Journal of Econometrics 239 (2024) 105420

Contents lists available at ScienceDirect

Journal of Econometrics

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

Stock co-jump networks

Yi Ding [a], Yingying Li [b,*], Guoli Liu [c], Xinghua Zheng [c]

[a] Faculty of Business Administration, University of Macau, Taipa, Macau
[b] Department of ISOM and Department of Finance, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[c] Department of ISOM, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 30 September 2021
Received in revised form 16 December 2022
Accepted 13 January 2023
Available online 22 March 2023

JEL classification:
C14
C38
C58
G17

Keywords:
Network
Community detection
Jumps
Co-jumps
Stock dependence
High-frequency data

ABSTRACT

We propose a Degree-Corrected Block Model with Dependent Multivariate Poisson edges (DCBM-DMP) to study stock co-jump dependence. To estimate the community structure, we extend the SCORE algorithm in Jin (2015) and develop a Spectral Clustering On Ratios-of-Eigenvectors for networks with Dependent Multivariate Poisson edges (SCORE-DMP) algorithm. We prove that SCORE-DMP enjoys strong consistency in community detection. Empirically, using high-frequency data of S&P 500 constituents, we construct two co-jump networks according to whether the market jumps and find that they exhibit different community features than GICS. We further show that the co-jump networks help in stock return prediction.

1. Introduction

1.1. Stock price jumps and co-jumps

Jumps, the sharp discontinuities in asset prices, have drawn considerable attention in recent years. They carry important information especially for financial crises. Bates (1991) shows that the jump component in an option pricing model can predict the stock market crash in 1987. Eraker (2004) finds that jump rates increased during the 1987 crash and the tech bubble burst in 2001–2002. Kou et al. (2017) further distinguish between positive and negative jumps and show that negative jump rates increased significantly during the 2008–2009 financial crisis, while the jump sizes changed little on average. Price jumps play important roles in volatility forecasting and asset pricing. Andersen et al. (2007) find that separating price jumps from continuous moves improves volatility forecasting. Bollerslev and Todorov (2011) show that risk compensation for jumps accounts for a large fraction of risk premia. Jumps are also used in modeling stock returns; see, for example, Cremers et al. (2015) and Bollerslev et al. (2015).

With the increasing availability of high-frequency data, important progress has been made in financial econometrics to understand jumps better. For example, Mancini (2009) proposes a threshold method for jump detection. Lee and Mykland

* Corresponding author.
E-mail addresses: yiding@um.edu.mo (Y. Ding), yyli@ust.hk (Y. Li), gliuaj@connect.ust.hk (G. Liu), xhzheng@ust.hk (X. Zheng).

https://doi.org/10.1016/j.jeconom.2023.01.026
0304-4076/

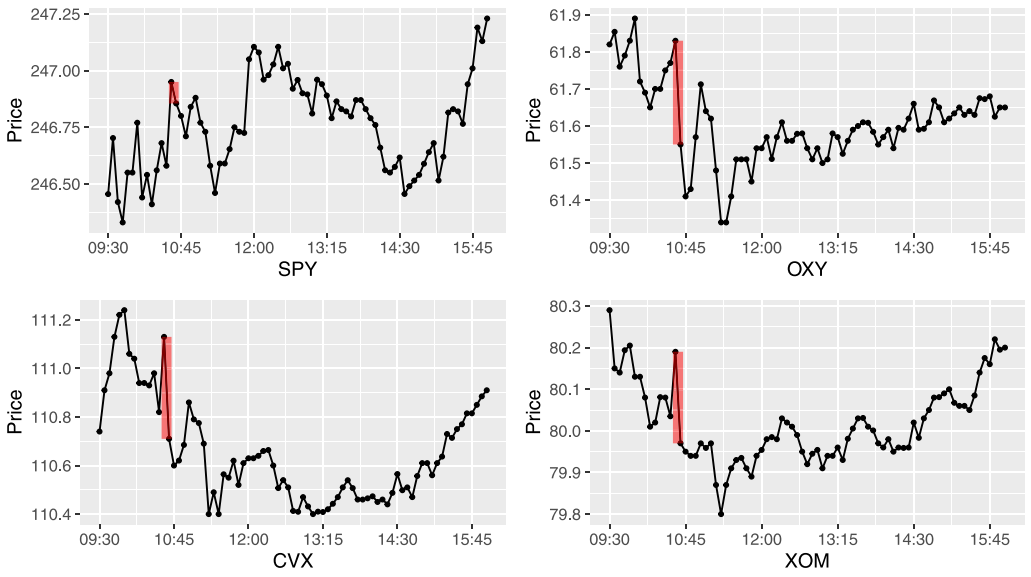

Fig. 1. Prices of SPDR S&P 500 ETF trust (SPY), Chevron (CVX), Occidental Petroleum (OXY) and Exxon Mobil (XOM) on August 9, 2017, sampled at 5-min frequency from 9:30 to 16:00.

(2008) develop a test to detect jump arrival times. Aït-Sahalia and Jacod (2009) establish a comprehensive theory about jump testing and propose a new test for the presence of jumps. Jump tests at the presence of microstructure noise are studied in Aït-Sahalia et al. (2012), Lee and Mykland (2012) and Li (2013).

There exists a growing literature on price jump dependence between stocks and the market. Co-jumps between stocks and the market have been documented by, for example, Bollerslev et al. (2008) and Gilder et al. (2014). Todorov and Bollerslev (2010) develop a theoretical framework to estimate continuous betas and jump betas of stocks with respect to a market portfolio. Bollerslev et al. (2016) find that the jump betas entail significant risk premia. Li et al. (2017a) propose an efficient estimator of the jump beta and develop a test for the stability of linear jump regression models. Recent studies further contribute to the theoretical development of jump dependence; see, for example, Li et al. (2017b, 2019). These studies about jump dependence mainly focus on co-jumps between stocks and the market.

Co-jumps among stocks also bear important information about stocks' cross-sectional dependence. We show an example of stock jumps in Fig. 1. We sampled the prices of SPY, a proxy for the market, CVX, OXY and XOM at 5-min frequency on August 9, 2017. From 10:35 a.m. to 10:40 a.m., the three stocks jumped simultaneously, while SPY did not jump. The example suggests that there exist co-jumps among stocks that are not driven by market jumps. To better understand the jump-implied dependence, it is important to study co-jumps among individual stocks even when the market does not jump.

1.2. Network models for community detection

Network models are powerful tools to describe pairwise relationships of a large number of objects. The seminal network model, the Stochastic Block Model (SBM), was proposed by Holland et al. (1983). An important extension of SBM is the Degree-Corrected Block Model (DCBM) introduced by Karrer and Newman (2011), which allows for degree heterogeneity.

Community detection is a fundamental problem in network study. In a network with a community structure, nodes can be partitioned into a small number of disjoint communities. The goal of community detection is to estimate the community labels. One widely used method in community detection is spectral clustering. Rohe et al. (2011) show consistency of spectral clustering for SBM. Jin (2015) proposes the Spectral Clustering On Ratios-of-Eigenvectors (SCORE) algorithm and shows that it enjoys consistency in community detection for DCBM. Jin et al. (2021a) show that SCORE attains an exponential rate of convergence for the grouping error and propose a SCORE+ algorithm for networks with weak signals. There are other developments in community detection under settings such as networks with mixed memberships (Jin et al., 2017) and topic modeling in textual analysis (Ke and Wang, 2017). Jin et al. (2020) develop a stepwise goodness-of-fit approach for estimating the number of communities. These studies about community detection focus on network models with independent edges.

1.3. Our contributions

We propose a network model, Degree-Corrected Block Model with Dependent Multivariate Poisson edges (DCBM-DMP). The proposed model reflects a dependent feature of edges in real-world networks. We apply our network model to study pairwise jump dependence among a large number of stocks, where the nodes of the network are stocks, and the edges are measured by pairwise co-jumps.

In terms of statistical theory, we extend the SCORE algorithm in Jin (2015) and develop a Spectral Clustering On Ratios-of-Eigenvectors for networks with Dependent Multivariate Poisson edges (SCORE-DMP). We prove strong consistency of SCORE-DMP for the proposed dependent networks. In particular, we prove that it achieves zero grouping error asymptotically.

Empirically, we analyze the co-jump networks of S&P 500 Index constituent stocks using high-frequency data from January 2003 to December 2020. Specifically, we use 5-min data to detect jumps in stocks and the market. According to whether the market jumps, we construct two co-jump networks and apply SCORE-DMP to estimate their community structures. We find that the co-jump networks contain important community information that cannot be explained by industrial classifications. Moreover, the two networks have different community features such as the number of communities, the magnitude of community co-jump intensities, and how the stocks are grouped.

Finally, we find that the detected communities in the co-jump network learned under the scenario when the market does not jump help predict stock returns. We study a return-prediction model which combines the community structure of the co-jump network with firm characteristics. We show that our model significantly improves the return forecasting accuracy both statistically and economically over the benchmark model which does not incorporate co-jump network information.

The rest of the paper is organized as follows. In Section 2, we present the main theoretical results. Section 3 is devoted to empirical analysis. Section 4 contains concluding remarks. Proofs are given in Appendix A of the online supplementary materials. Simulations and additional empirical results are collected in Appendices B – E of the online supplementary materials.

The following notation is used throughout the paper. For any vector $\mathbf{x}$, let x_i be the ith entry of $\mathbf{x}$. The ℓ_2 norm of $\mathbf{x}$ is defined as $\|\mathbf{x}\| = \sqrt{\sum x_i^2}$. For any matrix $\mathbf{M} = (M_{ij})$, its spectral norm is defined as $\|\mathbf{M}\| = \max_{\|\mathbf{x}\|\le 1}\|\mathbf{M}\mathbf{x}\|$; the Frobenius norm is defined as $\|\mathbf{M}\|_F = \sqrt{\sum_{i,j} M_{ij}^2}$; the maximum norm is defined as $\|\mathbf{M}\|_{\max} = \max_{ij}|M_{ij}|$. Denote by $\mathbf{M}_{i\cdot}$ the ith row of $\mathbf{M}$, $\lambda_k(\mathbf{M})$ the kth largest eigenvalue in absolute value of $\mathbf{M}$, and $\lambda_{\min}(\mathbf{M})$ the minimum eigenvalue in absolute value of $\mathbf{M}$. We use $\xrightarrow{p}$ and $\xrightarrow{d}$ to represent convergence in probability and convergence in distribution, respectively. For two sequences of positive real numbers a_n and b_n, we write $a_n \asymp b_n$ if $b_n/c \le a_n \le cb_n$ for some constant $c \ge 1$. Finally, c stands for a generic positive constant that may change in various scenarios.

2. Setting and main results

2.1. Continuous-time log price processes with jumps

Consider N log-price processes $\{(Y_i(t))\}_{i=1}^N$. For each i, suppose that the log-price process is an Itô semimartingale:

$$dY_i(t) = \mu_i(t)dt + \sigma_i(t)dB_i(t) + \kappa_i(t)dJ_i(t),$$

where $\mu_i(t)$ is the price drift, $\sigma_i(t)$ is the spot volatility, $B_i(t)$ is a standard Brownian motion, $(J_i(t) : t \ge 0)$ is a Poisson point process, and $\kappa_i(t)$ is the jump size. Define $\Delta Y_i(t) = Y_i(t) - Y_i(t-)$, which is nonzero only if there is a jump at time t.

Suppose that we sample stock prices at frequency M. Denote the sth intraday log return on day t for stock i as

$$r_{i,t,s} = Y_i\left(t-1+\frac{s}{M}\right) - Y_i\left(t-1+\frac{s-1}{M}\right),\ s = 1, 2, \ldots, M.$$

The Realized Volatility (RV) of stock i on day t is $RV_{i,t} = \sum_{s=1}^M r_{i,t,s}^2$. It is known that

$$\lim_{M\to\infty} RV_{i,t} \xrightarrow{p} \int_{t-1}^t \sigma_i^2(u)du + \sum_{t-1<s\le t} \big(\Delta Y_i(s)\big)^2,$$

where $\int_{t-1}^t \sigma_i^2(u)du$ and $\sum_{t-1<s\le t}\big(\Delta Y_i(s)\big)^2$ are the integrated volatility and quadratic variation due to the jump component, respectively. Barndorff-Nielsen and Shephard (2004) introduce the Realized Bipower Variation (BV):

$$BV_{i,t} = \frac{\pi}{2}\frac{M}{M-1}\sum_{s=1}^{M-1} |r_{i,t,s+1}||r_{i,t,s}|.$$

They show that BV is consistent in estimating integrated volatility:

$$\lim_{M\to\infty} BV_{i,t} \xrightarrow{p} \int_{t-1}^t \sigma_i^2(u)du.$$

In order to detect price jumps, we adopt the thresholding method (Mancini, 2009). Specifically, following Li et al. (2017a, 2019), for each stock i, we set a daily threshold level

$$v_{i,t} = \tau \times \left(\frac{1}{M}\right)^{0.49} \times \sqrt{\min\left(RV_{i,t}, BV_{i,t}\right)},$$

where τ is a tuning parameter. We set $\tau = 4$ as in Li et al. (2017a). Under some regularity conditions on $\mu_i(t)$ and $\sigma_i(t)$, as M goes to infinity, jumps can be consistently detected; see, Mancini (2009). Denote the total number of detected jumps for stock i from day 1 to day t as

$$J_{i,t} = \sum_{1\le d\le t, 1\le s\le M} \mathbb{1}_{\{|r_{i,d,s}|>v_{i,d}\}},$$

where $\mathbb{1}_{\{\cdot\}}$ is an indicator function. We follow Gilder et al. (2014) and define the co-jumps between any pair of stocks as the intersection of their jumps. Denote the total number of co-jumps between stocks i and j from day 1 to day t as

$$CJ_{ij,t} = \sum_{1\le d\le t, 1\le s\le M} \mathbb{1}_{\{|r_{i,d,s}|>v_{i,d}, |r_{j,d,s}|>v_{j,d}\}}.$$

2.2. Characteristics of stock jumps and co-jumps

The stock jumps and co-jumps have several notable features.

First, stock jumps are dependent cross-sectionally, which is why we observe co-jumps.

Second, stock co-jumps are also dependent cross-sectionally. For example, if at time t, stocks A and B co-jump, stocks B and C co-jump, then stocks B and C co-jump as well.

Third, as we will see in Figs. 3 and 5 in the empirical study, communities exist in the co-jumps, and stocks co-jump more frequently within communities than between communities. This is in line with empirical findings in the literature. For example, Fan et al. (2016) show that there is a block-diagonal pattern in the idiosyncratic correlation matrix of the S&P 500 Index constituent stocks after sorting them into industry groups.

Fourth, in Figs. 2 and 4, we will see that degree heterogeneity exists in individual stock jump intensities. Co-jump intensity is not only determined by the community information but also by individual stock jump intensities. In other words, even for stocks that are in the same group, the co-jump intensities can be different.

Fifth, as shown in Fig. 1, individual stock jumps may not be driven by the market jumps. It is interesting and informative to study separately the two scenarios according to whether the market jumps. Moreover, as we will see in Figs. 3 and 5 in the empirical study, the magnitudes of community co-jump intensities for the scenario when the market jumps are higher than the scenario when the market does not jump. When the market jumps, many stocks co-jump with the market; whereas when the market does not jump, fewer stocks co-jump as the stock jumps are mainly driven by individual firm events.

2.3. Modeling co-jump network

Consider a co-jump network $\mathcal{N} = (V, E)$, where $V = \{1, 2, \ldots, N\}$ is the set of stocks, and E is the set of edges measured by pairwise co-jumps. Suppose that the network has a community structure such that the N stocks can be divided into K disjoint communities:

$$V = V^{(1)} \cup \cdots \cup V^{(K)}.$$

Denote an $N \times 1$ vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)^T$, where θ_i is the stock jump rate. Define a $K \times K$ symmetric and positive matrix $\mathbf{P} = (P_{k\ell})_{1\le k,\ell\le K}$ such that $P_{k\ell}$ represents the community co-jump intensity between communities $V^{(k)}$ and $V^{(\ell)}$. In a given time interval $[0, t]$, we use the following dependent Poisson point processes to describe stock jump and co-jump processes:

- There is a latent Poisson point process $(L(t) : t \ge 0)$ with rate θ_{latent}, which can be interpreted as the rate of at least one jump in all stocks.
- Each individual stock jump process $(J_i(t) : t \ge 0)$ is a thinned point process: $J_i(t) = \sum_{h=1}^{L(t)} x_{i,h}$, where $\{x_{i,h}\} \overset{\text{i.i.d.}}{\sim}$ Bernoulli $(\theta_i/\theta_{latent})$ and are independent of $(L(t) : t \ge 0)$. $(J_i(t) : t \ge 0)$ is a Poisson point process with rate θ_i.
- The co-jump process between different stocks i and j, $\left(CJ_{ij}(t) : t \ge 0\right)$, is given by $CJ_{ij}(t) = \sum_{h=1}^{L(t)} x_{i,h}x_{j,h}$, where $\left(x_{i,h}, x_{j,h}\right)$ follows a bivariate Bernoulli distribution with $\mathbb{P}\left(x_{i,h} = 1, x_{j,h} = 1\right) = \mathbb{P}\left(x_{i,h} = 1\right)\mathbb{P}\left(x_{j,h} = 1\right)\theta_{latent}P_{k\ell} = \theta_i\theta_jP_{k\ell}/\theta_{latent}$. We have

  $$\mathrm{Cov}\left(x_{i,h}, x_{j,h}\right) = \mathbb{P}\left(x_{i,h} = 1\right)\mathbb{P}\left(x_{j,h} = 1\right)\left(\theta_{latent}P_{k\ell} - 1\right).$$

  If $\theta_{latent}P_{k\ell} > (<,$ respectively) 1, $x_{i,h}$ and $x_{j,h}$ are positively (negatively, respectively) correlated. If $\theta_{latent}P_{k\ell} = 1$, $x_{i,h}$ and $x_{j,h}$ are independent. It follows that $\left(CJ_{ij}(t) : t \ge 0\right)$ is a Poisson point process with rate $\theta_i\theta_jP_{k\ell}$. The co-jump intensities depend not only on community labels but also on individual stock jump intensities.

Given observations during a time interval $[0, T]$, define an $N \times N$ co-jump matrix $\mathbf{A}(T)$ as follows:

$$\Big(\mathbf{A}(T)\Big)_{ii} = \frac{J_i(T)}{T} \text{ for } 1 \le i \le N, \text{ and } \Big(\mathbf{A}(T)\Big)_{ij} = \frac{CJ_{ij}(T)}{T} \text{ for } 1 \le i \ne j \le N. \tag{2.1}$$

For ease of notation, we will drop T and denote the co-jump matrix as $\mathbf{A}$. We have that $\mathbb{E}(A_{ii}) = \theta_i$, and $\mathbb{E}(A_{ij}) = \theta_i\theta_j P_{k\ell}$ if $i \in V^{(k)}$, $j \in V^{(\ell)}$ and $i \ne j$.

We call our proposed model the Degree-Corrected Block Model with Dependent Multivariate Poisson edges (DCBM-DMP). This model captures important features in stock jumps and co-jumps, such as cross-sectional dependence, community structure and degree heterogeneity in individual jump intensities described in Section 2.2.

There are a couple of major differences between DCBM-DMP and SBM/DCBM. First, the edges of DCBM-DMP follow dependent Poisson point processes. In contrast, the edges of SBM/DCBM are assumed to be independent. Second, we keep the individual stock jumps in the diagonal of $\mathbf{A}$. In SBM/DCBM, by convention, all diagonal entries of the adjacency matrix are set to 0.

2.4. Algorithm for community detection in the co-jump network

In order to detect communities in our proposed co-jump network, we extend the SCORE method of Jin (2015) and propose the following algorithm, which we call *Spectral Clustering On Ratios-of-Eigenvectors for networks with Dependent Multivariate Poisson edges (SCORE-DMP)*:

- Step I Construct the co-jump matrix $\mathbf{A}$ defined by (2.1), in which the off-diagonal entries are the pairwise co-jumps and the diagonal entries are the individual jumps.
- Step II Get the largest K eigenvalues in absolute value of $\mathbf{A}$, $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K$, arranged in descending order, and the corresponding eigenvectors, $\widehat{\eta}_1, \ldots, \widehat{\eta}_K$. The number of communities K is determined by the method in Liu et al. (2019).
- Step III Obtain an $N \times (K-1)$ matrix $\widehat{\mathbf{R}}$ with $\widehat{R}_{ik} = (\widehat{\eta}_{k+1})_i/(\widehat{\eta}_1)_i$ for $1 \le i \le N$ and $1 \le k \le K-1$.
- Step IV Apply k-means++ algorithm (Arthur and Vassilvitskii, 2006) to the rows of $\widehat{\mathbf{R}}$, and get the estimated communities $V = \widehat{V}^{(1)} \cup \cdots \cup \widehat{V}^{(K)}$.

In community detection, estimating the number of communities is fundamental. For SBM, Lei (2016) proposes an approach based on a goodness-of-fit test, and Li et al. (2020) develop an Edge Cross-Validation method. Recently, Jin et al. (2020) develop a method to estimate the number of communities for DCBM with severe degree heterogeneity. The method applies to networks with independent edges. In this paper, we use the scree-plot method in Liu et al. (2019), which applies to our dependent case.

Our proposed SCORE-DMP differs from SCORE in the following ways. First, we keep the individual stock jumps in the diagonal of $\mathbf{A}$. Similar ideas can be found in Ji et al. (2021) that apply SCORE to the adjacent matrix plus a multiple of the identity matrix and find that it leads to better performance in community detection. Second, we use k-means++ for clustering. Using k-means++ facilitates the theoretical derivation of Theorem 2.2. Numerically, the performance of k-means++ and k-means are similar.

2.5. Theoretical properties of SCORE-DMP

We impose the following assumptions.

Assumption 1. There exists $c > 0$ such that $c \le \theta_i \le \theta_{latent} \le 1/c$ for all $1 \le i \le N$. The matrix $\mathbf{P}$ is a fixed $K \times K$ positive matrix, and $|\lambda_{\min}(\mathbf{P})| > c$.

Remark 2.1. There are studies that allow $\mathbf{P}$ to vary with N; see, for example, Jin et al. (2017, 2020), and Jin et al. (2021b). Specifically, Jin et al. (2021b) study the scenario when the eigenvalues of $\mathbf{P}$ satisfy $\|\boldsymbol{\theta}\||\lambda_K(\mathbf{P})|/\lambda_1(\mathbf{P}) \to \infty$ and show that the proposed approach enjoys consistency in community detection. For our proposed network model with dependent edges, the assumption about $\mathbf{P}$ can be relaxed to $\lambda_1(\mathbf{P}) \asymp 1$ and $|\lambda_K(\mathbf{P})| \gg 1/\sqrt{N}$. When $\lambda_K(\mathbf{P}) \to 0$, we need stronger conditions on the asymptotic sizes of N and T. Specifically, N can grow polynomially with T but not exponentially as in Assumption 4.

Remark 2.2. About the assumption that K is fixed, for the consistency of community detection, the assumption can be relaxed to $(K^{9/2}\sqrt{\log(N)/T} + K^3/N) \to 0$. The proof will be modified slightly by keeping the multiplicative polynomial terms involving K in the error bounds such as (A.12), (A.31) and (A.38).

For $k = 1, 2, \ldots, K$, let $n_k = |V^{(k)}|$ be the size of the kth community.

Assumption 2. $n_k \asymp n_\ell$ for all $1 \le k, \ell \le K$.

Write $\mathbf{D} = (D_{kk})_{1\le k\le K}$, where $D_{kk} = \big(\sum_{i\in V^{(k)}} \theta_i^2\big)^{1/2}/\|\boldsymbol{\theta}\|$.

Assumption 3. $\lambda_1(\mathbf{DPD}) - |\lambda_2(\mathbf{DPD})| \geq c$ for some constant $c > 0$.

Assumption 4. N and T both go to infinity with $\log(N) = o(T)$.

Denote $\boldsymbol{\Omega} = \left(\Omega_{ij}\right)_{1\leq i,j\leq N}$, where $\Omega_{ij} = \theta_i\theta_j P_{k\ell}$ if $i \in V^{(k)}$ and $j \in V^{(\ell)}$. Let the largest K eigenvalues in absolute value and the associated eigenvectors of $\boldsymbol{\Omega}$ be $\lambda_1, \ldots, \lambda_K$, and $\eta_1, \ldots, \eta_K$, respectively. Define $\mathbf{R} = (R_{ik})_{1\leq i\leq N, 1\leq k\leq K-1}$, where $R_{ik} = (\eta_{k+1})_i/(\eta_1)_i$.

Theorem 2.1. *Under Assumptions 1–4, for any constant $q > 0$, there exist a $(K-1)\times(K-1)$ orthogonal matrix $\mathbf{H}$ and $c > 0$ such that*

$$\mathbb{P}\left(\|\widehat{\mathbf{R}}\mathbf{H} - \mathbf{R}\|_{\max} > c\cdot\left(\sqrt{\frac{\log(N)}{T}} + \frac{1}{N}\right)\right) = o\left(\frac{1}{N^q}\right). \tag{2.2}$$

Remark 2.3. For DCBM with independent edges, the error rate of $\max_{1\leq i\leq N}\|(\widehat{\mathbf{R}}\mathbf{H})_{i\cdot} - \mathbf{R}_{i\cdot}\|$, which is equivalent to the error rate of $\|\widehat{\mathbf{R}}\mathbf{H} - \mathbf{R}\|_{\max}$ since K is fixed, is $\sqrt{\log(N)/N}$; see, for example, Jin et al. (2017). For DCBM-DMP, the error rate of $\|\widehat{\mathbf{R}}\mathbf{H} - \mathbf{R}\|_{\max}$ is $\sqrt{\log(N)/T} + 1/N$, and one needs both N and T to go to infinity in order to consistently estimate $\mathbf{R}$. The rate in (2.2) is sharp. Specifically, the term $\sqrt{\log(N)/T}$ comes from the random error of using the eigenvectors of $\mathbf{A}$ to estimate those of $\mathbb{E}(\mathbf{A})$, and the term $1/N$ comes from the difference in the leading eigenvectors between $\mathbb{E}(\mathbf{A})$ and $\boldsymbol{\Omega}$. Both error rates are sharp; see Remark A.1 for more details.

Denote the true labels as $\mathbf{g} = (g_1, g_2, \ldots, g_N)^T$, and the estimated labels from SCORE-DMP as $\widehat{\mathbf{g}}^{sc} = \left(\widehat{g}_1^{sc}, \widehat{g}_2^{sc}, \ldots, \widehat{g}_N^{sc}\right)^T$, where

$$g_i = \sum_{k=1}^{K} k\cdot\mathbb{1}_{\{i\in V^{(k)}\}}, \quad \text{and} \quad \widehat{g}_i^{sc} = \sum_{k=1}^{K} k\cdot\mathbb{1}_{\{i\in\widehat{V}^{(k)}\}}. \tag{2.3}$$

Following the literature on community detection, we measure the number of mismatched labels by the Hamming error:

$$\mathrm{Hamm}_N(\mathbf{g}, \widehat{\mathbf{g}}^{sc}) = \min_{\pi(\mathbf{g})}\sum_{i=1}^{N}\mathbb{1}_{\{(\pi(\mathbf{g}))_i\neq\widehat{g}_i^{sc}\}}, \tag{2.4}$$

where π is a permutation of the set $\{1, 2, \ldots, K\}$.

Theorem 2.2. *Under Assumptions 1–4, for any constant $q > 0$, we have*

$$\mathbb{P}\left(Hamm_N\left(\mathbf{g}, \widehat{\mathbf{g}}^{sc}\right) > 0\right) = o\left(\frac{1}{N^q}\right). \tag{2.5}$$

Remark 2.4. For DCBM with independent edges, one only needs N to go to infinity to achieve consistent community detection. For DCBM-DMP, besides requiring $N\to\infty$, we also need $T\to\infty$ to consistently detect the communities. To see this, note that all the jump processes are dominated by a single latent point process $(L(t))$, hence for any fixed T, there is a positive probability that there is no jump at all, in which case it is certainly impossible to detect communities.

Theorem 2.2 states that applying our proposed algorithm to the co-jump network model achieves strong consistency property (Zhao et al., 2012). Specifically, as N and T go to infinity, SCORE-DMP attains zero Hamming error with high probability.

We can also estimate the individual jump intensities $\boldsymbol{\theta}$ and the community co-jump intensity matrix $\mathbf{P}$. We estimate the individual jump intensity θ_i by A_{ii} for $1\leq i\leq N$ and estimate $\mathbf{P}$ by $\widehat{\mathbf{P}} = (\widehat{P}_{k\ell})$, where

$$\widehat{P}_{k\ell} = \frac{\sum_{i\in\widehat{V}^{(k)},\ j\in\widehat{V}^{(\ell)},\ i\neq j} A_{ij}}{\sum_{i\in\widehat{V}^{(k)},\ j\in\widehat{V}^{(\ell)},\ i\neq j} A_{ii}A_{jj}}. \tag{2.6}$$

Proposition 2.1. *Under Assumptions 1–4, for any constant $q > 0$, there exists $c > 0$ such that*

$$\mathbb{P}\left(\max_{1\leq i\leq N}|A_{ii} - \theta_i| > c\cdot\sqrt{\frac{\log(N)}{T}}\right) = o\left(\frac{1}{N^q}\right), \text{ and}$$
$$\mathbb{P}\left(\max_{1\leq k,\ell\leq K}|\widehat{P}_{k\ell} - P_{k\ell}| > c\cdot\sqrt{\frac{\log(N)}{T}}\right) = o\left(\frac{1}{N^q}\right).$$

We can further make inference about $\mathbf{P}$. Define a length-$K(K+1)$ vector $\boldsymbol{\xi}_h = (y_{k\ell,h}, z_{k\ell,h})_{1\le k\le \ell\le K}^T$ as follows:

$$y_{k\ell,h} = \frac{1}{|I_{(k,\ell)}|}\sum_{(i,j)\in I_{(k,\ell)}} x_{i,h}x_{j,h}, \quad z_{k\ell,h} = \frac{1}{|I_{(k,\ell)}|}\sum_{(i,j)\in I_{(k,\ell)}} (x_{i,h}\theta_j + x_{j,h}\theta_i), \text{ and}$$
$$I_{(k,\ell)} = \{(i,j) : i \in V^{(k)},\ j \in V^{(\ell)}, 1 \le i \ne j \le N\}.$$

Assumption 5. $\lambda_{\min}\big(\mathrm{Cov}(\boldsymbol{\xi}_h)\big) > c$ for some constant $c > 0$.

Remark 2.5. Assumption 5 guarantees that there is no strong collinearity among the aggregated group-wise co-jumps and individual jumps.

Proposition 2.2. *Under Assumptions 1–5,*

$$\sqrt{T}\,(\boldsymbol{\Sigma}_P)^{-1/2}\left(\widehat{\mathbf{P}}_u - \mathbf{P}_u\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}), \tag{2.7}$$

where $\widehat{\mathbf{P}}_u = (\widehat{P}_{11}, \widehat{P}_{12}, \ldots, \widehat{P}_{k\ell}, \ldots, \widehat{P}_{KK})_{1\le k\le \ell\le K}^T$, *and* $\mathbf{P}_u = (P_{11}, P_{12}, \ldots, P_{k\ell}, \ldots, P_{KK})_{1\le k\le \ell\le K}^T$, *and* $\boldsymbol{\Sigma}_P$ *is given in (A.50)–(A.52), and (A.59)–(A.61) of Appendix A.*

In practice, the covariance matrix $\boldsymbol{\Sigma}_P$ can be consistently estimated by (A.63)–(A.66) in Appendix A.

We perform simulation studies to evaluate the performance of the proposed estimators and present the results in Appendix B. The simulation results show that for large N and T, SCORE-DMP has low Hamming errors, $\widehat{\mathbf{P}}$ is close to $\mathbf{P}$, and the empirical distribution of the standardized errors of $\widehat{\mathbf{P}}$ matches well with the standard normal distribution.

3. Empirical studies

We perform two empirical studies. First, we use high-frequency data to build two co-jump networks according to whether the market jumps. We then investigate the community structures of the co-jump networks and compare them with the Global Industry Classification Standard (GICS).[1] Second, we utilize the community information to predict stock returns.

3.1. Community detection for co-jump networks

3.1.1. Data for co-jump networks

We obtain high-frequency data from the Trade and Quote (TAQ) database. We select the stocks that remained in the S&P 500 Index between January 2003 and December 2020 and use SPDR S&P 500 ETF trust (SPY) as our proxy for the market. To strike a balance between reducing the impact of microstructure noise and using as much data as possible, we sample the prices at 5-min frequency. After excluding the stocks with more than 5% missing data during the period, we end up with 190 stocks. We obtain the GICS codes of all the stocks under evaluation from the Compustat database.

To reduce the impact of noisy observations at the market opening and closing, we focus on intraday 5-min returns from 9:35 a.m. to 3:55 p.m., resulting in 76 log returns for each stock in a day. Jumps for stocks and the market are detected by the thresholding method described in Section 2.1. Following Li et al. (2019), we filter the jumps by the following two criteria: (1) remove the jumps that occur consecutively within 10 min and have opposite directions; (2) remove the jumps with sizes less than ten basis points. Criterion (1) removes bouncebacks,[2] and Criterion (2) removes small jumps when the day has a low volatility. For SPY, we originally detect 718 jumps over the 2003–2020 period. After filtering, there are 682 jumps left.

Besides the market factor, we also analyze the co-jump networks according to whether other factors jump. We provide one example in Appendix D.

3.1.2. Scenario I: When the market jumps

We first investigate the scenario when the market jumps (Scenario I). As mentioned earlier, during the 2003–2020 period, the market jumped 682 times. We construct a co-jump network using the pairwise stock co-jumps that occur simultaneously with the market jumps. Fig. 2 shows the distribution of the estimated individual stock jump intensities. We find that the estimated individual stock jump intensities exhibit substantial heterogeneity.

We then use the SCORE-DMP described in Section 2.4 to estimate the communities. The number of communities is estimated to be five. The clustering results are illustrated in Fig. 3 with details collected in Table 3 of Appendix C.

[1] The Global Industry Classification Standard (GICS) is an industry classification system. It defines eleven economic sectors: Energy (EN), Materials (MT), Industrials (ID), Consumer Discretionary (CD), Consumer Staples (CS), Health Care (HC), Financials (FN), Information Technology (IT), Communication Services (CM), Utilities (UT) and Real Estate (RE).

[2] Aït-Sahalia et al. (2011) define the bounceback to be log returns for two consecutive transactions that are larger than 1% but have opposite signs. They state that the 'roundtrips' are not meaningful transactions.

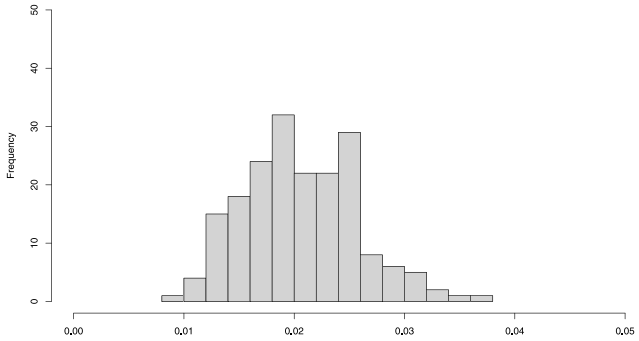

Fig. 2. The distribution of estimated individual stock jump intensities when the market jumps.

The left panel in Fig. 3 shows the heat map of the clustering. The red squares along the diagonal indicate the estimated communities. Shades of the blocks represent the estimated community co-jump intensities. We rearrange the estimated community co-jump intensity matrix $\widehat{\mathbf{P}}^{I}$ such that the diagonal entries of $\widehat{\mathbf{P}}^{I}$ are in descending order. We find that the diagonal entries of $\widehat{\mathbf{P}}^{I}$ are higher than the off-diagonal entries, and the test based on Proposition 2.2 suggests that the differences are mostly statistically significant. The results suggest a strong community structure in the co-jump network.

The right panel in Fig. 3 shows the constituents of each group. We use different colors to represent different GICS sectors. The order of the groups corresponds to the heat map, where Group 1 is in the lower-left corner, and Group 5 is in the upper-right corner.

Compared with GICS, we find that the communities in the stock co-jump network have the following interesting features.

The first two groups match largely with GICS but are different from GICS in the following ways. First, Group 1 mainly comprises the stocks from Utilities. Besides Utilities, Group 1 also includes Newmont (NEM) from Materials, which extracts natural gas. The Utilities companies such as Duke Energy Corporation (DUK) are mainly engaged in electricity and natural gas services. Thus, for stocks in Group 1, there is a supply chain relationship between natural gas extraction and use. Second, the majority of the Financials stocks are in Group 2, and the remaining stocks are split into Groups 3–5.

The remaining three groups have mixed sector constituents. There are interesting supply chain relationships for the stocks within each group that cannot be explained by GICS. In Group 3, the Materials companies, International Flavors and Fragrances (IFF) and Ecolab (ECL), are major suppliers of flavors and water, which are the raw materials for the food companies that account for a large proportion of Consumer Staples in Group 3. In Group 4, the Energy companies such as ConocoPhillips (COP) are suppliers of oil and gas products, which are heavily consumed by the airline and aerospace manufacturer like Southwest Airlines (LUV), and Northrop Grumman Corporation (NOC) from Industrials. In Group 5, the Information Technology companies such as Qualcomm (QCOM) and Xilinx (XLNX) are the main suppliers of semiconductors, which are used by the analytical instruments manufacturing companies like Agilent (A) from Health Care.

3.1.3. Scenario II: When the market does not jump

Next, we consider the scenario when the market does not jump (Scenario II). We construct a co-jump network under this scenario. Fig. 4 shows the distribution of estimated individual stock jump intensities. Again, we see that under Scenario II, individual jump intensities show substantial heterogeneity.

The number of communities is estimated to be seven. The clustering results are illustrated in Fig. 5. The diagonal entries of the community co-jump intensity matrix $\widehat{\mathbf{P}}^{II}$ are again mostly statistically significantly higher than the off-diagonal entries. The specific classifications for all stocks are collected in Table 4 of Appendix C.

We have the following observations about the communities in the co-jump network under Scenario II.

The first two groups with the strongest community co-jump intensities match the Energy and Utilities sectors.

The remaining five groups are different from GICS. The Financials sector stocks are mainly in Group 3. Group 4 includes the Information Technology stocks. Besides the Information Technology sector, Group 4 also includes the e-commerce company eBay (EBAY) from Consumer Discretionary. Group 5 is the retail group. It contains retail companies such as Best Buy (BBY) from Consumer Discretionary and Walmart (WMT) from Consumer Staples. Industrials and Materials are merged into Group 6. These stocks are naturally connected in manufacturing processes using materials such as steel. Group 7 mainly comprises stocks that are less sensitive to the economy, for example, stocks from Consumer Staples and Health Care.

3.1.4. Comparison between Scenario I and Scenario II

From Sections 3.1.2 and 3.1.3, we see that the networks under Scenario I and Scenario II have distinct features.

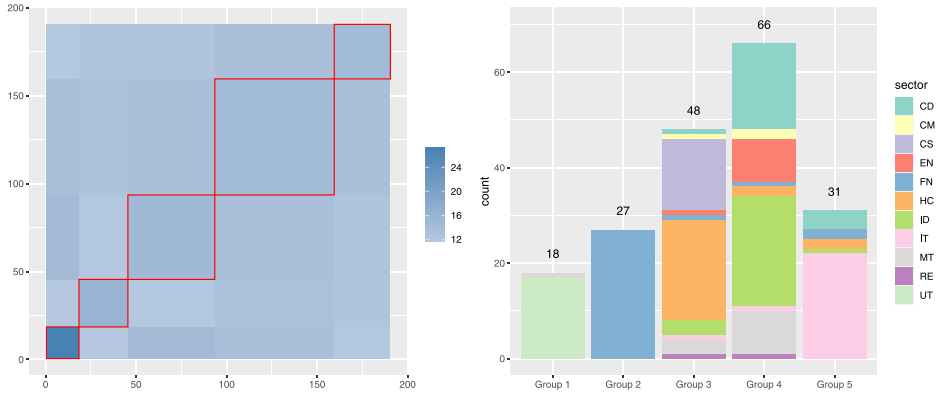

Fig. 3. Clustering results when the market jumps. Left panel: the heat map of the clustering. Shades of the blocks represent the estimated community co-jump intensities. The estimated five groups are ordered according to their estimated community co-jump intensities from the highest (Group 1) to the lowest (Group 5). Right panel: the group labels (x-axis) and the number of stocks in each group (y-axis). Different colors represent different GICS sectors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

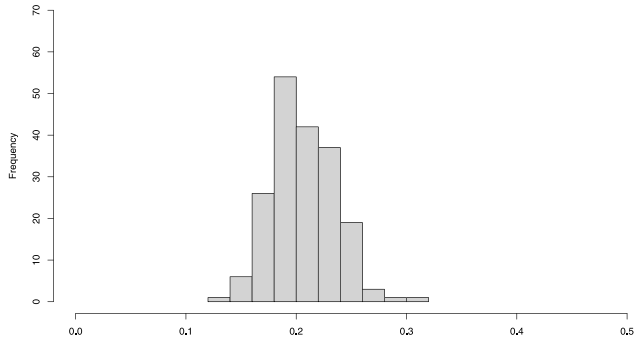

Fig. 4. The distribution of the estimated individual stock jump intensities when the market does not jump.

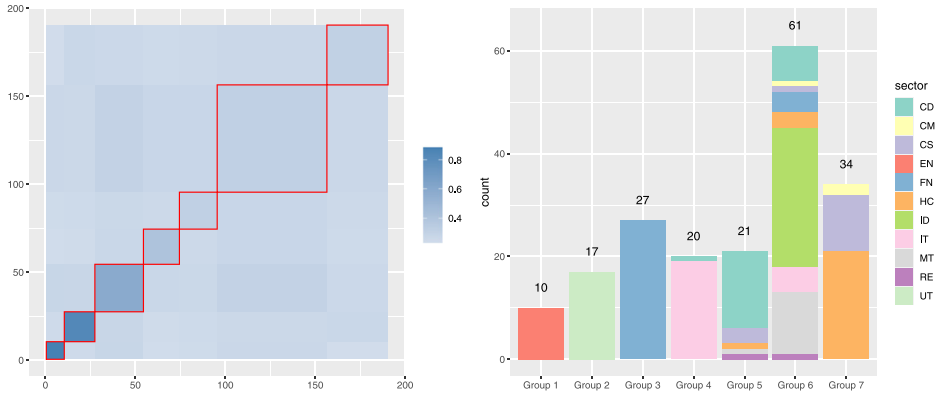

Fig. 5. Clustering results when the market does not jump. Left panel: the heat map of the clustering. Shades of the blocks represent the estimated community co-jump intensities. The estimated five groups are ordered according to their estimated community co-jump intensities from the highest (Group 1) to the lowest (Group 7). Right panel: the group labels (x-axis) and the number of stocks in each group (y-axis). Different colors represent different GICS sectors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

First, the estimated numbers of communities are different. When the market jumps, the co-jump network has five communities. When the market does not jump, there are seven communities. The difference suggests that the market jumps weaken the community structure.

Second, the magnitudes of the estimated community co-jump intensities are different. The entries of $\widehat{\mathbf{P}}^{I}$ range from 11 to 27, while the entries of $\widehat{\mathbf{P}}^{II}$ are smaller and range from 0.23 to 0.88. This can be due to that when the market jumps, stocks are more likely to co-jump with the market.

Third, the stocks are grouped differently. For example, when the market jumps, the stocks in the Energy sector are merged with other sectors. When the market does not jump, they form one group. The difference suggests that the co-jumps under the two scenarios are generated by different mechanisms.

3.2. Stock return prediction

3.2.1. Model

Several studies in the literature such as Lewellen (2015) model stock returns with lagged firm-level characteristics:

$$\mathbf{r}_t = \gamma_0 \mathbf{1} + \mathbf{Z}_{t-1}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_t, \tag{3.1}$$

where $\mathbf{r}_t \in \mathbb{R}^{N\times 1}$ is the excess returns of N stocks, $\gamma_0 \in \mathbb{R}$ is the intercept term, $\mathbf{1}$ is a length-N vector with all entries being one, $\mathbf{Z}_{t-1} \in \mathbb{R}^{N\times p}$ contains p firm-level characteristics, $\boldsymbol{\gamma} \in \mathbb{R}^{p\times 1}$ is the coefficient vector of characteristics, and $\boldsymbol{\varepsilon}_t \in \mathbb{R}^{N\times 1}$ is a random error term that is independent of $\mathbf{Z}_{t-1}$.

The analysis in Section 3.1 reveals that the stock co-jump networks have community structures. We study the following return-prediction model, which combines the community information in the co-jump network with firm-level characteristics,

$$\mathbf{r}_t = \gamma_0 \mathbf{1} + \underbrace{\gamma_1 \boldsymbol{\Gamma}\mathbf{P}\boldsymbol{\Gamma}^T\mathbf{x}_{t-1}}_{\text{network effect}} + \mathbf{Z}_{t-1}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_t, \tag{3.2}$$

where $\boldsymbol{\Gamma}$ is an $N \times K$ membership matrix, $(\boldsymbol{\Gamma})_{ik} = \mathbb{1}_{\{i\in V^{(k)}\}}$, $\mathbf{x}_{t-1} \in \mathbb{R}^{N\times 1}$ is a firm-level characteristic vector on which the network effect acts, and $\gamma_1 \in \mathbb{R}$ is the network effect parameter. Because the majority of stock jumps are non-market related, we use the co-jump network under the scenario when the market does not jump. The network effect component describes how cross-sectional information influences each stock. Specifically, $\boldsymbol{\Gamma}\mathbf{P}$ represents the community effect for each stock, and $\boldsymbol{\Gamma}^T\mathbf{x}_{t-1}$ transforms the firm-level characteristics into a community-level characteristic vector. The network effect component of (3.2) is similar to that of the network vector autoregressive model with community structure (CNAR) studied in Chen et al. (2020). There are two main differences between model (3.2) and CNAR. First, CNAR applies network effect to $\mathbf{r}_{t-1}$. In contrast, (3.2) can apply network effect to any firm-level characteristic. Second, in (3.2), $\mathbf{P}$ is the community co-jump intensity matrix from our proposed co-jump network, while in CNAR, it is treated as a parameter in the regression model that needs to be estimated.

3.2.2. Data for return prediction

We use monthly returns of S&P 500 Index constituent stocks from January 2003 to December 2021. The data are obtained from CRSP. To calculate the excess returns, we use the Treasury-bill rate as the risk-free rate. We obtain the high-frequency data of the stocks from TAQ for network construction. For firm-level characteristics, we select three characteristics that have been documented to be helpful in predicting stock returns: log market capitalization (*Size*), Book-to-Market ratio (*B2M*) and one-month cumulative return (*Mom*); see, for example, Jegadeesh and Titman (1993), Lewellen (2015) and Green et al. (2017). We get monthly *Size* and *Mom* data from CRSP and annual *B2M* data from Compustat. To avoid forward-looking bias, we lag the annual *B2M* data by six months. For each month, we winsorize each characteristic data at 1% and 99% and standardize them to have zero mean and unit standard deviation in the cross-section. For missing characteristics, we replace them with 0.

The benchmark model (3.1) only uses *Size*, *B2M* and *Mom* as predictors. For model (3.2), besides the predictors used in the benchmark model, we apply the community effect to *Size* (Network-Size).

3.2.3. Parameter estimation and return prediction

We perform return forecasting with a rolling window scheme. The training sample size is 120 months, and the out-of-sample testing period is from January 2013 to December 2021. The models are re-estimated every year. For each year, we select the stocks that remained in the S&P 500 Index in the past 120 months with less than 5% missing high-frequency data. The parameters in model (3.1) are estimated by the ordinary least squares method. For model (3.2), as the rolling window moves, we re-estimate the number of communities, use SCORE-DMP to estimate $\boldsymbol{\Gamma}$, and use (2.6) to estimate $\mathbf{P}$. Then we use the ordinary least squares method to estimate the remaining parameters, γ_0, γ_1 and $\boldsymbol{\gamma}$ in (3.2). At the end of each month, we get one-month-ahead stock return forecasts.

3.2.4. Performance evaluation metrics and prediction results

We use the following two metrics to evaluate the prediction performance of the Network-Size model and the benchmark model: out-of-sample R^2 and Sharpe ratio of zero-net portfolio based on predicted returns; see Appendix E.1 for a more detailed description of the performance evaluation metrics.

We have the following findings. First, the Network-Size model improves the out-of-sample R^2 compared with the benchmark model (2.673% vs. 2.637%). The improvement is statistically significant based on the Diebold–Mariano

test (Diebold and Mariano, 2002). Second, the Sharpe ratio of the zero-net Network-Size portfolio is about 0.3 higher than the benchmark portfolio (1.159 vs. 0.839). The above results demonstrate the superior prediction performance of the network-based model.

The detailed analysis of the return prediction results is collected in Appendix E.2.

4. Conclusion

We propose a model, DCBM-DMP, to study the stock co-jump dependence. The proposed model differs from the standard DCBM in that it has dependent edges. To detect the communities, we extend the SCORE algorithm in Jin (2015) and propose a SCORE-DMP algorithm. We prove that SCORE-DMP achieves strong consistency in detecting communities for our proposed model. Empirically, we construct two co-jump networks according to whether the market jumps. The results show that the co-jump networks have important community information that cannot be explained by GICS. The co-jump networks also provide a novel perspective to study the dependence among a large number of stocks. We further study a return-prediction model which combines the community structure of the co-jump network with firm characteristics. We find that adding the co-jump network community information improves return prediction accuracy.

Acknowledgments

We thank the Editor, Elie Tamer, the Guest Editor, Runze Li, and three anonymous referees for their constructive suggestions. We also thank the participants of SoFiE 2022 Annual Conference and 2022 Asian Meeting of the Econometric Society for valuable discussions and comments. Research was partially supported by National Science Foundation of China grants NSFC71922902 and NSFC72101226, Research Grant Council of Hong Kong SAR grants GRF15302321, GRF16503419, GRF16502118, GRF16304521, GRF16304019, T31-604/18-N and T31-603/21-N.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2023.01.026.

References

Aït-Sahalia, Y., Jacod, J., 2009. Testing for jumps in a discretely observed process. Ann. Statist. 184–222.
Aït-Sahalia, Y., Jacod, J., Li, J., 2012. Testing for jumps in noisy high frequency data. J. Econometrics 168 (2), 207–222.
Aït-Sahalia, Y., Mykland, P.A., Zhang, L., 2011. Ultra high frequency volatility estimation with dependent microstructure noise. J. Econometrics 160 (1), 160–175.
Andersen, T.G., Bollerslev, T., Diebold, F.X., 2007. Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. Rev. Econ. Stat. 89 (4), 701–720.
Arthur, D., Vassilvitskii, S., 2006. K-Means++: The Advantages of Careful Seeding. Technical report, Stanford.
Barndorff-Nielsen, O.E., Shephard, N., 2004. Power and bipower variation with stochastic volatility and jumps. J. Financ. Econom. 2 (1), 1–37.
Bates, D.S., 1991. The crash of '87: was it expected? the evidence from options markets. J. Finance 46 (3), 1009–1044.
Bollerslev, T., Law, T.H., Tauchen, G., 2008. Risk, jumps, and diversification. J. Econometrics 144 (1), 234–256.
Bollerslev, T., Li, S.Z., Todorov, V., 2016. Roughing up beta: Continuous versus discontinuous betas and the cross section of expected stock returns. J. Financ. Econ. 120 (3), 464–490.
Bollerslev, T., Todorov, V., 2011. Tails, fears, and risk premia. J. Finance 66 (6), 2165–2211.
Bollerslev, T., Todorov, V., Xu, L., 2015. Tail risk premia and return predictability. J. Financ. Econ. 118 (1), 113–134.
Chen, E.Y., Fan, J., Zhu, X., 2020. Community network auto-regression for high-dimensional time series. arXiv preprint arXiv:2007.05521.
Cremers, M., Halling, M., Weinbaum, D., 2015. Aggregate jump and volatility risk in the cross-section of stock returns. J. Finance 70 (2), 577–614.
Diebold, F.X., Mariano, R.S., 2002. Comparing predictive accuracy. J. Bus. Econom. Statist. 20 (1), 134–144.
Eraker, B., 2004. Do stock prices and volatility jump? reconciling evidence from spot and option prices. J. Finance 59 (3), 1367–1403.
Fan, J., Furger, A., Xiu, D., 2016. Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. J. Bus. Econom. Statist. 34 (4), 489–503.
Gilder, D., Shackleton, M.B., Taylor, S.J., 2014. Cojumps in stock prices: Empirical evidence. J. Bank. Financ. 40, 443–459.
Green, J., Hand, J.R., Zhang, X.F., 2017. The characteristics that provide independent information about average us monthly stock returns. Rev. Financ. Stud. 30 (12), 4389–4436.
Holland, P.W., Laskey, K.B., Leinhardt, S., 1983. Stochastic blockmodels: First steps. Social Networks 5 (2), 109–137.
Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. J. Finance 48 (1), 65–91.
Ji, P., Jin, J., Ke, Z.T., Li, W., 2021. Co-citation and co-authorship networks of statisticians. J. Bus. Econom. Statist. 1–17.
Jin, J., 2015. Fast community detection by SCORE. Ann. Statist. 43 (1), 57–89.
Jin, J., Ke, Z.T., Luo, S., 2017. Estimating network memberships by simplex vertex hunting. arXiv preprint arXiv:1708.07852.
Jin, J., Ke, Z.T., Luo, S., 2021a. Improvements on score, especially for weak signals. Sankhya A 1–36.
Jin, J., Ke, Z.T., Luo, S., 2021b. Optimal adaptivity of signed-polygon statistics for network testing. Ann. Statist. 49 (6), 3408–3433.
Jin, J., Ke, Z.T., Luo, S., Wang, M., 2020. Estimating the number of communities by stepwise goodness-of-fit. arXiv preprint arXiv:2009.09177.
Karrer, B., Newman, M.E., 2011. Stochastic blockmodels and community structure in networks. Phys. Rev. E 83 (1), 016107.
Ke, Z.T., Wang, M., 2017. A new svd approach to optimal topic estimation. arXiv preprint arXiv:1704.07016.
Kou, S., Yu, C., Zhong, H., 2017. Jumps in equity index returns before and during the recent financial crisis: A bayesian analysis. Manage. Sci. 63 (4), 988–1010.
Lee, S.S., Mykland, P.A., 2008. Jumps in financial markets: A new nonparametric test and jump dynamics. Rev. Financ. Stud. 21 (6), 2535–2563.
Lee, S.S., Mykland, P.A., 2012. Jumps in equilibrium prices and market microstructure noise. J. Econometrics 168 (2), 396–406.
Lei, J., 2016. A goodness-of-fit test for stochastic block models. Ann. Statist. 44 (1), 401–424.
Lewellen, J., 2015. The cross-section of expected stock returns. Crit. Financ. Rev. 4 (1), 1–44.

Li, J., 2013. Robust estimation and inference for jumps in noisy high frequency data: A local-to-continuity theory for the pre-averaging method. Econometrica 81 (4), 1673–1693.
Li, T., Levina, E., Zhu, J., 2020. Network cross-validation by edge sampling. Biometrika 107 (2), 257–276.
Li, J., Todorov, V., Tauchen, G., 2017a. Jump regressions. Econometrica 85 (1), 173–195.
Li, J., Todorov, V., Tauchen, G., 2017b. Robust jump regressions. J. Amer. Statist. Assoc. 112 (517), 332–341.
Li, J., Todorov, V., Tauchen, G., 2019. Jump factor models in large cross-sections. Quant. Econ. 10 (2), 419–456.
Liu, Y., Hou, Z., Yao, Z., Bai, Z., Hu, J., Zheng, S., 2019. Community detection based on the ℓ_∞ convergence of eigenvectors in dcbm. arXiv preprint arXiv:1906.06713.
Mancini, C., 2009. Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. Scand. J. Statist. 36 (2), 270–296.
Rohe, K., Chatterjee, S., Yu, B., 2011. Spectral clustering and the high-dimensional stochastic blockmodel. Ann. Statist. 39 (4), 1878–1915.
Todorov, V., Bollerslev, T., 2010. Jumps and betas: A new framework for disentangling and estimating systematic risks. J. Econometrics 157 (2), 220–235.
Zhao, Y., Levina, E., Zhu, J., 2012. Consistency of community detection in networks under degree-corrected stochastic block models. Ann. Statist. 40 (4), 2266–2292.