR	26.57	6.48	4.10	16.91	5.10	3.32
	SRSK	42.09	8.86	4.75	16.97	5.10	3.33
$GRACE_2$	M	0.06	8.59	0.01	−2.09	8.94	−0.46
	MV	4.55	10.57	0.43	0.45	8.80	−0.18
	MVSK	6.66	8.50	0.78	2.26	8.79	0.03
	SR	4.53	9.11	0.50	3.00	3.51	0.28
	SRSK	5.90	8.19	0.72	3.56	3.51	0.45
$GARCH_1$	M	−1.02	7.38	−0.14	−4.03	9.00	−0.45
	MV	0.76	8.85	0.09	−3.87	8.97	−0.43
	MVSK	4.38	8.44	0.52	−2.14	8.97	−0.23
	SR	5.32	10.22	0.52	−3.06	7.34	−0.41
	SRSK	6.03	10.46	0.57	−2.77	7.22	−0.38
$GARCH_2$	M	−2.04	8.43	−0.24	−11.20	8.97	−1.25
	MV	0.27	8.44	0.03	−9.87	8.97	−1.10
	MVSK	2.63	8.44	0.31	−8.88	8.97	−0.99
	SR	1.46	6.46	0.23	−6.26	5.11	−1.22
	SRSK	3.99	6.46	0.62	−5.70	5.11	−1.12
Market		24.85	9.69	2.56	12.51	6.74	1.85

MVSK (or SR and SRSK) portfolios. The aforementioned distinction in two markets may attribute to the fact that the NYSE has a relatively more "normal" environment than the NASDAQ, so that the function of higher-order conditional moments (particularly the conditional skewness and kurtosis) is relatively weaker for portfolio selection.

(ii) For the $GRACE_1$ and $GRACE_2$ methods, both of them perform worse than the GRACE method, regardless of the choice of performance measure. The advantage of GRACE method over $GRACE_1$ method is exceptionally significant for M and MV portfolios in NASDAQ, since the M and MV portfolios selected by GRACE method have 108% and 115% higher value of SR than those selected by $GRACE_1$ method, respectively. This finding shows that incorporating asset pricing information is more important for portfolio selection in NASDAQ than NYSE, especially when the conditional skewness and kurtosis are not taken into account. Moreover, the value of SR for the best portfolio selected by the GRACE method is 236% and 21% higher than that selected by the benchmark method (i.e., the $GRACE_1$ method with the performance measure M) in NASDAQ and NYSE, respectively. In all cases, the $GRACE_2$ method has a much worse performance than other two methods. This is not unexpected, because the simple model settings in the $GRACE_2$ method cannot capture the effects of features on stock returns adequately.

(iii) For the $GARCH_1$ and $GARCH_2$ methods, they have a much worse performance than the GRACE method, and their selected portfolios have values of SR less than 0.63 in NASDAQ while even less than zero in NYSE. The inferior performance of $GARCH_1$ and $GARCH_2$ methods is probably because they overlook the spatial structure of stocks and fail to capture the temporal dynamics of stock returns adequately.

(iv) For the market portfolios, they earn SR values of 2.67 and 1.85 in NASDAQ and NYSE, respectively. Notably, only five portfolios from the GRACE method and three portfolios (with respect to MVSK, SR, and SRSK) from the $GRACE_1$ method outperform the market portfolio in NASDAQ, and only all of portfolios from the GRACE and $GRACE_1$ methods perform better than the market portfolio in NSYE. These results further emphasize that when considering a transaction cost of 30 basis points, our GRACE method has the superiority over the other methods by beating the market largely.

Overall, the above findings clearly demonstrate the importance of higher-order conditional moments and asset pricing knowledge in portfolio selection through the GRACE method.

According to some additional results in the supplementary materials, we find that the advantage of our GRACE method remains across different choices of α and K and different levels of transaction cost. From a statistical viewpoint, the exceptional performance of GRACE method is probably due to its capacity to well estimate and predict the conditional moments of $r_{i,t}$, as demonstrated by our statistical performance evaluation in the supplementary materials.

3.3.2. Impact of $\boldsymbol{E}_{stock}$

Since all considered three graph-based methods depend on the S2S relation set $\boldsymbol{E}_{stock}$, a natural question is: What kind of S2S relation set is more informative for portfolio selection? To answer this question, we alter all three methods by choosing $\boldsymbol{E}_{stock} = \boldsymbol{E}^{si}_{stock}$

Table 4
Out-of-sample SRs of long-short portfolios across different choices of $\boldsymbol{E}_{stock}$.

Method	Measure	NASDAQ			NYSE		
		$\boldsymbol{E}^{si}_{stock}$	$\boldsymbol{E}^{wiki}_{stock}$	$\boldsymbol{E}^{all}_{stock}$	$\boldsymbol{E}^{si}_{stock}$	$\boldsymbol{E}^{wiki}_{stock}$	$\boldsymbol{E}^{all}_{stock}$
GRACE	M	2.74	2.98	2.98	3.04	3.10	3.12
	MV	2.86	3.72	3.74	3.04	3.17	3.17
	MVSK	3.98	4.39	4.40	3.04	3.25	3.25
	SR	3.81	4.15	4.18	3.04	3.47	3.47
	SRSK	4.52	4.81	4.81	3.13	3.48	3.48
$GRACE_1$	M	1.58	1.43	1.52	2.76	2.88	2.87
	MV	1.60	1.73	1.55	2.77	2.88	2.87
	MVSK	2.94	4.06	4.02	2.77	2.88	2.88
	SR	2.14	4.10	4.01	3.19	3.32	3.32
	SRSK	3.19	4.75	4.65	3.19	3.32	3.32
$GRACE_2$	M	−0.17	0.01	−1.24	−1.88	−0.46	−2.81
	MV	0.12	0.43	−1.02	−1.77	−0.18	−2.80
	MVSK	0.66	0.78	−0.55	−1.71	0.03	−2.78
	SR	0.43	0.50	−1.34	−1.56	0.28	−2.63
	SRSK	0.61	0.72	−1.29	−1.33	0.45	−2.58

or $\boldsymbol{E}^{all}_{stock}$ while keeping other settings as for Table 3 unchanged, where $\boldsymbol{E}^{all}_{stock}$ is the union set of $\boldsymbol{E}^{wiki}_{stock}$ and $\boldsymbol{E}^{si}_{stock}$. Table 4 reports the values of SR for portfolios selected from three different choices of $\boldsymbol{E}_{stock}$. From Table 4, we find that the value of SR for the $\boldsymbol{E}^{si}_{stock}$-based portfolio is smaller than that for the $\boldsymbol{E}^{wiki}_{stock}$-based portfolio in all cases, except for the M portfolios selected by the $GRACE_1$ method in NASDAQ. Particularly, the advantage of $\boldsymbol{E}^{wiki}_{stock}$-based portfolio over $\boldsymbol{E}^{si}_{stock}$-based portfolio is much more substantial for the $GRACE_1$ method with the performance measures MVSK, SR, and SRSK in NASDAQ. This finding indicates that the S2S relations in $\boldsymbol{E}^{si}_{stock}$ could be less informative than those in $\boldsymbol{E}^{wiki}_{stock}$ to learn higher-order conditional moments, especially when the asset pricing knowledge is absent.

Moreover, we find from Table 4 that using a richer S2S relation set $\boldsymbol{E}^{all}_{stock}$ to replace the single S2S relation set $\boldsymbol{E}^{wiki}_{stock}$ gives no change or little change to the values of SR in the GRACE and $GRACE_1$ methods, and this replacement even makes the portfolios have smaller values of SR for many cases in the $GRACE_1$ method. The reason is probably that the long-term correlations between stocks are largely driven by the factors through the F2S relations in $\boldsymbol{E}^{ff5}_{factor}$, and they could be wrongly captured by the S2S relations in $\boldsymbol{E}^{si}_{stock}$ when the asset pricing knowledge is absent. For example, the stocks belonging to the sector "Basic Industries" tend to have large market capitalization, while the SMB factor in Fama and French (2015) represents the outperformance of small-cap stocks over large-cap ones during a long-term. Hence, the comovement of stocks in the sector "Basic Industries" is more properly captured by the F2S relations with respect to the SMB factor in $\boldsymbol{E}^{ff5}_{factor}$ rather than the sector-industry relations in $\boldsymbol{E}^{si}_{stock}$. The unsatisfactory performance from the use of $\boldsymbol{E}^{all}_{stock}$ becomes more evident in the $GRACE_2$ method. This conveys the information that it is inappropriate to ignore the type of S2S relation as done by models (28)–(29), when the domain knowledge on multiple types of S2S relation is available.

In sum, we could reach a general conclusion that $\boldsymbol{E}^{si}_{stock}$ is less informative than $\boldsymbol{E}^{wiki}_{stock}$ for portfolio selection. Hence, if $\boldsymbol{E}^{wiki}_{stock}$ is accessible, we recommend it for practical use.[3]

3.4. Empirical results on expanded dataset

In this subsection, for the sake of robustness, we examine the performance of our GRACE method on an expanded dataset that contains daily stock data from January 2, 2013 to December 30, 2022. Although the relationships among stocks are likely to remain unchanged over a long period, using up-to-date relation data is expected to yield improved performance for our GRACE method. However, acquiring such data is challenging for us at this stage. Hence, as a compromise, we continue to employ the relation data provided by Feng et al. (2019) for the expanded dataset beyond December 8, 2017.

As for Table 3, we apply a moving-window analysis procedure to all considered methods using the expanded dataset. To be specific, we employ a 5-year moving window and divide the data within this window period into three disjoint samples: training, validation, and testing samples, which contain the data in the first 4 years, middle 1 year, and final 1 year, respectively. Table 5 reports the out-of-sample SRs of portfolios selected by all considered methods for each year from 2017 to 2022. From this table, we can obtain the similar findings as in Table 3, except that the performance of each portfolio selected by GRACE or $GRACE_1$ method generally has a decline trend in the value of SR overtime. The reason for this decline is probably that we do not use the up-to-date relation data. It is worth noting that the portfolio selected by the benchmark in Feng et al. (2019) is even worse than the market portfolio during years 2017–2022 (or year 2022) in NASDAQ (or NYSE), whereas our MVSK and SRSK portfolios selected by the GRACE method can perform much better than the market portfolios in both markets during years 2017–2022. These findings indicate the necessity of performance measures using higher-order moments in portfolio selection.

[3] In the supplementary materials, we obtain the top five influential relations across the GRACE, $GRACE_1$, and $GRACE_2$ methods, showing the robustness of the GRACE method to the choice of graph structure.

Table 5
Out-of-sample SRs of long-short portfolios during years 2017–2022.

Method	Measure	NASDAQ					NYSE				
		2017	2018	2019	2021	2022	2017	2018	2019	2021	2022
GRACE	M	2.98	0.43	2.23	1.11	1.04	3.10	0.08	1.92	0.31	0.99
	MV	3.72	1.72	2.56	1.18	1.06	2.17	0.68	2.03	0.43	1.25
	MVSK	4.39	2.35	2.90	1.46	1.42	3.25	1.13	2.55	0.91	1.68
	SR	4.15	2.02	2.84	1.70	1.32	3.47	0.37	2.64	0.75	1.30
	SRSK	4.81	2.49	3.04	1.75	1.46	3.48	1.17	2.96	1.02	1.68
$GRACE_1$	M	1.43	−0.46	1.42	0.57	0.70	2.87	−0.35	1.82	0.73	1.06
	MV	1.73	0.08	1.51	0.85	0.74	2.88	0.16	2.38	0.69	1.21
	MVSK	4.06	0.29	2.63	1.28	1.37	2.88	0.69	2.44	0.66	1.27
	SR	4.10	1.73	3.11	1.31	1.34	3.32	0.36	2.46	0.66	1.45
	SRSK	4.75	2.12	3.52	1.32	1.44	3.33	0.92	2.82	0.98	1.62
$GRACE_2$	M	0.01	−0.77	0.18	0.48	−0.17	−0.46	−2.42	0.01	−0.83	0.18
	MV	0.43	−0.74	0.65	0.34	0.45	−0.18	−2.02	0.54	−0.58	−0.08
	MVSK	0.78	−0.62	1.10	0.32	0.68	0.03	−2.07	0.89	−0.63	0.44
	SR	0.50	−0.36	0.44	0.19	0.53	0.28	−1.79	0.77	−0.48	0.15
	SRSK	0.72	−0.31	0.50	0.63	0.19	0.45	−1.80	1.07	−0.29	0.29
$GARCH_1$	M	−0.14	−0.82	−0.12	−0.31	−0.16	−0.45	−1.46	0.10	−0.67	0.13
	MV	0.09	−0.69	0.73	0.18	0.30	−0.43	−1.87	0.41	−0.45	−0.30
	MVSK	0.52	−0.57	0.31	0.56	0.65	−0.23	−1.58	0.77	−0.61	0.24
	SR	0.52	−0.28	0.61	0.06	−0.03	−0.41	−1.73	0.64	−0.65	−0.35
	SRSK	0.57	−0.29	0.51	0.28	0.37	−0.38	−1.65	0.63	−0.82	−0.35
$GARCH_2$	M	−0.24	−0.75	0.52	−0.35	0.11	−1.25	−2.29	0.18	−1.25	−0.39
	MV	0.03	−0.61	0.26	0.13	−0.26	−1.10	−2.15	0.28	−1.05	−0.20
	MVSK	0.31	−0.44	0.48	0.43	−0.07	−0.99	−2.16	0.24	−0.63	0.07
	SR	0.23	−0.49	0.93	0.51	0.12	−1.22	−2.40	−0.22	−1.36	0.04
	SRSK	0.62	−0.59	0.58	0.11	−0.02	−1.12	−1.90	−0.17	−1.24	0.19
Market		2.56	−0.14	1.93	1.16	1.10	1.85	−0.81	1.77	0.27	1.27

4. Concluding remarks

This paper proposes a new GRACE method for big portfolio selection under different performance measures that are defined by four conditional moments of stock returns. The GRACE method builds on the FTGCN and the QCM method: The former embeds the factor-augmented hypergraph within a graph neural network to obtain the estimates of mean and quantiles, and the latter transforms the estimates of quantiles into those of higher-order moments. The most attractive feature of the GRACE method is its capacity to estimate conditional variance, skewness, and kurtosis of high-dimensional stock returns, so the big portfolios under four performance measures MV, MVSK, SR, and SRSK can be constructed from thousands of stocks or even more. In the empirical studies on NASDAQ and NYSE, we find that regardless of the performance measure, the GRACE method can construct portfolios with more stable and larger values of out-of-sample SR than the competing methods, across different settings of hyperparameter and transaction cost.

In future, our GRACE method can be extended in several directions. First, it is interesting to design some new hypergraphs to incorporate subjective information from the experiences and beliefs of investors or unstructured data information from financial news and social media contents (Ke et al., 2019; Zhou et al., 2024). Second, it is intriguing to examine whether other observed factors or data-driven factors (Giglio et al., 2022) are useful for specifying the F2S relations. Third, it is worthwhile to propose a similar QCM method to learn the covariance of stock returns, so that the portfolios can be formed with the optimal weight in the mean–variance criterion (Markowitz, 1952).

CRediT authorship contribution statement

Zhoufan Zhu: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Ningning Zhang:** Methodology, Formal analysis, Data curation, Conceptualization. **Ke Zhu:** Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jempfin.2024.101533.

References

Ando, T., Bai, J., 2020. Quantile co-movement in financial markets: A panel quantile model with unobserved heterogeneity. J. Amer. Statist. Assoc. 115 (529), 266–279.
Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econometrics 31 (3), 307–327.

Burt, A., Hrdlicka, C., 2021. Where does the predictability from sorting on returns of economically linked firms come from? J. Financ. Quant. Anal. 56 (8), 2634–2658.
Chen, L., Dolado, J.J., Gonzalo, J., 2021. Quantile factor models. Econometrica 89 (2), 875–910.
Christoffersen, P.F., 1998. Evaluating interval forecasts. Internat. Econom. Rev. 39 (4), 841–862.
Cohen, L., Frazzini, A., 2008. Economic links and predictable returns. J. Finance 63 (4), 1977–2011.
Cornish, E.A., Fisher, R.A., 1938. Moments and cumulants in the specification of distributions. Revue de l'Institut international de Statistique 5 (4), 307–320.
Dittmar, R.F., 2002. Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns. J. Finance 57 (1), 369–403.
Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50 (4), 987–1007.
Engle, R.F., Ferstenberg, R., Russell, J., 2012. Measuring and modeling execution cost and risk. J. Portf. Manag. 38 (2), 14–28.
Engle, R.F., Ledoit, O., Wolf, M., 2019. Large dynamic covariance matrices. J. Bus. Econom. Statist. 37 (2), 363–375.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33 (1), 3–56.
Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. J. Financ. Econ. 116 (1), 1–22.
Fama, E.F., French, K.R., 2018. Choosing factors. J. Financ. Econ. 128 (2), 234–252.
Feng, F., He, X., Wang, X., Luo, C., Liu, Y., Chua, T.-S., 2019. Temporal relational ranking for stock prediction. ACM Trans. Inf. Syst. (TOIS) 37, 1–30.
Galvao, A.F., Kato, K., 2016. Smoothed quantile regression for panel data. J. Econometrics 193 (1), 92–112.
Ghosh, P., Neufeld, A., Sahoo, J.K., 2022. Forecasting directional movements of stock prices for intraday trading using LSTM and random forests. Finance Res. Lett. 46 (Part A), 102280.
Giglio, S., Kelly, B., Xiu, D., 2022. Factor models, machine learning, and asset pricing. Annu. Rev. Finan. Econ. 14 (1), 337–368.
Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Rev. Financ. Stud. 33 (5), 2223–2273.
Gu, S., Kelly, B., Xiu, D., 2021. Autoencoder asset pricing models. J. Econometrics 222 (1), 429–450.
Hansen, B.E., 1994. Autoregressive conditional density estimation. Internat. Econom. Rev. 35 (3), 705–730.
Härdle, W.K., Wang, W., Yu, L., 2016. TENET: Tail-event driven network risk. J. Econometrics 192 (2), 499–513.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Hou, K., Karolyi, G.A., Kho, B.-C., 2011. What factors drive global stock returns? Rev. Financ. Stud. 24 (8), 2527–2574.
Jondeau, E., Rockinger, M., 2003. Conditional volatility, skewness, and kurtosis: existence, persistence, and comovements. J. Econom. Dynam. Control 27 (10), 1699–1737.
Kato, K., Galvao, A.F., Montes-Rojas, G.V., 2012. Asymptotics for panel quantile regression models with individual effects. J. Econometrics 170 (1), 76–91.
Ke, Z.T., Kelly, B.T., Xiu, D., 2019. Predicting returns with text data. Working paper.
Kim, H.Y., Won, C.H., 2018. Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. Expert Syst. Appl. 103, 25–37.
Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: International Conference on Learning Representations.
Koenker, R., 2004. Quantile regression for longitudinal data. J. Multivariate Anal. 91 (1), 74–89.
Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46 (1), 33–50.
Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. J. Deriv. 3 (2), 73–84.
Lee, C.M., Sun, S.T., Wang, R., Zhang, R., 2019. Technological links and predictable returns. J. Financ. Econ. 132 (3), 76–96.
León, Á., Ñíguez, T.M., 2020. Modeling asset returns under time-varying semi-nonparametric distributions. J. Bank. Financ. 118, 105870.
León, Á., Rubio, G., Serna, G., 2005. Autoregresive conditional volatility, skewness and kurtosis. Q. Rev. Econ. Finance 45 (4–5), 599–618.
Lettau, M., Pelger, M., 2020. Estimating latent asset-pricing factors. J. Econometrics 218 (1), 1–31.
Livingston, M., 1977. Industry movements of common stocks. J. Finance 32 (3), 861–874.
Ma, S., Linton, O., Gao, J., 2021. Estimation and inference in semiparametric quantile factor models. J. Econometrics 222 (1), 295–323.
Markowitz, H.M., 1952. Portfolio selection. J. Finance 7 (1), 77–91.
Pakel, C., Shephard, N., Sheppard, K., Engle, R.F., 2021. Fitting vast dimensional time-varying covariance models. J. Bus. Econom. Statist. 39 (3), 652–668.
Scott, R.C., Horvath, P.A., 1980. On the direction of preference for moments of higher order than the variance. J. Finance 35 (4), 915–919.
Sharpe, W.F., 1994. The sharpe ratio. J. Portf. Manag. 21 (1), 49–58.
Socher, R., Chen, D., Manning, C.D., Ng, A., 2013. Reasoning with neural tensor networks for knowledge base completion. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems. Vol. 26, Curran Associates, Inc.
Xu, X., Wang, W., Shin, Y., Zheng, C., 2024. Dynamic network quantile regression model. J. Bus. Econom. Statist. 42 (2), 407–421.
Yang, X., Zhu, Z., Li, D., Zhu, K., 2024. Asset pricing via the conditional quantile variational autoencoder. J. Bus. Econom. Statist. 42 (2), 681–694.
Zhang, N., Zhu, K., 2023. Quantiled conditional variance, skewness and kurtosis by cornish-Fisher expansion. Working paper.
Zheng, Z., Chen, K., Sun, G., Zha, H., 2007. A regression framework for learning ranking functions using relative relevance judgments. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 287–294.
Zhou, Y., Fan, J., Xue, L., 2024. How much can machines learn finance from Chinese text data? Manag. Sci. (in press).
Zhu, X., Pan, R., Li, G., Liu, Y., Wang, H., 2017. Network vector autoregression. Ann. Statist. 45 (3), 1096–1123.
Zhu, X., Wang, W., Wang, H., Härdle, W.K., 2019. Network quantile autoregression. J. Econometrics 212 (1), 345–358.