On the Ordering of Dynamic Principal Components and the Implications for Portfolio Analysis*

Giovanni Bonaccolto[a] and Massimiliano Caporin[b]

[a]School of Economics and Law, Kore University of Enna, Italy

[b]Department of Statistical Sciences, University of Padova, Italy

This version: April 21, 2022

Abstract

When principal component analysis (PCA) is used on a rolling or conditional setting, ordering and incoherence issues may emerge. We provide empirical evidence supporting this claim and introduce an algorithm that allows dynamic re-ordering of the principal components (PCs). We provide additional results that shed light on the consequences of incoherence when analyzing the link between PCs and macroeconomic risk factors, with a focus on the COVID-19 pandemic period. When PCs are optimally re-ordered, the role of factors emerges more clearly, with relevant implications for risk management.

Keywords: principal component analysis, dynamic principal component GARCH, risk factors, principal components ordering.

JEL codes: C38, C32, C58, G11, G17.

*We thank [to be added] and the participants of the 9-th International Conference on Risk Analysis [other to be added] for their stimulating comments and suggestions on a preliminary draft of this paper. The second author acknowledges financial support from the Project of Excellence 2018–2022 "Statistical Methods and Models for Complex Data" awarded by Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) to the Department of Statistical Sciences of the University of Padova, and from the MIUR-PRIN2017 project "HiDEA: Advanced Econometrics for High-frequency Data, 2017RSMPZZ".

1 Introduction

Principal component analysis (PCA) is widely used in finance. Among its many possible applications are in the identification of risk factors and the subsequent analysis of pricing (Collin-Dufresn et al., 2001; Pelger, 2020; Shukla and Trzcinka, 1990), the evaluation of the integration of markets or countries (Aziakpono et al., 2012; Billio et al., 2017; Donadelli and Paradiso, 2014; Volosovych, 2011), in dimension reduction problems within predictive models (Aye et al., 2015; Zhang and Wang, 2022; Zhong and Enke, 2017), in the analysis of commonality in trading (Fung and Hsieh, 1997; Hasbrouck and Seppi, 2001; Korajczyk and Sadka, 2008; Mancini et al., 2013; Panayi et al., 2015), and finally, in volatility prediction (Carol and Chibumba, 1996; Müller et al., 2011).

When focusing on the application of PCA to financial returns, there is a general consensus on the interpretation of the first principal component as a proxy of the market factor. Such a view is supported by the behaviour of this component and by the fact that all its loadings are generally positive (Fraser et al., 2004; Meyers, 1973). PCA is usually applied based on a time-invariant covariance matrix, leading to what we call an *unconditional* PCA. Nevertheless, it is also acknowledged in literature that the market correlation structure is subject to changes, which proves the need for a *conditional* PCA. In this case, a simple strategy would build on a rolling or moving window evaluation of the covariance matrix,[1] leading to a time-varying PCA. The latter is associated with time variation in the loadings (the eigenvectors) and in the variance of the latent components (the eigenvalues). When one is moving in this direction, a common practice is to arrange the components according to their variance, from the most volatile (the first, the market proxy) to the least volatile.

Literature is silent, however, on the possible existence of coherence issues in the principal components (PCs) identified from different samples. Let us use a motivating example, considering the returns of N assets at time t, which are included in the $N \times 1$ vector $\boldsymbol{y}_t = [y_{1,t} \cdots y_{N,t}]'$. We observe these returns in two different samples that span the time intervals $[t+1, t+M]$ and $[t+1+h, t+M+h]$, respectively, with $h << M$, and thus, with the samples partially overlapping. Starting from the first sample, we estimate the covariance matrix of the asset returns, denoted as $\boldsymbol{\Sigma}_1$, and perform a PCA, from which the following equality holds: $\boldsymbol{\Sigma}^{(1)} = \boldsymbol{L}^{(1)}\boldsymbol{D}^{(1)}\left(\boldsymbol{L}^{(1)}\right)'$, where $\boldsymbol{L}^{(1)}$ is the matrix of eigenvectors and $\boldsymbol{D}^{(1)}$ is the diagonal matrix of eigenvalues. The corresponding PCs are computed as $\boldsymbol{u}_t^{(1)} = \left(\boldsymbol{L}_1^{(1)}\right)'\boldsymbol{y}_t$. Likewise, we obtain $\boldsymbol{\Sigma}^{(2)} = \boldsymbol{L}^{(2)}\boldsymbol{D}^{(2)}\left(\boldsymbol{L}^{(2)}\right)'$ from the second sample. If the PCA is coherent in both samples, the following should be true: (i) the loadings of the PCs would be (almost) time-invariant and orthogonal—that is, given the loading matrices $\boldsymbol{L}^{(1)}$ and $\boldsymbol{L}^{(2)}$, we should have $\left(\boldsymbol{L}^{(1)}\right)'\boldsymbol{L}^{(2)} \approx \boldsymbol{I}_N$, with $\boldsymbol{I}_N$ being an $N \times N$ identity matrix; and (ii) the ranking of the PCs based on their variance would not be subject to changes. To be more specific, let us assume that the

[1] An alternative to the use of rolling application of sample moment estimation is represented by the estimation of a Multivariate Generalized AutoRegressive Conditional Heteroskedastic (MGARCH) model and the subsequent estimation of the spectral decomposition of the estimated conditional covariances. We discuss this approach in the following section.

loadings do not change between the two samples, so that we have the following structure for the first three columns of $\boldsymbol{L}^{(1)}$ and $\boldsymbol{L}^{(2)}$:

$$\left[\boldsymbol{L}^{(1)}\right]_{.,1:3} = \left[\boldsymbol{L}^{(2)}\right]_{.,1:3} = \begin{bmatrix} \boldsymbol{\mathcal{L}}_a & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\mathcal{L}}_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\mathcal{L}}_c \end{bmatrix}, \tag{1}$$

where $\boldsymbol{\mathcal{L}}_a$, $\boldsymbol{\mathcal{L}}_b$, and $\boldsymbol{\mathcal{L}}_c$ are the orthonormal vectors with lengths satisfying the equality $N_a + N_b + N_c = N$, whereas $\mathbf{0}$ is a zero vector.[2]

Let now assume that the variance of the assets changes from sample 1 to sample 2 and that this change is completely absorbed by a change in the variance of the PCs in such a way that $\mathbb{V}\left(\left[\boldsymbol{L}^{(2)}\right]_{.,3}\boldsymbol{y}_t\right) > \mathbb{V}\left(\left[\boldsymbol{L}^{(2)}\right]_{.,2}\boldsymbol{y}_t\right)$. We are thus postulating that in h observations that are not common to the two samples, a shift is realized in the market, and the variance of the PCs is altered. If this change is realized, the standard practice of ordering PCs on the basis of their variance would return a matrix $\boldsymbol{L}^{(2)}$ with switched columns, as follows:

$$\left[\breve{\boldsymbol{L}}^{(2)}\right]_{.,1:3} = \begin{bmatrix} \boldsymbol{\mathcal{L}}_a & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\mathcal{L}}_b \\ \mathbf{0} & \boldsymbol{\mathcal{L}}_c & \mathbf{0} \end{bmatrix}, \tag{2}$$

thus leading to $\left(\boldsymbol{L}^{(1)}\right)' \breve{\boldsymbol{L}}^{(2)} \neq \boldsymbol{I}_N$ and generating incoherence.

This problem is particularly relevant when PCs are derived from dynamic conditional settings (either within a rolling scheme or starting from a dynamic model), as the presence of such incoherence might impact on the interpretation of PCs. In our study, we shed some light on the empirical occurrence of this problem. We also take a step further and analyze the coherence problems of PCc by exploiting the dynamic structure provided in a Multivariate Generalized AutoRegressive Conditional Heteroskedastic (MGARCH) model belonging to the Orthogonal GARCH (OGARCH) family. We refer to the dynamic principal component (DPC) of Aielli and Caporin (2015), an MGARCH specification that allows computing of *conditional* PCs, where ordering issues may easily arise.[3] In the DPC model, both eigenvectors and eigenvalues have a dynamic evolution; that is, the loadings will change over time, as the variance of the PCs, unlike in OGARCH, where only the variance of the PCs changes over time. In addition, unlike in other MGARCH models, in the DPC model, eigenvectors have a proper dynamic, and their temporal evolution is not a by-product of the conditional covariance or correlation dynamic.

When loadings are dynamic, the interpretation of the latent components will be affected by the

[2] To simplify the notation, we do not report the dimension of the zero vector, which changes according to the dimensions of the loading vectors.

[3] Ordering issues may arise in all frameworks where a model for the conditional covariance or conditional correlation is first estimated in the data and then used to recover conditional eigenvectors.

incoherence of the PCs. For instance, the appropriate use of PCs as risk factors will be prevented, as well as an accurate evaluation of the links between PCs and economic drivers (Alexander and Kaeck, 2008; Collin-Dufresn et al., 2001), or their use in risk management (Roncalli and Weisang, 2015; Sabelli et al., 2017; Topaloglou et al., 2002).

In this paper, we provide empirical evidence of the occurrence of ordering incoherence for conditional PCs, working with the constituents of the Dow Jones Industrial Average (DJIA) index over the period of March 2019 to December 2020. Our sample includes the data from the coronavirus disease 2019 (COVID-19) outbreak and represents an interesting case study to identify how both the role and the relevance of PCs change over time. After showing the existence of incoherence, which clearly emerges when assets are clustered by economic sector, we provide a second contribution, on the methodological side, by introducing an algorithm for re-ordering the conditional PCs. We then support the appropriateness of our algorithm with further analyses. We start by highlighting the strong impact of the COVID-19 shock on the reordered components. For some components, the effect of COVID-19 is transient, and the relevant peaks observed in the first stage of the pandemic disappear after a few weeks. For other components, the effects are more persistent and last until the end of 2020, marking a clear change of regime between the pre- and the post-COVID-19 periods. Moreover, we dynamically estimate a regression model, in which we employ a large set of factors, to provide information on the risks and performance of the equity, bond, commodity and currency markets, as well as on the stress and uncertainty beyond the financial system. We deal with the large dimensionality of the resulting model by using well-known machine learning techniques. Interestingly, we find that the impact of a subset of factors becomes more relevant when looking at the reordered components, compared to the original DPC estimates. Finally, we show the improvements provided by our method within a portfolio framework. Specifically, after decomposing the variance of the minimum variance portfolio into a systematic and an idiosyncratic component, we find that the former significantly reacts to the outbreak of the COVID-19 pandemic when adopting the reordered components. In contrast, the impact of the COVID-19 shock is less evident when the original DPC and the unconditional components are used.

This paper proceeds as follows. Section 2.1 gives an overview of the DPC model, and Section 2.2 presents the details of our reordering algorithm. Section 3 describes the dataset that we used in our study and the empirical setup. Section 4 analyzes the empirical properties of the reordered eigenvectors, and Section 5 highlights the implications of our approach in terms of regression and portfolio exposure. Section 6 concludes this paper.

2 Methods

2.1 An overview of the dynamic principal components model

Let $\boldsymbol{y}_t = [y_{1,t} \cdots y_{N,t}]'$ denote an $N \times 1$ vector of returns yielded by N stocks at time t, for $t = 1, \ldots, T$. We assume that the expected value of $\boldsymbol{y}_t$ conditional on the information set available at time $t-1$ is equal to zero; that is, $\mathbb{E}_{t-1}[\boldsymbol{y}_t] = \boldsymbol{0}$, where $\boldsymbol{0}$ is an $N \times 1$ zero vector. Therefore, the conditional covariance matrix of $\boldsymbol{y}_t$ is defined as $\boldsymbol{H}_t = \mathbb{E}_{t-1}[\boldsymbol{y}_t \boldsymbol{y}_t']$ and can be expressed using the following spectral decomposition equation:

$$\boldsymbol{H}_t = \boldsymbol{L}_t \boldsymbol{D}_t \boldsymbol{L}_t', \tag{3}$$

where $\boldsymbol{L}_t$ and $\boldsymbol{D}_t$ are both $N \times N$ matrices; the diagonal entries of $\boldsymbol{D}_t = \operatorname{diag}(d_{1,t} \cdots d_{N,t})$ are the eigenvalues of $\boldsymbol{H}_t$, and the columns of $\boldsymbol{L}_t = \left[l_t^{(i,j)}\right]$ are the corresponding eigenvectors.

The conditional PCs of $\boldsymbol{y}_t$ are computed as follows:

$$\boldsymbol{u}_t = \boldsymbol{L}_t' \boldsymbol{y}_t, \tag{4}$$

so that the asset returns can be written as:

$$\boldsymbol{y}_t = \boldsymbol{L}_t \boldsymbol{u}_t. \tag{5}$$

The components computed from Equation (4) are conditional on $\mathcal{I}_{t-1}$ (i.e., the information set available at time $t-1$). They are conditionally orthogonal and have $\boldsymbol{D}_t$ as their conditional covariance matrix. We estimate $\boldsymbol{L}_t$ and $\boldsymbol{D}_t$ in Equation (3) using the dynamic principal components (DPC) model introduced by Aielli and Caporin (2015), where the term 'dynamic' points out the time-varying nature of the linear mapping from the PCs to the corresponding asset returns. We briefly describe the DPC model as follows.

In its first step, the DPC model focuses on the estimation of the $\boldsymbol{L}_t$ matrix. We stress that $\boldsymbol{L}_t$ is orthonormal (i.e., $\boldsymbol{L}_t \boldsymbol{L}_t' = \boldsymbol{L}_t' \boldsymbol{L}_t = \boldsymbol{I}_N$, where $\boldsymbol{I}_N$ is the N-dimensional identity matrix) due to the properties of the spectral decomposition of its positive definite matrix (Gruber, 2013). This estimation builds on an auxiliary process, the so-called *loading-driving process*, which reproduces the features of the underlying loading dynamics under the required orthonormality constraints. Aielli and Caporin (2015) modeled this auxiliary process by resorting to the following BEKK specification (Ding and Engle, 2001; Engle and Kroner, 1995; Engle and Mezrich, 1996):

$$\boldsymbol{Q}_t = (1 - a - b)\, \boldsymbol{S} + a \boldsymbol{y}_{t-1} \boldsymbol{y}_{t-1}' + b \boldsymbol{Q}_{t-1}, \tag{6}$$

where $(\boldsymbol{y}_0, \boldsymbol{Q}_0) \in \mathcal{I}_0$, whereas the scalars a and b, along with the $\boldsymbol{S}$ matrix, are parameters to be estimated.

Under the assumptions $a \geq 0$, $b \geq 0$, $a + b < 1$, $\boldsymbol{S} \succ \boldsymbol{0}$ and $\boldsymbol{Q}_0 \succ \boldsymbol{0}$ (i.e., $\boldsymbol{S}$ and $\boldsymbol{Q}_0$ are positive definite matrices), $\boldsymbol{Q}_t$ is positive definite, so that its spectral decomposition exists for each t. After obtaining $\boldsymbol{Q}_t$ from Equation (6), it is possible to retrieve the matrix of conditional loadings $\boldsymbol{L}_t$, which is defined as the eigenvector matrix of $\boldsymbol{Q}_t$:

$$\boldsymbol{Q}_t = \boldsymbol{L}_t \boldsymbol{G}_t \boldsymbol{L}_t', \tag{7}$$

where $\boldsymbol{G}_t = \operatorname{diag}(g_{1,t}, \ldots, g_{N,t})$, with $g_{1,t}, \ldots, g_{N,t}$ being the eigenvalues of $\boldsymbol{Q}_t$.

The DPC model requires the following conditions to ensure the uniqueness of the spectral decomposition in Equation (7) and the existence of a unique loadings sequence for a given dataset: (i) the eigenvalues of $\boldsymbol{Q}_t$ are arranged in strictly decreasing order; (ii) the sign of the corresponding eigenvectors is such that the diagonal elements of $\boldsymbol{L}_t'\boldsymbol{L}$ are positive, where $\boldsymbol{L}$ is computed from the following spectral decomposition of $\boldsymbol{S}$: $\boldsymbol{S} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}'$, with $\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_N)$; and (iii) for a given covariance-stationary process, the magnitude of $\boldsymbol{S}$ is restricted by satisfying the equality $\operatorname{tr}(\boldsymbol{S}) = \operatorname{tr}\left(\overline{\boldsymbol{S}}\right)$, where $\overline{\boldsymbol{S}} = \mathbb{E}[\boldsymbol{y}_t \boldsymbol{y}_t']$ is the unconditional covariance matrix of $\boldsymbol{y}_t$ (Aielli and Caporin, 2015).

In the next step, the DPC model estimates the conditional variances of the PCs using a univariate GARCH model (Bollerslev, 1986) with variance targeting (Engle and Mezrich, 1996):

$$\mathbb{E}_{t-1}\left[u_{i,t}^2\right] = d_{i,t} = (1 - \alpha_i - \beta_i)\,\gamma_i + \alpha_i u_{i,t-1}^2 + \beta_i d_{i,t-1}, \tag{8}$$

where $(u_{i,0}, d_{i,0}) \in \mathcal{I}_0$, $\gamma_i = d_i$, $\alpha_i \geq 0$, $\beta_i \geq 0$, $\alpha_i + \beta_i < 1$ and $d_{i,0} > 0$, for $i = 1, \ldots, N$.[4]

Since $d_i > 0$, it follows that $d_{i,t} > 0$, so $\boldsymbol{H}_t = \boldsymbol{L}_t \boldsymbol{D}_t \boldsymbol{L}_t'$ is positive definite. Furthermore, the condition $\alpha_i + \beta_i < 1$ ensures that the PCs are covariance-stationary, with the unconditional second moment equal to $\mathbb{E}\left[u_{i,t}^2\right] = \mathbb{E}[d_{i,t}] = d_i$ (Aielli and Caporin, 2015; Bollerslev, 1986). Under the assumption that the sequence $d_1, \ldots, d_N$ is arranged in strictly decreasing order, it follows that $\mathbb{E}\left[u_{1,t}^2\right] > \mathbb{E}\left[u_{2,t}^2\right] > \cdots > \mathbb{E}\left[u_{N,t}^2\right]$. As a result, the components are arranged in decreasing order according to their corresponding unconditional variances (Aielli and Caporin, 2015).

Aielli and Caporin (2015) referred the model resulting from Equations (5)—(7) and (8) as the scalar DPC model, where the term 'scalar' refers to the scalar BEKK recursion in Equation (6). Aielli and Caporin showed that under weak stationary conditions, the loading process of the scalar DPC model is identified. Moreover, under usual conditions, the DPC model possesses the loading targeting property (Aielli and Caporin, 2015); that is, the columns of $\boldsymbol{L}$ are the eigenvectors of the stationary second moment of $\boldsymbol{y}_t$, as shown in the following equation:

$$\overline{\boldsymbol{S}} = \boldsymbol{L}\overline{\boldsymbol{D}}\boldsymbol{L}', \tag{9}$$

[4] In addition to the GARCH(1,1) specification, we can flexibly employ any other univariate GARCH specification, with the possibility of including exogenous regressors and leverage effects (Aielli and Caporin, 2015).

where $\overline{\boldsymbol{D}}$ is a diagonal matrix.

The loading targeting property can also be expressed by means of the following relationship:

$$\overline{\boldsymbol{u}}_t = \boldsymbol{L}'\boldsymbol{y}_t, \tag{10}$$

where $\overline{\boldsymbol{u}}_t$ is the vector of the unconditional PCs of $\boldsymbol{y}_t$ (Aielli and Caporin, 2015).

The scalar BEKK recursion defined in Equation (6) leads to the simplest specification of the DPC model (i.e., the scalar DPC model). Aielli and Caporin (2015) also proposed a more general specification, which provides greater flexibility in the loading dynamics. This additional specification builds on the Full BEKK(1,1,1) recursion, which is defined as follows:

$$\boldsymbol{Q}_t = \left(\boldsymbol{S} - \boldsymbol{\mathcal{A}}\boldsymbol{S}\boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{B}}\boldsymbol{S}\boldsymbol{\mathcal{B}}'\right) + \boldsymbol{\mathcal{A}}\boldsymbol{y}_{t-1}\boldsymbol{y}_{t-1}'\boldsymbol{\mathcal{A}}' + \boldsymbol{\mathcal{B}}\boldsymbol{Q}_{t-1}\boldsymbol{\mathcal{B}}', \tag{11}$$

where

$$\boldsymbol{\mathcal{A}} = \boldsymbol{L}\boldsymbol{V}\boldsymbol{L}', \quad \boldsymbol{\mathcal{B}} = \boldsymbol{L}\boldsymbol{B}\boldsymbol{L}', \tag{12}$$

$$\boldsymbol{V} = \operatorname{diag}\left(\sqrt{v_1}, \ldots, \sqrt{v_N}\right) \text{ and } \boldsymbol{B} = \operatorname{diag}\left(\sqrt{b_1}, \ldots, \sqrt{b_N}\right). \tag{13}$$

Under the conditions $\boldsymbol{Q}_0 \succ \boldsymbol{0}$, $\boldsymbol{S} \succ \boldsymbol{0}$, $v_i \geq 0$, $b_i \geq 0$ and $v_i + b_i < 1$, for $i = 1, \ldots, N$, the intercept of $\boldsymbol{Q}_t$ is positive definite, which ensures that $\boldsymbol{Q}_t$ is also positive definite (Engle and Kroner, 1995) and that the spectral decompositions of $\boldsymbol{Q}_t$ exist (Aielli and Caporin, 2015). Aielli and Caporin identified the model resulting from Equations (5), (7), (8) and (11)—(13) as the DPC model.

2.2 Economic sector and eigenvector reordering

We estimated the conditional eigenvector matrix $\boldsymbol{L}_t$ given in Equation (3) using the DPC model (Aielli and Caporin, 2015). We denoted the resulting estimate as $\widehat{\boldsymbol{L}}_t$, with $t = 1, \ldots, T$. In this study, we investigate whether the ordering of the columns of $\widehat{\boldsymbol{L}}_t$ is affected by the sector classification of these companies. If there is a connection between the sector classification and the positions of the columns of $\widehat{\boldsymbol{L}}_t$, we can rearrange the columns of $\widehat{\boldsymbol{L}}_t$ to increase the stability of the estimates in the time interval $[1, T]$.

Specifically, for each column j of $\widehat{\boldsymbol{L}}_t$ (that corresponds to the j-th eigenvector, with $j = 1, \ldots, N$) and for each time period t (with $t = 1, \ldots, T$), we aggregated the N companies into K economic sectors. Then, we measured the relevance of each sector within each eigenvector and for each time period by computing the weight:

$$\widehat{w}_{k,t,j} = \sum_{i=1}^{N} \left(\hat{l}_t^{(i,j)}\right)^2 \mathbb{I}_{i \in k}, \tag{14}$$

for $k = 1, \ldots, K$, $t = 1, \ldots, T$ and $j = 1, \ldots, N$, where $\mathbb{I}_{i \in k}$ is an indicator function that has a value of 1 if the i-th company belongs to the k-th sector, and a value of 0 otherwise.

To evaluate the dynamics of $\widehat{w}_{k,t,j}$ over time, we built the following $K \times T$ matrix:

$$\widehat{\boldsymbol{A}}_j = \begin{bmatrix} \widehat{w}_{1,1,j} & \cdots & \widehat{w}_{1,T,j} \\ \vdots & \ddots & \vdots \\ \widehat{w}_{K,1,j} & \cdots & \widehat{w}_{K,T,j} \end{bmatrix} \tag{15}$$

for $j = 1, \ldots, N$.

In case of perfect stability along the time interval $[1, T]$, the entries placed on each row of $\widehat{\boldsymbol{A}}_j$ take the same value; that is, $\widehat{w}_{k,1,j} = \widehat{w}_{k,2,j} = \cdots = \widehat{w}_{k,T,j}$. However, we typically observed fluctuations due to the dynamics of the financial markets. In addition to these typical and expected fluctuations, a different source of instability may be the 'unnatural' positioning of some eigenvectors along the time interval $[1, T]$. This instability implies sudden and transient shifts of entire columns of $\widehat{\boldsymbol{A}}_j$, which are significantly distant from the dominant cluster of $\widehat{\boldsymbol{A}}_j$; that is, the cluster which includes the majority of the columns of $\widehat{\boldsymbol{A}}_j$ that exhibited a similar behaviour along the time interval $[1, T]$. We clarify this point with an example in Figure 1, where we display the area plots of the $\widehat{\boldsymbol{A}}_{23}$ [panel (a)] and $\widehat{\boldsymbol{A}}_{24}$ [panel (c)] matrices obtained from our dataset.[5] These matrices have $K = 8$ rows, corresponding to the economic sectors of the analyzed companies, and $T = 450$ columns, corresponding to the number of periods of each time series. We clearly observed the presence of two clusters for the $\widehat{\boldsymbol{A}}_{24}$ matrix. A first and dominant cluster collected the majority of the columns of $\widehat{\boldsymbol{A}}_{24}$, and a second cluster included the columns of $\widehat{\boldsymbol{A}}_{24}$ from 193 to 198 and from 200 to 223. Furthermore, we saw an evident correspondence of $\widehat{\boldsymbol{A}}_{24}$ with the $\widehat{\boldsymbol{A}}_{23}$ matrix in a subset of their columns. Indeed, columns 193—198 and 200—223 of $\widehat{\boldsymbol{A}}_{23}$ do not belong to the dominant cluster of $\widehat{\boldsymbol{A}}_{23}$ but to the dominant cluster of $\widehat{\boldsymbol{A}}_{24}$. Therefore, we could increase the stability of the DPC estimates by replacing columns 193—198 and 200—223 of $\widehat{\boldsymbol{A}}_{24}$ with the same columns of $\widehat{\boldsymbol{A}}_{23}$, and vice versa. This graphical example also helped us identify a clear relationship between the sector classification of the companies included in our dataset and $\widehat{\boldsymbol{L}}_t$.

In this study, we rearranged the $\widehat{\boldsymbol{A}}_1, \ldots, \widehat{\boldsymbol{A}}_N$ matrices by implementing the following reordering algorithm.

1. We determined the number of clusters within each $\widehat{\boldsymbol{A}}_j$ matrix using three alternative measures: (i) the Davies-Bouldin index (Davies and Bouldin, 1979); (ii) the Calinski-Harabasz index (Calinski and Harabasz, 1974); and (iii) the silhouette values (Rousseeuw, 1987). Then, for each $\widehat{\boldsymbol{A}}_j$ matrix, we computed the mode, denoted as m_j, of the number of clusters suggested by the three stated approaches, for $j = 1, \ldots, N$. By doing so, the choice of the number of clusters was more robust, as we used the information retrieved from the three different methods.

2. We implemented the k-means clustering method for each $\widehat{\boldsymbol{A}}_j$ matrix using m_j as the optimal number of clusters and the squared Euclidean distance metric for $j = 1, \ldots, N$. Then, we

[5] We describe the dataset employed in our study in Section 3.

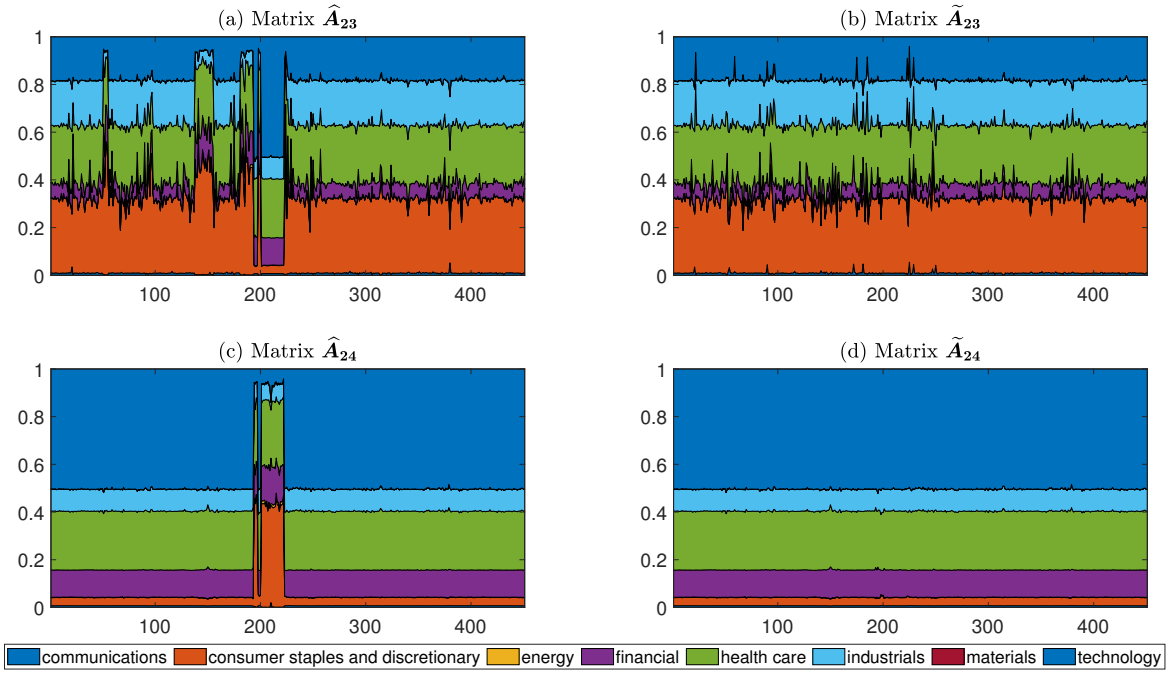


Figure 1: Graphical representation of the $\widehat{\boldsymbol{A}}_{23}$, $\widehat{\boldsymbol{A}}_{24}$, $\widetilde{\boldsymbol{A}}_{23}$ and $\widetilde{\boldsymbol{A}}_{24}$ matrices.

allocated the T columns of $\widehat{\boldsymbol{A}}_j$ to the m_j clusters identified for the same matrix and determined the dominant cluster of $\widehat{\boldsymbol{A}}_j$, that is, the cluster of $\widehat{\boldsymbol{A}}_j$ that recorded the greatest frequency (i.e., the greatest number of columns of $\widehat{\boldsymbol{A}}_j$). We denoted the frequency of the dominant cluster of $\widehat{\boldsymbol{A}}_j$ as f_j, and its relative frequency, as $p_j = f_j/T$. We also computed the sum of the point-to-centroid distances within the dominant cluster of $\widehat{\boldsymbol{A}}_j$, denoted as s_j, from which we obtained its average value $\delta_j = s_j/f_j$. From δ_j and p_j, we calculated the ratio $r_j = \delta_j/p_j$. Finally, we obtained a $T \times 1$ vector $\boldsymbol{x}_j = [x_{j,1} \cdots x_{j,T}]'$, where $x_{j,1}, \ldots, x_{j,T}$ are the distances from each point (i.e., each column) to the centroid of the dominant cluster of $\widehat{\boldsymbol{A}}_j$, for $j = 1, \ldots, N$.

3. We identified the $\widehat{\boldsymbol{A}}_j$ matrix with the lowest value of r_j, denoted as $\widehat{\boldsymbol{A}}_{j^\star}$, with $1 \leq j^\star \leq N$. $\widehat{\boldsymbol{A}}_{j^\star}$ exhibited a more persistent and stable structure of the dominant cluster than the other $\widehat{\boldsymbol{A}}_j$ matrices, with $1 \leq j \leq N$ and $j \neq j^\star$.

4. For each $t = 1, \ldots, T$ and $j = 1, \ldots, N$, we replaced the t-th column of $\widehat{\boldsymbol{A}}_{j^\star}$ with the t-th column of $\widehat{\boldsymbol{A}}_j$ and denoted the resulting matrix as $\widehat{\boldsymbol{A}}_{j^\star}^{(j,t)}$. Therefore, $\widehat{\boldsymbol{A}}_{j^\star}^{(j,t)}$ differs from $\widehat{\boldsymbol{A}}_{j^\star}$ only in the t-th column when $j \neq j^\star$. Then, we re-implemented the k-means clustering method on $\widehat{\boldsymbol{A}}_{j^\star}^{(j,t)}$ using the original number of clusters of $\widehat{\boldsymbol{A}}_{j^\star}$ (i.e., $m_{j^\star}$) determined in Step 1. Among the different measures defined in Step 2, we focused on $x_{j^\star,t}^{(j)}$: the distance between the t-th column and the centroid of the dominant cluster of $\widehat{\boldsymbol{A}}_{j^\star}^{(j,t)}$. We highlight that the replacement of the t-th column of $\widehat{\boldsymbol{A}}_{j^\star}$ with the t-th column of $\widehat{\boldsymbol{A}}_j$ is appropriate if and only if $x_{j^\star,t}^{(j)} < x_{j^\star,t}$, where $x_{j^\star,t}$ was calculated in Step 2. We iterated the procedure described above for $t = 1, \ldots, T$ and

$j = 1, \ldots, N$ and obtained the $T \times N$ matrix as follows:

$$\boldsymbol{X}_{j^\star} = \begin{bmatrix} x^{(1)}_{j^\star,1} & \cdots & x^{(N)}_{j^\star,1} \\ \vdots & \ddots & \vdots \\ x^{(1)}_{j^\star,T} & \cdots & x^{(N)}_{j^\star,T} \end{bmatrix}. \tag{16}$$

5. For $t = 1, \ldots, T$, we identified the minimum value of the t-th row of $\boldsymbol{X}_{j^\star}$ and denoted the position of the corresponding row and column as (t, j^*), with $1 \leq j^* \leq N$. If $j^\star \neq j^*$, we replace the t-th column of $\widehat{\boldsymbol{A}}_{j^\star}$ with the t-th column of $\widehat{\boldsymbol{A}}_{j^*}$ and vice versa, for $t = 1, \ldots, T$. We then obtained a modified version of $\widehat{\boldsymbol{A}}_{j^\star}$, which we denoted as $\widetilde{\boldsymbol{A}}_{j^\star}$.

6. We removed $\widetilde{\boldsymbol{A}}_{j^\star}$ from our algorithm to conclude the current iteration.

7. We repeated Steps 1—6 until the $(N-1)$-th iteration, focusing, from time to time, on the residual $\widehat{\boldsymbol{A}}_j$ matrices.

The described algorithm provides the new reordered matrices $\widetilde{\boldsymbol{A}}_1, \ldots, \widetilde{\boldsymbol{A}}_N$. Our method increased the stability in the dynamics of the original $\widehat{\boldsymbol{A}}_1, \ldots, \widehat{\boldsymbol{A}}_N$ matrices along the time interval $[1, T]$. An example is given in Figure 1, where we compare the original $\widehat{\boldsymbol{A}}_{23}$ [panel (a)] and $\widehat{\boldsymbol{A}}_{24}$ [panel (c)] matrices with their corresponding counterparts $\widetilde{\boldsymbol{A}}_{23}$ [panel (b)] and $\widetilde{\boldsymbol{A}}_{24}$ [panel (d)]. Figure 1 shows a graphical example of the improvements generated by our reordering algorithm on $\widehat{\boldsymbol{A}}_{23}$ and $\widehat{\boldsymbol{A}}_{24}$, which were retrieved from the 23-th and 24-th columns of $\widehat{\boldsymbol{L}}_t$, respectively. In addition, we calculated a measure that resembles the turnover within a portfolio framework (see, e.g., Kang et al., 2018; Pun and Wang, 2020), which allows for a complete assessment of the stability of all eigenvectors. For this purpose, for a given time point t, we collected the weights of the K economic sectors, as defined in Equation (14), within a new $N \times K$ matrix $\widehat{\boldsymbol{W}}_t$, whose j-th row is equal to the vector $[\widehat{w}_{1,t,j} \cdots \widehat{w}_{K,t,j}]$, for $j = 1, \ldots, N$. $\widehat{\boldsymbol{W}}_t$ was obtained from the original DPC eigenvectors and was compared to $\widetilde{\boldsymbol{W}}_t$; that is, the counterpart retrieved from the rearranged matrices $\widetilde{\boldsymbol{A}}_1, \ldots, \widetilde{\boldsymbol{A}}_N$. Then, we defined the following stability indicator:

$$\widehat{z}_t = \frac{\mathbf{1}'_K \operatorname{abs}\left(\widehat{\boldsymbol{W}}'_t - \widehat{\boldsymbol{W}}'_{t-1}\right) \mathbf{1}_N}{2K} \tag{17}$$

for $t = 2, \ldots, T$, where $\operatorname{abs}\left(\widehat{\boldsymbol{W}}'_t - \widehat{\boldsymbol{W}}'_{t-1}\right)$ is a $K \times N$ matrix whose elements are the absolute values of the pairwise differences in $\widehat{\boldsymbol{W}}'_t - \widehat{\boldsymbol{W}}'_{t-1}$, and $\mathbf{1}_K$ and $\mathbf{1}_N$ are $K \times 1$ and $N \times 1$ unit vectors, respectively.

We normalized the indicator defined in Equation (17) by $2K$ so that it would range within the interval $[0, 1]$. In particular, $\widehat{z}_t = 0$ when the weights of the K economic sectors have the maximum stability along the time interval $[1, T]$. Likewise, we defined the indicator $\widetilde{z}_t$ by replacing $\widehat{\boldsymbol{W}}_t$ with $\widetilde{\boldsymbol{W}}_t$ in Equation (17). We display the trend of the $\widetilde{z}_t$ and $\widehat{z}_t$ indicators in Figure 2. Figure 2 provides further evidence of the stability improvements achieved with our reordering algorithm. Compared with

$\widehat{z}_t$, $\widetilde{z}_t$ has, on average, lower values and is less volatile. Indeed, the means of $\widetilde{z}_t$ and $\widehat{z}_t$ are equal to 0.4117 and 0.4876, with variances of 0.0090 and 0.0210, respectively.

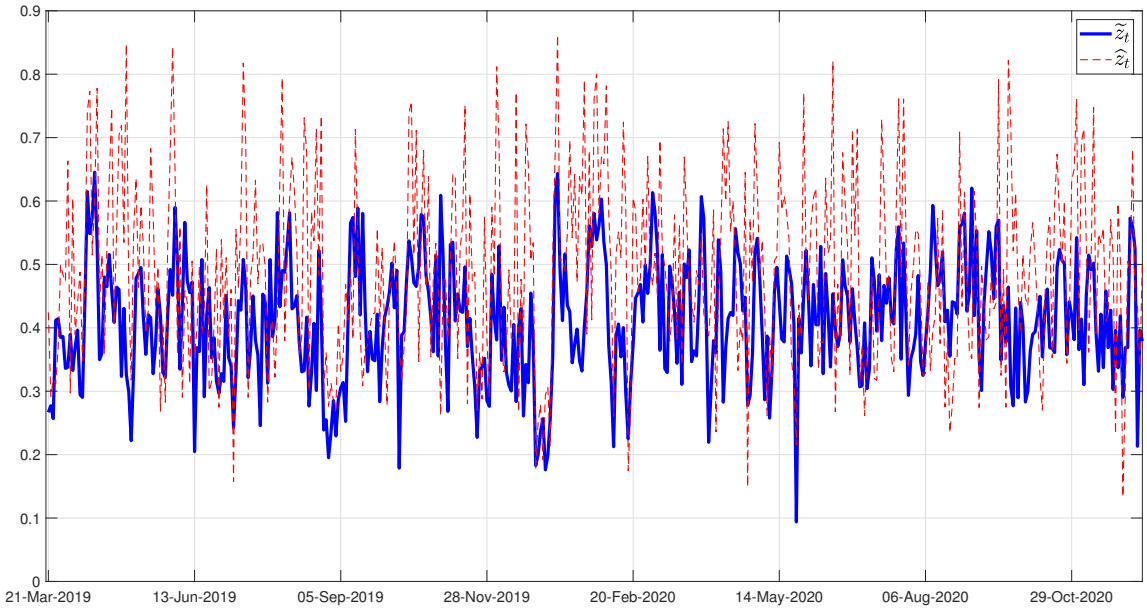


Figure 2: Trend of the $\widetilde{z}_t$ and $\widehat{z}_t$ indicators.

Given the one-to-one correspondence between the t-th column of $\widehat{\boldsymbol{A}}_j$ and the j-th column of $\widehat{\boldsymbol{L}}_t$, we also rearranged $\widehat{\boldsymbol{L}}_t$, for $t = 1, \ldots, T$ and $j = 1, \ldots, N$ and obtained the new matrices of eigenvectors $\widetilde{\boldsymbol{L}}_1, \ldots, \widetilde{\boldsymbol{L}}_T$. Specifically, for a given $t \in \{1, \ldots, T\}$ and a given $j \in \{1, \ldots, N\}$, we identified which of the N columns of $\widehat{\boldsymbol{L}}_t$ generated the t-th column of $\widetilde{\boldsymbol{A}}_j$. Then, we used this selected column as the j-th column of $\widetilde{\boldsymbol{L}}_t$. After obtaining $\widetilde{\boldsymbol{L}}_t$, we computed the associated vector of the conditional PCs as $\widetilde{\boldsymbol{u}}_t = [\widetilde{u}_{1,t} \cdots \widetilde{u}_{N,t}]' = \widetilde{\boldsymbol{L}}_t' \boldsymbol{y}_t$.

3 Data description and empirical setup

Our dataset included the daily returns of the companies that belonged to the basket of the DJIA index from March 20, 2019 to December 9, 2020, so we had $T = 450$ trading days for each time series.[6] Therefore, we focused on the period characterised by the outbreak of the COVID-19 pandemic. The companies included in our dataset were classified into eight sectors: communications (2 companies), consumer staples and discretionary (7), energy (1), financials (4), health care (4), industrials (4), materials (1) and technology (7). Appendix A lists the 30 companies included in our dataset.

In Section 5, we evaluate the relevance of our method within a portfolio framework, on the basis of a regression model in which we use a 25×1 vector $\boldsymbol{F}_t$. Specifically, $\boldsymbol{F}_t$ includes the regressors described as follows. The first five variables are the Fama-French research factors: (i) MKT-RF (the

[6]The data were recovered from Refinitiv Eikon.

excess market return); (ii) SMB (Small Minus Big, the average return on the small stock portfolios minus the average return on the big stock portfolios); (iii) HML (High Minus Low, the average return on the value portfolios minus the average return on the growth portfolios); (iv) RMW (Robust Minus Weak, the average return on the robust operating profitability portfolios minus the average return on the weak operating profitability portfolios); and (v) CMA (Conservative Minus Aggressive, the average return on the conservative investment portfolios minus the average return on the aggressive investment portfolios). Then, we considered the following: (vi) RF (the change in the risk-free rate) and (vii) MOM (the Momentum Factor, the average return on the high-prior-return portfolios minus the average return on the low-prior-return portfolios).[7] In addition, we used (viii) DJII (the return of the Dow Jones Industrials Index); (ix) OIL (the change in the Crude Oil-WTI Spot price); (x) DOLLAR (the return of the US Dollar Index); (xi) GOLD (the return of the Gold Bullion LBM); (xii) EFFR (the change in the Effective Federal Funds Rate); (xiii) T10Y3M (the change in the 10-Year Treasury Constant Maturity Minus 3-Month Treasury Constant Maturity); (xiv) BAA-AAA (the change in the spread between Moody's Seasoned Aaa and Baa Corporate Bond Yield Relative to the Yield on the 10-Year Treasury Constant Maturity); (xv) VIX (the change in the CBOE SPX Volatility VIX Index); and (xvi) VOIL (the change in the CBOE Crude Oil Volatility Index).[8] We further emphasize the impact of risks considering the following stress indexes: (xvii) FSI (the change in the OFR Financial Stress Index, which incorporates five categories of indicators: credit, equity valuation, funding, safe assets and volatility); (xviii) US FSI (the change in the FSI relative to the US area); (xix) OAE FSI (the change in the FSI relative to the advanced economies area) and (xx) EM FSI (the change in the FSI relative to the emerging markets area).[9] We also considered the uncertainty arising from newspaper articles with the following variables: (xxi) EPU (the change in the Economic Policy Uncertainty Index) and (xxii) EMU (the change in the Equity Market Uncertainty Index).[10] Finally, we captured the direct impact of the COVID-19 pandemic with the following variables: (xxiii) RECOV (the change in the number of US recovered citizens); (xxiv) CONFIRM (the change in the number of US confirmed citizens); and (xxv) DEATHS (the change in the number of US deaths due to COVID-19).[11] Each of the time series included in $\boldsymbol{F}_t$ were standardized, so we had a set of 25 factors expressed in the same scale.

[7]The data on the variables from (i) to (vii) were recovered from the Kenneth R. French library at https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

[8]The data on the variables from (viii) to (xi), (xv) and (xvi) were recovered from Refinitiv Eikon and the data on the variables from (xii) to (xiv) were recovered from the Federal Reserve Bank of St. Louis at https://fred.stlouisfed.org.

[9]We recovered the data on the variables from (xvii) to (xx) from the Office of Financial Research of the US Department of the Treasury at https://www.financialresearch.gov/financial-stress-index/.

[10]We recovered the data on the variables in (xxi) and (xxii) from https://www.policyuncertainty.com.

[11]We recovered the data on the variables from (xxiii) to (xxv) from the COVID-19 Data Repository of the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University at https://github.com/CSSEGISandData/COVID-19.

4 Empirical properties of the reordered eigenvectors

First, we evaluated the dynamics of the conditional principal components computed from the reordered eigenvectors; that is, the sequence $\widetilde{u}_{j,1}, \ldots, \widetilde{u}_{j,T=450}$, for $j = 1, \ldots, N = 30$. Specifically, we started by implementing the Engle's (1982) AutoRegressive Conditional Heteroskedasticity (ARCH) test for conditional heteroscedasticity. Then, we identified the components for which the null hypothesis of no ARCH effects was rejected with a 10% significance level by using either 5 or 15 lags.[12] We found that 11 components did not satisfy this: components 12, 15, 16, 19, 20, 21, 24, 26, 28, 29 and 30. For these components, we computed their respective unconditional sample variances. In contrast, we estimated an exponential GARCH (EGARCH) model (Nelson, 1991) with a degree of 1 for each GARCH, ARCH and leverage polynomial for the remaining 19 components. Nevertheless, the EGARCH optimization provided local and suboptimal solutions for a subset of these components (i.e., components 6, 7, 8 and 14). As a solution, we estimated a GARCH model (Bollerslev, 1986) with a degree of 1 for each GARCH and ARCH polynomial for components 6, 7, 8 and 14 that yielded more accurate estimates.

We display the variances of the 30 components on the right sides of Figures 3—5. From Figure 3 to Figure 5, we sort the positions of the underlying components according to the sample mean of the associated variances, from the highest to the lowest. For instance, the first component (on the first row of Figure 3) records, on average, the greatest variance during the period March 20, 2019—December 9, 2020, equal to 0.00706. In contrast, the 30-th component (on the last row of Figure 5) generated the lowest variance, equal to 0.00004. We stress that the horizontal lines on the right columns of Figures 3—5 correspond to the sample variances of the components for which the null hypothesis of no ARCH effects was not rejected. We also display on the left side of Figures 3—5 the area plots of the $\widetilde{\boldsymbol{A}}_1, \ldots, \widetilde{\boldsymbol{A}}_{30}$ matrices. First, it is interesting to observe that the dynamics of $\widetilde{\boldsymbol{A}}_j$ became more volatile as the variance of the corresponding component increased. Second, we highlight that nine out of 10 components were affected by heteroskedasticity, as shown in Figure 3. This number is reduced to three in Figure 5. Therefore, the greater the volatility in the dynamics of $\widetilde{\boldsymbol{A}}_j$ is, the greater is the probability of observing heteroskedasticity for the corresponding $\widetilde{u}_{j,t}$ time series. Third, the variances displayed on the right side of Figures 3—5 point out a strong impact of the COVID-19 shock. Indeed, the variance of many components significantly increased around March 2020. This evidence is more strongly seen in Figures 3 and 4. For some components, the effect of COVID-19 was transient, and the relevant peaks observed in March 2020 disappeared after a few weeks. This is the case, for instance, with components 1, 14 and 11, as seen in Figure 3. For other components, the effect of the COVID-19 pandemic was more persistent and lasted until the end of 2020, marking a clear change of regime between the pre- and the post-COVID-19 periods. This is the case, for example, with components 7, 2 and 10.

[12] We found that this setup fits well the dynamics of the estimated components.

Figure 3: $\widetilde{\boldsymbol{A}}_j$ matrices (left side) and variances (right side) of the j-th components (first group). From top to bottom, we adopt the following order of j: 1, 14, 7, 2, 10, 11, 3, 9, 4, 12.

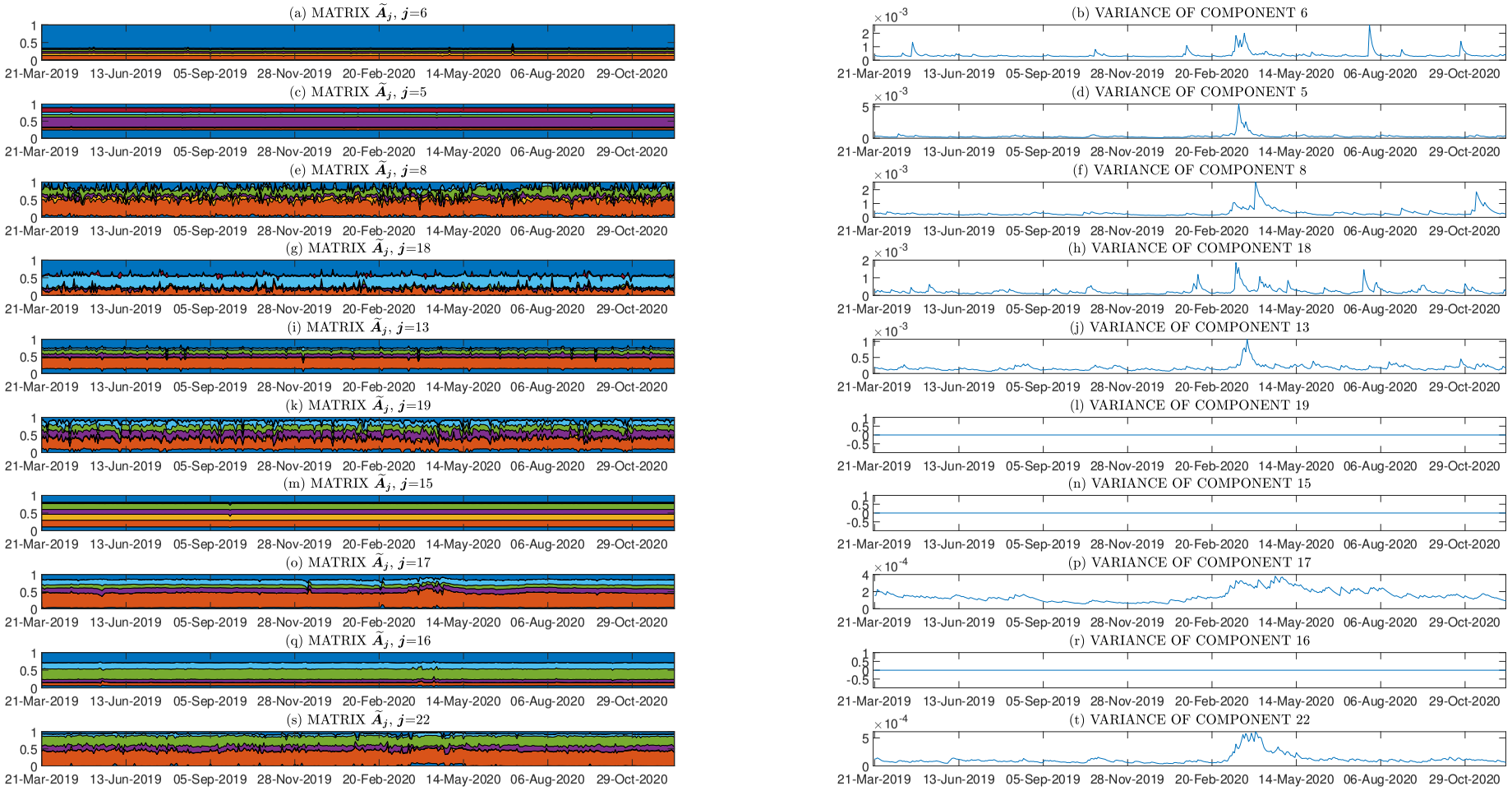


Figure 4: $\widetilde{\boldsymbol{A}}_j$ matrices (left side) and variances (right side) of the j-th components (second group). From top to bottom, we adopt the following order of j: 6, 5, 8, 18, 13, 19, 15, 17, 16, 22.

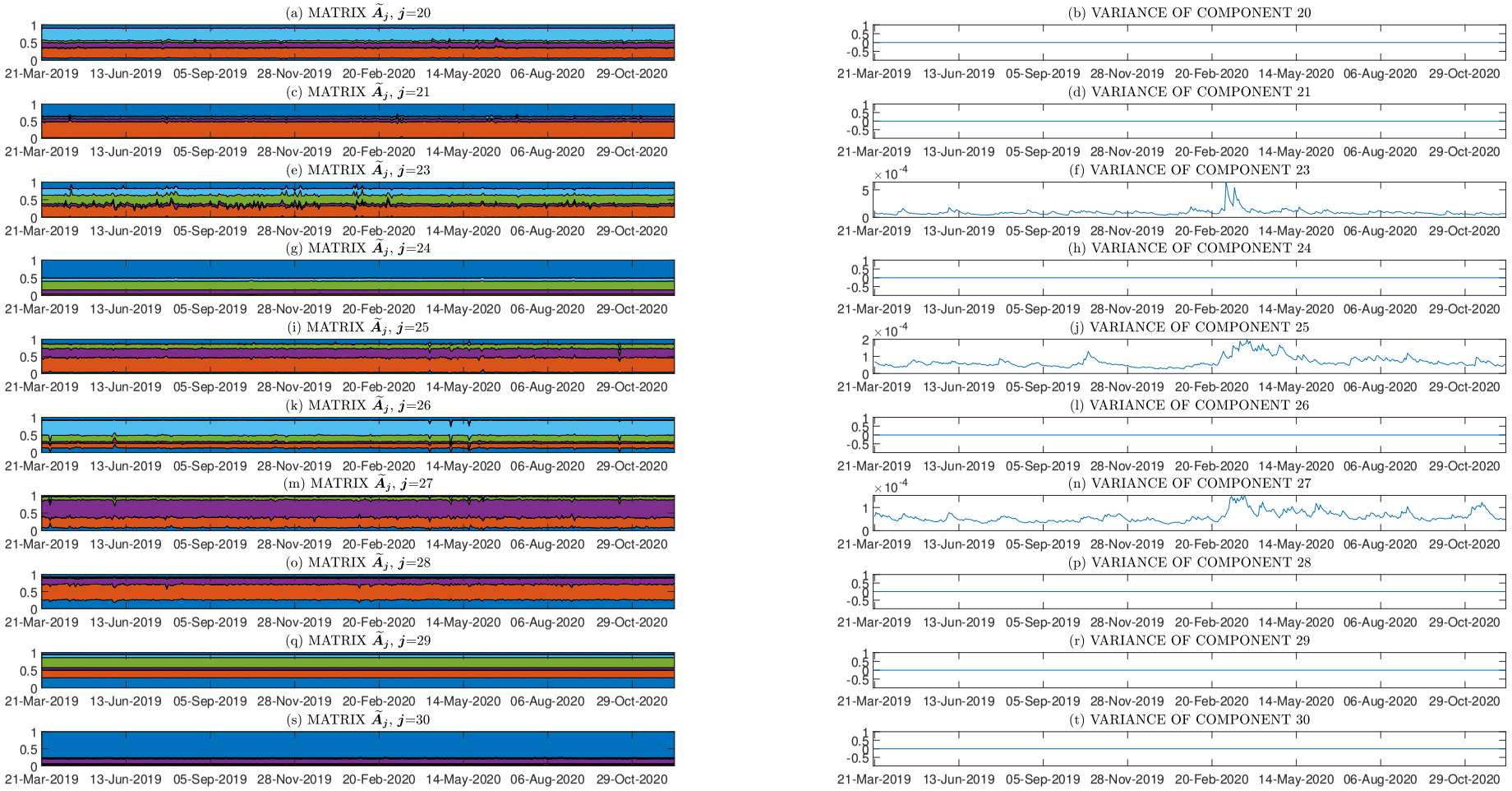


Figure 5: $\widetilde{\boldsymbol{A}}_j$ matrices (left side) and variances (right side) of the j-th components (third group). From top to bottom, we adopt the following order of j: 20, 21, 23, 24, 25, 26, 27, 28, 29, 30.

Figure 6 ranks the 30 conditional PCs according to their daily variances. For instance, on the first day (represented as the first column of Figure 6), the first component had the highest variance, 0.00702, and therefore, had rank 1. The 14-th component had rank 2, as it recorded the second highest variance on the first day, 0.00159. As a result, the 14-th component is on the second row of the first column. We used the same criterion to fill the remaining rows and columns of Figure 6, where the color is proportional to the position of the j-th component, from blue (first component) to yellow (30-th component). We can see from Figure 6 that the outbreak of the COVID-19 pandemic affected the ranking of the estimated components. Some components, typically classified as low-ranking, became more important from March 2020. For instance, component 23 had rank 7 on March 5, 2020, rank 10 on March 9, 2020 and rank 9 on March 13, 2020 (see Figure 6). Similar considerations hold for component 22. In contrast, other components (e.g., component 12) became less important during the COVID-19 pandemic. This phenomenon is evident in Figure 7, which shows the different impact of the COVID-19 pandemic on components 12 and 22.

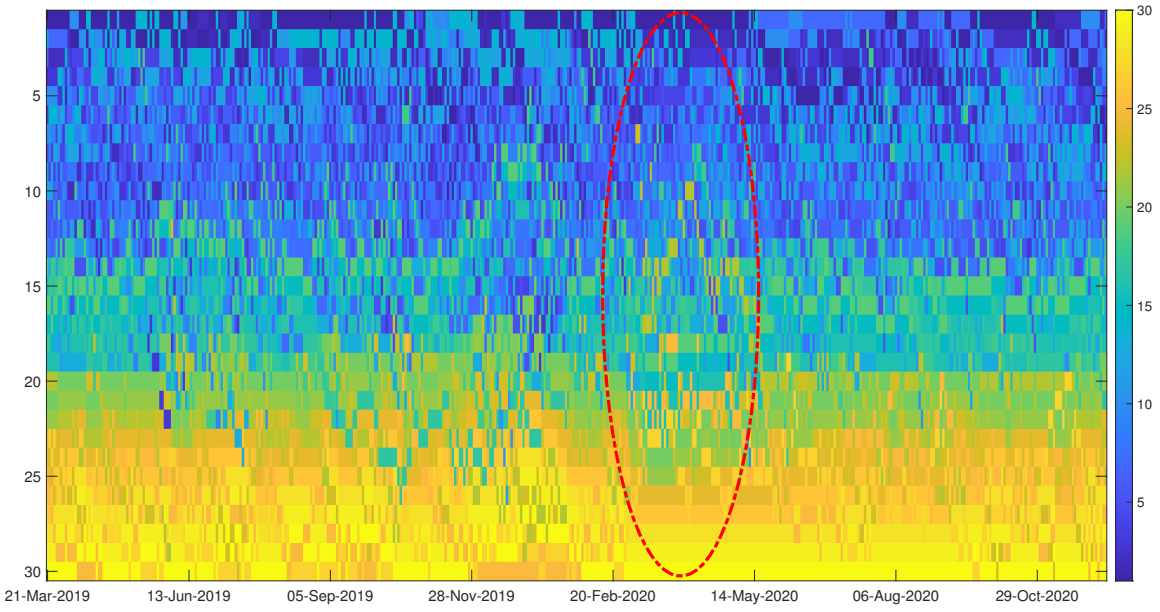


Figure 6: Ranking of the 30 components according to their variances.

In PCA, it is important to evaluate the contribution of the individual components in explaining the features of the data. A measure of how well the first $n \leq N$ principal components explain variation in the data is given by the ratio between the sum of the first n eigenvalues and the sum of all eigenvalues. A typically used graphical representation of these proportions is the scree plot. From PCA, the j-th eigenvalue coincides with the variance of the j-th PC (Härdle and Simar, 2015). Likewise, in our framework, we were able to evaluate the relevance of the first n components by building on their variances displayed in Figures 3—5. Following the same decreasing order in Figures 3—5, we sorted the 30 components according to their average variance. Then, we computed two weights: (i) the

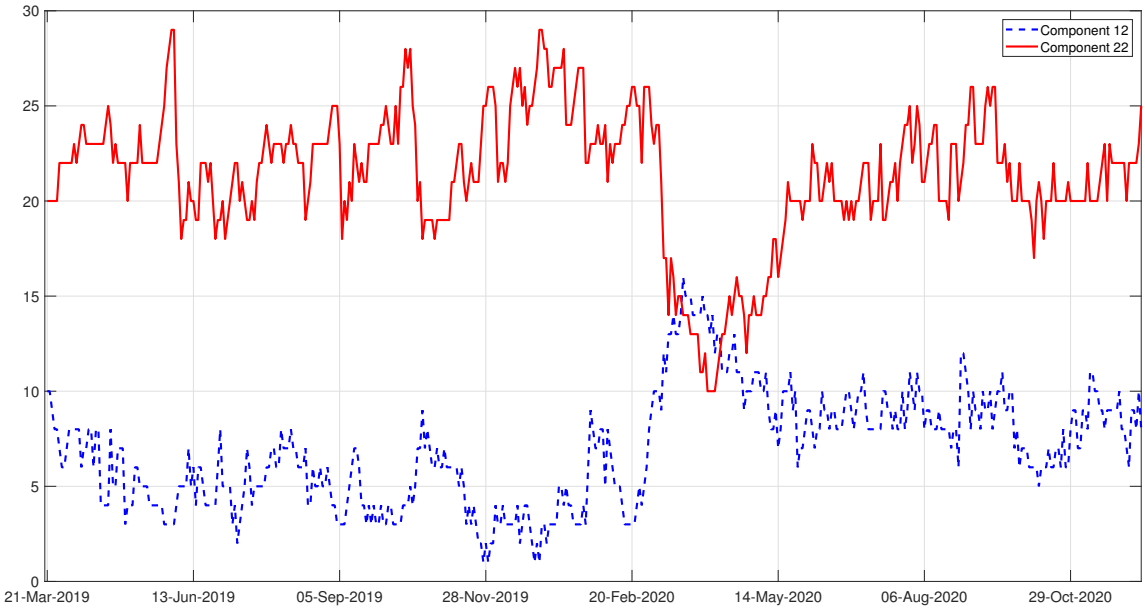


Figure 7: Ranking of components 12 and 22.

average variance of each component divided by the sum of the average variances of all the components and (ii) the ratio between the sum of the average variances of the first n components and the sum of the average variances of all the components. These weights are shown in Figure 8. We can see that the first nine sorted components (i.e., components 1, 14, 7, 2, 10, 11, 3, 9 and 4) explained 81.77% of the overall variation, whereas this proportion increased to 90.97% with the first 14 components (by adding components 12, 6, 5, 8 and 18).

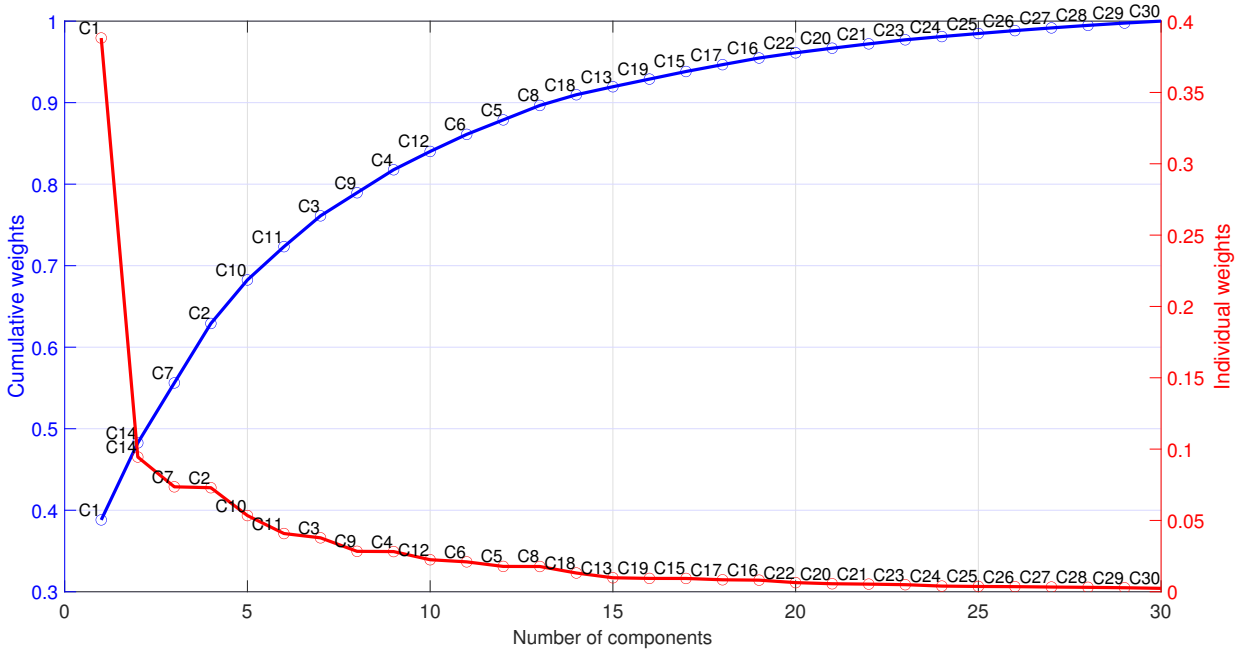


Figure 8: Relevance of the principal components according to their average variance.

In addition to the proportions shown in Figure 8, which were computed from average variances, we assessed the relevance of the estimated components for each day from March 20, 2019 to December 9, 2020. By doing so, we were able to evaluate the dynamics of the daily proportions over time. We still followed the order of the PCs given in Figures 3—5. Nevertheless, in contrast to the previous analysis, we now computed the weight of the j-th component as the variance recorded by this component on day t, divided by the sum of the variances of all the components on the same day, for $t = 1, \ldots, 450$. We also computed the cumulative weight as the ratio between the sum of the variances of the first n components on day t and the sum of the variances of all the components on the same day. The results are shown in Figure 9. The first rows of columns (a) and (b) in Figure 9 represent the first component, which had, on average, the greatest variance during the period March 20, 2019—December 9, 2020. The second row represents the 14-th component, and so on, until the last row, which represents the 30-th component, consistent with the order given in Figures 3—5. We highlight in Figure 9 the strong impact of the COVID-19 shock. First, panel (a) highlights the relevant weight of the first component during the early stage of the COVID-19 pandemic. From March 10, 2020 to March 19, 2020, this weight averaged 87.04%, with a peak of 93.17% on March 17, 2020. Furthermore, the degree of concentration significantly increased after the COVID-19 outbreak. For instance, the first (sorted) seven components (i.e., components 1, 14, 7, 2, 10, 11 and 3) had a daily average cumulative weight of 73.20% from February 26, 2020 to December 9, 2020 and 53.80% from March 20, 2019 to February 25, 2020 [see panel (b) of Figure 9].

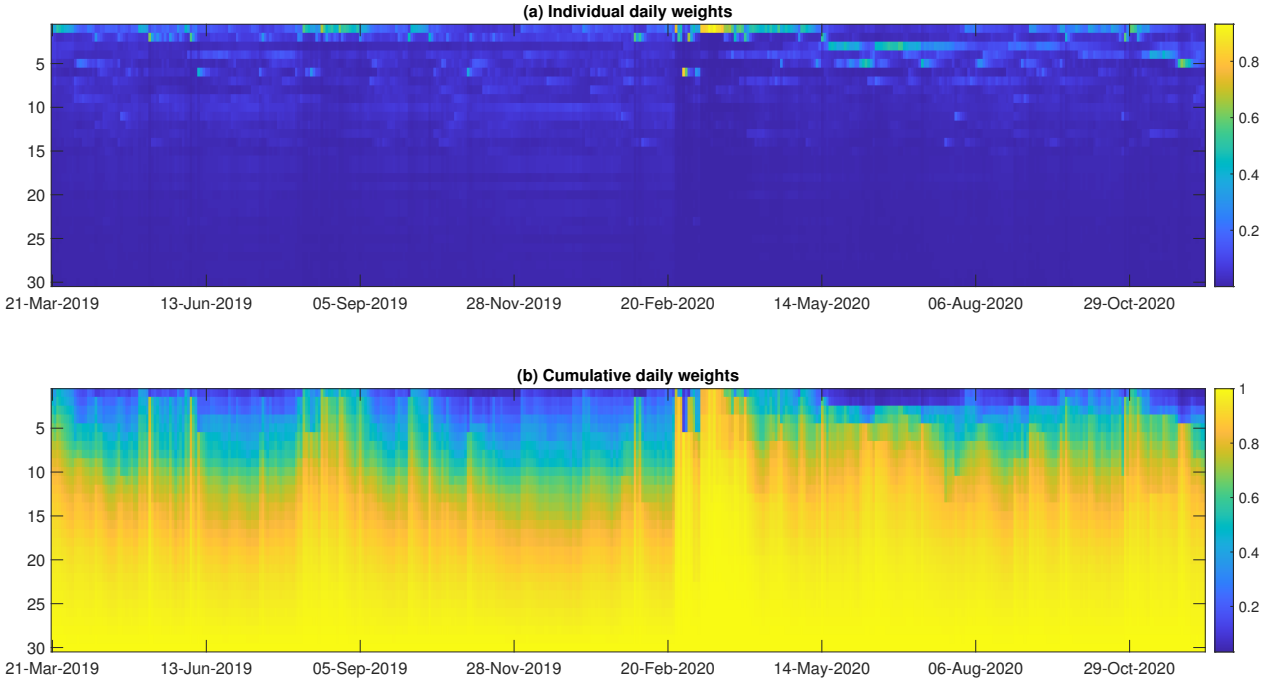


Figure 9: Daily relevance of the principal components.

Building on Figure 9, we also show in Figure 10 three scree plots related to three different days,

which reproduced the conditions observed before, during and after the COVID-19 shock, respectively. Specifically, we focused on the following days: June 24, 2019; March 13, 2020; and November 30, 2020, which correspond to the 68-th, 257-th and 443-rd columns in Figure 9, respectively. First, we highlight the importance of component 1, which significantly increased with the outbreak of the COVID-19 pandemic. Indeed, the weight of component 1 achieved the value of 92.57% on March 13, 2020, which is significantly greater than on other days: 11.22% on June 24, 2019 and 4.95% on November 30, 2020. Component 2 had the greatest weight on June 24, 2019, whereas component 10 had the greatest impact on November 30, 2020. Therefore, the importance of component 1 decreased during stable periods, as it strongly reflected the uncertainty or turbulence in financial markets. Second, we again highlight the greater concentration among a few components after the outbreak of the COVID-19 pandemic. For instance, 37.83% of the variability on June 24, 2019 is attributable to components 1, 14, 7, 2 and 10. In contrast, the same components achieved the levels of 94.94% and 72.67% on March 13, 2020 and November 30, 2020, respectively.

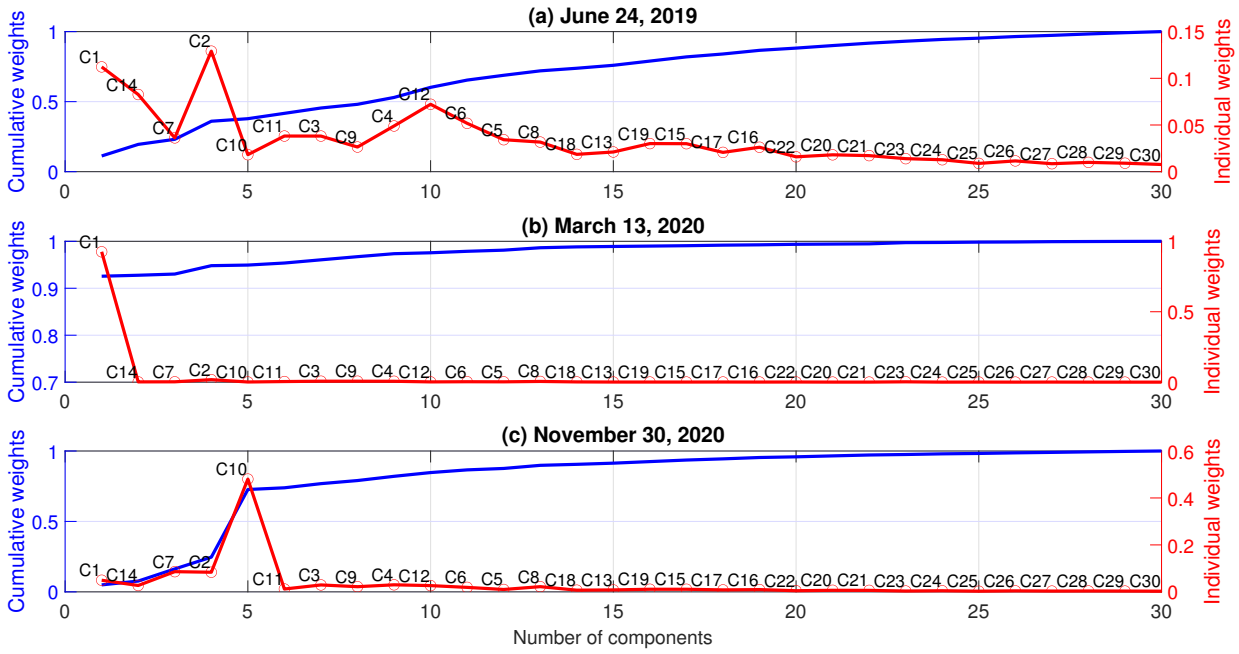


Figure 10: Relevance of the principal components before, during and after the outbreak of the COVID-19 pandemic.

We now compare the conditional PCs computed with the reordered eigenvectors (i.e., $\widetilde{\boldsymbol{u}}_t = [\widetilde{u}_{1,t} \cdots \widetilde{u}_{30,t}]' = \widetilde{\boldsymbol{L}}_t' \boldsymbol{y}_t$) with the conditional PCs obtained using the original DPC eigenvectors (i.e., $\widehat{\boldsymbol{u}}_t = [\widehat{u}_{1,t} \cdots \widehat{u}_{30,t}]' = \widehat{\boldsymbol{L}}_t' \boldsymbol{y}_t$). First, we analyze the correlations between each pair of components. We display the correlations between the 30 entries of $\widetilde{\boldsymbol{u}}_t$, numbered from 1 to 30, on the top-left panel of Figure 11. There, we see that the off-diagonal entries of the correlation matrix have low values, within the interval $[-0.2344, 0.1941]$. Similar considerations apply to the bottom-right panel of Figure 11, which shows the correlation matrix of the 30 entries of $\widehat{\boldsymbol{u}}_t$, numbered from 31 to 60. In contrast, the top-right (or, equivalently, the bottom-left) panel of Figure 11 displays the correlation matrix between the elements

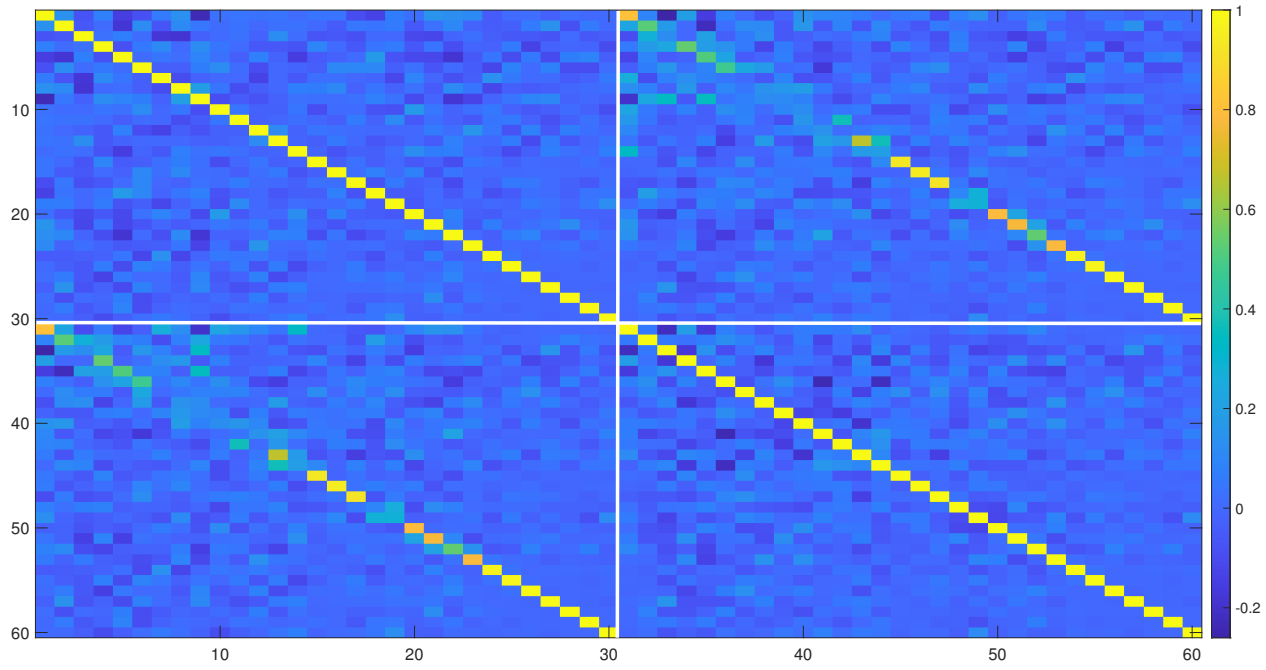


Figure 11: Correlation matrix of the rearranged components (numbered from 1 to 30) and the original components (numbered from 31 to 60).

of $\widetilde{\boldsymbol{u}}_t$ and $\widehat{\boldsymbol{u}}_t$. There, we see that $\widetilde{u}_{j,t}$ and $\widehat{u}_{j,t}$ have a correlation equal to 1 when $j = 28, 29, 30$. This evidence is clearer in Figure 12, which shows the values of the main diagonal of the bottom-left panel of Figure 11. These results are due to the stability of matrices $\widehat{\boldsymbol{A}}_{28}$, $\widehat{\boldsymbol{A}}_{29}$ and $\widehat{\boldsymbol{A}}_{30}$, which did not change after the implementation of our reordering algorithm, so that they coincided with $\widetilde{\boldsymbol{A}}_{28}$, $\widetilde{\boldsymbol{A}}_{29}$ and $\widetilde{\boldsymbol{A}}_{30}$, respectively.

We note that the components from 15 to 17 and from 24 to 27 also had high correlations, greater than 0.9 (see Figure 12). Interestingly, among the first 14 components, the pair $(\widetilde{u}_{1,t}, \widehat{u}_{1,t})$ recorded the highest correlation, 0.81. Similar to the other panels, the off-diagonal elements of the correlation matrix displayed on the bottom-left panel of Figure 11 had low values.

For completeness, we also studied the relationships between the variances of the entries of $\widetilde{\boldsymbol{u}}_t$ and $\widehat{\boldsymbol{u}}_t$. Similar to the described analysis, we estimated the variances of the series in $\widehat{\boldsymbol{u}}_t$. Specifically, we computed the sample variances for components 15, 16, 24, 26, 28, 29 and 30 (for which the null hypothesis of no ARCH effects was not rejected). In contrast, the variances of the remaining entries of $\widehat{\boldsymbol{u}}_t$ were estimated using the EGARCH model, with a degree of 1 for each GARCH, ARCH and leverage polynomial. The correlations of these variances are shown in Figure 13. Here, for simplicity, we set the correlation of the constant variances (i.e., the sample variances computed for those components for which we have evidence of no ARCH effects) at 0, so the corresponding cell is blue. In contrast to Figure 11, the off-diagonal elements of each panel in Figure 13 have greater values. It is interesting to note that the variances of the first 10 components had, on average, higher values when the original

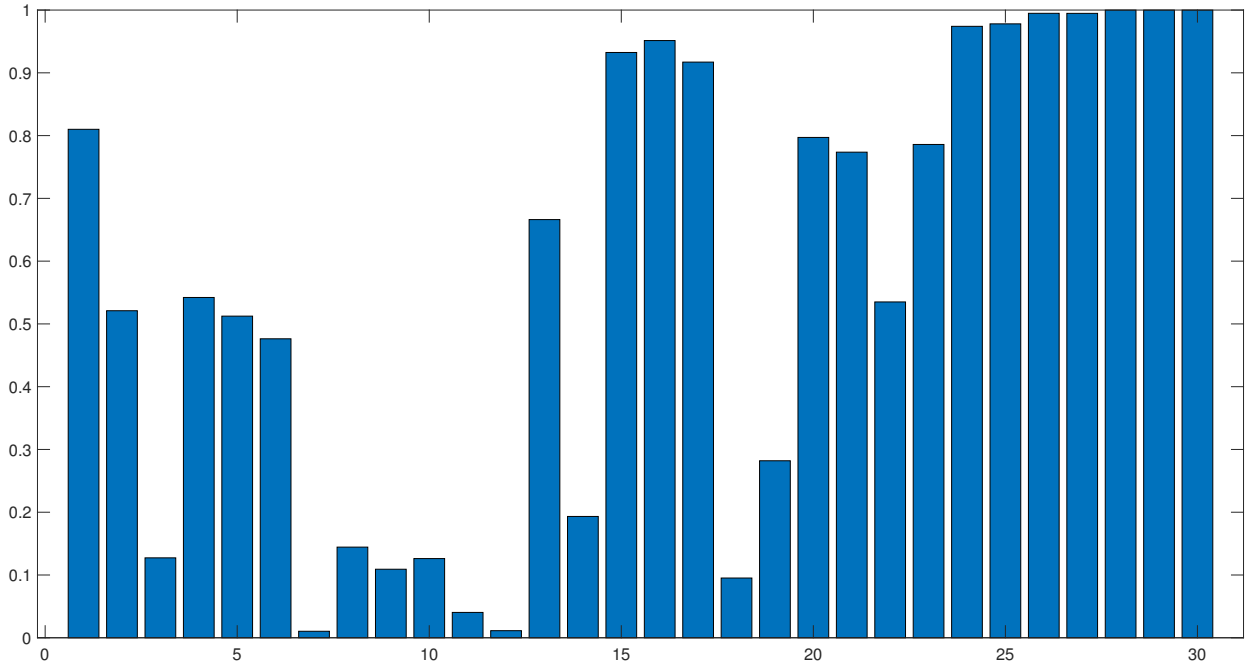


Figure 12: Main diagonal of the bottom-left (or top-right) panel of Figure 11.

eigenvectors were used. Indeed, the mean value of the first 10 rows and 10 columns of the bottom-right panel (corresponding to the original eigenvectors) in Figure 13 is 0.6823, greater than the mean of 0.4074 calculated for the first 10 rows and 10 columns of the top-left panel (corresponding to the rearranged eigenvectors) of Figure 13. We also highlight the strong correlation between the variances of the original and rearranged components, especially when focusing on the first positions. For instance, the mean of the first five values placed on the main diagonal of the top-right (or, equivalently, bottom-left) panel in Figure 13 is 0.7733.

5 Implications in terms of portfolio exposure

In this section, we assess the relevance of our method, starting from the estimation of a regression model that lays the foundation for a portfolio analysis. For this purpose, we adopt the 25 variables included in the $\boldsymbol{F}_t$ vector defined in Section 3. These covariates provide information on the risks and performance not only of the equities market, but also of the bonds, commodities and currencies markets. Moreover, we use different indexes of stress and uncertainty to emphasize the role of risk in our study, where variance plays a central role. Finally, we also use health variables that are directly linked to the COVID-19 pandemic, which, as we saw from our empirical findings discussed in Section 4, had a relevant impact on the dynamics of the conditional components of $\boldsymbol{y}_t$. Therefore, $\boldsymbol{F}_t$ provides a rich information set that allows us to better analyze the relationships between the components $\widetilde{u}_{1,t}, \ldots, \widetilde{u}_{30,t}$ and a large set of factors that drive the overall system, highlighting the differences from

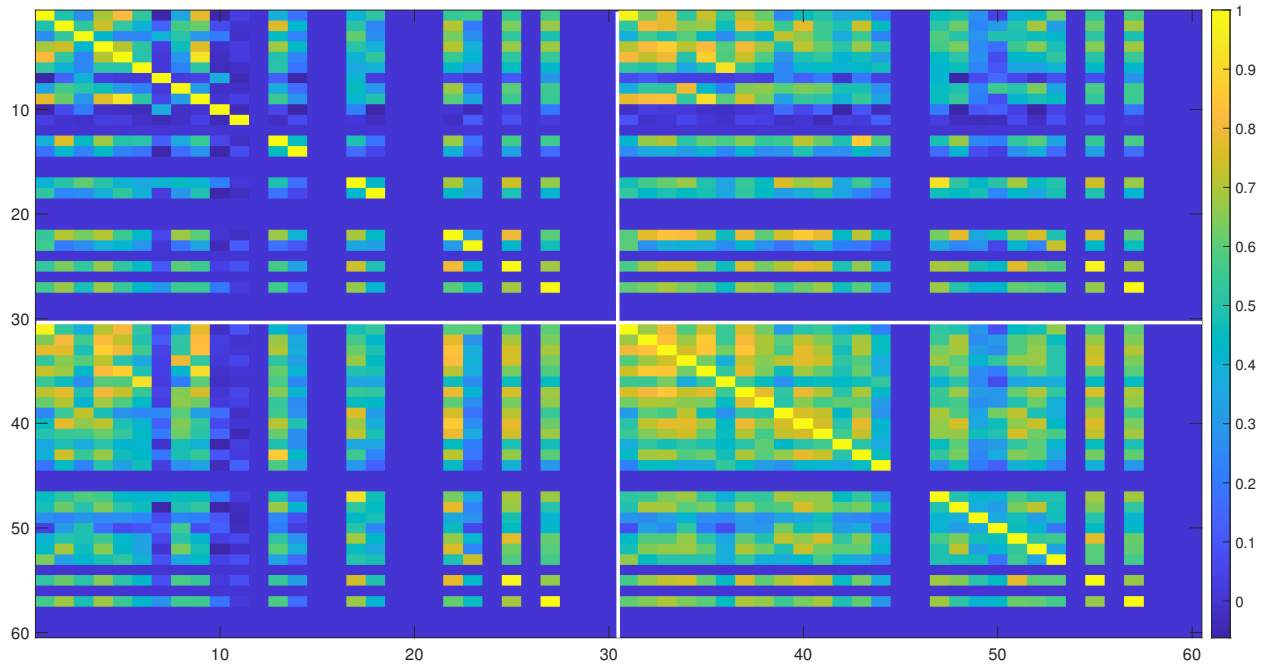


Figure 13: Correlation matrix of the variances of the rearranged components (numbered from 1 to 30) and the original components (numbered from 31 to 60).

the original DPC estimates.

We first compared the sensitivity of $\widetilde{u}_{j,t}$ and $\widehat{u}_{j,t}$ to each factor in $\boldsymbol{F}_t$, for $j = 1, \ldots, 30$. We dynamically estimated these relationships by implementing a rolling window procedure, with a window size of 150 observations. By doing so, we were able to capture the impact of the outbreak of the COVID-19 pandemic, which, as we saw in Section 4, affected the dynamics and stability of the estimated eigenvectors from March 20, 2019 to December 9, 2020. Specifically, we divided our dataset into 300 equally-sized subsamples, each of which included 150 observations for each time series. Therefore, starting from the adjusted components of $\boldsymbol{y}_t$ and using rolling subsamples that spanned the intervals $[\tau, \tau + 150 - 1]$, with $\tau = 1, \ldots, 300$, we iteratively estimated the following model:

$$\widetilde{u}_{j,t} = \left(\boldsymbol{\theta}_j^{(\tau)}\right)' \boldsymbol{F}_t + \eta_{j,t}^{(\tau)}, \tag{18}$$

where $\boldsymbol{\theta}_j^{(\tau)}$ is a 25×1 coefficient vector and $\eta_{j,t}^{(\tau)}$ is the error term, for $j = 1, \ldots, 30$.

We estimated $\boldsymbol{\theta}_j^{(\tau)}$ using the post-LASSO (least absolute shrinkage and selection operator) method described as follows. First, we considered a potential correlation between the entries of $\boldsymbol{F}_t$, the dimensionality of which could imply issues in the accumulation of estimation errors. We did not know *a priori* which of the covariates in $\boldsymbol{F}_t$ were relevant to the explanation of the conditional PCs of $\boldsymbol{y}_t$. On the one hand, a large number of regressors could imply overfitting. On the other hand, *ad hoc* omissions of some regressors introduced substantial bias in the estimates. We dealt with this issue by using

the LASSO method Tibshirani (1996). This is a machine learning technique that allows automatic selection of the relevant factors in $\boldsymbol{F}_t$ by minimizing the following loss function:

$$\mathcal{L}\left(\boldsymbol{\theta}_j^{(\tau)}\right) = \sum_{t=\tau}^{\tau+150-1}\left(\widetilde{u}_{j,t} - \left(\boldsymbol{\theta}_j^{(\tau)}\right)'\boldsymbol{F}_t\right)^2 + \lambda_j^{(\tau)}\left\|\boldsymbol{\theta}_j^{(\tau)}\right\|_1, \tag{19}$$

where $\left\|\boldsymbol{\theta}_j^{(\tau)}\right\|_1$ is the ℓ_1-norm penalty of the parameter vector $\boldsymbol{\theta}_j^{(\tau)}$ and $\lambda_j^{(\tau)} \geq 0$ is the tuning parameter, for $\tau = 1, \ldots, 300$ and $j = 1, \ldots, 30$.

The intensity of the penalization in Equation (19) depends on $\lambda_j^{(\tau)}$: the greater $\lambda_j^{(\tau)}$ is, the larger the number of elements in the estimate of $\boldsymbol{\theta}_j^{(\tau)}$ that approach zero is, providing sparser solutions. For each j and each τ, we selected the optimal value of $\lambda_j^{(\tau)}$ by using the five-fold cross-validation technique, which is commonly used in applied machine learning (James et al., 2013). We then LASSO-selected, in this first step, the relevant drivers of $\widetilde{u}_{j,t}$; that is, the factors in $\boldsymbol{F}_t$ whose coefficients—derived from the minimization of the loss function in Equation (19)—were not zero.

Second, we considered that although LASSO has appealing properties in terms of variable selection, it typically provides biased estimates for the retained variables, overshrinking the magnitude of their impact, in absolute value, on the response variable (Fan and Li, 2001). In contrast, we obtained more accurate estimates using the post-LASSO approach. Specifically, in the second step, we inserted the LASSO-selected factors into a new vector, $\boldsymbol{F}_t^{(s)}$. Then, we re-estimated the coefficients of the relevant factors (i.e., those that belonged to $\boldsymbol{F}_t^{(s)}$) by minimizing the loss function defined in Equation (19) with $\lambda_j^{(\tau)} = 0$ (so that the resulting solution coincided with that provided by the standard ordinary least squares method) and replacing $\boldsymbol{F}_t$ with $\boldsymbol{F}_t^{(s)}$. In contrast, the coefficients of the factors that were not LASSO-selected in the first step were set at zero.

We denote the vector of the coefficients obtained with the post-LASSO method, as described, as $\widetilde{\boldsymbol{\theta}}_j^{(\tau)}$, which accurately reflects the impact of each factor in $\boldsymbol{F}_t$ on the j-th adjusted component $\widetilde{u}_{j,t}$. The residual of the model defined in Equation (18), computed by replacing $\boldsymbol{\theta}_j^{(\tau)}$ with $\widetilde{\boldsymbol{\theta}}_j^{(\tau)}$, is denoted as $\widetilde{\eta}_{j,t}^{(\tau)}$, for $t = \tau, \ldots, \tau + 150 - 1$. We repeated the procedure described using the original DPC components as the response variables (i.e., by using $\widehat{u}_{j,t}$ in place of $\widetilde{u}_{j,t}$). We denote the corresponding coefficient vector and the residual term as $\widehat{\boldsymbol{\theta}}_j^{(\tau)}$ and $\widehat{\eta}_{j,t}^{(\tau)}$, respectively.

Interestingly, by comparing $\widetilde{\boldsymbol{\theta}}_j^{(\tau)}$ with $\widehat{\boldsymbol{\theta}}_j^{(\tau)}$, we discovered that the impact of a subset of factors in $\boldsymbol{F}_t$ becomes more relevant when looking at the adjusted components. This is the case, for instance, with RF, as shown in Figure 14. Panel (a) in the Figure displays the impact, in absolute value, of RF on the components $\widetilde{u}_{1,t}, \ldots, \widetilde{u}_{30,t}$ (sorted by row) for each of the 300 rolling subsamples (sorted by column). Here, we see a relevant impact on the first and 14-th components, which became stronger after the outbreak of the COVID-19 pandemic. We also detect some effect on the other components as well as in panel (b) of Figure 14, which is, however, less evident. Another interesting case is observed concerning GOLD. We see a clear effect of the COVID-19 pandemic in both panels (a) and (b) of

Figure 15. However, the effects on the rearranged components emerged earlier than the effects on the original DPC estimates. As for the health variables related to the COVID-19 pandemic, the effects were more transient and concentrated in the initial stage of the shock. An example of such effects is shown in Figure 16, in terms of DEATHS.

We now assess whether the differences in the $\widetilde{\boldsymbol{\theta}}_j^{(\tau)}$ and $\widehat{\boldsymbol{\theta}}_j^{(\tau)}$ vector coefficients imply relevant consequences in terms of portfolio exposure. For this purpose, among the different efficient portfolios we can build on the 30 constituents of the DJIA index, we focus on the minimum variance portfolio (MVP), given the central role of volatility in our study. Specifically, for each $\tau \in \{1, \ldots, 300\}$, we iteratively estimate the MVP weights as follows:

$$\boldsymbol{\phi}^{(\tau)} = \frac{\boldsymbol{\Omega}^{(\tau)} \mathbf{1}_{30}}{(\mathbf{1}_{30})' \boldsymbol{\Omega}^{(\tau)} \mathbf{1}_{30}}, \tag{20}$$

where $\mathbf{1}_{30}$ is a 30×1 unit vector, and $\boldsymbol{\Omega}^{(\tau)}$ is the inverse of the 30×30 sample covariance matrix of the stock returns, observed from τ to $\tau + 150 - 1$; that is, from the data $\boldsymbol{y}_\tau, \ldots, \boldsymbol{y}_{\tau+150-1}$.

Therefore, for each τ-th rolling subsample, we compute 150 returns of MVP, denoted as $\xi_t^{(\tau)} = \left(\boldsymbol{\phi}^{(\tau)}\right)' \boldsymbol{y}_t$, for $t = \tau, \ldots, \tau + 150 - 1$. We can rewrite $\xi_t^{(\tau)}$ as follows:

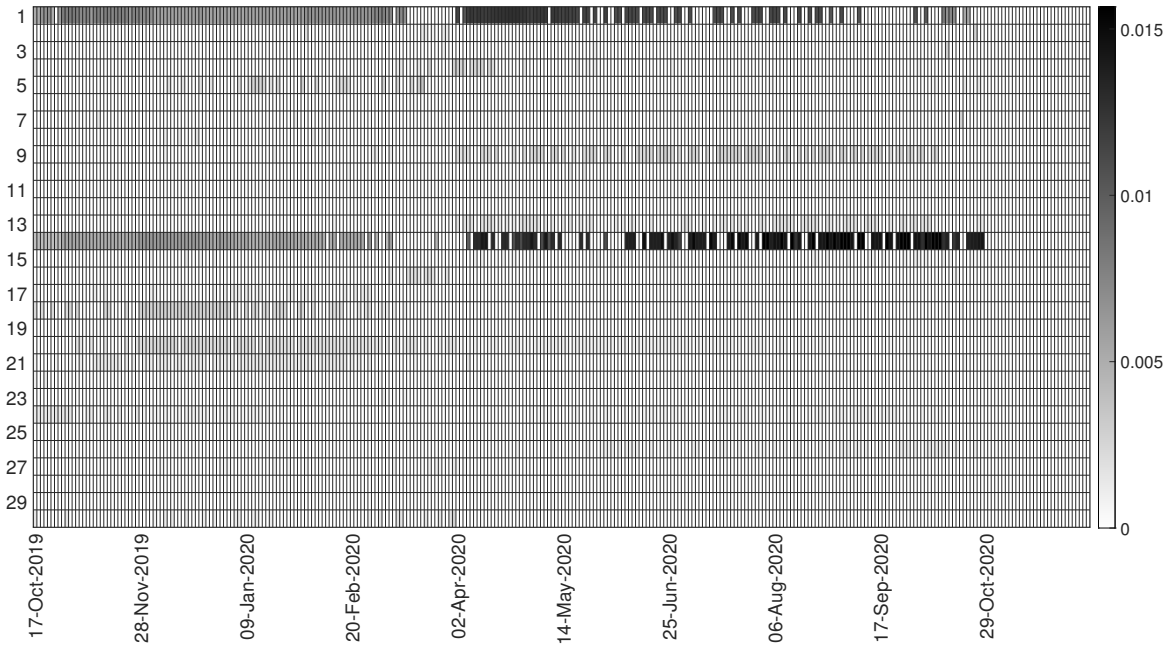
$$\begin{aligned} \xi_t^{(\tau)} &= \left(\boldsymbol{\phi}^{(\tau)}\right)' \boldsymbol{y}_t = \left(\boldsymbol{\phi}^{(\tau)}\right)' \widetilde{\boldsymbol{L}}_t \widetilde{\boldsymbol{u}}_t = \left(\boldsymbol{\phi}^{(\tau)}\right)' \widetilde{\boldsymbol{L}}_t \left[\left(\widetilde{\boldsymbol{\Theta}}^{(\tau)}\right)' \boldsymbol{F}_t + \widetilde{\boldsymbol{\eta}}_t^{(\tau)}\right] \\ &= \underbrace{\left(\boldsymbol{\phi}^{(\tau)}\right)' \widetilde{\boldsymbol{L}}_t \left(\widetilde{\boldsymbol{\Theta}}^{(\tau)}\right)' \boldsymbol{F}_t}_{\widetilde{\xi}_{F,t}^{(\tau)}} + \underbrace{\left(\boldsymbol{\phi}^{(\tau)}\right)' \widetilde{\boldsymbol{L}}_t \widetilde{\boldsymbol{\eta}}_t^{(\tau)}}_{\widetilde{\xi}_{\eta,t}^{(\tau)}}, \end{aligned} \tag{21}$$

where $\widetilde{\boldsymbol{u}}_t = [\widetilde{u}_{1,t} \cdots \widetilde{u}_{30,t}]'$, $\widetilde{\boldsymbol{\eta}}_t^{(\tau)} = \left[\widetilde{\eta}_{1,t}^{(\tau)} \cdots \widetilde{\eta}_{30,t}^{(\tau)}\right]'$, and $\widetilde{\boldsymbol{\Theta}}^{(\tau)}$ is a 25×30 matrix, the j-th column of which is the $\widetilde{\boldsymbol{\theta}}_j^{(\tau)}$ vector, for $j = 1, \ldots, N$, $\tau = 1, \ldots, 300$ and $t = \tau, \ldots, \tau + 150 - 1$.

We can see from Equation (21) that it is possible to decompose the portfolio return $\xi_t^{(\tau)}$ into a systematic component—i.e., $\widetilde{\xi}_{F,t}^{(\tau)}$—and an idiosyncratic component—i.e., $\widetilde{\xi}_{\eta,t}^{(\tau)}$. Let $\left(\sigma_\xi^{(\tau)}\right)^2$ and $\left(\widetilde{\sigma}_F^{(\tau)}\right)^2$ be the variances of $\xi_t^{(\tau)}$ and $\widetilde{\xi}_{F,t}^{(\tau)}$, respectively, computed from each rolling subsample; that is, for $\tau = 1, \ldots, 300$. Building on this decomposition, we can evaluate the contribution of the systematic component to the variability of the MVP return based on the ratio $\widetilde{\varphi}^{(\tau)} = \left(\widetilde{\sigma}_F^{(\tau)}\right)^2 / \left(\sigma_\xi^{(\tau)}\right)^2$. The greater $\widetilde{\varphi}^{(\tau)}$ is, the greater the capability of our method to capture the dynamics and fluctuations of the overall system, channelled through the adjusted eigenvectors. It is interesting to contrast the trend of $\widetilde{\varphi}^{(\tau)}$ with the same ratio computed from the original DPC eigenvectors, which we denote as $\widehat{\varphi}^{(\tau)}$. Therefore, we obtained $\widehat{\varphi}^{(\tau)}$ by replacing $\widetilde{\boldsymbol{L}}_t$, $\widetilde{\boldsymbol{\Theta}}^{(\tau)}$ and $\widetilde{\boldsymbol{\eta}}_t^{(\tau)}$ in Equation (21) with $\widehat{\boldsymbol{L}}_t$, $\widehat{\boldsymbol{\Theta}}^{(\tau)}$ and $\widehat{\boldsymbol{\eta}}_t^{(\tau)}$, respectively. For completeness, we repeated the described calculations by adopting the unconditional eigenvectors of the sample covariance matrix estimated from $\boldsymbol{y}_t, \ldots, \boldsymbol{y}_{450}$ (i.e. our entire dataset). We denote the corresponding ratio as $\overline{\varphi}^{(\tau)}$.

We display the trend of $\widetilde{\varphi}^{(\tau)}$, $\widehat{\varphi}^{(\tau)}$ and $\overline{\varphi}^{(\tau)}$, for $\tau = 1, \ldots, 300$, in Figure 17. Here, we see that

(a) Adjusted conditional principal components



(b) Original conditional principal components

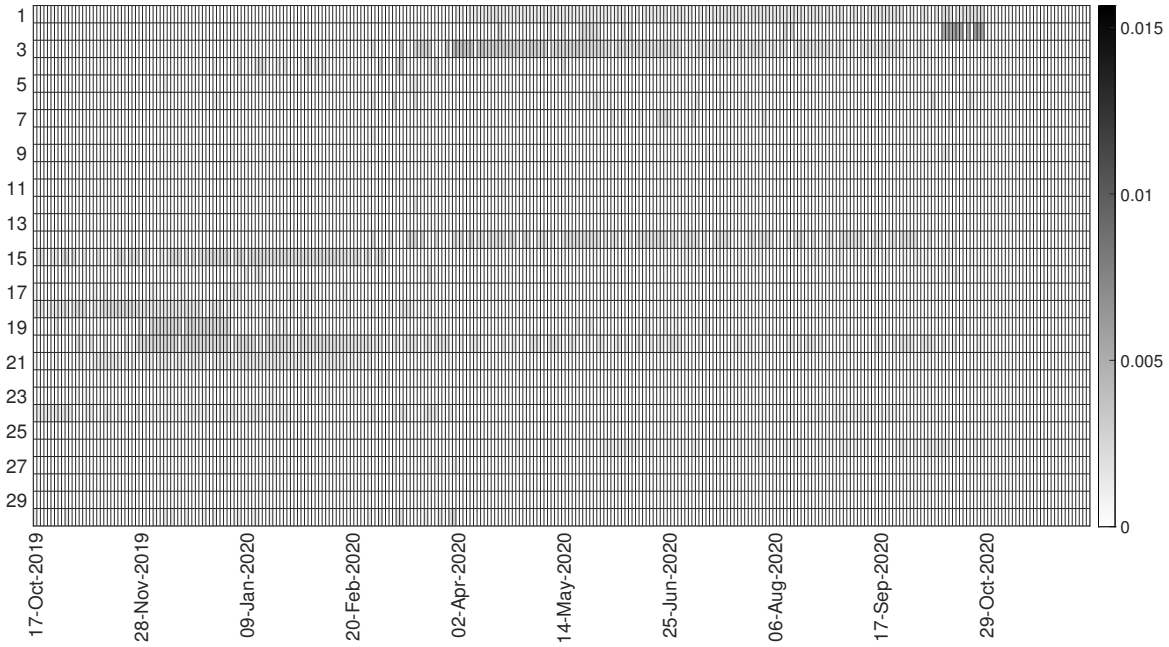


Figure 14: Impact (in absolute value) of the risk-free rate (RF) on the adjusted [panel (a)] and original [panel (b)] conditional components $\widehat{u}_{j,t}$ and $\widetilde{u}_{j,t}$, respectively, along the rolling subsamples, for $j = 1, \ldots, 30$.

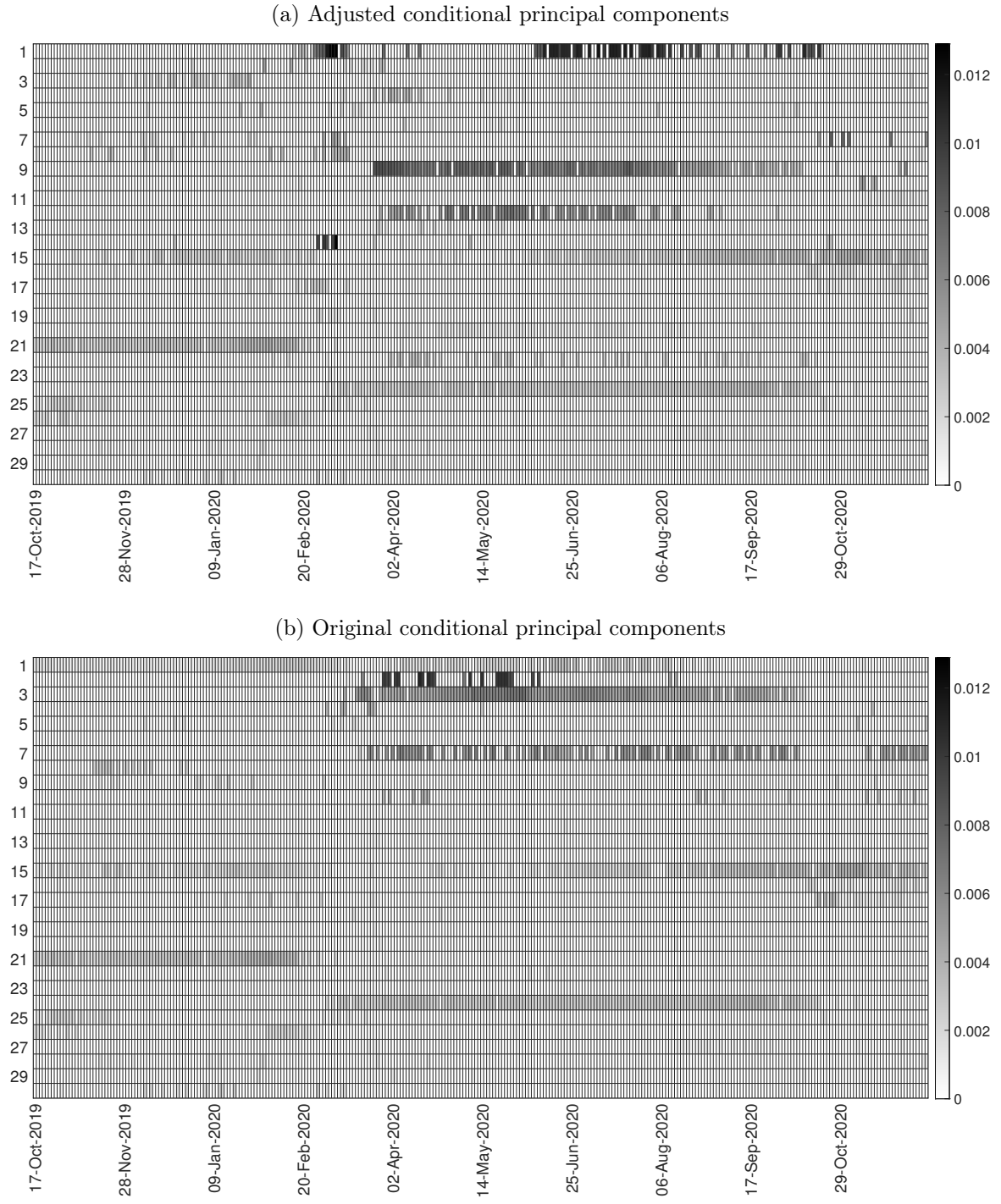
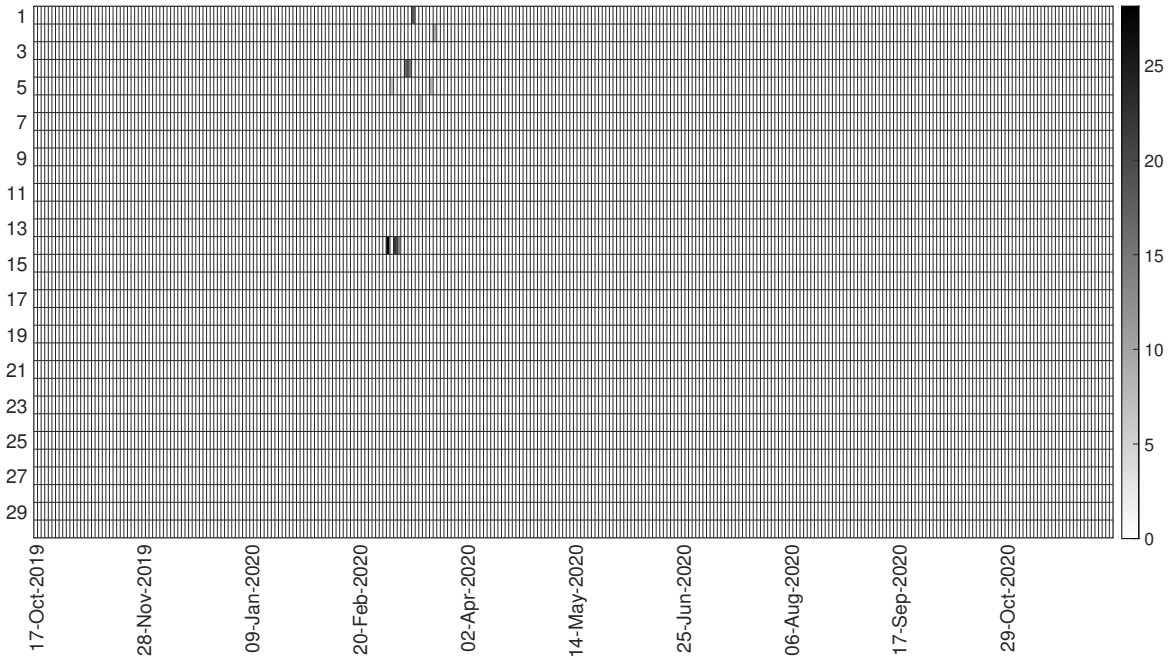


Figure 15: Impact (in absolute value) of the Gold Bullion LBM (GOLD) on the adjusted [panel (a)] and original [panel (b)] conditional components $\widehat{u}_{j,t}$ and $\widetilde{u}_{j,t}$, respectively, along the rolling subsamples, for $j = 1, \ldots, 30$.

(a) Adjusted conditional principal components



(b) Original conditional principal components

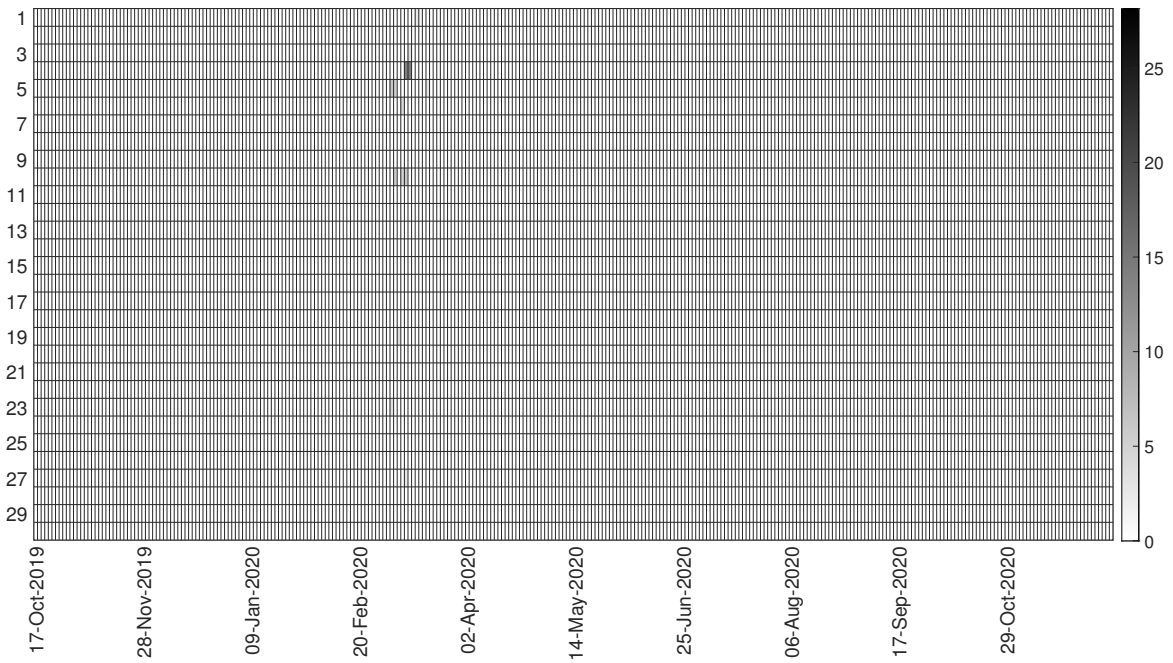


Figure 16: Impact (in absolute value) of the deaths due to COVID-19 (DEATHS) on the adjusted [panel (a)] and original [panel (b)] conditional components $\widehat{u}_{j,t}$ and $\widetilde{u}_{j,t}$, respectively, along the rolling subsamples, for $j = 1, \ldots, 30$.

$\widetilde{\varphi}^{(\tau)}$ significantly reacted to the outbreak of the COVID-19 pandemic. In contrast, the impact of the COVID-19 shock is not evident when looking at $\widehat{\varphi}^{(\tau)}$ and $\overline{\varphi}^{(\tau)}$. Furthermore, it is interesting to highlight the inversion in the ranking before and after the COVID-19 shock. Before this shock, the original DPC method led to a greater contribution of the systematic component to the portfolio risk, followed by the unconditional eigenvectors. After the COVID-19 shock, our method almost always pointed out a greater contribution to the portfolio variance.

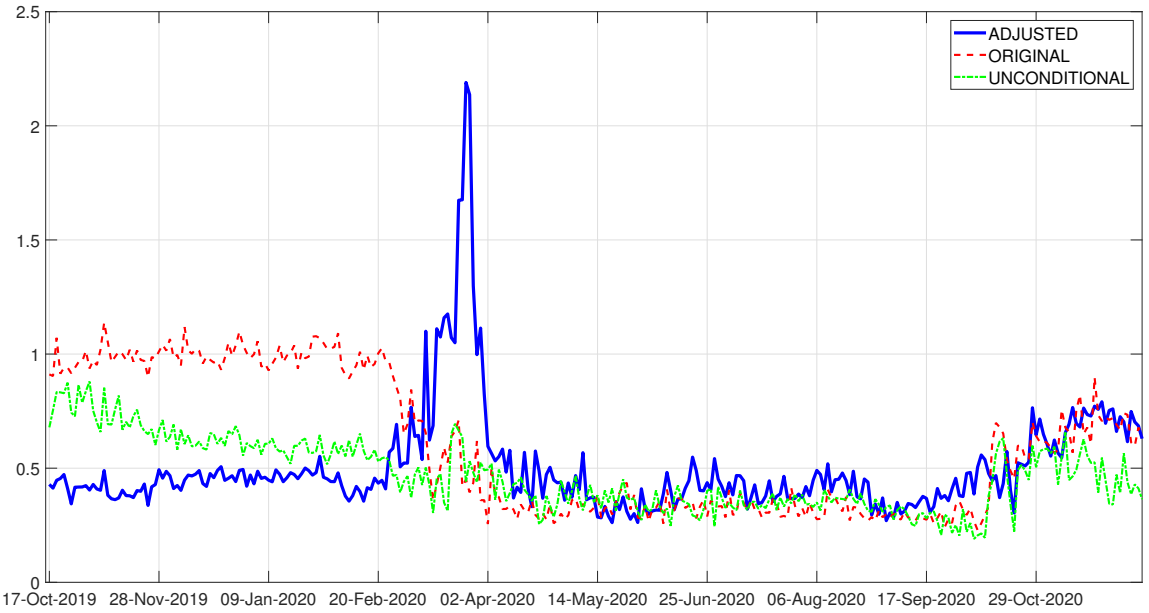


Figure 17: Contribution of the systematic component to the overall minimum variance portfolio, using the adjusted, original and unconditional eigenvectors.

6 Concluding remarks

In this study, we first presented the possible existence of ordering issues when PCA is used in a rolling or conditional setting. Then, using a specific MGARCH model, we provided empirical evidence of the occurrence of ordering issues and incoherence among PCs and then introduced an algorithm for optimal re-ordering of the PCs. We applied our proposed algorithm to data that included those during the outbreak of the COVID-19 pandemic. We found an advantage of using coherent PCs: a clearer interpretation of the link between PCs and risk factors.

Our results may pave the way for further applications in risk management and asset allocation, areas where PCA is widely used.

References

Gian Piero Aielli and Massimiliano Caporin. Dynamic Principal Components: a new class of multivariate GARCH models. Technical report, "Marco Fanno" Working Papers 0193, feb 2015. Available at SSRN: https://ssrn.com/abstract=2559758.

Carol Alexander and Andreas Kaeck. Regime dependent determinants of credit default swap spreads. *Journal of Banking & Finance*, 32(6):1008–1021, jun 2008. doi: 10.1016/j.jbankfin.2007.08.002.

Goodness Aye, Rangan Gupta, Shawkat Hammoudeh, and Won Joong Kim. Forecasting the price of gold using dynamic model averaging. *International Review of Financial Analysis*, 41:257–266, oct 2015. doi: 10.1016/j.irfa.2015.03.010.

Meshach Aziakpono, Stefanie Kleimeier, and Harald Sander. Banking market integration in the SADC countries: evidence from interest rate analyses. *Applied Economics*, 44(29):3857–3876, oct 2012. doi: 10.1080/00036846.2011.583219.

Monica Billio, Michael Donadelli, Antonio Paradiso, and Max Riedel. Which market integration measure? *Journal of Banking & Finance*, 76:150–174, mar 2017. doi: 10.1016/j.jbankfin.2016.12.002.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, apr 1986. doi: 10.1016/0304-4076(86)90063-1.

T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101.

Alexander Carol and Aubrey Chibumba. Multivariate orthogonal factor GARCH. Technical report, University of Sussex, Mimeo, 1996.

Pierre Collin-Dufresn, Robert S. Goldstein, and J. Spencer Martin. The determinants of credit spread changes. *The Journal of Finance*, 56(6):2177–2207, dec 2001. doi: 10.1111/0022-1082.00402.

David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, apr 1979. doi: 10.1109/tpami.1979.4766909.

Zhuanxin Ding and Robert F. Engle. Large scale conditional covariance matrix modeling, estimation and testing. Technical report, NYU Working Paper No. FIN-01-029, may 2001. Available at SSRN: https://ssrn.com/abstract=1294569.

Michael Donadelli and Antonio Paradiso. Is there heterogeneity in financial integration dynamics? Evidence from country and industry emerging market equity indexes. *Journal of International Financial Markets, Institutions and Money*, 32:184–218, sep 2014. doi: 10.1016/j.intfin.2014.06.003.

Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987, jul 1982. doi: 10.2307/1912773.

Robert F. Engle and Kenneth F. Kroner. Multivariate simultaneous generalized ARCH. *Econometric Theory*, 11(1):122–150, feb 1995. doi: 10.1017/s0266466600009063.

Robert F. Engle and J. Mezrich. GARCH for groups. *Risk*, 9(8):36–40, 1996.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, dec 2001. doi: 10.1198/016214501753382273.

Patricia Fraser, Foort Hamelink, Martin Hoesli, and Bryan Macgregor. Time-varying betas and the cross-sectional return–risk relation: evidence from the UK. *The European Journal of Finance*, 10 (4):255–276, aug 2004. doi: 10.1080/13518470110053407.

William Fung and David A. Hsieh. Empirical characteristics of dynamic trading strategies: The case of hedge funds. *Review of Financial Studies*, 10(2):275–302, apr 1997. doi: 10.1093/rfs/10.2.275.

Marvin H. J. Gruber. *Matrix Algebra for Linear Models*. John Wiley & Sons, December 2013. ISBN 978-1-118-59255-7.

Wolfgang Karl Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis*. Springer-Verlag GmbH, February 2015. ISBN 978-3-662-45171-7.

Joel Hasbrouck and Duane J. Seppi. Common factors in prices, order flows, and liquidity. *Journal of Financial Economics*, 59(3):383–411, mar 2001. doi: 10.1016/s0304-405x(00)00091-x.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer-Verlag GmbH, June 2013. ISBN 978-1-4614-7138-7.

Zhilin Kang, Xun Li, Zhongfei Li, and Shushang Zhu. Data-driven robust mean-CVaR portfolio selection under distribution ambiguity. *Quantitative Finance*, 19(1):105–121, jun 2018. doi: 10.1080/14697688.2018.1466057.

Robert A. Korajczyk and Ronnie Sadka. Pricing the commonality across alternative measures of liquidity. *Journal of Financial Economics*, 87(1):45–72, jan 2008. doi: 10.1016/j.jfineco.2006.12.003.

Loriano Mancini, Angelo Ranaldo, and Jan Wrampelmeyer. Liquidity in the foreign exchange market: measurement, commonality, and risk premiums. *The Journal of Finance*, 68(5):1805–1841, sep 2013. doi: 10.1111/jofi.12053.

Stephen L. Meyers. A re-examination of market and industry factors in stock price behavior. *The Journal of Finance*, 28(3):695–705, jun 1973. doi: 10.1111/j.1540-6261.1973.tb01390.x.

Hans-Georg Müller, Rituparna Sen, and Ulrich Stadtmüller. Functional data analysis for volatility. *Journal of Econometrics*, 165(2):233–245, dec 2011. doi: 10.1016/j.jeconom.2011.08.002.

Daniel B. Nelson. Conditional heteroskedasticity in asset returns: a new approach. *Econometrica*, 59 (2):347, mar 1991. doi: 10.2307/2938260.

Efstathios Panayi, Gareth W. Peters, and Ioannis Kosmidis. Liquidity commonality does not imply liquidity resilience commonality: a functional characterisation for ultra-high frequency cross-sectional LOB data. *Quantitative Finance*, 15(10):1737–1758, sep 2015. doi: 10.1080/14697688.2015.1071075.

Markus Pelger. Understanding systematic risk: a high-frequency approach. *The Journal of Finance*, 75(4):2179–2220, may 2020. doi: 10.1111/jofi.12898.

Chi Seng Pun and Lei Wang. A cost-effective approach to portfolio construction with range-based risk measures. *Quantitative Finance*, 21(3):431–447, jul 2020. doi: 10.1080/14697688.2020.1781237.

T. Roncalli and G. Weisang. Risk parity portfolios with risk factors. *Quantitative Finance*, 16(3): 377–388, jul 2015. doi: 10.1080/14697688.2015.1046907.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, nov 1987. doi: 10.1016/0377-0427(87)90125-7.

Chiara Sabelli, Michele Pioppi, Luca Sitzia, and Giacomo Bormetti. Multi-curve HJM modelling for risk management. *Quantitative Finance*, 18(4):563–590, aug 2017. doi: 10.1080/14697688.2017.1355104.

Ravi Shukla and Charles Trzcinka. Sequential tests of the arbitrage pricing theory: A comparison of principal components and maximum likelihood factors. *The Journal of Finance*, 45(5):1541–1564, dec 1990. doi: 10.1111/j.1540-6261.1990.tb03727.x.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, jan 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Nikolas Topaloglou, Hercules Vladimirou, and Stavros A. Zenios. CVaR models with selective hedging for international asset allocation. *Journal of Banking & Finance*, 26(7):1535–1561, jul 2002. doi: 10.1016/s0378-4266(02)00289-3.

Vadym Volosovych. Measuring financial market integration over the long run: Is there a U-shape? *Journal of International Money and Finance*, 30(7):1535–1561, nov 2011. doi: 10.1016/j.jimonfin.2011.07.011.

Yaojie Zhang and Yudong Wang. Forecasting crude oil futures market returns: a principal component analysis combination approach. *International Journal of Forecasting*, feb 2022. doi: 10.1016/j.ijforecast.2022.01.010.

Xiao Zhong and David Enke. Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, 67:126–139, jan 2017. doi: 10.1016/j.eswa.2016.09.027.

Appendix

A List of the companies included in our dataset

Table 1: Companies and economic sector

NUMBER	COMPANY	SECTOR
1	VERIZON COMMUNICATIONS	communications
2	WALT DISNEY	communications
3	COCA COLA	consumer staples and discretionary
4	HOME DEPOT	consumer staples and discretionary
5	MCDONALD'S	consumer staples and discretionary
6	NIKE	consumer staples and discretionary
7	PROCTER & GAMBLE	consumer staples and discretionary
8	WALMART	consumer staples and discretionary
9	WALGREENS BOOTS ALLIANCE	consumer staples and discretionary
10	CHEVRON	energy
11	AMERICAN EXPRESS	financials
12	JP MORGAN CHASE & CO.	financials
13	GOLDMAN SACHS	financials
14	TRAVELERS COS.	financials
15	AMGEN	health care
16	JOHNSON & JOHNSON	health care
17	MERCK & COMPANY	health care
18	UNITEDHEALTH GROUP	health care
19	HONEYWELL INTL.	industrials
20	BOEING	industrials
21	CATERPILLAR	industrials
22	3M	industrials
23	DOW ORD SHS	materials
24	INTERNATIONAL BUS.MCHS.	technology
25	APPLE	technology
26	CISCO SYSTEMS	technology
27	SALESFORCE.COM	technology
28	INTEL	technology
29	MICROSOFT	technology
30	VISA	technology