Equity Clusters through the Lens of Realized Semicorrelations

This version: December 8, 2021

Tim Bollerslev[a,*], Andrew J. Patton[b], Haozhe Zhang[c]

[a] *Department of Economics, Duke University, NBER and CREATES*
[b] *Department of Economics, Duke University*
[c] *Department of Economics, Duke University*

Abstract

We rely on newly-developed realized semicorrelations constructed from high-frequency returns together with hierarchical clustering and cross-validation techniques to identify groups of individual stocks that share common features. Implementing the new procedures based on intraday data for the S&P 100 constituents spanning 2019-2020, we uncover distinct changes in the "optimal" groupings of the stocks coincident with the onset of the COVID-19 pandemic. Many of the clusters estimated with data post-January 2020 evidence clear differences from conventional industry type classifications. They also differ from the clusters estimated with standard realized correlations, underscoring the advantages of "looking inside" the correlation matrix through the lens of the new realized semicorrelations.

Keywords: Clustering; Stock returns; High-frequency data; Semicorrelations; COVID-19.

JEL: C32; C38; C58; G10

*We thank Jia Li and George Tauchen and other participants in the Duke Financial Econometrics workshop for helpful comments.

*Corresponding author: Department of Economics, Duke University, 213 Social Sciences Building, Box 90097, Durham, NC 27708-0097, United States. Email: boller@duke.edu.

1. Introduction

The realized volatility concept, and empirical applications thereof, based on the use of intraday returns for the construction of more accurate ex-post volatility measurements, easily ranks among the most active areas of research in econometrics over the past two decades (the introductory chapter in Andersen and Bollerslev (2018) provides a review). Extending the basic multivariate realized volatility concept, Bollerslev, Li, Patton and Quaedvlieg (2020) (henceforth BLPQ) recently proposed a decomposition of the realized covariance matrix into separate so-called "realized semicovariances" based on the signs of the high-frequency returns. In this paper, we show how the corresponding realized semicorrelations may be used for timely and meaningful clustering of individual stocks into groups that react similarly to new information.

Our empirical analysis is based on high-frequency intraday returns for the S&P 100 constituents spanning January 2019 to December 2020. We rely on hierarchical clustering methods for grouping each of the stocks into clusters based on the similarities in their daily realized semicorrelations, together with a novel cross-validation procedure for determining the number of clusters and the single break date that allows for the lowest overall cluster assignment errors. Our analysis points to January 31, 2020 as the "optimal" break date, coincident with the World Health Organization (WHO) first declaring the coronavirus outbreak a health emergency of international concern. Three distinct new clusters of "Good Covid," "Bad Covid" and "Very Bad Covid" stocks also emerge at that time, with each of the clusters comprised of stocks from very different industries.

Our use of clustering methods for categorizing individual stocks isn't new to the literature. Ahn, Conrad and Dittmar (2009) have previously relied on clustering methods based on rolling sample correlations from monthly returns for sorting stocks into portfolios. A series of more recent papers, as exemplified by Patton and Weller (2021), Lucas, Schaumburg and Schwaab (2020), Jensen, Kelly and Pedersen (2021) and Lumsdaine, Okui and Wang (2021), also rely on clustering techniques to help better understand various commonalities in equity returns. To the best of our knowledge, however, clustering techniques have not

hitherto been used in conjunction with high-frequency-based realized volatility measures, let alone the realized semicorrelation measures that we rely on here.

Our paper is also related to the rapidly-growing recent literature on the economic impact of the COVID-19 pandemic, on the cross-section of stock returns in particular. This includes the work by Bretscher, Hsu, Simasek and Tamoni (2020) on stock returns and local COVID-19 transmission rates, Albuquerque, Koskinen, Yang and Zhang (2020) on the effect of firms' ESG ratings, Ding, Levine, Lin and Xie (2021) on corporate immunity, Papanikolaou and Schmidt (2021) on industry-level exposures to COVID-19, and Pagano, Wagner and Zechner (2021) on companies' resilience to social distancing.[1] In contrast to all of these studies, however, which rely on daily or more coarsely-sampled returns together with additional information cleaned from other sources, we demonstrate how the information inherent in the high-frequency-based realized semicorrelation measures may be effectively used for assessing changes in the cross-sectional dependencies observed as a results of the pandemic.

The rest of the paper is organized as follows. Section 2 formally defines the realized semicorrelation measures, Section 3 provides a brief overview of the clustering and cross-validation procedures that we use in our empirical analyses, Section 4 presents our key empirical findings, Section 5 concludes.

2. Realized Semicorrelations

The traditional realized covariance matrix (e.g., Barndorff-Nielsen and Shephard (2004)), defined by the summation of squares or cross-products of finely sampled returns within a fixed time-interval, is effectively "blind" to the sign of the underlying returns. The realized semicovariance measures recently proposed by BLPQ extend the basic measures to distinguish between "good" and "bad" covariation by conditioning on the signs of the high-frequency returns.

Formally, let $\mathbf{X}_{t,i} = (X_{1,t,i}, \ldots, X_{k,t,i})^\top$ denote the log-price vector for all of the k stocks in the sample measured at time i on day t. For simplicity, assume that prices are observed n

[1] In addition to these specific studies, several special journal issues devoted to the financial impact of COVID-19 have also appeared, including *Review of Asset Pricing Studies*, Vol. 10, No. 4, 2020, and *Review of Financial Studies*, Vol. 34, No. 11, 2021.

times each day on a regular time grid of width $\Delta = 1/(n-1)$. Denote the ith return vector on day t by $\Delta_i^n \mathbf{X}_t \equiv \mathbf{X}_{t,i\Delta_n} - \mathbf{X}_{t,(i-1)\Delta_n}$. Further, let $p(\mathbf{x}) \equiv \max\{\mathbf{x}, 0\}$ and $n(\mathbf{x}) \equiv \min\{\mathbf{x}, 0\}$ denote the component-wise positive and negative elements of the vector $\mathbf{x}$. The positive, negative, and mixed realized semicovariance matrices are then defined as,[2]

$$\begin{aligned} \widehat{P}_t &\equiv \sum_{i=1}^{1/\Delta_n} p(\Delta_i^n \mathbf{X}_t)\, p(\Delta_i^n \mathbf{X}_t)^\top, \quad \widehat{N}_t \equiv \sum_{i=1}^{1/\Delta_n} n(\Delta_i^n \mathbf{X}_t)\, n(\Delta_i^n \mathbf{X}_t)^\top, \\ \widehat{M}_t &\equiv \sum_{i=1}^{1/\Delta_n} \left(p(\Delta_i^n \mathbf{X}_t)\, n(\Delta_i^n \mathbf{X}_t)^\top + n(\Delta_i^n \mathbf{X}_t)\, p(\Delta_i^n \mathbf{X}_t)^\top \right). \end{aligned} \tag{1}$$

Note, the traditional realized covariance matrix is simply obtained as $\widehat{C}_t \equiv \widehat{P}_t + \widehat{N}_t + \widehat{M}_t$. The relative magnitudes of the two concordant $\widehat{P}_t$ and $\widehat{N}_t$ semicovariance matrices and the discordant $\widehat{M}_t$ semicovariance matrix are directly related to the strength of the association between the stocks. In particular, one might naturally expect the relative importance of $\widehat{P}_t + \widehat{N}_t$ versus $\widehat{M}_t$ to increase during times of market turmoil and stronger overall cross-stock linkages. The in-fill asymptotic theory developed in BLPQ further identifies three distinct channels through which differences in $\widehat{P}_t$ and $\widehat{N}_t$ can occur, namely common jumps, common drifts and/or a dynamic leverage effect. It is not our goal to separately identify these channels. Instead, we seek to demonstrate how the information in each of these effects, as captured by the new realized semicovariance measures, can help clarify the way in which different stocks respond to new information and in turn solicit groups of stocks that share similar features.

To facilitate this clustering, rather than working with the measures defined in (1), it is more convenient to work with the scale-invariant realized semicorrelation matrices obtained by scaling the original measures by the realized volatilities. In particular, we define the positive realized semicorrelation matrix as $\widehat{R}_t^P \equiv \mathrm{dg}(\hat{\mathbf{v}}_t)^{-1} \cdot \widehat{P}_t \cdot \mathrm{dg}(\hat{\mathbf{v}}_t)^{-1}$, where $\mathrm{dg}(\hat{\mathbf{v}}_t)$ denotes a diagonal matrix with the square-root of the diagonal elements in $\widehat{C}_t$ along the diagonal. The negative and mixed semicorrelation matrices, $\widehat{R}_t^N$ and $\widehat{R}_t^M$, are defined anal-

[2]The "mixed" semicovariance matrix is comprised of two components. However, since the ordering of the individual stocks is arbitrary, it is natural to sum the two components, as in equation (1).

ogously. While $\widehat{R}_t^P$ and $\widehat{R}_t^N$ are both guaranteed to be positive semidefinite, the mixed $\widehat{R}_t^M$ semicovariance matrix has diagonal elements identically equal to zero and so is necessarily indefinite. The use of an indefinite "feature matrix" presents a problem for some clustering algorithms. However, the approach discussed next that we adopt here does not require the feature matrices to posses any special properties.[3]

3. Hierarchical Clustering and Cross-Validation

The increased availability of "big data" in many areas of economics combined with increased computing power have spurred a rapid growth in the use of machine learning techniques in economics (see, e.g., the review in Athey and Imbens (2019)). This includes so-called unsupervised learning procedures designed to partition a given sample into sub-samples, or clusters, comprised of members that share similar features.

The specific hierarchical clustering algorithm that we rely on here traces back to Ward (1963). To convey the main intuition, let $\hat{\mathbf{S}}_{\mathbf{t,i}}$ denote the vector of day t realized semicorrelations for stock i with all of the other stocks in the sample; i.e., the ith rows in the above-defined $\widehat{R}_t^P$, $\widehat{R}_t^N$ and $\widehat{R}_t^M$ matrices. Let the cluster assignment of stock i be $\gamma_i \in \{1, 2, ..., G\}$, where G denotes the number of clusters and $N_g \equiv \sum_i \mathbf{1}(\gamma_i = g)$ the number of members in cluster g. Starting with each stock in a cluster of its own, the algorithm works by iteratively merging the two closest clusters, with the distance between clusters g and h defined as,

$$\begin{aligned}
\Delta(g,h) \;\equiv\; & \sum_t \sum_i \mathbf{1}(\gamma_i \in \{g,h\}) \cdot d\left(\hat{\mathbf{S}}_{\mathbf{t,i}}, \frac{1}{N_g + N_h} \sum_j \mathbf{1}(\gamma_j \in \{g,h\}) \hat{\mathbf{S}}_{\mathbf{t,j}}\right) \\
& - \sum_t \left[\sum_i \mathbf{1}(\gamma_i = g) \cdot d\left(\hat{\mathbf{S}}_{\mathbf{t,i}}, \frac{1}{N_g} \sum_j \mathbf{1}(\gamma_j = g) \hat{\mathbf{S}}_{\mathbf{t,j}}\right) \right. \\
& \left. + \sum_i \mathbf{1}(\gamma_i = h) \cdot d\left(\hat{\mathbf{S}}_{\mathbf{t,i}}, \frac{1}{N_h} \sum_j \mathbf{1}(\gamma_j = h) \hat{\mathbf{S}}_{\mathbf{t,j}}\right) \right],
\end{aligned}$$

where $d(\cdot, \cdot)$ refers to the usual Euclidean distance between two vectors.

[3] More traditional principal components-based procedures and factor analysis, of course, also require the input matrices to be positive definite.

While this provides an unambiguous sequential approach for grouping the stocks, the algorithm remains silent about the number of clusters and when to halt the iterations. This is a common difficulty shared by most clustering procedures. Correspondingly, in many economic applications the number of clusters is simply chosen *a priori* based on some heuristic arguments. Instead, we rely on a data-driven approach, in which we utilize the time-series dimension of the data, exploiting the fact that the estimation error in realized measures is serially independent across days. Specifically, splitting the sample into odd and even days, we generate separate assignments for each of the two samples (Jegadeesh, Noh, Pukthuanthong, Roll and Wang (2019) employ a similar odd-even months sample split in empirical asset pricing). We then use the odd (even) days to generate one-day-ahead forecasts for the even (odd) days based on the average within-cluster semicorrelations. Concretely, the forecast for stock i on day $t+1$ is $\hat{\mathbf{S}}_{\mathbf{t+1|t,i}} \equiv N_{\gamma_i}^{-1} \sum_j \mathbf{1}(\gamma_j = \gamma_i) \hat{\mathbf{S}}_{\mathbf{t,j}}$. Averaging the resulting prediction errors $d(\hat{\mathbf{S}}_{\mathbf{t+1|t,i}}, \hat{\mathbf{S}}_{\mathbf{t+1,i}})$ over all of the days and stocks in the sample, we determine the number of clusters that minimizes this overall predictive loss. This approach amounts to a two-fold cross-validation procedure with the odd and even days alternating between being in- and out-of-sample (see, e.g., the review in Arlot and Celisse (2010)).

We further utilize this same idea to search for a possible break date in the sample. Assuming day t^* to be a break day, we estimate separate cluster assignments for the $t \leq t^*$ and $t > t^*$ samples, relying on the same cross-validation approach for each of the two periods. Mirroring traditional structural break procedures (e.g., Chow (1960) and Bai (1997)), we then determine the optimal break point as the day t^* that minimizes the combined pre- and post-break predictive losses (see also the related discussion in Lumsdaine, Okui and Wang (2021)).

4. Covid Clusters of Stocks

Our empirical analysis is based on data for the S&P 100 constituents over the period from January 2, 2019 to December 31, 2020, obtained from the New York Stock Exchange's

Trades and Quotes (TAQ) database.[4] The 2019-2020 sample period was deliberately chosen to study changes in the dependencies around the onset of the pandemic. To help mitigate the effect of asynchronous trading and market microstructure noise, we follow common practice in the literature (see, e.g., Liu, Patton and Sheppard (2015)) and compute the daily realized measures at a five-minute sampling frequency. To enhance the efficiency of the five-minute estimates, we follow Zhang, Mykland and Aït-Sahalia (2005) in applying a subsampling approach, whereby we calculate the realized measures starting at five different one-minute marks, averaging the resulting five measures to obtain our final daily realized measures.

Our hierarchical clustering procedure based on daily realized semicorrelations together with our odd-even-day cross-validation technique indicates a structural break in the cluster assignments on January 31, 2020. Interestingly, this day coincides with the World Health Organization (WHO) first declaring the coronavirus outbreak a health emergency of international concern.[5] By comparison, applying the same clustering and structural break approach based on the conventional daily realized correlations $\widehat{R}_t \equiv \widehat{R}_t^P + \widehat{R}_t^N + \widehat{R}_t^M$, the break date for the cluster assignments does not occur until March 18, 2020, following the precipitous decline in the overall market from early February through mid-March 2020.

Looking at the estimated cluster assignments for the pre-Covid period, reported in the top panel of Table 1, along with the corresponding Global Industry Classification Standard (GICS) codes, the clusters labelled A-D are naturally identified as financials (GIC 40), consumer staples (30), health care (35), and information technology (45). The rest of the clusters are somewhat more mixed. However, each of the clusters are still mostly comprised of stocks from a few specific sectors, with E containing mostly industrials (20) and information technology stocks (45), F mostly consumer discretionary (25) and communication services stocks (50), G primarily energy stocks (10), and H mostly utilities (55).

Turning to the post-Covid clusters, reported in the bottom panel of Table 1, our cross-validation procedure suggests the need for one additional cluster to best accommodate the

[4] We rely on the procedures of Barndorff-Nielsen, Hansen, Lunde and Shephard (2009) to clean the data. We leave out DowDuPont (DOW), which completed its split into three separate companies on June 1, 2019.

[5] World Health Organization, "Novel Coronavirus (2019-nCoV): situation report," 11, January 31, 2020.

Pre-Covid

	A	B	C	D	E	F	G	H
1	AIG (40)	CL (30)	ABBV (35)	ACN (45)	AAPL (45)	BA (20)	COP (10)	AMT (60)
2	BAC (40)	KO (30)	ABT (35)	ADBE (45)	ALL (40)	BKNG (25)	CVS (35)	DUK (55)
3	BK (40)	MCD (25)	AMGN (35)	CRM (45)	AMZN (25)	CHTR (50)	CVX (10)	EXC (55)
4	BLK (40)	MDLZ (30)	BIIB (35)	MA (45)	AXP (40)	CMCS (50)	DD (15)	NEE (55)
5	C (40)	MO (30)	BMY (35)	MSFT (45)	BRK (40)	COST (30)	F (25)	SO (55)
6	COF (40)	PEP (30)	DHR (35)	PYPL (45)	CAT (20)	DIS (50)	GM (25)	SPG (60)
7	GS (40)	PG (30)	GILD (35)	V (45)	CSCO (45)	FB (50)	KHC (30)	
8	JPM (40)	PM (30)	JNJ (35)		EMR (20)	GE (20)	KMI (10)	
9	MET (40)	SBUX (25)	LLY (35)		FDX (20)	HD (25)	SLB (10)	
10	MS (40)	T (50)	MDT (35)		GD (20)	LOW (25)	WBA (30)	
11	USB (40)	VZ (50)	MRK (35)		GOOG (50)	NFLX (50)	XOM (10)	
12	WFC (40)		PFE (35)		HON (20)	QCOM (45)		
13			TMO (35)		IBM (45)	TGT (25)		
14			UNH (35)		INTC (45)	TSLA (25)		
15					LMT (20)	WMT (30)		
16					MMM (20)			
17					NKE (25)			
18					NVDA (45)			
19					ORCL (45)			
20					RTX (20)			
21					TXN (45)			
22					UNP (20)			
23					UPS (20)			

Post-Covid

	Good Covid	Bad Covid	Very Bad Covid	A	B	C	D	E	F
1	BIIB (35)	AIG (40)	SLB (10)	AXP (40)	CL (30)	ABBV (35)	AAPL (45)	ALL (40)	ACN (45)
2	BKNG (25)	BA (20)	SPG (60)	BAC (40)	CMCS (50)	ABT (35)	ADBE (45)	BLK (40)	CSCO (45)
3	CHTR (50)	COP (10)		BK (40)	DUK (55)	AMGN (35)	AMZN (25)	BRK (40)	DIS (50)
4	GILD (35)	CVX (10)		C (40)	EXC (55)	AMT (60)	CRM (45)	CAT (20)	HD (25)
5	NFLX (50)	DD (15)		COF (40)	KO (30)	BMY (35)	FB (50)	EMR (20)	INTC (45)
6	TSLA (25)	F (25)		GS (40)	MDLZ (30)	COST (30)	GOOG (50)	FDX (20)	LOW (25)
7		GE (20)		JPM (40)	MO (30)	CVS (35)	MSFT (45)	GD (20)	MA (45)
8		GM (25)		MET (40)	PEP (30)	DHR (35)	NVDA (45)	HON (20)	MCD (25)
9		KMI (10)		MS (40)	PG (30)	JNJ (35)	PYPL (45)	IBM (45)	NKE (25)
10		RTX (20)		USB (40)	PM (30)	KHC (30)		LMT (20)	ORCL (45)
11		XOM (10)		WFC (40)	SO (55)	LLY (35)		MMM (20)	QCOM (45)
12					T (50)	MDT (35)		UNP (20)	SBUX (25)
13					VZ (50)	MRK (35)		UPS (20)	TXN (45)
14						NEE (55)			V (45)
15						PFE (35)			
16						TGT (25)			
17						TMO (35)			
18						UNH (35)			
19						WBA (30)			
20						WMT (30)			

Table 1: **Stock Clusters**. The top (bottom) panel reports the cluster assignments of the S&P 100 stocks before (after) the January 31, 2020 break date. GICS sector codes are given in parentheses.

cross-stock commonalities in the semicorrelations. Interestingly, even though the specific clusters labelled A-D for the post-Covid period obviously differ from the similar named pre-Covid clusters, the dominant sectors for each of the identical named clusters remain the same. Also, the post-Covid E cluster is mostly comprised of industrials, while the F cluster

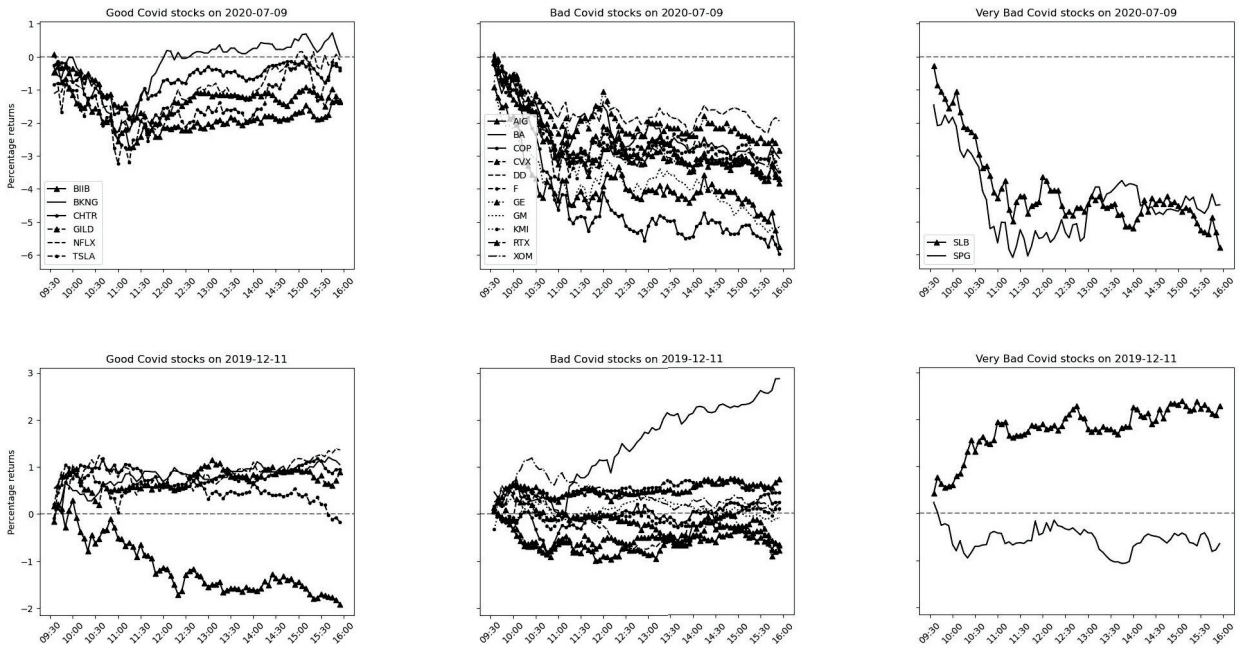

Figure 1: **Intraday Returns**. The figure shows the intraday (normalized to zero at the beginning of the day) logarithmic prices at five-minute intervals on July 7, 2020 (top panels) and December 12, 2019 (bottom panels) for each of the stocks included in the "Good Covid," "Bad Covid" and "Very Bad Covid" clusters. The scales of the y-axes for a given date are all the same.

contains mostly consumer discretionary and information technology stocks. Meanwhile, three distinct new clusters also emerge: a "Good Covid," "Bad Covid" and "Very Bad Covid" cluster of stocks. In contrast to clusters A-E, which seem to adhere fairly closely to traditional industry type classifications, these three new clusters defy such identification. The "Good Covid" cluster, for instance, contains three "online" companies (Netflix (NFLX), Booking.com (BKNG) and Spectrum (CHTR)), an electronic car maker (Tesla (TSLA)), together with two pharmaceutical companies (Biogen (BIIB) and Gilead (GILD)). Three of these stocks (NFLX, CHTR and TSLA) performed very well in 2020, easily beating the S&P 500 for the year, while the annual returns on the three other stocks (BKNG, BIIB and GILD) all fell short of the S&P 500. Instead, what sets these six stocks apart from the other stocks in the S&P 100 are their dynamic co-dependencies, and the similar ways in which they responded to news about the pandemic, as imbued in the realized semicorrelation measures.

To further illustrate this point, the top panel in Figure 1 shows the within day five-minute logarithmic prices (normalized to zero at the beginning of the day) on July 9, 2020, for each of the stocks in the three different "covid clusters." That particular trading day began with reports of sharply increasing coronavirus cases in many parts of the U.S., followed by a report later in the day of falling unemployment insurance claims, which in turn may have helped alleviate some of the worst fears about the adverse economic consequences of the pandemic. The similarities in the intraday price paths for the otherwise very different companies included in each of the three clusters, together with the cross-cluster differences, are quite striking. While all of the stocks performed poorly over the earlier part of the day, the "Good Covid" stocks ended up with fairly modest losses for the day, while the "Bad Covid" stocks, and the "Very Bad Covid" stocks, in particular, all experienced substantial end-of-day losses.

For comparison, the bottom three panels in Figure 1 show the intraday price paths for the same three "covid clusters" of stocks on December 11, 2019, before there was any awareness of the impending pandemic. On that day the Federal Reserve announced its intention to keep interest rates at current levels. As a result most, albeit not all, stocks ended up higher for the day. However, in sharp contrast to the prices on July 9, 2020, depicted in the top three panels, where the worst performing stocks are all from the two "bad" covid clusters, there is no such differentiation evident for the December 11, 2019 daily returns. Moreover, some of the December 11, 2019 intraday price paths for the stocks included in the same "covid cluster" also appear quite disparate. Of course, the "covid clusters" didn't manifest until the end of January, 2020, and as such it is not necessarily surprising that some of the different stocks only subsequently grouped together as either "good" or "bad" covid stocks reacted very differently to economic news before the pandemic.

To alternatively illuminate these changes, we apply Principal Component Analysis (PCA) directly to the high-frequency returns used in the construction of the realized measures (Aït-Sahalia and Xiu (2019)). PCA, of course, is primarily designed for dimension reduction and doesn't in and of itself provide for any clustering, let alone the identification of structural

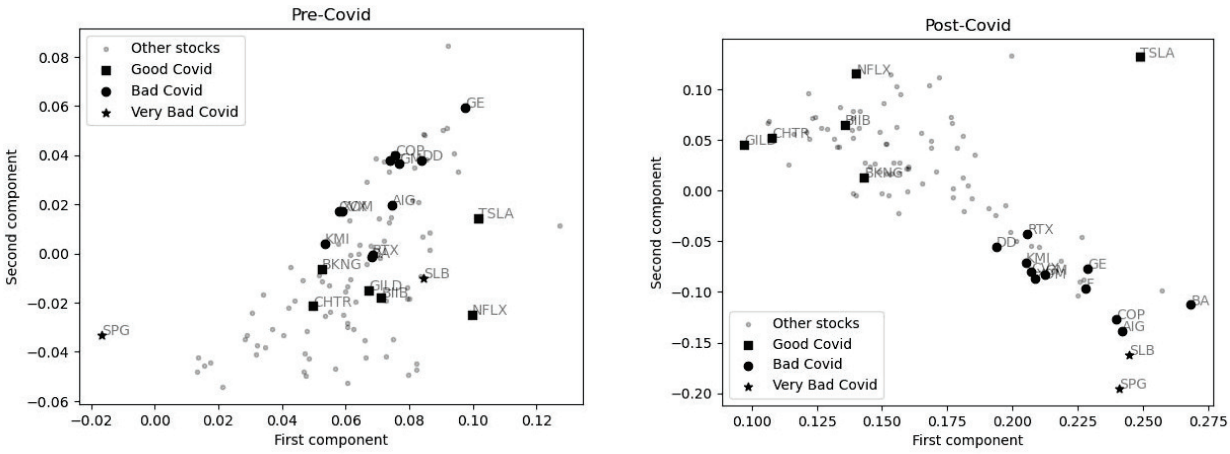

Figure 2: **Principal Component Analysis**. The figure shows the estimated loadings for the first and second principal components for each of the stocks in the S&P 100. The left (right) panel shows the estimates based on 15-minute returns before (after) January 31, 2020.

breaks in cluster assignments. As such, we rely on the January 31, 2020 break date identified by the semicorrelation-based clustering for the estimation of separate pre- and post-covid Principal Components (PCs). Figure 2 shows the resulting biplots of the loadings for the first two PCs. Looking first at the right panel pertaining to the post-covid sample, there is a clear tendency for the "bad" covid stocks to load strongly on the first PC, which tend to mimic the return on the aggregate market, and load negatively on the second PC. On the other hand, the "good" covid stocks load strongly on the second PC, and (with the exception of TSLA) are much less affected by the first PC. By contrast, looking at the results in the left panel pertaining to the PCA for the pre-covid sample, no obvious such patterns seem to exist for the "bad" versus "good" stocks. Taken together this points to the possible emergence of a new systematic risk factor around the time of the structural break in the clusters identified by the realized semicorrelations. Relying on very different procedures involving the pricing of disaster risk and measures of firms' resilience to social distancing, Pagano, Wagner and Zechner (2021) have also recently argued for the emergence of a new pandemic-related priced risk factor.

5. Conclusion

We demonstrate how new "directional" realized semicorrelation measures, constructed from high-frequency intraday returns, may be used for grouping individual stocks into clusters of stocks that respond similarly to new information. Our empirical results indicate the emergence of new and distinct "covid clusters" of stocks at the onset of the pandemic. It would be interesting to further explore the economic mechanisms behind these new clusters, and whether they may be traced to a new priced pandemic-risk factor. Relatedly, it would be interesting to more thoroughly assess the economic gains that may be available from uses of the semicorrelation-based clusters for better asset pricing, investment and risk management decisions. It would also be interesting to extend the clustering procedures developed here to explicitly allow for the possibility of more than one break point. We leave further work along these lines for future research.

References

Ahn, D.-H., Conrad, J. and Dittmar, R. F. (2009). Basis assets. *The Review of Financial Studies*, **22** (12), 5133–5174.

Aït-Sahalia, Y. and Xiu, D. (2019). Principal component analysis of high frequency data. *Journal of the American Statistical Association*, **114** (525), 287–303.

Albuquerque, R., Koskinen, Y., Yang, S. and Zhang, C. (2020). Resiliency of environmental and social stocks: An analysis of the exogenous covid-19 market crash. *Review of Corporate Finance Studies*, **9** (3), 594–621.

Andersen, T. G. and Bollerslev, T. (2018). Volatility: Introduction. In T. G. Andersen and T. Bollerslev (eds.), *Volatility*, Cheltenham, UK: Edward Elgar Publishing, pp. xiiv–xliii.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statical Surveys*, **4**, 40–79.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know. *Annual Review of Economics*, **11**, 685–725.

Bai, J. (1997). Estimation of a change point in multiple regression models. *Journal of Time Series Analysis*, **79** (4), 551–563.

Barndorff-Nielsen, O., Hansen, P., Lunde, A. and Shephard, N. (2009). Realized kernels in practice: trades and quotes. *Econometrics Journal*, **12**, C1–C32.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Econometric analysis of realized covariation: High-frequency based covariance, regression, and correlation in financial economics. *Econometrica*, **72** (3), 885–925.

Bollerslev, T., Li, J., Patton, A. J. and Quaedvlieg, R. (2020). Realized semicovariances. *Econometrica*, **88** (4), 1515–1551.

Bretscher, L., Hsu, A., Simasek, P. and Tamoni, A. (2020). Covid-19 and the cross-section of equity returns: Impact and transmission. *Review of Asset Pricing Studies*, **10** (4), 705–741.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28** (3), 591–605.

Ding, W., Levine, R., Lin, C. and Xie, W. (2021). Corporate immunity to the covid-19 pandemic. *Journal of Financial Economics*, **141**, 802–830.

Jegadeesh, N., Noh, J., Pukthuanthong, K., Roll, R. and Wang, J. (2019). Empirical tests of asset pricing models with individual assets: Resolving the errors-in-variables bias in risk premium estimation. *Journal of Financial Economics*, **133**, 273–298.

Jensen, T. I., Kelly, B. T. and Pedersen, L. H. (2021). Is there a replication crisis in finance? *Journal of Finance*, **forthcoming**.

Liu, L. Y., Patton, A. J. and Sheppard, K. (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple assets. *Journal of Econometrics*, **187**, 293–311.

Lucas, A., Schaumburg, J. and Schwaab, B. (2020). Dynamic clustering of multivariate panel data. *Working Paper, Free University Amsterdam.*

Lumsdaine, R. L., Okui, R. and Wang, W. (2021). Estimation of panel group structure models with structural breaks in group memberships and coefficients. *Working Paper, American University.*

Pagano, M., Wagner, C. and Zechner, J. (2021). Disaster resilience and asset prices. *Working Paper, University of Vienna.*

Papanikolaou, D. and Schmidt, L. D. W. (2021). Working remotely and the supply-side impact of covid-19. *Working Paper, MIT Sloan School of Management.*

Patton, A. J. and Weller, B. W. (2021). Risk price variation: The missing half of empirical asset pricing. *Review of Financial Studies*, **forthcoming**.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58** (301), 236–244.

Zhang, L., Mykland, P. A. and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, **100**, 1394–1411.