Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Clustering high-frequency financial time series based on information theory

Haitao Liu[1] | Jian Zou*[2] | Nalini Ravishanker[3]

[1]Data Science Program, Worcester Polytechnic Institute, MA, USA
[2]Department of Mathematical Sciences, Worcester Polytechnic Institute, MA, USA
[3]Department of Statistics, University of Connecticut, CT, USA

Correspondence
*Jian Zou, Department of Mathematical Sciences, Worcester Polytechnic Institute, MA, USA. Email: jzou@wpi.edu

Abstract

Clustering large financial time series data enables pattern extraction that facilitates risk management. The knowledge gathered from unsupervised learning is useful for improving portfolio optimization and making stock trading recommendations. Most methods available in the literature for clustering financial time series are based on exploiting linear relationships between time series. However, prices of different assets (stocks) may have non-linear relationships which may be quantified using information based measures such as mutual information (MI). To estimate the empirical mutual information between time series of stock returns, we employ a novel kernel density estimator (KDE) based jackknife mutual information estimation (JMI), and compare it with the widely used binning method. We then propose an average distance gradient change algorithm (ADGC) and an algorithm based on the average silhouette criterion that use pairwise and groupwise MI of high-frequency financial stock returns. Through numerical studies, we provide insights into the impact of the clustering on asset allocation and risk management based on the nonlinear information structure of the US stock market.

KEYWORDS:
clustering, high-frequency, mutual information, time series

1 | INTRODUCTION

High-frequency transaction-by-transaction financial data are readily available to investors and researchers who seek to analyze these data to understand patterns that will facilitate risk management[1,2]. Clustering stocks based on properties such as intra-day returns is an attractive approach to understanding interdependencies between them. Under a satisfactory clustering scheme, we would expect "similar" stocks that stochastically move together to be grouped into the same cluster. The clusters can then provide insight into properties of the stocks and investors can hedge volatility risk by allocating investments among the stocks in different clusters[3,4]. Clustering financial time series has been widely used in the finance domain to extract patterns from large data sets[5,6], and the extracted information has been used to improve portfolio optimization[7,8,9,10] or generate stock trading recommendations[11].

Clustering approaches for financial time series have mainly focused on using measures that capture linear relationships between different stocks, such as Euclidean distance or pairwise Pearson correlation. For instance, Brown and Goetzmann[12] proposed a K-means algorithm based on Euclidean distances for clustering time series of monthly returns of mutual funds, while Pattarin et al[5] presented a study on clustering mutual funds into two types (growth or value) based on their pairwise Pearson correlations. Onnela et al[6] described statistical network analysis of companies based on the pairwise Pearson correlations between

daily stock returns. However, a linear measure fails to capture nonlinear relationships between financial time series. To address this issue, Bastos and Caiado[13] studied a distance based on multidimensional scaling (MDS) to cluster financial time series using variance ratio test statistics. Harvill et al[4] proposed spectral and bispectral based quasi-distances for clustering nonlinear and nonstationary financial time series.

Another useful nonlinear measure is mutual information (MI), which is related to the joint entropy of two random variables and measures the dependence between them[14]. Kinney and Atwal[15] showed that MI satisfies the "self-equitability" property, i.e., it places equal importance on linear and nonlinear dependence between random variables. Brillinger[16] employed the concept of MI to carry out data analysis in various application domains including sports analytics, neuroscience, and forest science. In the context of clustering interdependent financial time series, an MI based distance captures relationships between stock returns that are not detected by the usual linear distance metrics such as Euclidean distances between return time series or their spectra, or Pearson correlations between the time series. For instance, Kim and Sayama[17] and Guo et al[18] have used MI to study the behavior of financial networks.

Estimation of MI has received considerable attention in the literature and there are three primary estimation approaches for continuous-valued data. One approach consists of discretizing the continuous data into different bins and estimating MI from the discretized data. Another approach is based on estimating the joint and marginal MI's using, for example, the kernel density estimator (KDE)[19,20,21], the B-spline estimator[22], or the wavelet estimator[23]. A third approach is based on invoking the relationship between entropy and MI and using k-nearest neighbor (KNN) distances[14,24]. To our knowledge, only the binning approach has often been used to estimate the MI between stock prices or log-returns[17,18]. However, the binning approach, as well as other KDE based approaches tend to provide biased estimates of MI. To remedy this, Zeng et al[21] proposed a "jackknife estimate" of MI which reduces the bias and frees the estimation from the number of bins and grouping criteria (equal number, frequency or quantile) selection. Their method facilitates the estimation of MI between a pair of variables X and Y as well as between two sets of random vectors $\mathbf{X}$ and $\mathbf{Y}$.

In this article, our goal is to employ the approach of Zeng et al[21] for estimating MI and using it to cluster stocks from the S&P 100 stock market index based on intra-day returns data downloaded from the Trade and Quote (TAQ) database of Wharton Research Data Services (WRDS). We then compare these results with those from clustering the same stocks using the binning method based MI, in order to see whether differences in the MI estimates lead to very different financial metrics such as the Sharpe ratio[25] or the diversification index (DI)[26]. We also apply two different algorithms to select the optimal number of clusters, viz., an average distance gradient change that we propose, and an approach based on the average silhouette criterion[27,28]. We study patterns of the comovement of clustered stocks over multiple trading days under the two clustering methods with the two different MI estimation approaches, and discuss similarities and differences among them. We carry out interesting post-clustering analyses related to portfolio allocations based on our clustering schemes compared with that directly based on the market without pre-clustering. We also analyze patterns within and between various industry sectors based on the jackknife estimate of MI between *groups* of stocks.

The remainder of the paper is organized as following. Section 2 reviews the definition and properties of MI. Section 3 describes the computation of the empirical MI between stock returns using the KDE based jackknife approach. Section 4 describes MI based clustering using two different approaches for selecting the optimal number of clusters, an average distance gradient change method, and a method based on the average silhouette criterion. Section 5 describes results for intra-day stock returns based on the clustering algorithms of Section 4. Section 6 presents an interesting post-hoc use of the cluster based information for portfolio allocation. Section 7 describes the computation of groupwise MI to analyze behavior between and within industry grouping provided by the Global Industry Classification Standard (GICS). Section 8 provides a discussion and summary.

2 | MUTUAL INFORMATION

The mutual information (MI) between two random variables is a generalization of the correlation between them and measures the dependence between the two variables[14]. MI is also known as "information gain", and has been related to the joint entropy of the random variables. Let $p(x)$ denote the probability mass function (pmf) of a discrete random variable X, and let $f(x)$ be the probability density function (pdf) of a continuous random variable. Likewise, let $p(x, y)$ and $f(x, y)$ denote the joint pmf and pdf respectively of two random variables X and Y. The entropy of X measures the quantity of information it contains and

is defined as

$$H(X) = \begin{cases} -\sum_x p(x)\log\left(p(x)\right) & \text{if } X \text{ is discrete,} \\ -\int f(x)\log\left(f(x)\right)dx & \text{if } X \text{ is continuous.} \end{cases} \tag{1}$$

The joint entropy of two random variables X and Y is

$$H(X,Y) = \begin{cases} -\sum_x \sum_y p(x,y)\log\left(p(x,y)\right) & \text{if } X \text{ and } Y \text{ are discrete,} \\ -\int\int f(x,y)\log\left(f(x,y)\right)dxdy & \text{if } X \text{ and } Y \text{ are continuous.} \end{cases} \tag{2}$$

The MI between X and Y measures the information shared by the two random variables and is defined as

$$\begin{aligned} MI(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= \begin{cases} \sum_x \sum_y p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right), & \text{discrete case} \\ \int\int f(x,y)\log\left(\frac{f(x,y)}{f(x)f(y)}\right)dxdy, & \text{continuous case.} \end{cases} \end{aligned} \tag{3}$$

Next, consider two groups of random variables, $\mathbf{X} = (X_1, X_2, \cdots, X_P)'$ and $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_Q)'$[21]. The MI between $\mathbf{X}$ and $\mathbf{Y}$ is

$$MI(\mathbf{X},\mathbf{Y}) = E\left\{\log\frac{f_{\mathbf{XY}}(\mathbf{X},\mathbf{Y})}{f_{\mathbf{X}}(\mathbf{X})f_{\mathbf{Y}}(\mathbf{Y})}\right\}. \tag{4}$$

Note that this definition is equivalent to the definition in equation (3), except that here, $\mathbf{X}$ and $\mathbf{Y}$ are respectively vectors of length P and Q. Since $MI(X,Y)$ is a special case of $MI(\mathbf{X},\mathbf{Y})$ when $P = 1$ and $Q = 1$, we use the notation $MI(\mathbf{X},\mathbf{Y})$ below.

A large value of $MI(\mathbf{X},\mathbf{Y})$ indicates that the two random variables are highly associated, while a low value indicates low association. If and only if $MI(\mathbf{X},\mathbf{Y}) = 0$, the two variables are independent. While the MI defined in equations (3) and (4) is not a distance metric, it can be converted to a distance metric:

$$d(\mathbf{X},\mathbf{Y}) = H(\mathbf{X},\mathbf{Y}) - MI(\mathbf{X},\mathbf{Y}), \tag{5}$$

which satisfies the non-negativity, symmetry and triangle inequality properties. The normalized distance is then defined as

$$D(\mathbf{X},\mathbf{Y}) = 1 - \frac{MI(\mathbf{X},\mathbf{Y})}{H(\mathbf{X},\mathbf{Y})}, \tag{6}$$

so that $0 \le D(\mathbf{X},\mathbf{Y}) \le 1$.

3 | ESTIMATION OF PAIRWISE/GROUPWISE MUTUAL INFORMATION BETWEEN STOCKS

To compute the MI between stock returns, we employ a jackknife version of the kernel estimate with equalized bandwidth and allow the bandwidth to vary over an interval. The estimated MI is the largest value among these kernel estimates[21] and we summarize its computation in the following steps.

Step 1: Consider two random vectors $\mathbf{X} = (X_1, X_2, \cdots, X_P)'$ and $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_Q)'$. Let $C = (\mathbf{x}_t, \mathbf{y}_t), t = 1, \cdots, T$ with $\mathbf{x}_t = (x_{t1}, x_{t2}, \cdots, x_{tP})'$ and $\mathbf{y}_t = (y_{t1}, y_{t2}, \cdots, y_{tQ})'$ denoting independent samples from $(\mathbf{X},\mathbf{Y})$.

Step 2: Consider the uniform transformation $\mathbf{U} = (U_1, U_2, \cdots, U_P)' = (F_{X_1}(X_1), F_{X_2}(X_2), \cdots, F_{X_P}(X_P))'$ and $\mathbf{V} = (V_1, V_2, \cdots, V_Q)' = (F_{Y_1}(Y_1), F_{Y_2}(Y_2), \cdots, F_{Y_Q}(Y_Q))'$, where $F_{X_p}(.)$ and $F_{Y_q}(.)$, are the cumulative distribution functions of X_p and Y_q respectively for $p = 1, 2, \cdots, P$ and $q = 1, 2, \cdots, Q$. The MI between $\mathbf{X}$ and $\mathbf{Y}$ is invariant to such transformations, i.e., $MI(\mathbf{U},\mathbf{V}) = MI(\mathbf{X},\mathbf{Y})$.

Step 3: We use $c_{\mathbf{U}}(\mathbf{u})$, $c_{\mathbf{V}}(\mathbf{v})$ and $c_{\mathbf{UV}}(\mathbf{u},\mathbf{v})$ to denote the copula density functions of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{UV}$ respectively. The corresponding transformation of the observed data is $(\mathbf{u}^\star{}_t, \mathbf{v}^\star{}_t)' = (u^\star_{t1}, \cdots, u^\star_{tP}, v^\star_{t1}, \cdots, v^\star_{tQ})' = (F_{X_1,T}(x_1), \cdots, F_{X_P,T}(x_P), F_{Y_1,T}(y_1), \cdots, F_{Y_Q,T}(y_Q))'$.

Step 4: Let $\mathbf{K}^P$ denote a P-dimensional symmetric density function with $\mathbf{K}^P(\mathbf{x}) = \Pi_{p=1}^P K(x_p)$. For a diagonal matrix $\mathbf{A}$, $\mathbf{K_A}^P(\mathbf{x}) = |\mathbf{A}|^{-1/2}\mathbf{K}^P(\mathbf{A}^{-1/2}\mathbf{x})$. The MI estimator with one common bandwidth h is

$$\hat{I}(h) = \frac{1}{T}\sum_{t=1}^{T}\log\frac{\hat{c}^{\setminus t}_{\mathbf{UV},h^2\mathbf{I}_P,h^2\mathbf{I}_Q}(\mathbf{u}^*{}_t, \mathbf{v}^*{}_t)}{\hat{c}^{\setminus t}_{\mathbf{U},h^2\mathbf{I}_P}(\mathbf{u}^*{}_t)\hat{c}^{\setminus t}_{\mathbf{V},h^2\mathbf{I}_Q}(\mathbf{v}^*{}_t)}, \tag{7}$$

where $\mathbf{I}_P$ and $\mathbf{I}_Q$ are $P \times P$ and $Q \times Q$ identity matrices, and

$$\hat{c}^{\setminus t}_{\mathbf{U},h^2\mathbf{I}_P}(\mathbf{u}) = \frac{1}{T-1}\sum_{t' \neq t}^{T} \mathbf{K}^P_{h^2\mathbf{I}_P}(\mathbf{u}^\star{}_{t'} - \mathbf{u});$$

$$\hat{c}^{\setminus t}_{\mathbf{V},h^2\mathbf{I}_Q}(\mathbf{v}) = \frac{1}{T-1}\sum_{t' \neq t}^{T} \mathbf{K}^Q_{h^2\mathbf{I}_Q}(\mathbf{v}^\star{}_{t'} - \mathbf{v});$$

$$\hat{c}^{\setminus t}_{\mathbf{UV},h^2\mathbf{I}_P,h^2\mathbf{I}_Q}(\mathbf{u},\mathbf{v}) = \frac{1}{T-1}\sum_{t' \neq t}^{n} \mathbf{K}^P_{h^2\mathbf{I}_P}(\mathbf{u}^\star{}_{t'} - \mathbf{u})\mathbf{K}^Q_{h^2\mathbf{I}_Q}(\mathbf{v}^\star{}_{t'} - \mathbf{v}).$$

Step 5: The Jacknife estimator of MI is

$$\widehat{\mathrm{JMI}}\,(\mathbf{X},\mathbf{Y}) = \max_{h>0}[\hat{I}(h)]. \tag{8}$$

These formulas apply directly to pairwise MI between two stocks when each group has one stock, i.e., $P = Q = 1$. Let $\mathcal{S} = \{1,\ldots,S\}$ denote a set of S stocks that are observed over T time intervals (say, one-minute intervals) over N trading days. On the nth trading day, we observe log returns $r_{i,t,n}$ in the tth time interval for the ith stock, for $t = 1,\ldots,T$ and $i = 1,\ldots,S$. The MI that measures the similarity between the ith stock and the i'th stock within trading day n is computed as

$$\widehat{\mathrm{JMI}}\,(s_i, s_i') = \max_{h>0}[\hat{I}(h)], \tag{9}$$

with $x_t = r_{i,t,n}$ and $y_t = r_{i',t,n}$ as inputs.

We can then compare the "jackknife estimate" (JMI) of the MI between two stocks with the MI estimate obtained from the binning method described in the Appendix. As discussed later, while the binning method is computationally much faster than the "jackknife estimate", JMI is a bias-reduced estimate of MI. Later in this paper, we will investigate whether the properties of these two different estimates of MI impact financial metrics such as the Sharpe ratio[25] or diversification index[26].

4 | CLUSTERING ALGORITHMS

This section describes clustering the stocks on each trading day using two different approaches. On trading day n, let K_n^* denote the optimum number of clusters, and $K_{\max}$ denote a pre-determined maximum number of possible clusters, and $C_{\kappa,n}$ denote cluster κ, for $\kappa = 1,\ldots,K_n^*$. Section 4.1 describes a new clustering algorithm that we propose based on an average distance gradient change (ADGC) criterion. It improves upon a clustering approach based on the average silhouette criterion[27,28] discussed in Section 4.2. Although the average silhouette criterion (see the Appendix) is widely used for selecting the number of clusters, our experience shows that it a) tends to favor a very small number, i.e., 2 clusters in many cases, and b) is computationally intensive.

Although a few steps in these two algorithms overlap (i.e., Step 1, Steps 2a-2c and Step 3), the two methods are inherently different. One main difference is that the average silhouette criterion requires the user to prespecify $K_{\max}$, and find the optimum number of clusters K_n^* with the largest average silhouette $\overline{Sil}_K$ for $K = 2,\ldots,K_{\max}$. Our proposed method detects a gradient change in a sequence of average distances of cluster seeds and only takes $\frac{1}{K_{\max}-1}$ computing time compared to the average silhouette method.

We will run each clustering algorithm with each type of MI estimation to see which combination gives the best financial results. Let $\mathcal{S} = \{1,\ldots,S\}$ denote a set of S stocks. Let $g = 2,3,\ldots$ denote the current number of clusters at any stage of the algorithm. Let $\mathcal{X}$ denote the corresponding set of seeds for the clusters, and let $|\mathcal{X}|$ denote its cardinality, i.e., the number of seeds in the set $\mathcal{X}$.

4.1 | Average Distance Gradient Change Algorithm

The average distance gradient change method consists of the following steps.

Step 1: Select seeds for the first two clusters.
Step 1a: Set $g = 2$.
Step 1b: Compute the pairwise mutual information based distances $\hat{D}(s_i, s_i')$ for two different stocks $s_i, s_i' \in \mathcal{S}$, $i \neq i'$.
Step 1c: Find the pair of stocks with the largest distance $\bar{D}^2_{\max} = \max_{s_i,s_i' \in \mathcal{S}} \hat{D}(s_i, s_i')$, and denote these as x_1 and x_2.

Step 1d: Select x_1 as the seed for the first cluster $C_{1,n}$ and x_2 as the seed for the second cluster $C_{2,n}$. Remove these "seed" stocks x_1 and x_2 from the "stocks set" S, and denote the current set of remaining stocks as S^g (with cardinality $S-2$). Set the current seed set as $\mathcal{X}^g = \{x_1, x_2\}$ (with cardinality 2).

Step 2: Determine number of clusters and corresponding seeds for successive clusters.
Step 2a: Compute $\hat{D}(s_i, x_i)$ for all $s_i \in S^g$ and $x_i \in \mathcal{X}^g$.
Step 2b: Compute the average distance between stock s_i and all the seeds in set $\mathcal{X}^g$ as

$$\bar{D}_{s_i} = \sum_{x_i \in \mathcal{X}^g} \frac{\hat{D}(s_i, x_i)}{|\mathcal{X}^g|}$$

where $|\mathcal{X}^g|$ denotes the number of seeds in the set $\mathcal{X}^g$.
Step 2c: Find the stock with the largest average distance from all current clusters, given by $\bar{D}^{g+1}_{\max} = \max_{s_i \in S^g}(\bar{D}_{s_i})$
Step 2d: Compare $\bar{D}^{g}_{\max}$ and $\bar{D}^{g+1}_{\max}$.

- If $\bar{D}^{g}_{\max} \geq \bar{D}^{g+1}_{\max}$, set x_{g+1} as the seed for a new $(g+1)$th cluster. Move x_{g+1} from the set S^g to the current seed set $\mathcal{X}^g$. Denote the updated seed set and stocks set as $\mathcal{X}^{g+1}$ and S^{g+1} respectively. Set $g = g+1$, and repeat from Step 2b to Step 2d until $\bar{D}^{g}_{\max} < \bar{D}^{g+1}_{\max}$.
- If $\bar{D}^{g}_{\max} < \bar{D}^{g+1}_{\max}$, conclude there are $K^*_n = g$ clusters.

Step 3: Assign stocks to the K^*_n clusters.
Step 3a: For each stock $s_i \in S^g$, calculate the distance $\hat{D}(s_i, s_f)$ for all $s_f \in C_{\kappa,n}$, $\kappa = 1, \cdots, K^*_n$.
Step 3b: Record the value $D_{s_i, C_{\kappa,n}} = \max_{s_f \in C_{\kappa,n}}(\hat{D}(s_i, s_f))$ as the distance between stock s_i and cluster $C_{\kappa,n}$.
Step 3c: Find the cluster that has smallest distance with stock s_i, i.e., $\min_{\kappa=1,\cdots,K^*_n}(D_{s_i, C_{\kappa,n}})$, denote this cluster as $C_{\kappa_i,n}$.
Step 3d: Find the stock that has the smallest distance with the cluster, i.e, s_j has the smallest distance, $\min_{i=1,\cdots,S}(C_{\kappa_i,n})$. Add stock s_j to cluster $C_{\kappa_j,n}$, remove stock s_j from S^g, and denote the updated stocks set as S^{g+1}. Set g=g+1.
Step 3e: Repeat from Step 3a to Step 3d until S_g is empty.

4.2 | Average Silhouette Criterion Method-Algorithm

The Appendix provides a brief discussion of the average silhouette criterion. The clustering approach consists of the following steps.

Step 1: The same as that under the ADGC algorithm.

Step 2: Set the number of clusters as K, and determine seeds for successive $K-2$ clusters.
Step 2a-2c: The same as that under the ADGC algorithm.
Step 2d: Set x_{g+1} as the seed for a new $(g+1)$th cluster. Move x_{g+1} from the set S^g to the current seed set $\mathcal{X}^g$. Denote the updated seed set and stocks set as $\mathcal{X}^{g+1}$ and S^{g+1} respectively. Set $g = g+1$.
Step 2e: Repeat from Step 2a to Step 2d until $g = K$.

Step 3: The same as that under the ADGC algorithm except using K to replace K^*_n.
Step 4: Compute the average silhouette value. Compute $Sil(s_{i,C_{\kappa,n}})$ for all the S stocks. Denote the average silhouette for all the S stocks as

$$\overline{Sil}_K = \sum_{\kappa=1}^{K} \sum_{i=1}^{S} \frac{Sil(s_{i,C_{\kappa,n}})}{S}.$$

Step 5: Find the optimal number of clusters. Vary the number of clusters $K = 2, \cdots, K_{\max}$, repeat Steps 2-6, find the value K with $\max_{K=2,\cdots,K_{\max}}(D_{s_i, C_{\kappa,n}}(\overline{Sil}_K)$ to determine the optimal number of clusters K^*_n.

5 | CLUSTERING INTRA-DAY STOCK RETURNS

We use the S&P 100 data downloaded from the Trade and Quote (TAQ) database of Wharton Research Data Services (WRDS). We only analyze companies for which the entire 2013 data is available. The stock names with their symbols according to the Global Industry Classification Standard (GICS) classification are shown in Table 1. The data set consists of prices of 94 stocks from the S&P100 index from 9:30 am to 4:00 pm for 249 trading days in 2013. We pre-average the stock prices to one-minute intervals within a day and then analyze the logarithms of the returns constructed from the one-minute averaged stock prices. The pre-processing of the data is similar to Liu et al[29].

5.1 | Clustering using MI

We implement each clustering algorithm described in Section 4 using both methods for estimating MI that we discussed in Section 3. The binning estimation of MI and the two clustering algorithms are implemented in Python, while we use the R package *JMI* for the JMI estimation. In the rest of the paper, we refer to the four methods as Bin-A, Bin-S, JMI-A and JMI-S, where

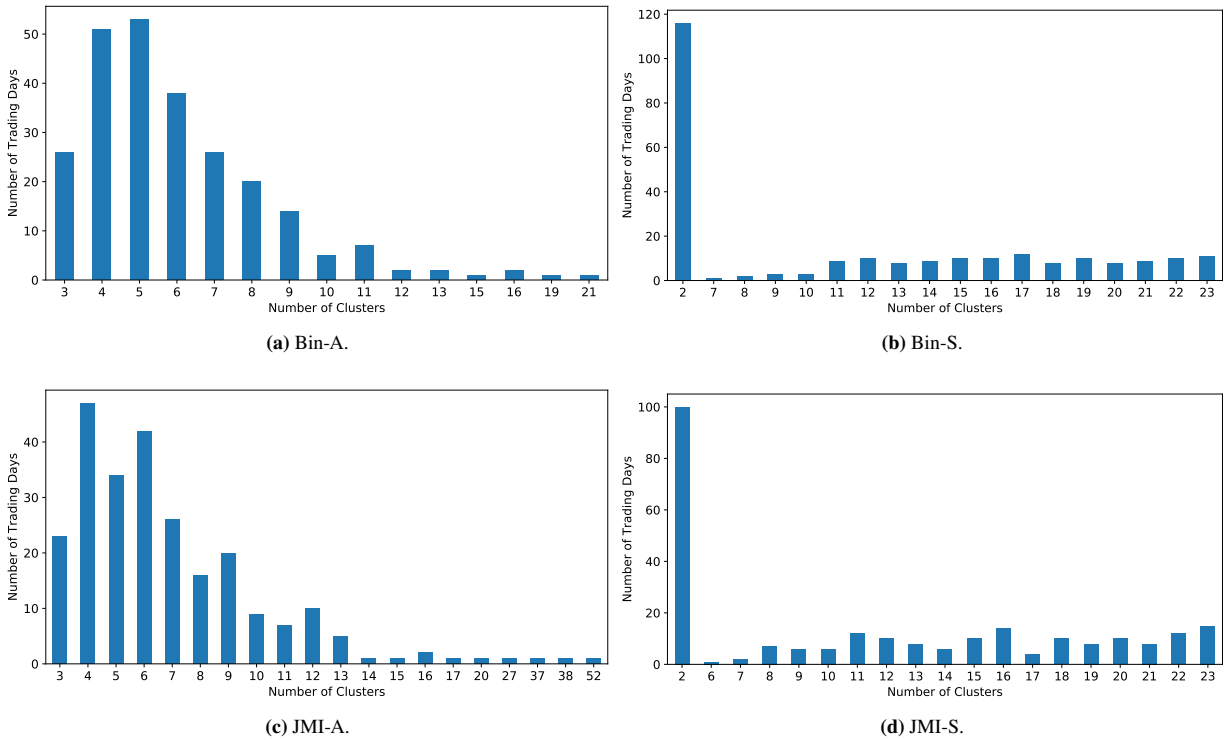
Bin-A: clusters by the ADGC method using the binning estimation of MI.

Bin-S: clusters by the Silhouette method using the binning estimation of MI.

JMI-A: clusters by the ADGC method using the JMI estimation of MI.

JMI-S: clusters by the Silhouette method using the JMI estimation of MI.

Figure 1 shows the distribution of the number of clusters over all the $N = 249$ trading days in 2013. For Bin-A, the number of clusters K_n^* ranges from 3 to 11 on most trading days, the mode being 4 or 5 clusters and the distribution having a long tail indicating a few trading days with more than 12 clusters. For JMI-A, the number of clusters is less than 13 for most trading days, and the distribution of K_n^* has a long tail as well. For both Bin-S and JMI-S, the algorithms choose two clusters on most trading days, while for the remaining trading days, K_n^* ranges from 10 to 23.



(a) Bin-A. **(b)** Bin-S.

(c) JMI-A. **(d)** JMI-S.

FIGURE 1 Distribution of K_n^* over trading days in 2013

TABLE 1 S&P 100 stock symbols and companies (Co.) or corporations (Corp.) in ten different GICS sectors

Sector	Symbol	Company	Sector	Symbol	Company
Consumer Discretionary	AMZN	Amazon.Com Inc	Consumer Staples	CL	Colgate-Palmolive Co.
	DIS	Walt Disney Co.		COST	Costco Wholesale Corp.
	F	Ford Motor Co.		CVS	CVS Corp.
	HD	Home Depot		KO	Coca-Cola Co.
	LOW	Lowe's Co.		MDLZ	Mondelez Intl. Inc. Cl A
	MCD	McDonald's Corp.		MO	Altria Group
	NKE	Nike Inc.		PEP	Pepsico Inc.
	SBUX	Starbucks Corp.		PG	Procter & Gamble Co.
	TGT	Target Corp.		PM	Philip Morris Intl. Inc.
	TWX	Time Warner Inc.		WAG	Walgreens Boots Alliance Inc.
				WMT	Wal-Mart Stores
Financials	ALL	Allstate Corp.	Health Care	ABBV	AbbVie Inc.
	AXP	American Express Co.		ABT	Abbott Laboratories
	BAC	Bank of America Corp.		AMGN	Amgen Inc.
	BK	Bank of New York Mellon Corp.		BAX	Baxter Intl. Inc.
	C	Citigroup Inc.		BMY	Bristol-Myers Squibb Co.
	COF	Capital One Financial Corp.		GILD	Gilead Sciences
	GS	Goldman Sachs Group		JNJ	Johnson & Johnson
	JPM	JP Morgan Chase & Co.		LLY	Eli Lilly and Co.
	MET	Metlife Inc.		MDT	Medtronic Inc.
	MS	Morgan Stanley		MRK	Merck & Co.
	SPG	Simon Property Group		PFE	Pfizer Inc.
	USB	U.S. Bancorp		UNH	United Health Care Group
	WFC	Wells Fargo & Co.			
Information Technology	AAPL	Apple Inc.	Industrials	BA	Boeing Co.
	ACN	Accenture PLC		CAT	Caterpillar Inc,
	CSCO	Cisco Systems		EMR	Emerson Electric Co.
	EBAY	Ebay Inc.		FDX	Fedex Corp.
	EMC	EMC Corp.		GD	General Dynamics Corp,
	GOOG	Google Inc.		GE	General Electric Co.
	HPQ	Hewlett-Packard Co.		HON	Honeywell Intl. Inc.
	IBM	Intl. Business Machines		LMT	Lockheed Martin Corp.
	INTC	Intel Corp.		MMM	3M Co.
	MA	Mastercard Inc.		NSC	Norfolk Southern Corp.
	MSFT	Microsoft Corp.		RTN	Raytheon Co.
	ORCL	Oracle Corp.		UNP	Union Pacific Corp.
	QCOM	Qualcomm Inc.		UPS	United Parcel Service
	TXN	Texas Instruments Inc.		UTX	United Technologies Corp.
	V	Visa Inc.			
Energy	AEP	American Electric Power	Materials	DD	E.I. DuPont De Nemours and Co.
	APA	Apache Corporation		DOW	DOW Chemical Co.
	APC	Anadarko Petroleum Corp		MON	Monsanto Co.
	COP	Conocophillips			
	CVX	Chevron Corp.			
	DVN	Devon Energy Corp.			
	FCX	Freeport-McMoRan			
	HAL	Halliburton Co.	Telecom	T	A T & T Inc.
	NOV	National Oilwell Varco		VZ	Verizon Communications Inc.
	OXY	Occidental Petroleum Corp.			
	SLB	Schlumberger N.V.			
	XOM	Exxon Mobil Corp			
Utilities	EXC	Exelon Corp.			
	SO	Southern Co.			

We have carried out a comparison of the computing times for the two MI estimation methods on the R platform by using the R package *infotheo* for the "binning" method. For 94 stocks each with time length $T = 389$, the computing time for all the pairwise MI under the binning method is 2.3 seconds, while that for the JMI approach is 9.5 minutes (the resampling process is expensive). The advantage of JMI is the bias reduction it offers, and we expect that distributed computing across many cores can significantly reduce the JMI computing time.

5.2 | Comovement Probability

Post clustering, we explore interesting patterns in selected m-tuples of stocks which fall within the same cluster in a calendar week w in 2013, for $w = 1, \ldots, W$. To do this, we employ the cumulative comovement empirical probability[29],

$$P(w_c) = \frac{\sum_{w=1}^{w_c} \tau(w)}{\sum_{w=1}^{w_c} T(w)}, \tag{10}$$

where for $w = 1, \ldots, W$, $T(w)$ denotes the number of trading days in calendar week w and $\tau(w)$ denotes the number of trading days for which the m-tuple of stocks falls within the same cluster in week w. We select stocks that belong to certain industry sectors as defined by the Global Industry Classification Standard (GICS). Table 2 shows the top $m = 2$ pairs of stocks (in each GICS sector) with the largest comovement empirical probability defined in equation (10). For instance, for two stocks in the Consumer Discretionary sector, DIS (Walt Disney Co) and EXC (TWX Group Holding Ltd), we see that the value is 0.4337. We also observe that the top $m = 2$ pairs of stocks in each GICS sector that are selected by the Bin-A, Bin-S, JMI-A and JMI-S methods are very similar, suggesting that the comovement behavior of stocks is robust to the MI estimation methods and the clustering methods we employed.

TABLE 2 Top $m = 2$ comovement pairs in each GICS sector

Sectors	Bin-A			Bin-S			JMI-A			JMI-S		
	S1	**S2**	**P(%)**	**S3**	**S4**	**P(%)**	**S5**	**S6**	**P(%)**	**S7**	**S8**	**P(%)**
Consumer	DIS	TWX	43.37	DIS	TWX	49.40	DIS	TWX	49.00	LOW	HD	47.79
Discretionary	SBUX	TWX	42.57	LOW	HD	46.59	HD	SBUX	46.18	DIS	TWX	46.99
Consumer	PEP	KO	44.18	MO	PM	57.03	PM	PEP	43.78	PM	MO	49.00
Staples	CL	PG	45.78	CL	PG	60.24	PM	MO	47.39	KO	PEP	50.20
Energy	SLB	HAL	59.44	COP	CVX	69.08	XOM	CVX	54.62	COP	CVX	66.67
	XOM	CVX	57.83	COP	APC	67.07	COP	APC	54.22	CVX	APC	64.66
Financials	USB	C	29.72	AXP	MA	34.54	WFC	GS	31.33	AXP	MA	35.34
	AXP	MA	28.11	BK	COF	33.33	JPM	USB	30.92	BK	ALL	33.33
Health	LLY	JNJ	40.56	LLY	JNJ	47.79	BMY	LLY	47.39	JNJ	PFE	51.00
	JNJ	BMY	39.76	LLY	MRK	45.78	JNJ	MDT	46.18	LLY	PFE	49.80
Information	IBM	ORCL	41.77	INTC	TXN	46.99	ORCL	IBM	42.97	INTC	TXN	48.19
Technology	ACN	ORCL	38.55	IBM	ORCL	44.98	IBM	EBAY	42.17	QCOM	ORCL	46.99
Industrials	UNP	NSC	47.39	HON	EMR	53.01	HON	MMM	47.79	HON	LMT	55.42
	MMM	HON	50.60	MMM	EMR	54.22	MMM	EMR	49.80	HON	MMM	55.42
Materials	DD	DOW	34.14	DD	DOW	45.38	DOW	DD	42.57	DOW	DD	44.18
	DD	MON	33.73	DOW	MON	39.36	DOW	MON	31.73	DD	MON	41.37
Telecom	VZ	T	30.52	VZ	T	44.58	T	VZ	32.13	T	VZ	43.37
Utilities	SO	EXC	41.77	EXC	SO	60.24	EXC	SO	42.17	SO	EXC	61.04

P is the cumulative comovement proportion (up to week 53) defined in equation (10). For Telecom and Utilities sectors, there are only two stocks in the S&P 100. Accordingly, only one pair of co-moving stocks is shown in this table for these sectors.

6 | PORTFOLIO ALLOCATION

Portfolio allocation refers to the problem of allocating an available investment outlay to different stocks. The most well-known solution is the "Modern Portfolio Theory" framework introduced by Markowitz[30].

Suppose a portfolio contains O stocks. Let $\mathbf{w} \in \mathcal{R}^O$ denote the vector of proportions of the O stocks in the portfolio. Let $\mathbf{r} \in \mathcal{R}^O$ denote the vector of market returns of the stocks in the portfolio and let $\mathbf{\Sigma} \in \mathcal{R}^{O\times O}$ denote the covariance matrix of these returns.

The Markowitz criterion for portfolio allocation states that the optimal allocation weights of the efficient (tangent) portfolio maximize

$$\begin{aligned} \mathbf{w}^* = \underset{\mathbf{w}\in[\mathbf{0},\mathbf{1}]}{\arg\max}\ & \frac{\mathbf{r}^T\mathbf{w}}{\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}} \\ \text{s.t. } & \mathbf{1}^T\mathbf{w} = 1. \end{aligned} \tag{11}$$

One shortcoming of this portfolio allocation approach is that it magnifies the estimation error[31] and also suffers from the curse of dimensionality[32]. Clustering is often employed to overcome such difficulties by grouping homogeneous stocks, thus reducing the dimensionality of the assets[33,34].

In Section 6.1, we describe our implementation of a post clustering portfolio allocation scheme, denoted by Method C. In Section 6.2, we compare the results from Method C with results from portfolio allocation using the Markowitz approach (Method M) and an equal weights approach (Method E).

6.1 | Post Clustering Portfolio Allocation Scheme

We describe an approach for portfolio allocation based on clustering the stocks.

Method C. For each trading day n, cluster S stocks into homogeneous groups using one of these approaches described in Section 4, i.e., Bin-A, Bin-S, JMI-A, or JMI-S. Recall that the number of clusters K^* is possibly different for these four methods. For simplicity, in what follows, we use K^* to denote the number of clusters in the chosen method.

Let C_κ denote cluster κ, for $\kappa = 1, 2, \ldots, K^*$. Let H_κ denote the number of stocks in cluster C_κ, for $j = 1, \cdots, H_\kappa$. Let $r_{\kappa,j} = \frac{1}{T}\sum_{t=1}^{T} r_{j,t}$ denote the average of the log returns for stock s_j in cluster C_κ, and let $\mathbf{r}_\kappa^T = (r_{\kappa,1}, r_{\kappa,2}, \cdots, r_{\kappa,H_\kappa})$. We first compute the within cluster weights $\mathbf{w}_\kappa^T = (w_{\kappa,1}, w_{\kappa,2}, \cdots, w_{\kappa,H_\kappa})$, where $w_{\kappa,j}$ denote the weight of stock s_j in cluster C_κ, then compute the cluster weights $\mathbf{w}_c^T = (w_1, w_2, \cdots, w_\kappa)$. Finally, the cluster adjusted weights $\mathbf{w}^{*T} = (w_1^*, w_2^*, \cdots, w_S^*)$ can be obtained by the product of within cluster weights and the cluster weights.

1. Portfolio allocation within each cluster. With the log returns $\mathbf{r}_\kappa$, we use equation (11) to decide the optimal weights $\mathbf{w}_\kappa$ for each stock within cluster C_κ.

2. Portfolio allocation across K^* clusters. With the within cluster weights $\mathbf{w}_\kappa$, we can compute the κth cluster (portfolio) return, $\bar{r}_\kappa = \mathbf{r}_\kappa^T\mathbf{w}_\kappa$, and $\mathbf{r}$ of all the K^* clusters, $\mathbf{r}^T = (\bar{r}_1, \bar{r}_2, \cdots, \bar{r}_{K^*})$. Using $\mathbf{r}$ as an input to equation (11), the cluster weights $\mathbf{w}_c$ are achieved.

3. Find the cluster adjusted weights. The cluster adjusted weight w_i^* of stock s_i in cluster C_κ is the product of the weight $w_{\kappa,j}$ of the stock within the cluster and the cluster weight w_κ, $w_i^* = w_{\kappa,j} \cdot w_\kappa$.

Repeat the above steps for each of Bin-A, Bin-S, JMI-A, and JMI-S, and denote C-Bin-A, C-Bin-S, C-JMI-A, and C-JMI-S as the above portfolio allocation strategy under these clustering methods.

6.2 | Comparison of Portfolio Allocation Schemes

We compare our post clustering portfolio allocation scheme with two well-known schemes, Markowitz approach (Method M) and an equal weights approach (Method E). We briefly describe the two methods below.

Method M. Let $\bar{r}_i = \frac{1}{T}\sum_t^T r_{i,t}$ denote the average log return of stock s_i, and let $\mathbf{r}^T = (\bar{r}_1, \bar{r}_2, \cdots, \bar{r}_S)$.

Using $\mathbf{r}$ as an input to equation (11), the optimal weights $\mathbf{w}^*$ are achieved for each stock.

Method E. Assign equal weights $\mathbf{w}^{*T} = (\frac{1}{S}, \frac{1}{S}, \cdots, \frac{1}{S})$ to all the stocks.

To evaluate the portfolio allocation performance, we employ two well-known metrics, the Sharpe ratio[25] and the diversification index (DI)[26]. The Sharpe ratio measures the return given a unit volatility and is defined as

$$SR = \frac{r_p - r_f}{\sigma_p}, \tag{12}$$

where r_p is the portfolio return, r_f is the risk-free rate and σ_p is the standard deviation of the portfolio return. For simplicity, we assume that the risk-free rate is constant during the year, and the Sharpe ratio is

$$SR = \frac{r_p}{\sigma_p}. \tag{13}$$

The second metric is the diversification index (DI), which measures how well the portfolio is diversified and is defined as

$$\mathrm{DI} = 1 - \sum_{s=1}^{O} W_s^2, \tag{14}$$

where W_s is the proportion of the portfolio market value invested in stock s and O is the number of stocks in the portfolio.

A comparison of the three portfolio allocation methods, Method C, Method M and Method E using the Sharpe ratio and the DI is shown in Table 3 for January 2013. Overall, the Sharpe ratios under Method C (denoted by C-Bin-A, C-Bin-S, C-JMI-A, and C-JMI-S) are close to the corresponding ratios under Method M, the market portfolio. However, the corresponding Sharpe ratios under Method E, the equal weights portfolio, are extremely low. For the DI, Table 3, Figure 2 and Figure 3 indicate that for several trading days, the diversification under Method C is much better than the corresponding diversification under Method M, but worse than under Method E. Method E may not be a good portfolio allocation strategy since the gains of a relatively larger DI over the other methods is much smaller than that of the loss of the Sharpe ratio unless the investors are extremely risk-averse. The means of the Sharpe ratios and the DI over different months are shown in Table 4 and indicate patterns similar to those discussed above.

Method C can overcome the curse of dimensionality from which Method M suffers. To evaluate the performance under Method C, we count the number of trading days of the maximum of C-Bin-A, C-Bin-S, C-JMI-A, and C-JMI-S for the Sharpe ratio and DI during the year 2013. Specifically, if we denote $SR_{C\_Bin\_A}, SR_{C\_Bin\_S}, SR_{C\_JMI\_A}, SR_{C\_JMI\_S}$ and SR_M as the Sharpe ratio for C-Bin-A, C-Bin-S, C-JMI-A, C-JMI-S strategy and the market, we use $D_{C\_Bin\_A} = (SR_M - SR_{C\_Bin\_A})/SR_M$ to measure the distance between the Sharpe ratio of strategy C-Bin-A and that of the market, similarly for C-Bin-S, C-JMI-A and C-JMI-S. If a value that is larger than one interquartile range (IQR), we think this Sharpe ratio is significantly large, and count 1, for example, for a trading day, if $D_{C\_Bin\_A} > IQR$, then we count 1 for C-Bin-A, otherwise 0. Similarly, we can have the counts for the DI. To combine the effect of the Sharpe ratio and the DI, we employ a homotopy of the counts of the Sharpe ratio and the DI, that is, $C = (1-\lambda) * C_{SR} + \lambda * C_{DI}$, if we denote C_{SR} as the counts for the Sharpe ratio and C_{DI} as the counts for the DI, and λ is a tuning parameter. A larger value of the counts C is desirable. Table 5 shows the counts C for C-Bin-A, C-Bin-S, C-JMI-A, C-JMI-S strategy with different value λ. C-JMI-A achieves the larger counts with relatively small value λ compare with C-Bin-A, where the Sharpe ratio is heavily weighted. No strategy is uniformly better than the others. The best strategy can be chosen by the investors according to their risk preferences. For the aggressive investors, a smaller λ is desirable with more weights on the Sharpe ratio than the DI, in contrast, the risk-averse investors may choose a larger value of λ, where the DI is heavily weighted.

7 | GROUPWISE MI OF CROSS AND WITHIN VARIOUS SECTORS

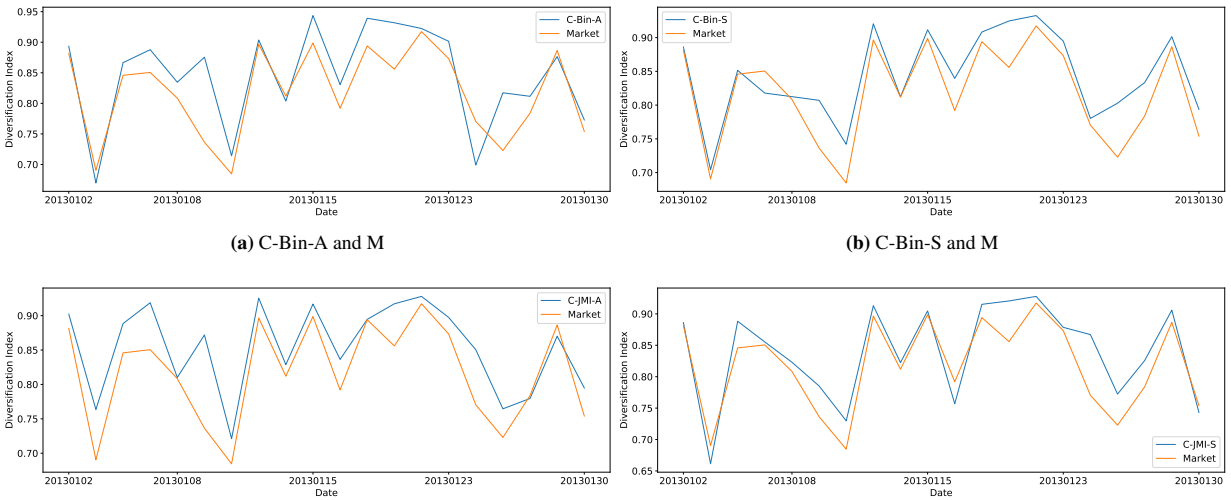
The Global Industry Classification Standard (GICS) is a standardized classification system of stocks, and it classifies the companies mainly based on their revenues, earnings, and market perception, however, such a classification system may miss many other factors, such as tax, location, and economic policies. It could be interesting to study whether such stocks within sectors hold together according to their MI, and the MI patterns of the cross and within various sectors based on the GICS groups under JMI estimation. In this section, we studied groupwise MI of cross and within various sectors.

Groupwise MI trend. Figure 4 shows 45 pairs of MI trends of ten groups. There are two clusters of the trends, the top clusters with most pairwise MI from 2 to 3, although there are several drops at some trading days, another cluster of the trends with most pairs MI below 1.5.

Network structure of sectors. Figure 5 shows the network structure of the ten sectors by averaging the MI over the year 2013. The size of each node indicates the total MI of this node with each of the rest nodes. The width of the edges indicates the

TABLE 3 Sharpe ratio and DI comparison of the three portfolio allocation methods

Date	Sharpe Ratio						DI					
	C-Bin-A	C-Bin-S	C-JMI-A	C-JMI-S	M	E	C-Bin-A	C-Bin-S	C-JMI-A	C-JMI-S	M	E
20130102	4.118	4.059	3.969	4.015	4.213	0.186	0.893	0.886	0.902	0.886	0.881	0.971
20130103	3.028	3.010	3.002	3.037	3.051	0.010	0.670	0.704	0.764	0.661	0.691	0.971
20130104	3.957	3.951	3.849	3.790	3.985	0.058	0.867	0.851	0.888	0.888	0.846	0.972
20130107	3.584	3.628	3.661	3.877	3.949	0.062	0.888	0.818	0.919	0.855	0.851	0.972
20130108	3.406	3.443	3.414	3.496	3.558	0.001	0.835	0.813	0.810	0.823	0.808	0.972
20130109	3.799	4.124	3.974	4.122	4.223	0.051	0.876	0.807	0.872	0.785	0.736	0.972
20130110	3.418	3.530	3.501	3.497	3.616	0.009	0.714	0.742	0.721	0.730	0.685	0.972
20130111	3.022	3.110	3.048	3.231	3.247	-0.012	0.904	0.920	0.926	0.913	0.897	0.972
20130114	3.747	3.838	3.799	3.801	3.943	-0.023	0.804	0.812	0.829	0.822	0.812	0.972
20130115	4.472	4.845	4.798	4.977	5.005	0.105	0.944	0.912	0.917	0.905	0.899	0.972
20130116	3.776	3.794	3.753	3.890	3.914	0.023	0.831	0.840	0.836	0.757	0.792	0.972
20130117	4.894	5.320	5.076	5.269	5.409	0.034	0.939	0.908	0.895	0.915	0.894	0.972
20130118	3.998	4.045	3.869	4.026	4.237	0.056	0.932	0.925	0.917	0.921	0.856	0.973
20130122	5.236	5.350	5.238	5.425	5.493	0.077	0.922	0.933	0.928	0.928	0.917	0.973
20130123	5.684	5.883	5.777	6.115	6.216	0.022	0.901	0.895	0.897	0.879	0.873	0.972
20130124	3.633	3.735	3.745	3.756	3.885	0.024	0.699	0.780	0.850	0.867	0.770	0.973
20130125	4.733	4.939	4.905	4.978	5.037	0.009	0.817	0.803	0.765	0.772	0.723	0.973
20130128	3.818	3.764	3.837	3.704	3.908	-0.037	0.812	0.833	0.780	0.826	0.784	0.973
20130129	5.489	5.552	5.431	5.593	5.672	0.089	0.876	0.901	0.870	0.906	0.886	0.973
20130130	1.709	1.729	1.672	1.735	1.755	-0.075	0.773	0.794	0.795	0.743	0.754	0.972



(a) C-Bin-A and M

(b) C-Bin-S and M

(c) C-JMI-A and M

(d) C-JMI-S and M

FIGURE 2 DI comparisons of Method C and the market portfolio (M)

amount of MI between the two nodes. The dark blue denotes the strong correlation (in terms of MI), the medium blue indicates the moderate correlation, and the light blue indicates the weak correlation. There are strong correlations among Industrials, Information Technology, Health Care, Financials, Energy, Consumer Staples and Consumer Discretionary. Materials sector has a weak correlation with Financials sector, and a moderate correlation with Industrials,Information Technology, Health, Energy, Consumer Staples and Consumer Discretionary. The correlation among Materials, Telecom and Utilities is extremely weak. Moreover, Telecom and Utilities have weak correlations with all the rest sectors. We have explored the network structure by averaging the MI over the month, week, or a specific weekday, the network is very robust.

TABLE 4 The means and standard deviations of the Sharpe ratio and the DI comparison by month

Date	Type	Sharpe Ratio						DI					
		C-Bin-A	C-Bin-S	C-JMI-A	C-JMI-S	M	E	C-Bin-A	C-Bin-S	C-JMI-A	C-JMI-S	M	E
Jan.	Mean	3.953	4.057	3.996	4.093	4.192	0.031	0.837	0.846	0.855	0.841	0.818	0.972
	SD	0.9137	0.9764	0.9427	0.9989	1.0086	0.0560	0.0869	0.0638	0.0619	0.0745	0.0712	0.0005
Feb.	Mean	3.733	3.797	3.742	3.822	3.927	-0.011	0.741	0.753	0.761	0.752	0.711	0.972
	SD	1.2823	1.3148	1.2561	1.3376	1.3802	0.1407	0.2181	0.2148	0.2195	0.2165	0.2107	0.0002
Mar.	Mean	4.106	4.197	4.105	4.172	4.323	0.011	0.784	0.765	0.772	0.750	0.743	0.972
	SE	1.1312	1.2048	1.1367	1.1980	1.2557	0.0989	0.1468	0.1555	0.1463	0.1588	0.1469	0.0003
Apr.	Mean	3.709	3.765	3.706	3.738	3.883	0.006	0.745	0.724	0.731	0.726	0.684	0.973
	SD	1.3348	1.3744	1.3298	1.3683	1.4261	0.1009	0.1566	0.1608	0.1822	0.1728	0.1668	0.0003
May.	Mean	3.883	3.927	3.887	3.964	4.048	0.020	0.727	0.721	0.717	0.711	0.700	0.972
	SD	1.3418	1.3614	1.3085	1.3942	1.4137	0.0762	0.2099	0.2106	0.2214	0.2048	0.2025	0.0003
Jun.	Mean	2.379	2.387	2.397	2.389	2.461	0.003	0.671	0.650	0.640	0.649	0.623	0.971
	SD	1.3114	1.3286	1.2840	1.2807	1.3489	0.0753	0.1794	0.1907	0.2052	0.2161	0.1996	0.0003
Jul.	Mean	3.370	3.452	3.386	3.454	3.543	0.033	0.798	0.784	0.791	0.771	0.745	0.971
	SD	0.9889	1.0611	1.0126	1.0576	1.1010	0.0424	0.1000	0.1068	0.1103	0.0994	0.1105	0.0003
Aug.	Mean	3.483	3.524	3.497	3.541	3.636	0.020	0.808	0.793	0.792	0.782	0.756	0.971
	SD	1.3479	1.3930	1.3623	1.4141	1.4489	0.0432	0.1091	0.1169	0.1185	0.1190	0.1209	0.0002
Sep.	Mean	3.359	3.467	3.400	3.464	3.552	0.014	0.748	0.719	0.739	0.738	0.693	0.971
	SD	1.1521	1.2639	1.1928	1.2649	1.3073	0.0563	0.2019	0.1853	0.1932	0.1807	0.1889	0.0002
Oct.	Mean	3.797	3.894	3.866	3.883	4.023	0.033	0.757	0.740	0.743	0.745	0.714	0.970
	SE	1.2854	1.3876	1.3416	1.3970	1.4536	0.0746	0.1843	0.1900	0.1903	0.1638	0.1903	0.0012
Nov.	Mean	3.949	4.069	3.966	4.027	4.101	0.037	0.728	0.730	0.713	0.696	0.686	0.969
	SD	1.3505	1.4564	1.3780	1.4329	1.5561	0.0676	0.1771	0.1733	0.1863	0.1847	0.1632	0.0002
Dec.	Mean	3.389	3.453	3.414	3.465	3.558	0.003	0.734	0.703	0.721	0.700	0.653	0.967
	SD	0.8799	0.9397	0.8995	0.9661	0.9896	0.0401	0.1232	0.1380	0.1282	0.1361	0.1602	0.0005

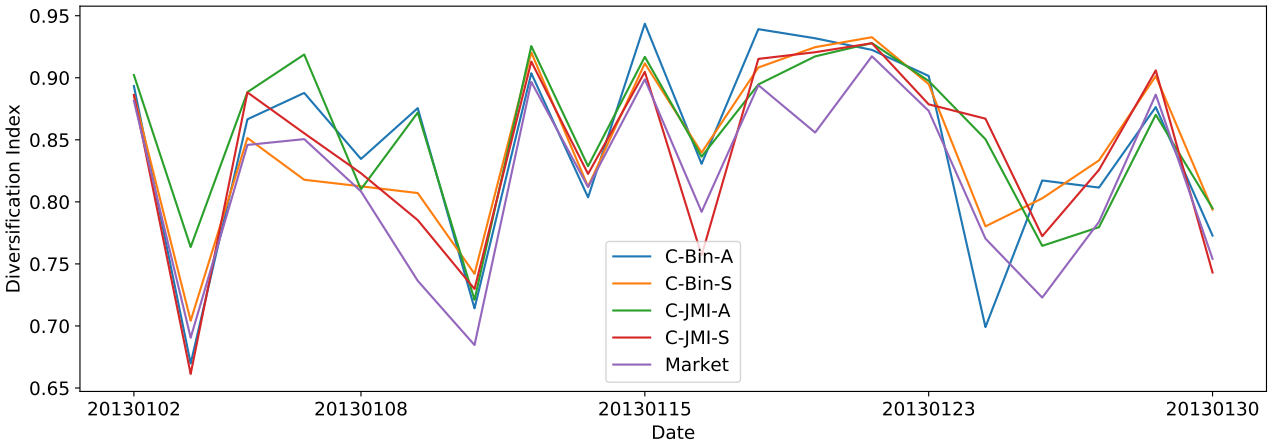


FIGURE 3 DI comparisons of Method C and the market portfolio (M) over all the trading days in January, 2013

Leading stocks within sector. To explore the leading stocks in each sector except Telecom and Utilities since there are only two stocks in these two sectors, we take each stock out of the sector and compute the MI between the stock and the rest of stocks (as a group) in the sector over the trading days of the year 2013. Figure 6 shows the distributions of the number of trading days that the stock having the largest MI with the rest of the group, for example, CVX (Chevron Corporation) has the largest MI with the rest of the group over more than 80 trading days.

Leading stocks cross sector. For leading stocks cross sectors, we only show the results of Energy and Health Care for brevity. Figure 7 shows the distributions of the smallest MI between Energy and Health Care sector by dropping each stock in Energy sector and dropping that in Health Care sector. For more than 175 trading days, Energy and Health Care sector are closer by

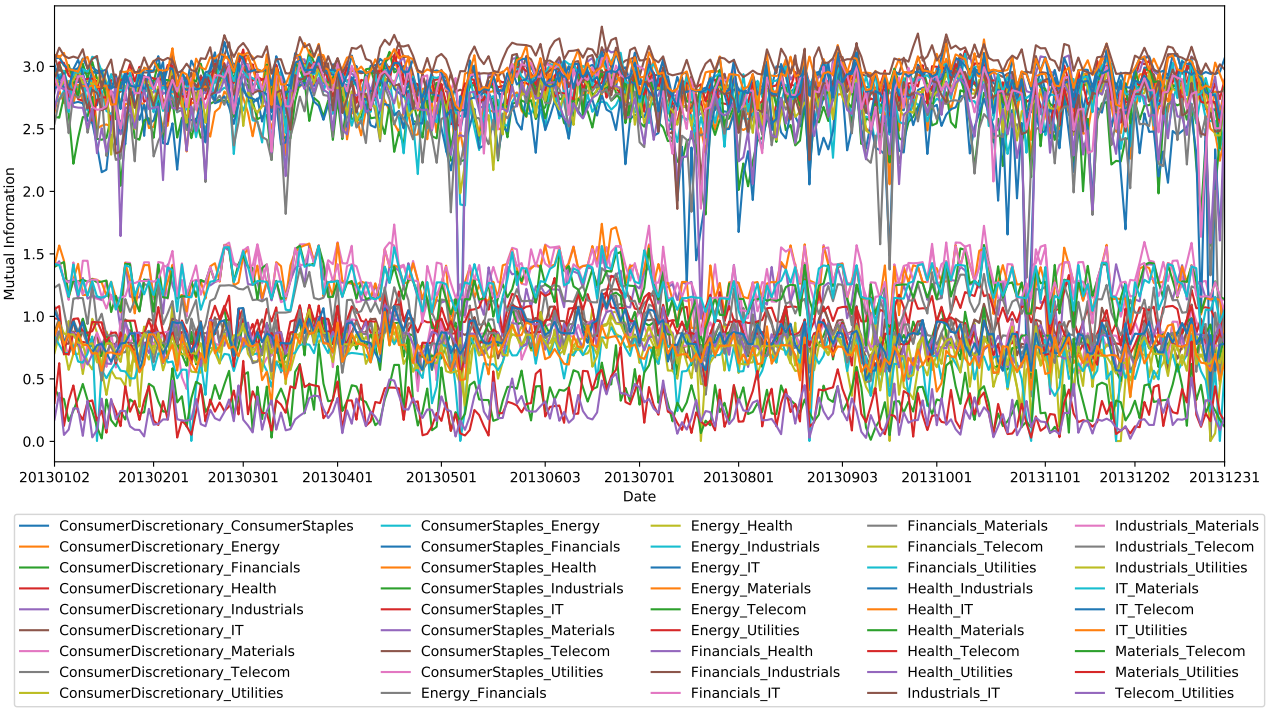


FIGURE 4 Groupwise MI trends of ten sectors for year 2013

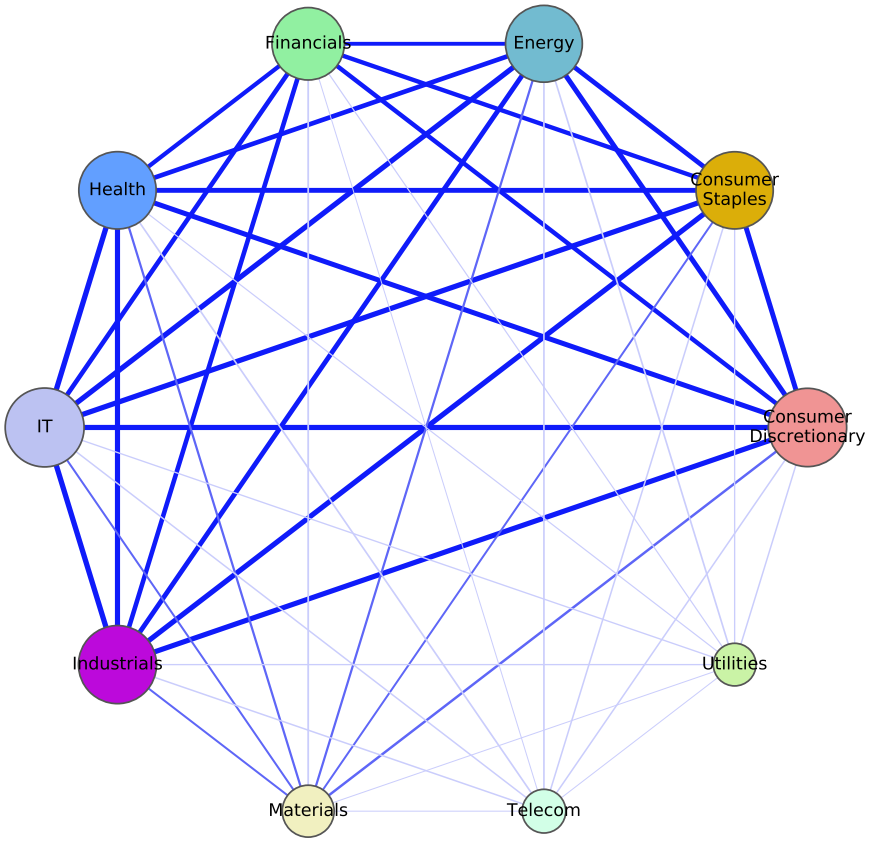


FIGURE 5 MI network for year 2013

TABLE 5 Counts of the weighted largest Sharpe ratio and DI under Method C for year 2013

	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$	$\lambda = 0.9$	$\lambda = 1$
C-Bin-A	86	86	86	86	86	87	87	87	87	87	87
C-Bin-S	166	156	147	137	127	118	108	98	88	79	69
C-JMI-A	110	106	101	97	92	88	83	79	70	74	65
C-JMI-S	167	156	144	133	121	110	98	87	75	64	52

removing AEP (American Electric Power), that is, AEP (American Electric Power) deviates from Energy sector. Similarly, UNH (UnitedHealth Group) deviates from Health Care sector for about 80 trading days.

Figure 8 shows the groupwise MI trend of Energy and Health Care sector with each stock dropped in Health Care/Energy sector. The groupwise MI of Energy and Health Care sector by dropping a stock decreases during most of the trading days. Moreover, for many tradings, the groupwise MI trend of Energy and Health Care sector decreases the most by dropping AEP (American Electric Power) and UNH (UnitedHealth Group) correspondingly.

8 | SUMMARY AND DISCUSSION

We have explored the JMI approach on the estimation of MI between high-frequency financial time series and compared such an approach with the traditional binning method under the two clustering methods we proposed for high-frequency financial time series. The two methods differ in the way in which they select the optimal number of clusters and their clustering mechanism. The ADGC method tends to generate a small number of clusters with relatively large cluster sizes. By contrast, the silhouette criterion method provides a larger number of smaller sized clusters. Based on the clustered stocks, we can estimate the comovement probabilities of any selected m-tuple of stocks. Moreover, we apply our clustering methods to overcome the curse of dimensionality of the mean-variance portfolio allocation. The Sharpe ratios of the clustering based portfolio are very close to those of the market portfolio, and for most trading days in 2013, the diversification of the clustering portfolio is better than that of the market portfolio. Lastly, we use groupwise MI to analyze behavior between and within the Global Industry Classification Standard (GICS) based groups of stocks.

We have also tried the KNN approach in the work of Kraskov et al[14] and Gao et al[24] for the MI estimation, and we obtained negative values for the MI in many cases due to (numerical) boundary issues. Therefore, this paper does not provide a comparison with the KNN approach.

9 | ACKNOWLEDGEMENT

This paper was based upon work partially supported by the National Science Foundation under Grant DMS-1638521 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Baron M, Brogaard J, Hagströmer B, Kirilenko A. Risk and return in high-frequency trading. *Journal of Financial and Quantitative Analysis* 2019; 54(3): 993–1024.

2. Diebold FX, Hahn J, Tay AS. Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. *Review of Economics and Statistics* 1999; 81(4): 661–673.

3. Kou G, Peng Y, Wang G. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences* 2014; 275: 1–12.

4. Harvill JL, Kohli P, Ravishanker N. Clustering nonlinear, nonstationary time series using BSLEX. *Methodology and Computing in Applied Probability* 2017; 19(3): 935–955.

5. Pattarin F, Paterlini S, Minerva T. Clustering financial time series: an application to mutual funds style analysis. *Computational Statistics & Data Analysis* 2004; 47(2): 353–372.

6. Onnela JP, Kaski K, Kertész J. Clustering and information in correlation based financial networks. *The European Physical Journal B* 2004; 38(2): 353–362.

7. Tola V, Lillo F, Gallegati M, Mantegna RN. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control* 2008; 32(1): 235–258.

8. Nanda S, Mahanty B, Tiwari M. Clustering Indian stock market data for portfolio management. *Expert Systems with Applications* 2010; 37(12): 8793–8798.

9. León D, Aragón A, Sandoval J, Hernández G, Arévalo A, Niño J. Clustering algorithms for risk-adjusted portfolio construction. *Procedia Computer Science* 2017; 108: 1334–1343.

10. Raffinot T. Hierarchical clustering-based asset allocation. *The Journal of Portfolio Management* 2017; 44(2): 89–99.

11. Nair BB, Kumar PS, Sakthivel N, Vipin U. Clustering stock price time series data to generate stock trading recommendations: An empirical study. *Expert Systems with Applications* 2017; 70: 20–36.

12. Brown SJ, Goetzmann WN. Mutual fund styles. *Journal of Financial Economics* 1997; 43(3): 373–399.

13. Bastos JA, Caiado J. Clustering financial time series with variance ratio statistics. *Quantitative Finance* 2014; 14(12): 2121–2133.

14. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. *Physical Review E* 2004; 69(6): 066138.

15. Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences* 2014; 111(9): 3354–3359.

16. Brillinger DR. Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics* 2004: 163–182.

17. Kim M, Sayama H. Predicting stock market movements using network science: An information theoretic approach. *Applied Network Science* 2017; 2(1): 35.

18. Guo X, Zhang H, Tian T. Development of stock correlation networks using mutual information and financial big data. *PloS one* 2018; 13(4): e0195941.

19. Singh S, Póczos B. Exponential concentration of a density functional estimator. In: ; 2014: 3032–3040.

20. Moon KR, Sricharan K, Hero AO. Ensemble estimation of mutual information. In: IEEE. ; 2017: 3030–3034.

21. Zeng X, Xia Y, Tong H. Jackknife approach to the estimation of mutual information. *Proceedings of the National Academy of Sciences* 2018; 115(40): 9956–9961.

22. Daub CO, Steuer R, Selbig J, Kloska S. Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data. *BMC Bioinformatics* 2004; 5(1): 118.

23. Peter AM, Rangarajan A. Maximum likelihood wavelet density estimation with applications to image and shape matching. *IEEE Transactions on Image Processing* 2008; 17(4): 458–468.

24. Gao W, Kannan S, Oh S, Viswanath P. Estimating mutual information for discrete-continuous mixtures. In: ; 2017: 5986–5997.

25. Sharpe WF. The sharpe ratio. *Journal of Portfolio Management* 1994; 21(1): 49–58.

26. Woerheide W, Persson D. An index of portfolio diversification. *Financial Services Review* 1993; 2(2): 73–85.

27. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 1987; 20: 53–65.

28. De Amorim RC, Hennig C. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences* 2015; 324: 126–145.

29. Liu H, Zou J, Ravishanker N. Multiple day biclustering of high-frequency financial time series. *Stat* 2018; 7(1): e176.

30. Markowitz H. PORTFOLIO SELECTION. *The Journal of Finance* 1952; 7(1): 77-91.

31. Michaud RO. The Markowitz optimization enigma: Is 'optimized' optimal?. *Financial Analysts Journal* 1989; 45(1): 31–42.

32. Dose C, Cincotti S. Clustering of financial time series with application to index and enhanced index tracking portfolio. *Physica A: Statistical Mechanics and its Applications* 2005; 355(1): 145–151.

33. Lemieux V, Rahmdel PS, Walker R, Wong B, Flood M. Clustering techniques and their effect on portfolio formation and risk analysis. In: ACM. ; 2014: 1–6.

34. Ren F, Lu YN, Li SP, Jiang XF, Zhong LX, Qiu T. Dynamic portfolio strategy using clustering approach. *PloS one* 2017; 12(1): e0169299.

APPENDIX

A BINNING APPROACH TO ESTIMATE MUTUAL INFORMATION

Let $\mathcal{S} = \{1, \dots, S\}$ denote a set of S stocks that are observed over T time intervals (say, one-minute intervals) over N trading days. On the nth trading day, we observe log returns $r_{i,t,n}$ in the tth time interval for the ith stock, for $t = 1, \dots, T$ and $i = 1, \dots, S$.

We divide the one-minute averaged log returns of each stock for each trading day, i.e., $r_{i,t,n}$, into Q bins. Let $\boldsymbol{V}(i, n)$ denote the T-dimensional vector representing the ordered values of $r_{i,t,n}$ (from lowest to highest) for the ith stock on the nth trading day. We set the lower and upper endpoints for each of the Q bins using the quantiles of the ordered log returns as follows:

(i) The lower endpoint of the first bin (when $q = 1$) is the first element of $\boldsymbol{V}(i, n)$, and its upper endpoint is the $(\frac{1}{Q} \times T)$th value of $\boldsymbol{V}(i, n)$.

(ii) The second bin (when $q = 2$) has lower bound equal to the $(\frac{1}{Q} \times T)$th value in $\boldsymbol{V}(i, n)$ and upper bound equal to the $(\frac{2}{Q} \times T)$th value of the vector.

(iii) Proceeding similarly, the lower and upper endpoints of the Qth bin are respectively the $(\frac{(Q-1)}{Q} \times T)$th value in $\boldsymbol{V}(i, n)$ and the last element of $\boldsymbol{V}(i, n)$.

We denote the bins for the ith stock on the nth trading day as $b_n(i, q)$, for $q = 1, \dots, Q$. Let $I_{t,n}(i, q)$ indicate whether or not the log return $r_{i,t,n}$ falls into the bin $b_n(i, q)$, i.e.,

$$I_{t,n}(i, q) = \begin{cases} 1 & \text{if } r_{i,t,n} \in b_n(i, q) \\ 0 & \text{if } r_{i,t,n} \notin b_n(i, q). \end{cases} \tag{A1}$$

The proportion (relative frequency) of log returns of the ith stock on day n that fall into the bin $b_n(i, q)$ is

$$\hat{p}_n(i, q) = \frac{1}{T} \sum_{t=1}^{T} I_{t,n}(i, q), \tag{A2}$$

and is an estimate of the marginal pmf $p(x)$ shown in (1) and (3). We compute the joint proportion of log returns of the ith stock falling into the bin $b_n(i, q)$ and the i'th stock falling into the bin $b_n(i', q')$ on day n as

$$\hat{m}_n(i, i', q, q') = \frac{1}{T} \sum_{t=1}^{T} I_{t,n}(i, q) I_{t,n}(i', q'). \tag{A3}$$

This gives an estimate of the joint pmf $p(x, y)$ in (2) and (3). The joint entropy that measures the uncertainty associated with the ith and i'th stocks within trading day n is computed as

$$\hat{H}_n(s_i, s_i') = -\sum_{q=1}^{Q}\sum_{q'=1}^{Q} \hat{m}_n(i, i', q, q') \log\left(\hat{m}_n(i, i', q, q')\right), \tag{A4}$$

and is an estimate of equation (2). The mutual information that measures the similarity between the ith stock and the i'th stock within trading day n is computed as

$$\hat{MI}_n(s_i, s_{i'}) = \sum_{q=1}^{Q}\sum_{q'=1}^{Q} \hat{m}_n(i, i', q, q') \log\left(\frac{\hat{m}_n(i, i', q, q')}{\hat{p}_n(i, q)\hat{p}_n(i', q')}\right), \tag{A5}$$

and estimates (3). The normalized mutual information based distance between s_i s_i' is estimated using (A4) and (A5) as

$$\hat{D}_n(s_i, s_i') = 1 - \frac{\hat{MI}_n(s_i, s_i')}{\hat{H}_n(s_i, s_i')}, \tag{A6}$$

and is an estimate of (6). The amount of information given by (A5) and (A6) is the same, although the former measures similarity between stocks while the latter measures dissimilarity.

B THE AVERAGE SILHOUETTE METHOD

The average silhouette method is used to determine the number of clusters. The silhouette value $Sil(\cdot)$ measures the cohesion of an object's own cluster and separation of other clusters. Denote $C_{\kappa,n}$ as the κth cluster in trading day n, and for the stock in cluster $C_{\kappa,n}$, denote $s_{i,C_{\kappa,n}}$, then denote

$$a(s_{i,C_{\kappa,n}}) = \frac{1}{|C_{\kappa,n}| - 1}\sum_{i \neq i'} D(s_{i,C_{\kappa,n}}, s_{i',C_{\kappa,n}})$$

where $D(s_{i,C_{\kappa,n}}, s_{i',C_{\kappa,n}})$ is the distance between stock $s_{i,C_{\kappa,n}}$ and $s_{i',C_{\kappa,n}}$ in the cluster $C_{\kappa,n}$, and $|C_{\kappa,n}|$ denotes the number of stocks in cluster $C_{\kappa,n}$, and $a(s_{i,C_{\kappa,n}})$ measures the quality of the cluster $C_{\kappa,n}$, a smaller $a(s_{i,C_{\kappa,n}})$ indicates stock $s_{i,C_{\kappa,n}}$ is close to the other stocks in the cluster $C_{\kappa,n}$.

Denote stocks not in cluster $C_{\kappa,n}$ as $s_{j,C_{\kappa',n}}$ and $\kappa \neq \kappa'$,

$$b(s_{i,C_{\kappa,n}}) = \min_{\kappa \neq \kappa'} \frac{1}{|C_{\kappa',n}|} \sum_{s_{j,C_{\kappa',n}} \notin C_{\kappa,n}} D(s_{i,C_{\kappa,n}}, s_{j,C_{\kappa',n}})$$

be the minimum average distance of stock $s_{i,C_{\kappa,n}}$ with all the stocks in any other cluster.

A silhouette value of stock $s_{i,C_{\kappa,n}}$ is

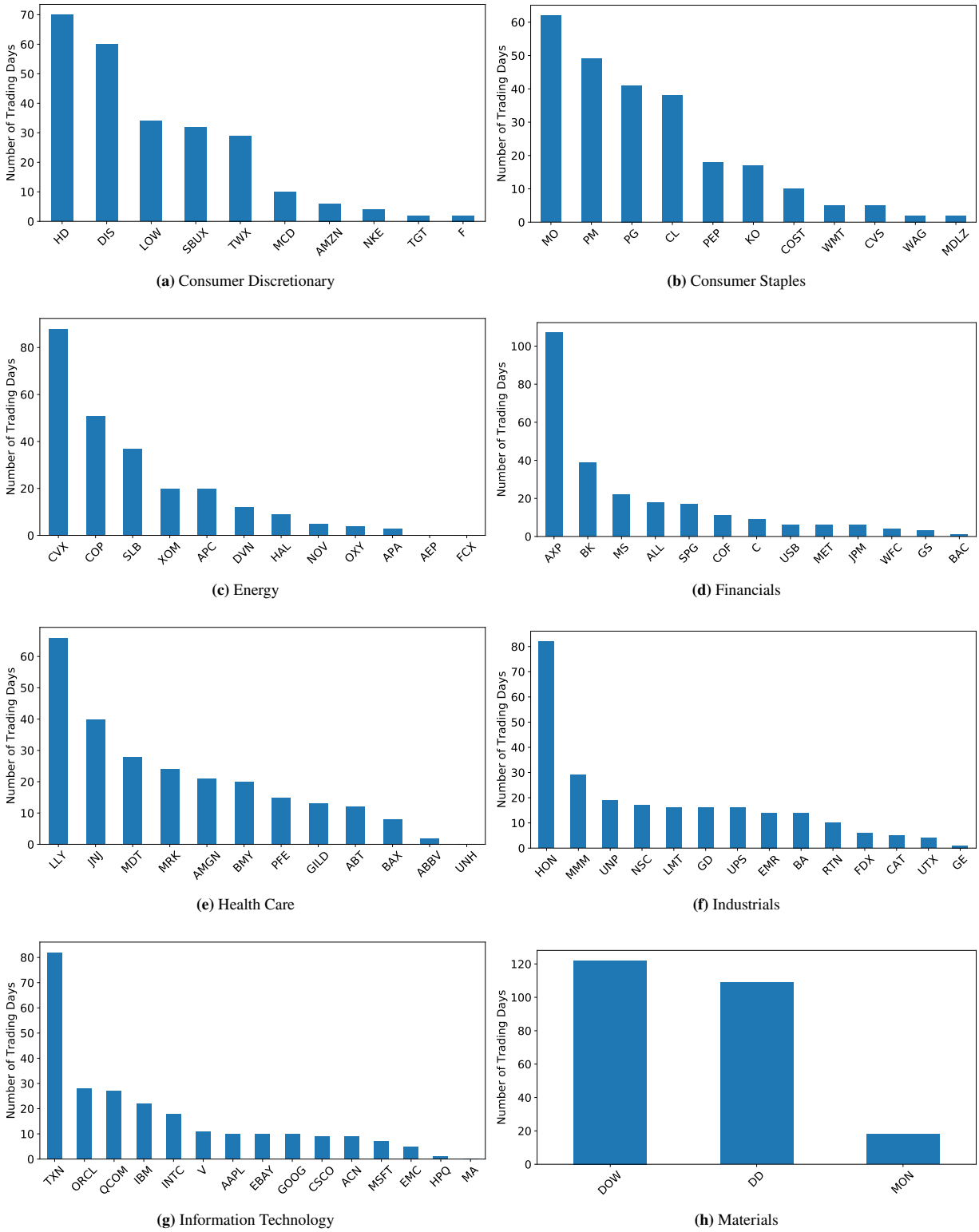
$$Sil(s_{i,C_{\kappa,n}}) = \frac{b(s_{i,C_{\kappa,n}}) - a(s_{i,C_{\kappa,n}})}{\max\{a(s_{i,C_{\kappa,n}}), b(s_{i,C_{\kappa,n}})\}}$$

Now the silhouette value is define

$$Sil(s_{i,C_{\kappa,n}}) = \begin{cases} 1 - a(s_{i,C_{\kappa,n}})/b(s_{i,C_{\kappa,n}}), & \text{if } a(s_{i,C_{\kappa,n}}) < b(s_{i,C_{\kappa,n}}) \\ 0, & \text{if } a(s_{i,C_{\kappa,n}}) = b(s_{i,C_{\kappa,n}}) \\ b(s_{i,C_{\kappa,n}})/a(s_{i,C_{\kappa,n}}) - 1, & \text{if } a(s_{i,C_{\kappa,n}}) > b(s_{i,C_{\kappa,n}}) \end{cases}$$

and

$$-1 \leq Sil(s_{i,C_{\kappa,n}}) \leq 1$$



(a) Consumer Discretionary

(b) Consumer Staples

(c) Energy

(d) Financials

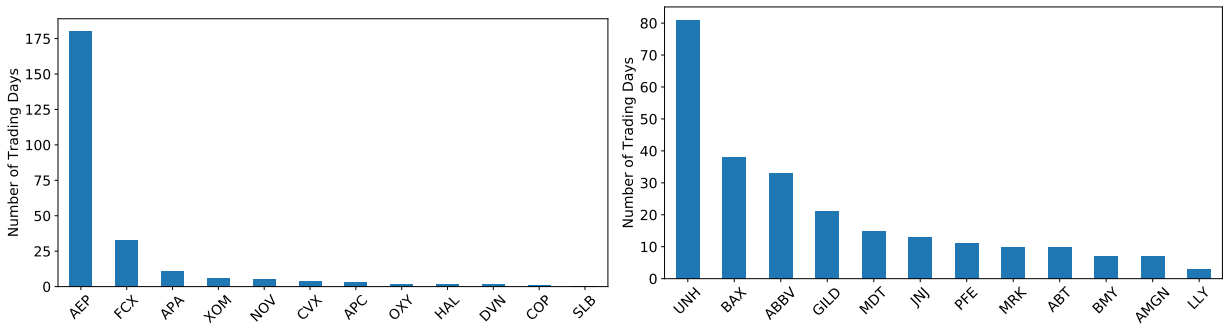
(e) Health Care

(f) Industrials

(g) Information Technology

(h) Materials

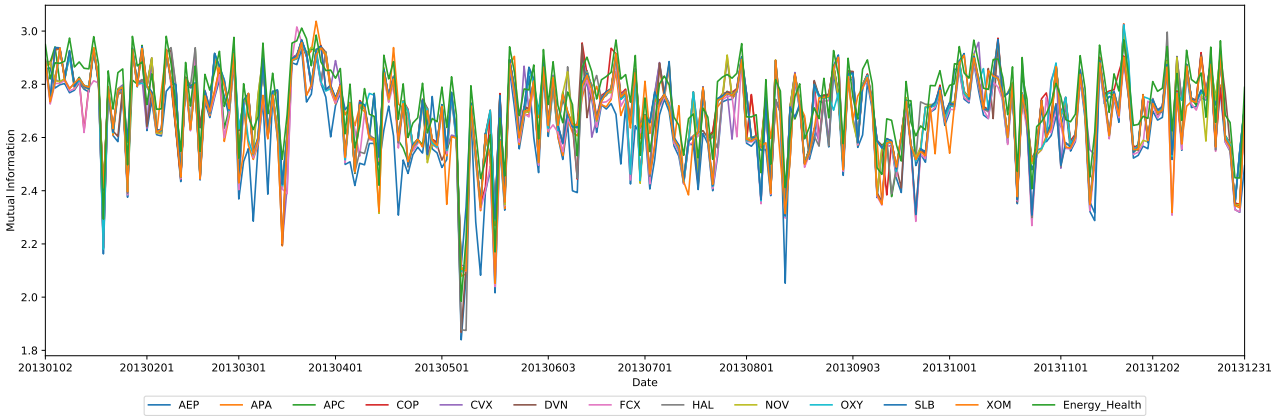
FIGURE 6 Distributions of the leading stocks within each sector



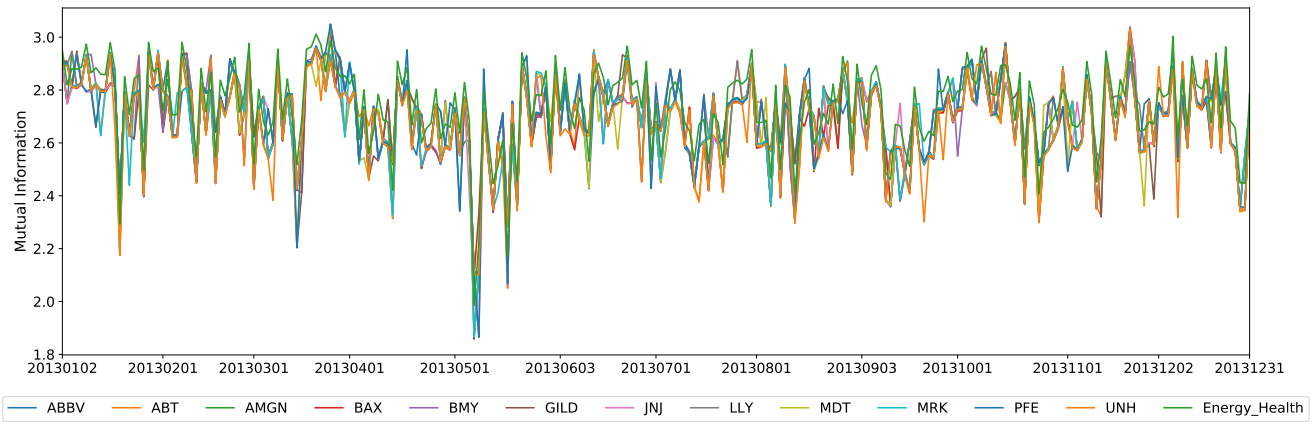
(a) MI between Energy and Health Care Sector by dropping each Stock in Energy sector

(b) MI between Energy and Health Care sector by dropping each stock in Health Care Sector

FIGURE 7 Distributions of the leading stocks across sectors



(a) MI Trend between Energy and Health Care sector by dropping each stock in Energy sector



(b) MI Trend between Energy and Health Care sector by dropping each stock in Health Care sector

FIGURE 8 The groupwise MI trend of Energy and Health Care sector with each stock dropped in the Health Care/Energy sector