Hierarchical PCA and Modeling Asset Correlations

Marco Avellaneda[1] and Juan Andrés Serur[2]

Abstract

Modeling return correlations between thousands of stocks poses great challenges, as empirical estimators tend to perform poorly when assets don't share common risk factors, such as country or industry sector. In this paper, we show the advantages of using Hierarchical Principal Component Analysis (HPCA) for modeling correlations, as opposed to the classic PCA. Furthermore, we propose a statistical clustering approach using orthogonal decomposition, which seems to work outstandingly well in identifying "synthetic sectors". To show the robustness of the results and make them fully reproducible, we conduct a comprehensive analysis of the stock markets of the US, Europe, China, and Emerging Markets.

Keywords — correlations; factor models; hierarchical PCA; statistical clusters

[1] New York University (NYU) - Courant Institute of Mathematical Sciences. Email: avellane@cims.nyu.edu.

[2] New York University (NYU) - Courant Institute of Mathematical Sciences. Email: juan.serur@nyu.edu.

1 Introduction and Literature Review

Correlation matrix modeling, one of the main elements in classical portfolio optimization, represents a great challenge as equity markets and their cross-dynamics have become increasingly complex. Traditional empirical estimators have several limitations. For example, as the size of the universe of stocks increases, more observations are needed to estimate the correlations; otherwise, when the number of stocks N is greater than the number of observations T, the matrix is singular. Overcoming this with longer estimation windows is generally not an option, as large observation samples are not available, and even if they were, investors prefer shorter estimation windows as stock returns behavior changes dramatically. Furthermore, even if the matrix is not singular, it is usually ill-conditioned with unstable off-diagonal elements. To overcome these problems, practitioners have focused their efforts on modeling correlations using factor models (hence model correlation matrices rather than empirical correlation matrices).

Quantitative factor analysis has become increasingly important in portfolio management. The well-known Capital Asset Pricing Model (CAPM) was one of the most promising breakthroughs in modern finance, proposed in [Sharpe, 1964]. In this paper, Sharpe proffered a market equilibrium model, where asset returns were explained by the assets' exposure to the market portfolio (systematic risk) plus a idiosyncratic risk component. Intending to generalize the CAPM and accustom it to real-world conditions, [Ross, 1976] coined the so-called Arbitrage Pricing Theory (APT). Unlike CAPM, this theory is based on arbitrage factor models and extends the equilibrium single-factor model to multi-factor models. In [Fabozzi, Focardi and Kolm, 2010] the authors provide a comprehensive compendium of quantitative methods for equity strategies, with the main focus on portfolio optimization and factor-based models intended to overcome the main shortcomings of the classical Modern Portfolio Theory. In [Avellaneda and Lee, 2010], the authors show that statistical arbitrage strategies can be formulated using statistical factors (via PCA) and sector ETFs, providing evidence that both approaches work remarkably well and that after augmenting the signals by the traded volume, factor models based on ETFs further enhance their performance.[3]

Developing systematic portfolio strategies is now regarded as the core intellectual activity of "quant funds" managing hundreds of billions of dollars in assets. Usually, practitioners choose between two types of factor models. On the one hand, there are models based on explicit factors such as momentum, value, size, quality, among others. In [Fama and French, 1992] and [Fama and French, 1993] the authors proposed a multi-factor model, extending the CAPM by adding the factors value and size. Empirical evidence has shown that this model provides a better characterization of the cross-section of the stocks' returns. More recently, Fama and French came up with a new model (see [Fama and French, 2015]), augmenting its previously three-

[3]For more literature on factor models and their application to risk and portfolio management, see [Fabozzi, Kolm and Focardi, 2006], [Fabozzi *et al*, 2007].

factor model with the factors profitability and investment, originating the so-called Fama and French five-factor model.

On the other hand, there are those models based on implicit factors like statistical features extracted from assets' returns using Principal Component Analysis (PCA), maximum likelihood, among others approaches. For example, in [Connor and Korajczyk, 1988] the authors, using asymptotic PCA, identified five factors that work remarkably better than CAPM and are able to capture the time-varying risk premium.[4]

One of the main advantages of implicit factor model is that they do not make assumptions about the drivers behind price movements. They rely on market data without additional information. However, like any technique in finance, they are not a panacea. These models must be treated carefully to avoid undesirable instabilities that lead to high estimation errors. In addition, there are several challenges. For example, setting up the number of K implicit factors –mainly in the context of PCA– is a crucial step. In this matter, various techniques have been proposed. One of the most famous approaches is based on Random Matrix Theory (RMT) (see [Cizeau *et al.*, 2000] and [Laloux *et al*, 2000]), intended to retain only a few significant eigenvectors and filter out the noisy ones, modeling them as random noise. In [Kakushadze, 2015], the author proposed the "minimization algorithm", choosing the number K based on the total risk attributed to idiosyncratic risk as $K \to N$, where N is the number of stocks. So the approach aims to minimize the absolute difference between function $g(K)$ and 1.[5] Likewise, the effective rank proposed in [Roy and Vetterli, 2007] is a versatile method based on the Shannon entropy, intended to measure the effective dimension of the matrix as it is generally lower than the number of positive eigenvalues due to the highly correlated components.

Throughout this paper, we implement a technique called Hierarchical PCA (HPCA), introduced by [Avellaneda, 2019] in the context of equity correlation matrices ordered hierarchically by the MSCI Global Industrial Classification Standard (GICS). To model this hierarchical structure, we use the GICS and countries. Furthermore, we present a novel statistical clustering technique that harnesses the power of PCA, which is based solely on returns data and is capable to overcome some drawbacks inherent in static-like clusters such as GICS and/or countries. The empirical analysis shows that the eigenvectors obtained by HPCA works outstandingly well, overcoming some issues of classical PCA. Finally, we provide some trading strategies ideas leveraging the modeled factors. To make results fully reproducible, we repeat the analysis to the US, European, Chinese, and Emerging stock markets.

[4]For more recent developments in this arena, see, for example, [Kakushadze and Yu, 2017], [Meucci, 2010], [Torun *et al.*, 2011].

[5]Here, the function $g(K) = \sqrt{\min(\hat{\xi}^2)} + \sqrt{\max(\hat{\xi}^2)}$, where $\hat{\xi}$ is the idiosyncratic risk.

2 PCA revisited

In a universe consisting of N stocks and T observations, we consider the $N \times N$ empirical correlation matrix,

$$C = \frac{1}{T} R R^t \tag{1}$$

where R is the $T \times N$ matrix of standardized returns.[6]

PCA calculates the eigenvalues and eigenvectors of the correlation matrix ranked in decreasing order by eigenvalues. Accordingly, the first eigenvector solves the variational problem

$$V^{(1)} = argmax \, \{V^t C V : ||V||_2 = 1\} \tag{2}$$

where $||.||_2$ represents the Euclidean space.[7] Higher-order eigenvectors satisfy a similar variational problem, restricting the problem to the orthogonal complement to the previous eigenvectors:

$$V^{(k)} = argmax \, \{V^t C V : ||V||_2 = 1, V^{(k)t} V^{(r)} = 0, \; 1 \leq r < k\} \tag{3}$$

The vectors $V^{(k)}$ satisfy $CV^{(k)} = \lambda^{(k)} V^{(k)}$, i.e. they are eigenvectors of C, which can be factored as

$$C = \mathcal{V} \Lambda \mathcal{V}^T. \tag{4}$$

Here $\mathcal{V} = [V^{(1)}, ..., V^{(N)}]$ is an orthogonal matrix. Its columns of which are the eigenvectors of C, and Λ is the diagonal matrix with the eigenvalues ordered from the highest to the lowest.[8,9]

2.1 Eigenvalue Analysis and Spectral Cutoffs

We see from the above that PCA is a "greedy" algorithm in the sense that the eigenvectors explaining the most variance are extracted successively. In particular, the amount of variance explained by the first k eigevectors (eigenportfolios). Figure 1 displays the top 50 eigenvectors of the correlation matrix of US stocks.

[6] As usual, standardized return $= \frac{r_t - \overline{r}}{\sigma(r)}$.

[7] The n-dimensional Euclidean norm in $\mathbb{R}^n$ is defined as $||X||_2 := \sqrt{X_1^2 + X_2^2 + X_3^2 + \cdots + X_m^2}$.

[8] Note also that the eigenvectors are defined up to sign. For more details, see [Jollife, 2002].

[9] In financial economics, the first-order optimality condition for a portfolio that maximizes the Sharpe Ratio over all competing portfolios investing in the same N stocks can be represented as $r - E(r) = \beta_r (F - E(F)) + \epsilon_r$. Thus, remarkably, the Principal Eigenportfolio is connected to the concept of the Market Portfolio in the sense of Modern Portfolio Theory [Avellaneda and Lee, 2010], and [Boyle, 2014].

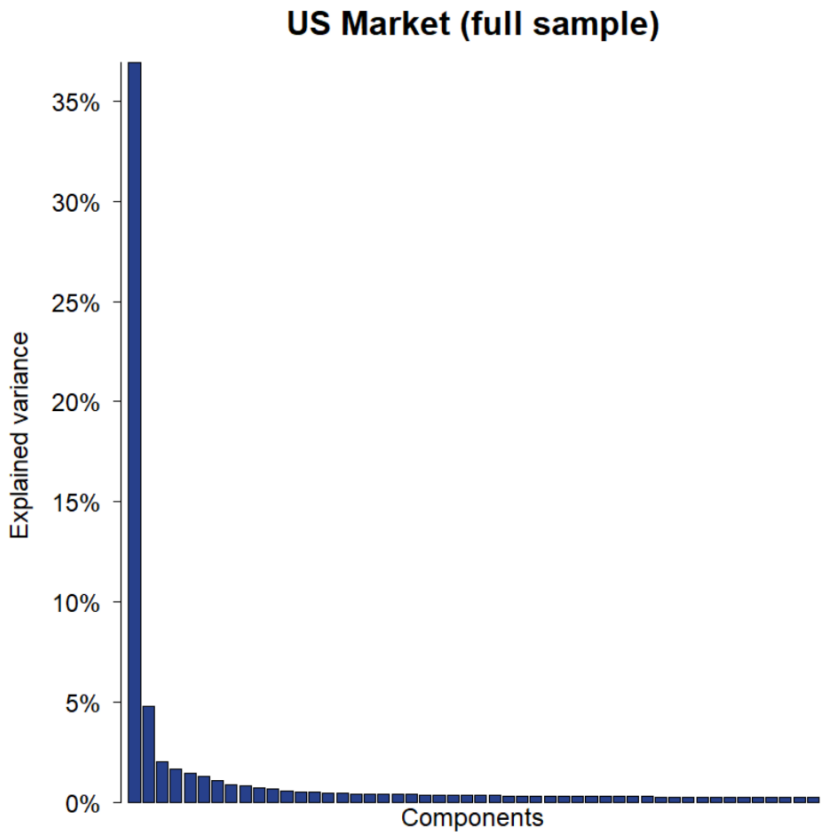

Figure 1: Top 50 eigenvalues for the US stocks with the data sample spanning from 2010 to 2019. The variance explained by the first eigenvalue is approximately $\lambda^{(1)}/N = 40\%$.

Figure 1 shows that the first eigenvalue accounts for almost 40% of the variance. However, it is well-known empirically that the amount of variance[10] explained by the principal eigenportfolios, or by the top eigenvectors, varies over time [Avellaneda and Lee, 2010]. In financial stress periods, the variance explained by the first eigenportfolio increases sharply, depicting the increase in correlations across different assets and attempting against diversification benefits. This is shown in Figure 2, which compares the "diversity level" with the 2-year Treasury Rate.

[10]This is actually measured as a percentage of the total trace.

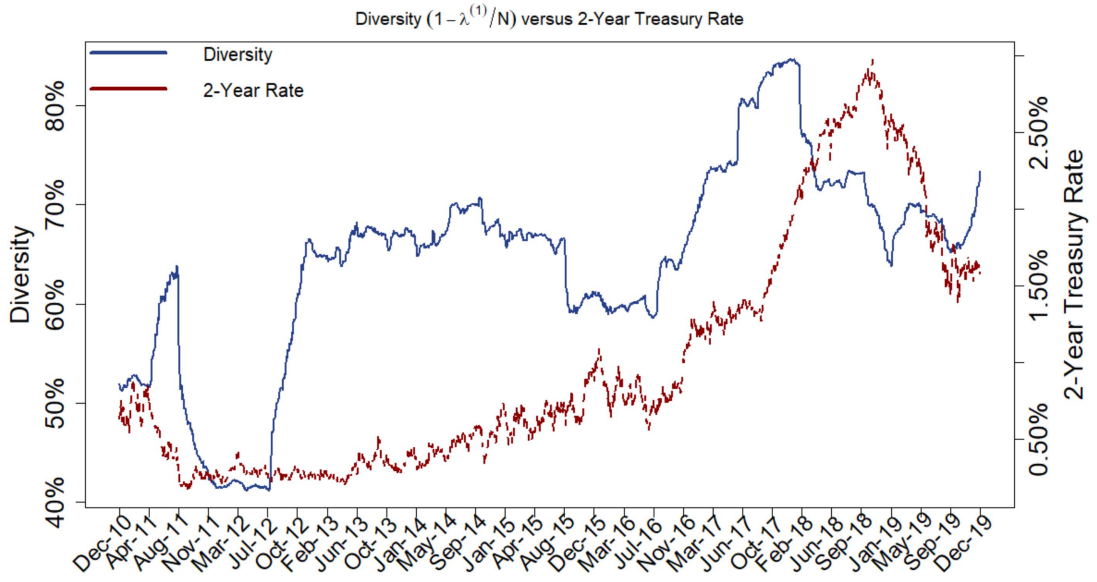

Figure 2: Diversity level $(1-\lambda^{(1)}/N)$ is calculated with the first eigenvalue using a 1-year rolling correlation matrix. The diversity level moves in the same direction as the 2-Year Treasury Constant Maturity Rate. The 2-Year CMT Rate was obtained from the FRED repository.

Ill-conditioning of correlation matrices for large groups of equities is a feature of the market: the ratio of the largest to the smallest eigenvalues (condition number) is very large. If we fixed the number of observation dates T, and gradually increase the number of stocks N, the more correlated stocks would appear to be.

In practice, short observation windows are necessary. Investment managers typically "refresh" correlation matrices and use a relatively short window for sampling data, such as 120 or 180 days. For large universes, this leads to degenerate "empirical" correlation matrices.

The correlation coefficients of stocks which are not linked, economically or otherwise, are unreliable, or at least suspicious. Practitioners often use the eigenvalues of C to determine what they believe are "significant factors", attempting to separate separate "signal" from "noise" and regularize degenerate or ill-conditioned correlation matrices. This leads naturally to dimension-reduction techniques, such as using spectral cutoffs that retain only a few "significant" eigevectors and model the orthogonal complement as random noise (see [Ledoit and Wolf, 2014]), some of which are based on RMT, introduced in Physics by Wigner in 1930 (see, for example, [Marčenko and Pastur, 1967], [Laloux *et al*, 2000]).

2.2 Higher-order eigenvectors and eigenportfolios: the identification problem

A particular issue with PCA is that beyond the first eigenportfolio –associated with the market mode– it is difficult to find an economic or financial explanation in higher-order eigenportfolios which are believed to be "significant" after applying

a spectral cutoff [Laloux *et al*, 2000]; see Figure 3. This is quite different from the standard situation in fixed-income, where changes in bond yields or forward rates can more easily be interpreted in terms of higher-order eigenportfolios (curve steepening/flattening, or changes in 2-versus-10 or 10-versus-30 yields [Litterman and Scheinkman, 1991]).

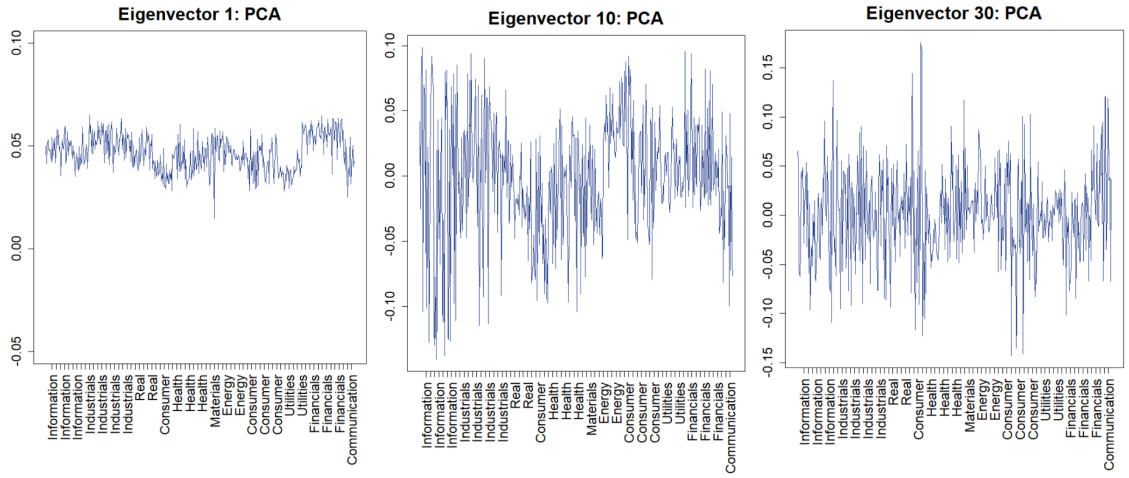

Figure 3: PCA eigenvectors for US markets. The first represents the market, but those of higher-order (tenth and thirtieth in this case) suffer the so-called identification problem since it is very difficult to find a meaningful economic intuition.

In Figure 3, the first eigenvector (and eigenportfolio) is straightforward to interpret. It has all the positive weights, and is a reasonable proxy for the "market portfolio", such as an S&P 500 index tracker. However, other eigenvectors/eigenportfolios are less straightforward to interpret. They have positive and negative weights across the spectrum, making it difficult to analyze for portfolio management purposes; see nevertheless [Plerou *et al.*, 2002] and [Avellaneda and Lee, 2010].

3 Hierarchical PCA

As a first step, we partition the stock universe into clusters which are believed to share common features, such as an industry sector or a country, or that stocks are otherwise clustered according to the modeler's beliefs. The exact specification of the clusters will be addressed later.

We assume that the modeler holds strong beliefs on the data but only for stock returns belonging to the same cluster. In contrast, the modeler trusts less the empirical correlations between stocks which belong to different clusters. To fix ideas, assume that there are b clusters.

In addition, assume that the modeler chooses b "benchmark portfolios", each of which is associated with a cluster. The benchmark could be, for instance, an ETF tracking a basket of stocks in the cluster. The modeler believes in the correlations

between the pairs of the benchmark portfolio returns. We denote the correlations by $\rho^{k,k'};\ k = 1, .., b, \quad k' = 1, .., b.$

Let F^k denote the standardized returns of the benchmark portfolio associated with cluster k. The regression coefficient of stock i on the return of the corresponding benchmark portfolio, F^k, is denoted by β_i. We assume that the modeler also believes in the β_i.

Consider the function $\mathbb{I}(i)$, which returns the sector of stock i: so $\mathbb{I}(i) = \mathbb{I}(j)$ if and only if asset i and the asset j belongs to the same sector.

A correlation matrix $\hat{C}$ which incorporates the modeler's beliefs in a parsimonious fashion is given by:

$$\hat{C}_{i,j} = \begin{cases} C_{i,j} & \text{if} \quad \mathbb{I}(i) = \mathbb{I}(j) \\ \beta_i \beta_j \hat{\rho}^{\mathbb{I}(i),\mathbb{I}(j)} & \text{otherwise.} \end{cases} \tag{5}$$

In fact, the reader can may easily check that $\hat{C}$ represents the correlation matrix of a Gaussian probability measure in N-dimensions.[11,12]

3.1 Selecting the benchmark portfolios as the first eigenportfolios for each cluster

A further simplification comes from making a specific choice: we assume that the benchmark portfolio for a given cluster is the first eigenportfolio of the cluster. The standardized return of the benchmark portfolio of sector k can be written in the form

$$F^k = \frac{1}{\sqrt{\lambda^{1,k}}} \sum_{i:\mathbb{I}(i)=k} V_i^k X_i \tag{6}$$

where X_i represents the standardized returns of stock i, V_i^k is the first column in the PCA factorization of the correlation matrix of sector k, and $\lambda^{1,k}$ is the first eigenvalue. Notice that the benchmark portfolios are standardized by definition (mean = 0, variance = 1).

From the orthogonality of the eigenportfolios in the same sector, we can derive a simple formula for the regression coefficients:

$$\beta_i = Corr(X_i, F^{\mathbb{I}(i)}) = \sqrt{\lambda^{1,\mathbb{I}(i)}}\ V_i^{\mathbb{I}(i)}. \tag{7}$$

Figure 4 displays the empirical correlation matrix C side-by-side with the model correlation matrix $\hat{C}$ (HPCA matrix). The HPCA matrix presents a more clear, distinct block structure. The blocks associated with inter-sector correlation are lighter,

[11] This probability measure is the maximum-entropy distribution, in which the modeler's beliefs are viewed as moment constraints [Golan *et al.*, 1996].

[12] Justifying the claim that the model is "parsimonious". It is readily seen that $\hat{C}$ is indeed non-negative definite with unit diagonal elements [Avellaneda, 2019], and thus corresponds to a correlation matrix.

suggesting that correlations between stocks in different clusters are less pronounced than in the empirical data.

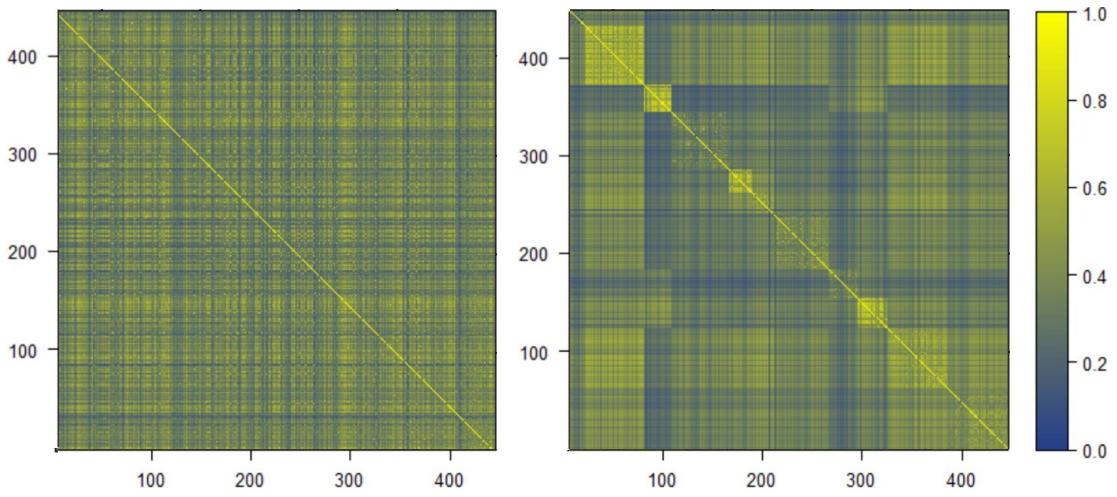

Figure 4: Original (left) and modified (right) correlation matrices estimated with the S&P 500 returns' constituents and GICS clusters, from 2010 to 2019. Dark areas mean low correlation, while lighter areas mean higher correlation.

4 Examples: Global Stocks

4.1 Data and Methodology

We analyze the four major global equity markets: United States, Europe, Emerging markets and China.[13] The data covers the period from January 2010 to November 2019. This period includes different macroeconomic events which affected world markets, such as the "flash crash" of 2010, the European financial crisis in 2011/2012, the downgrade by Standard and Poor's of the U.S. Treasury, the invasion of Ukraine, the Ebola virus outbreak, the downgrading of the Chinese Yuan in 2015, Brexit in 2016, the U.S. elections of 2016, and the Trade Wars of the end of the decade.

[13]The data consists in end-of-day prices adjusted for dividends and splits extracted from Reuters Market Data System. To ensure homogeneity of the asset returns and avoid differences due to the currency of each country, all asset prices were converted to US dollars before calculating log-returns.

Sector (GICS)	USA	Europe	Emerging Mkts.	China
Communication	24	42	59	10
Consumer Discretionary	64	66	114	84
Consumer Staples	32	43	92	23
Energy	28	22	56	5
Financials	64	109	277	19
Health Care	60	54	54	50
Industrials	69	115	89	97
Information Technology	69	33	121	88
Materials	28	49	121	81
Real Estate	31	33	44	22
Utilities	28	30	46	19
Total	**497**	**596**	**1049**	**498**

Table 1: Numbers of companies considered in the study by GICS sectors and regions.

Table 1 shows the main 11 sectors and the number of companies in each sector by geography. The sectors correspond to the GICS. The study comprises 2640 companies across the world. To give perspective on the countries included in the European and Emerging markets groups, Table 10 in Appendix shows the number of companies belonging to the main countries for these regions.

In the following sections, we apply HPCA for the four regions mentioned above. As stated in the introduction, the standard (static) clustering method is based on the GICS. In addition, Emerging and European markets are also clustered by countries. Finally, we propose a statistical clustering technique that is based solely on asset returns and not on exogenous information.

4.2 US Stocks

The stocks analyzed correspond to the S&P 500 constituents, clustered according to the GICS metric. The main GICS are Information Technology, Industrials, Financials, Consumer Discretionary, and Health Care, accounting for more than 65% of the US stocks.

The reason to use the GICS is to capture the economic link between each sector, which is appealing if we consider that companies belonging to the same industries share common factors that can be identified and provide insight on their behavior.

4.2.1 Eigenvalue Analysis: clusters based on GICS

Figure 5 shows that the curve of cumulative explained variance of PCA rises faster than the curve of HPCA due to the nature of HPCA algorithm. This is considered good since it indicates a lower concentration in a few components. In other words, HPCA is less "greedy" than PCA.

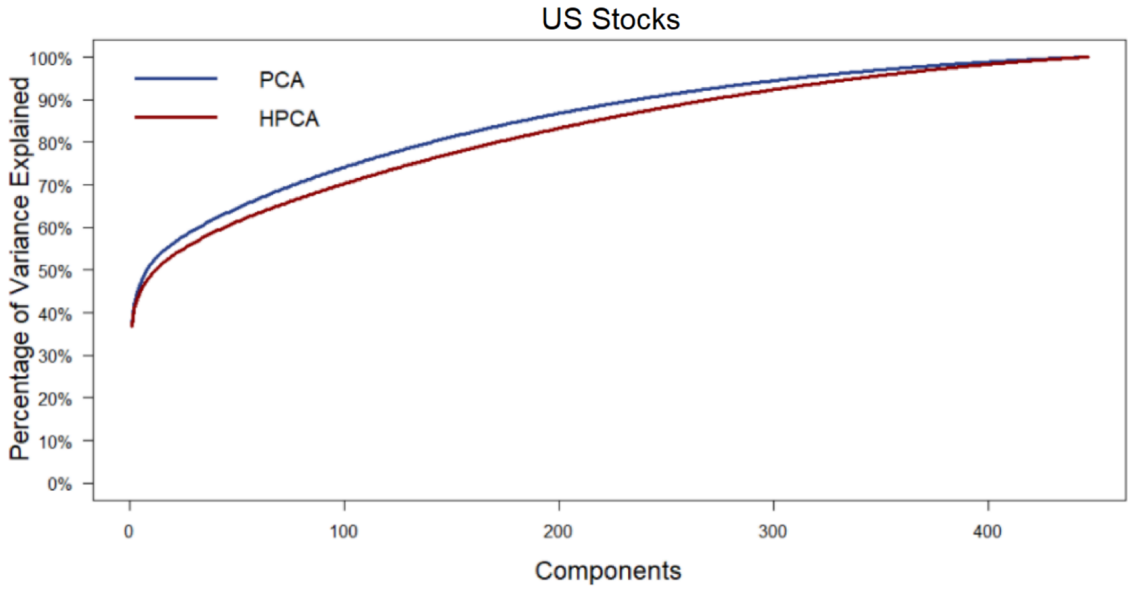

Figure 5: Cumulative variance explained by eigenvalues of PCA and HPCA.

Table 2 depicts the same as shown in the previous figure. HPCA has lower eigenvalues than PCA across the spectrum.

US Stocks					
Eigen	PCA	HPCA	PCA (%)	HPCA (%)	Eigenportfolio
Eigen 1	164.65	163.55	37.01%	36.67%	Multi-sector
Eigen 2	21.24	18.55	4.77%	4.16%	Multi-sector
Eigen 3	8.89	7.10	1.99%	1.59%	Multi-sector
Eigen 4	7.40	6.18	1.66%	1.39%	Multi-sector
Eigen 5	6.43	5.41	1.44%	1.21%	Multi-sector
Eigen 6	5.61	4.27	1.26%	0.96%	Multi-sector
Eigen 7	4.81	3.92	1.08%	0.88%	Multi-sector
Eigen 8	3.88	3.04	0.87%	0.68%	Multi-sector
Eigen 9	3.51	2.81	0.78%	0.63%	Consumer Disc.
Eigen 10	3.13	2.70	0.70%	0.61%	Multi-sector
Eigen 11	2.82	2.56	0.63%	0.57%	Financials
Eigen 12	2.44	2.53	0.55%	0.57%	Inf. Technology
Eigen 13	2.25	2.19	0.51%	0.49%	Health Care
Eigen 14	2.13	2.16	0.48%	0.48%	Health Care
Eigen 15	2.06	2.07	0.46%	0.46%	Industrials

Table 2: First 15 HPCA and PCA eigenvalues clustered by GICS for the US market. In the Eigenportfolio column, we show if a given eigenportfolio is constituted by multiple or single sectors.

Furthermore, we identified two types of eigenportfolios: those localized in mul-

tiple sectors and those concentrated in a single sector. Table 2 shows that almost all eigenportfolios from the first to the fifteenth are multi-sector.

4.2.2 Eigenvector Analysis: clusters based on GICS

Figure 6 shows that the first eigenvector has all the coefficients positive and that the higher-order eigenvectors are concentrated in a narrow range of components, which represents specific sectors or group of sectors.[14]

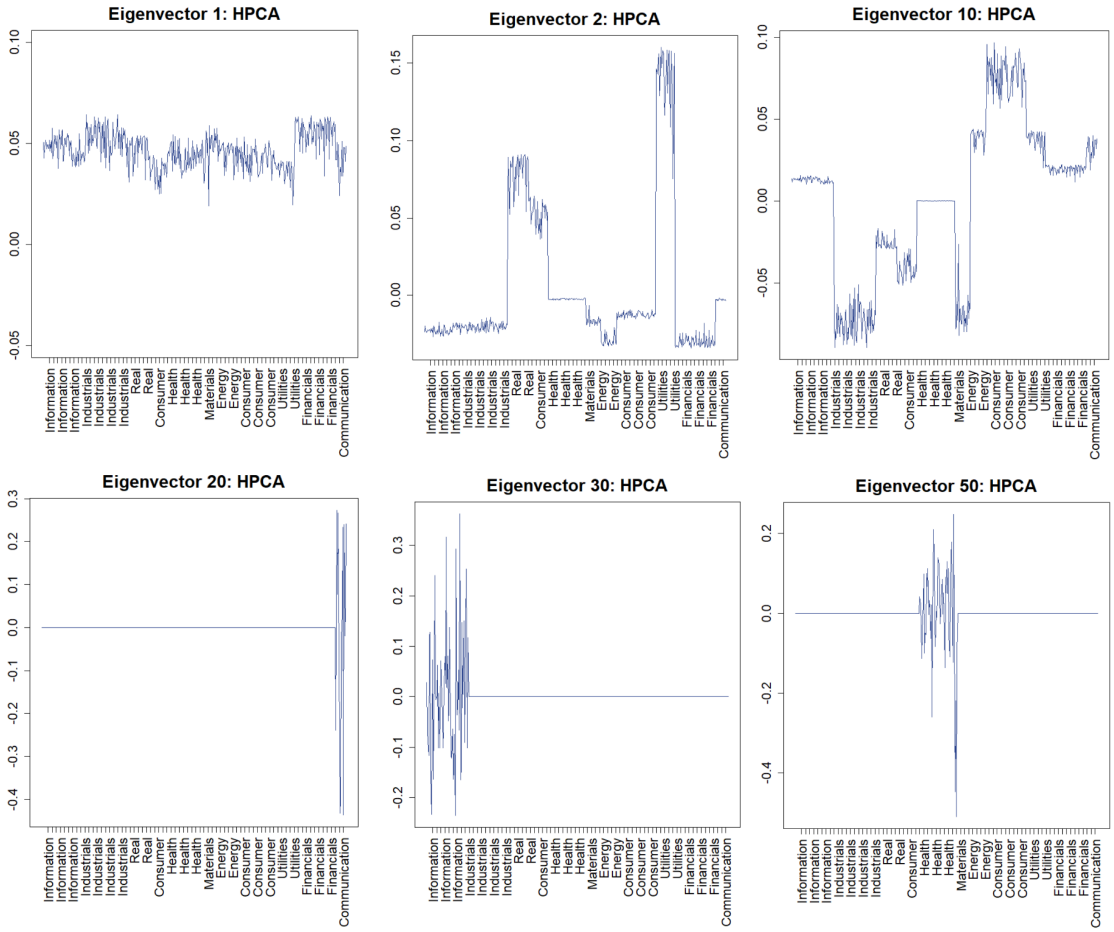

Figure 6: Higher-order eigenvectors of HPCA for US markets clustered by GICS. Higher-order HPCA eigenvectors are localized in one or a few a sectors.

Based on this analysis, HPCA can be used to clean the correlation matrix, build corresponding factor models and apply them to portfolio management. The technique mitigates the identification problem by associating higher-order eigevectors to a specific cluster or a group of clusters.

[14]We only take a few eigenvectors to demonstrate how HPCA works.

4.3 China

The stock universe is constituted by those belonging to the CSI 500 index. The main GICS are Industrials, Information Technology, Consumer Discretionary and Materials, accounting for more than 70% of the total constituents.

4.3.1 Eigenvalue Analysis: clusters based on GICS

Again, the curve of cumulative explained variance of PCA rises faster than the curve of HPCA. The difference between the two curves is more pronounced here than in the US stocks. One of the explanations is that China presents a lower level of diversity.

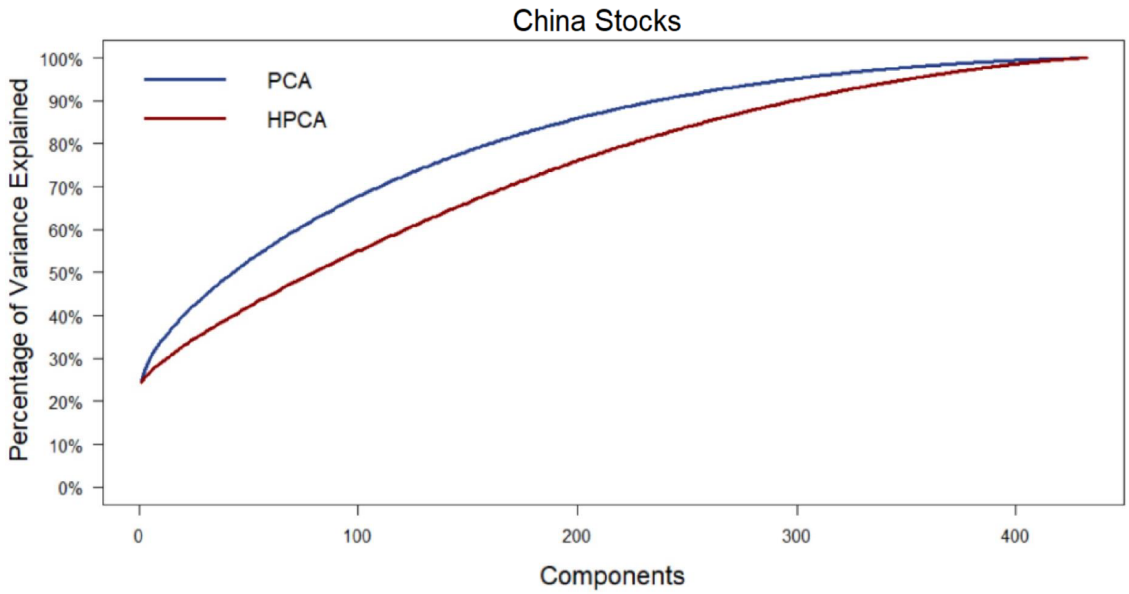

Figure 7: Cumulative variance explained by eigenvalues of PCA and HPCA.

Table 3 shows the same as above. HPCA has lower eigenvalues and the difference in the variance explained by each eigenvalue is substantially higher than in the case of the US stocks.

Chinese Stocks					
Eigen	PCA	HPCA	PCA (%)	HPCA (%)	Eigenportfolio
Eigen 1	106.03	105.45	24.54%	24.41%	Multi-sector
Eigen 2	7.44	3.43	1.72%	0.79%	Multi-sector
Eigen 3	5.94	2.32	1.38%	0.54%	Multi-sector
Eigen 4	5.29	2.09	1.22%	0.48%	Multi-sector
Eigen 5	4.58	2.05	1.06%	0.47%	Multi-sector
Eigen 6	4.08	1.99	0.94%	0.46%	Multi-sector
Eigen 7	3.57	1.97	0.83%	0.46%	Multi-sector
Eigen 8	3.27	1.96	0.76%	0.45%	Multi-sector
Eigen 9	3.09	1.87	0.72%	0.43%	Multi-sector
Eigen 10	2.92	1.86	0.68%	0.43%	Multi-sector
Eigen 11	2.84	1.76	0.66%	0.41%	Multi-sector
Eigen 12	2.79	1.73	0.65%	0.40%	Multi-sector
Eigen 13	2.67	1.72	0.62%	0.40%	Multi-sector
Eigen 14	2.62	1.70	0.61%	0.39%	Multi-sector
Eigen 15	2.57	1.67	0.59%	0.39%	Multi-sector

Table 3: First 15 HPCA and PCA eigenvalues clustered by GICS for Chinese markets.

4.3.2 Eigenvector Analysis: clusters based on GICS

Figure 8 shows different eigenvectors for the case of China. Again, the first eigenvector has all the positive coefficients, representing a good proxy for the market portfolio. Higher-order eigenvectors are concentrated in a narrow range of components, representing specific sectors or group of sectors (i.e., portfolios of eigenportolios), as depicted in Table 3.

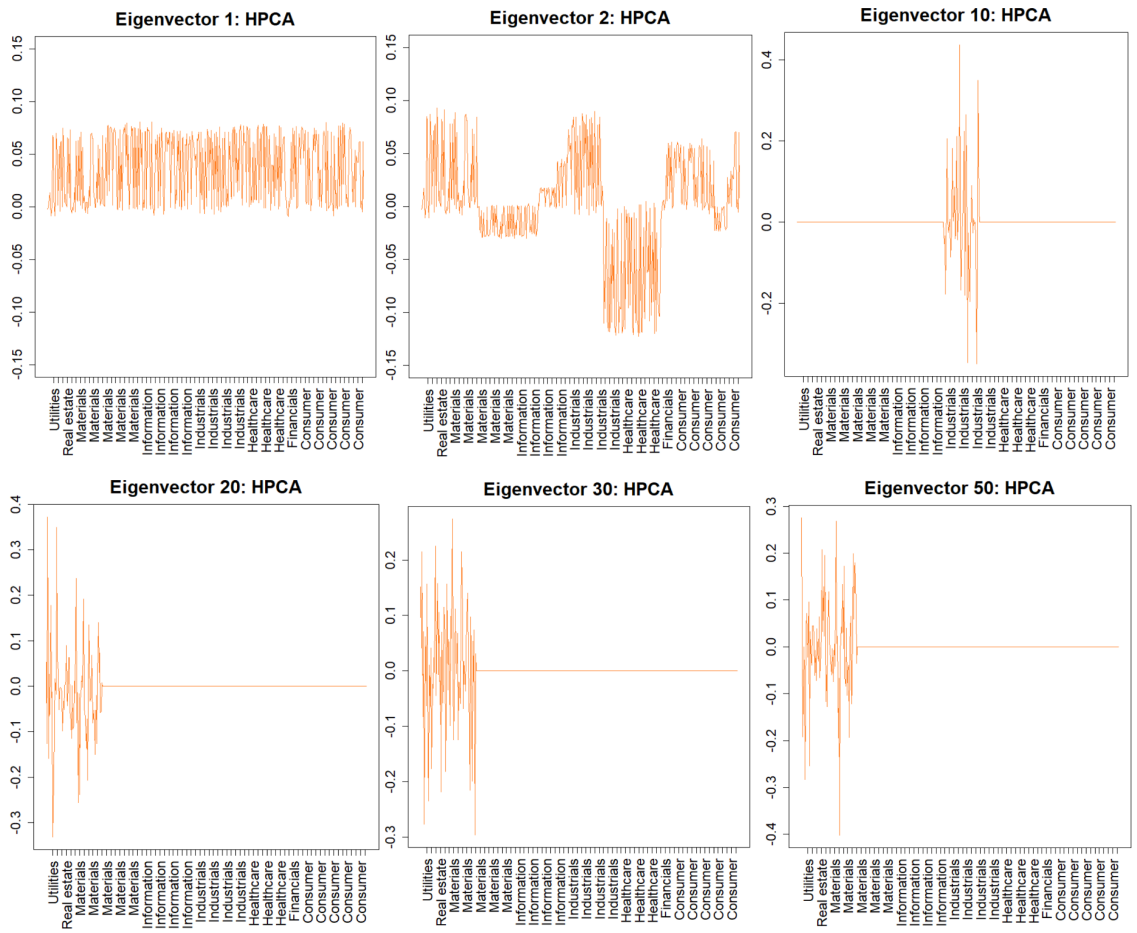

Figure 8: Higher-order eigenvectors of HPCA for China markets clustered by GICS. Higher-order HPCA eigenvectors are localized in one or a few sectors.

4.4 European Stocks

In this subsection, we repeat the previous analysis for European stocks, but here, in addition to the GICS clustering, we extend the analysis to clusters based on European countries.

The stocks analyzed belongs to the STOXX Europe 600 index, one of the main stock indexes of Europe. The main GICS are Industrials, Financials, Consumer Discretionar and Health Care. The main countries are Great Britain, France, Switzerland, Germany, Netherlands, Sweden, Spain, Italy, among others.

4.4.1 Eigenvalue Analysis: clusters based on GICS

As expected, the curve of the PCA cumulative explained variance increases faster than the HPCA curve. The behavior here is similar to that of the US market.

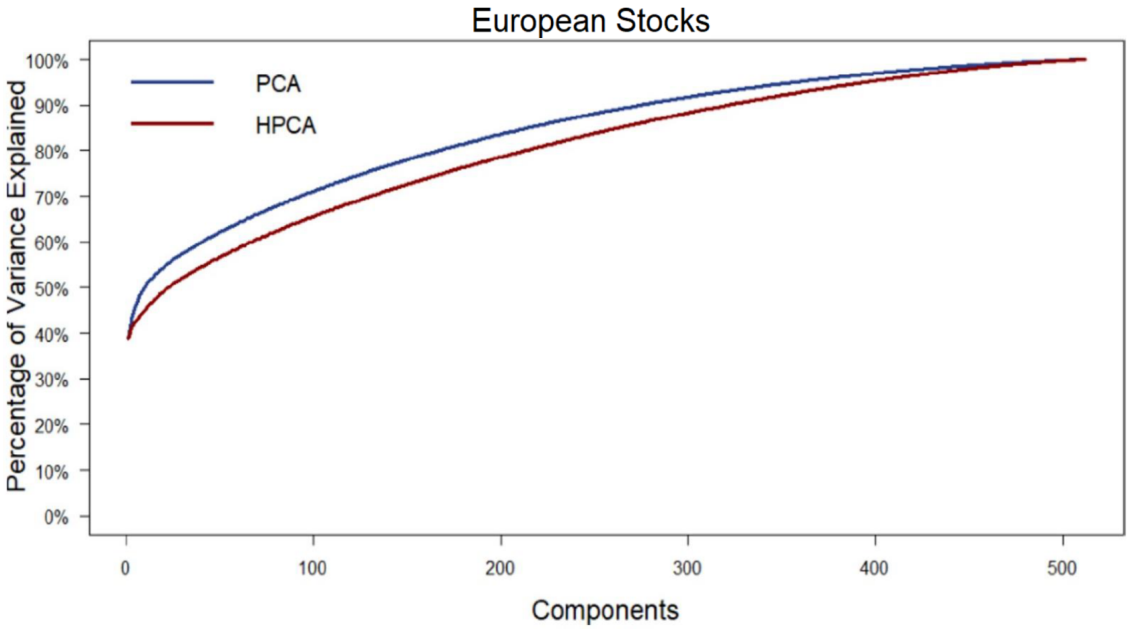

Figure 9: Cumulative variance explained by eigenvalues of PCA and HPCA.

Table 4 shows that HPCA has slightly lower eigenvalues. We note that the difference between PCA and HPCA is smaller here than in the Chinese stocks, as the market is more diverse.

European Stocks					
Eigen	PCA	HPCA	PCA (%)	HPCA (%)	Eigenportfolio
Eigen 1	197.67	194.94	38.76%	38.20%	Multi-sector
Eigen 2	11.42	7.45	2.24%	1.46%	Multi-sector
Eigen 3	10.13	5.82	1.99%	1.14%	Multi-sector
Eigen 4	9.24	3.96	1.81%	0.78%	Multi-sector
Eigen 5	6.28	3.18	1.23%	0.63%	Financials
Eigen 6	5.63	2.94	1.10%	0.58%	Multi-sector
Eigen 7	4.80	2.66	0.94%	0.52%	Consumer Disc.
Eigen 8	3.88	2.54	0.76%	0.50%	Industrials
Eigen 9	3.64	2.47	0.71%	0.48%	Financials
Eigen 10	3.40	2.43	0.67%	0.48%	Multi-sector
Eigen 11	3.05	2.28	0.60%	0.45%	Multi-sector
Eigen 12	2.68	2.26	0.53%	0.44%	Multi-sector
Eigen 13	2.49	2.17	0.49%	0.43%	Materials
Eigen 14	2.37	2.16	0.47%	0.42%	Multi-sector
Eigen 15	2.14	2.12	0.42%	0.41%	Multi-sector

Table 4: First 15 HPCA and PCA eigenvalues clustered by GICS for the European market.

4.4.2 Eigenvector Analysis: clusters based on GICS

Figure 10 shows that the first eigenvector represents a long-only portfolio. Higher-order eigenvectors are concentrated in a narrow range of components.

For example, the twentieth eigenportfolio is concentrated with long and short coefficients in the Communication and Financial sectors. The thirteenth eigenportfolio focuses on Industrial and Real Estate and the fifteenth has its coefficients almost entirely in the Financial sector. The second and the tenth are less intuitive, with significant coefficients across the spectrum. Table 4 shows which components represent specific sectors or groups of sectors.

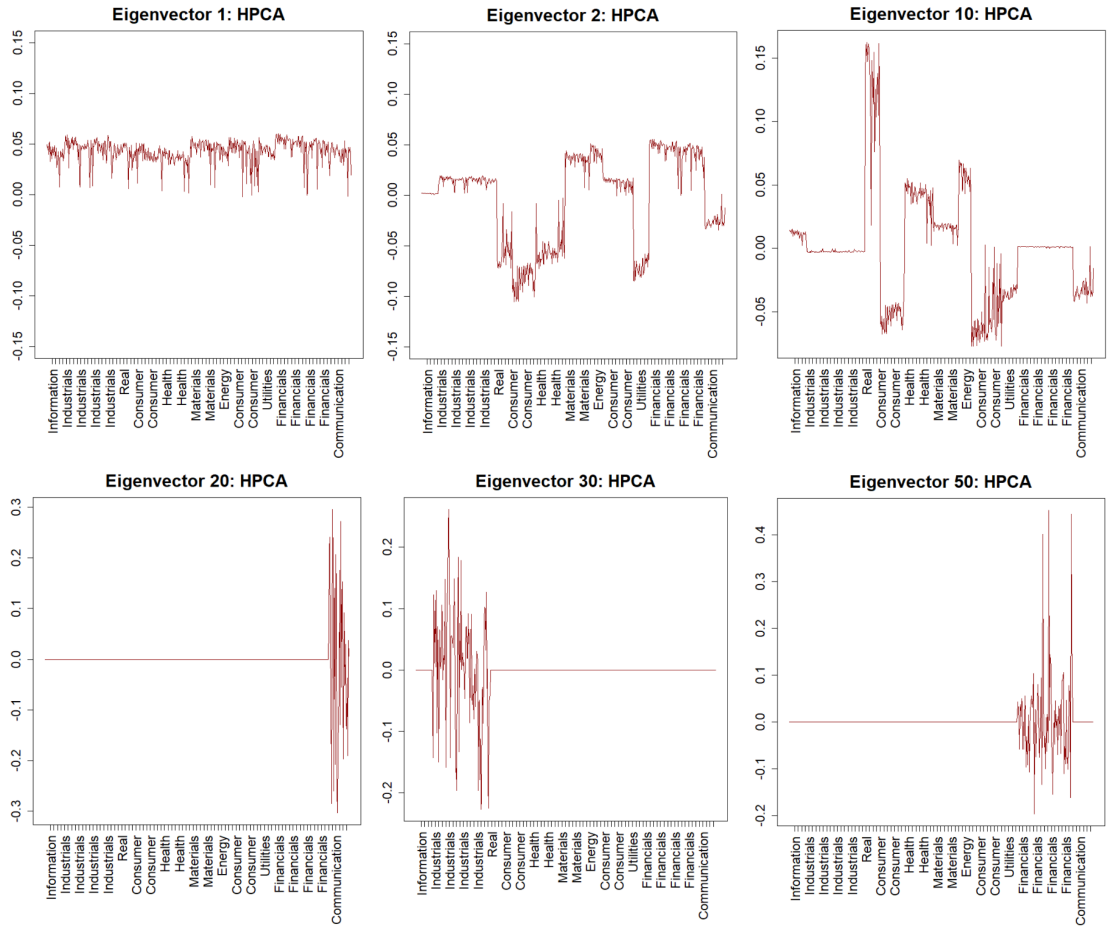

Figure 10: Higher-order eigenvectors of HPCA for European markets clustered by GICS. Higher-order HPCA eigenvectors are localized in one or a few sectors.

4.4.3 Eigenvalue Analysis: clusters based on Countries

Here we conducted HPCA based on the main countries belonging to the European market (see Table 10 in the Appendix for more details). As in the GICS-based analysis, the curve of cumulative explained variance of PCA increases faster than

the HPCA counterpart. In addition, compared to HPCA for the European stocks based on GICS, here the first eigenvalues explain a greater variation, meaning a higher level of concentration in a few components. For example, the first GICS-based eigenvalue of HPCA explains 33.36% of the total variance, while the country-based first eigenvalue of HPCA explains 38.76%.

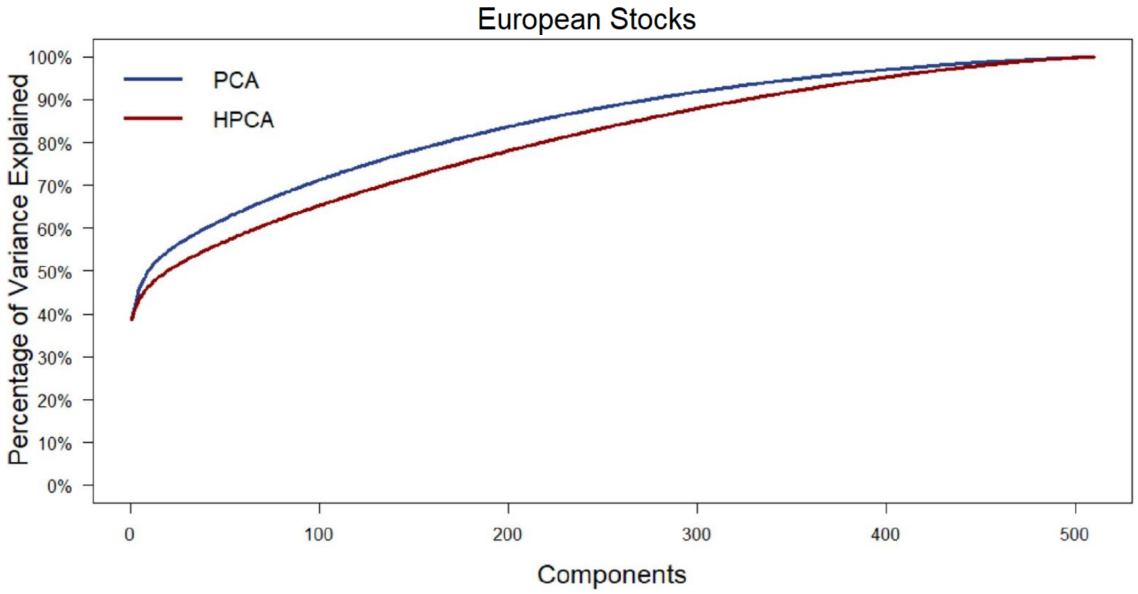

Figure 11: Cumulative variance explained by eigenvalues of PCA and HPCA based on country clustering.

Table 5 shows that HPCA has slightly lower eigenvalues than PCA. As mentioned earlier, the level concentration is higher than in the case of GICS-based HPCA. Also, here the difference between PCA and HPCA is smaller. Some components represent multiple countries (portfolios of eigenportfolios) while others are concentrated in only one country. For example, the sixth and the seventh eigenportfolios are concentrated in United Kingdom, while the thirteenth and fourteenth represent Switzerland and Germany, respectively.

European Stocks					
Eigen	PCA	HPCA	PCA (%)	HPCA (%)	Eigenportfolio
Eigen 1	197.67	197.18	38.76%	38.66%	Multi-country
Eigen 2	11.42	9.14	2.24%	1.79%	Multi-country
Eigen 3	10.13	5.65	1.99%	1.11%	Multi-country
Eigen 4	9.24	5.39	1.81%	1.06%	Multi-country
Eigen 5	6.28	4.33	1.23%	0.85%	Multi-country
Eigen 6	5.63	3.86	1.10%	0.76%	United Kingdom
Eigen 7	4.80	3.36	0.94%	0.66%	United Kingdom
Eigen 8	3.88	2.78	0.76%	0.54%	Multi-country
Eigen 9	3.64	2.74	0.71%	0.54%	Multi-country
Eigen 10	3.40	2.48	0.67%	0.49%	Multi-country
Eigen 11	3.05	2.19	0.60%	0.43%	Multi-country
Eigen 12	2.68	2.16	0.53%	0.42%	Multi-country
Eigen 13	2.49	2.13	0.49%	0.42%	Switzerland
Eigen 14	2.37	1.96	0.47%	0.38%	Germany
Eigen 15	2.14	1.86	0.42%	0.37%	Multi-country

Table 5: First 15 HPCA and PCA eigenvalues clustered by country for European markets.

4.4.4 Eigenvector Analysis: clusters based on Countries

The eigenvectors here are unequivocal, showing the power of HPCA. The first eigenvector represents a market portfolio across European countries. Higher-order portfolios are concentrated in a few countries. For example, the tenth, the twentieth and the thirtieth portfolios are the three a combination of France and Germany. The fifteenth eigenportfolio is concentrated with almost all the positive coefficients in Norway and Finland. See Table 5 for more details.

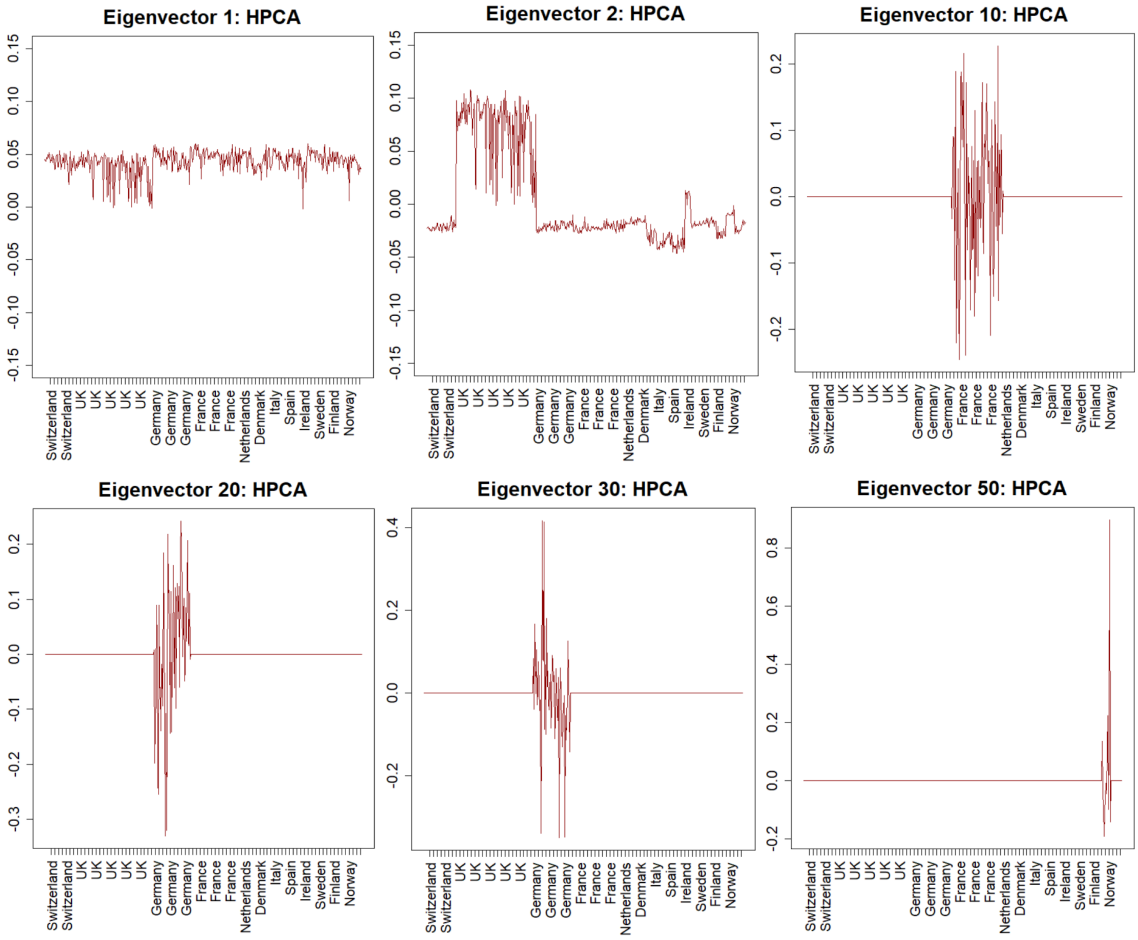

Figure 12: Higher-order eigenvectors of HPCA for European markets clustered by countries. Higher-order HPCA eigenvectors are localized in one or a few countries.

4.5 Emerging Markets

As in the case of European stocks, here we cluster the stocks based on GICS and countries. The stocks analyzed belongs to the MSCI Emerging Markets Index. Almost 60% of the stocks are concentrated in Financials, Information Technology, Materials, and Consumer Discretionary. The main countries are China, Korea, Taiwan, India, Brazil, South Africa, Russia, Mexico, and Thailand.

4.5.1 Eigenvalue Analysis: clusters based on GICS

The cumulative explained variance curve of PCA increases faster than the HPCA curve. The behavior here is similar to that of the Chinese market, i.e., the difference in the explained variance between PCA and HPCA is wider than in the case of the US and the European markets.

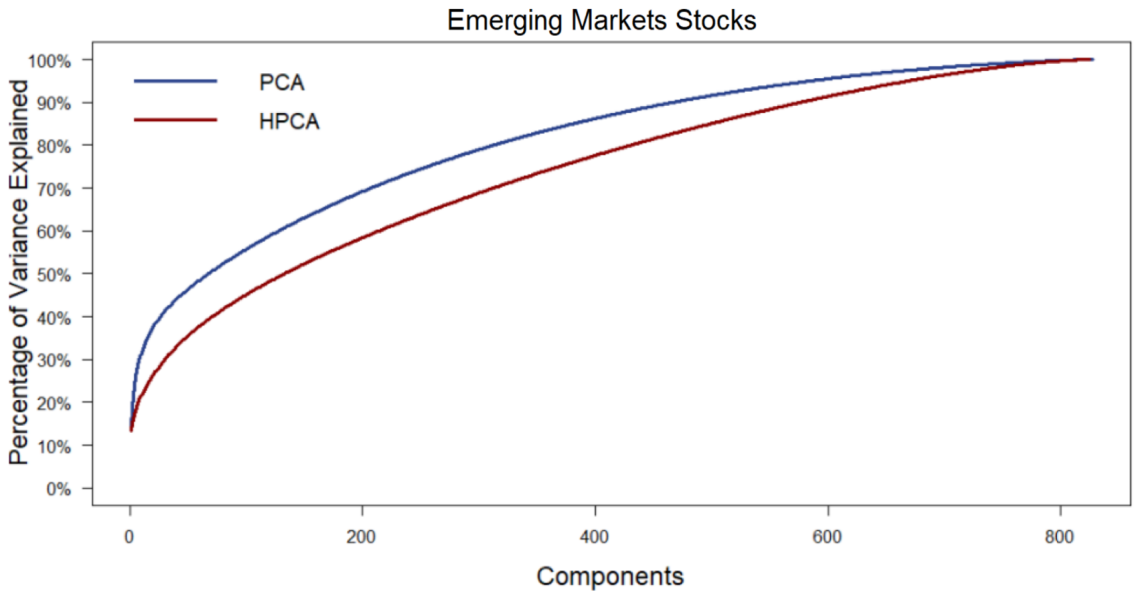

Figure 13: Cumulative variance explained by eigenvalues of PCA and HPCA.

Table 6 shows that HPCA has lower eigenvalues than PCA, repeating the pattern observed in all the previous cases analyzed. Furthermore, the first eigenvalue explains significantly less variance than in all other cases. To put it in perspective, the first eigenvalue of PCA and HPCA accounts approximately for 13% of the total variance, whereas in the US, Chinese and European markets represent approximately 37%, 24% and 33% of the total variance, respectively.

Emerging Stocks					
Eigen	PCA	HPCA	PCA (%)	HPCA (%)	Eigenportfolio
Eigen 1	113.67	109.19	13.83%	13.24%	Multi-sector
Eigen 2	39.44	14.72	4.80%	1.78%	Financials
Eigen 3	23.77	8.98	2.89%	1.09%	Multi-sector
Eigen 4	20.68	8.10	2.52%	0.98%	Multi-sector
Eigen 5	14.87	6.97	1.81%	0.85%	Inf. Technology
Eigen 6	11.27	6.95	1.37%	0.84%	Multi-sector
Eigen 7	10.81	6.63	1.32%	0.80%	Multi-sector
Eigen 8	9.30	6.19	1.13%	0.75%	Multi-sector
Eigen 9	8.68	5.75	1.06%	0.70%	Industrials
Eigen 10	7.14	4.50	0.87%	0.55%	Consumer Stap.
Eigen 11	6.35	4.38	0.77%	0.53%	Financials
Eigen 12	6.07	4.33	0.74%	0.53%	Financials
Eigen 13	5.94	4.21	0.72%	0.51%	Consumer Stap.
Eigen 14	5.37	3.89	0.65%	0.47%	Multi-sector
Eigen 15	4.81	3.87	0.59%	0.47%	Multi-sector

Table 6: First 15 HPCA and PCA eigenvalues clustered by GICS for Emerging markets.

4.5.2 Eigenvector Analysis: clusters based on GICS

Figure 14 shows that the first eigenvector represents a proxy for a market portfolio. Higher-order eigenvectors are concentrated in a narrow range of components.

For example, the second eigenvector is concentrated exclusively in the Financial sector, the tenth in Consumer Staples, the twentieth eigenportfolio is concentrated with long and short coefficients in the Material and Energy sectors. The thirteenth eigenportfolio focuses on Financial and Utilities sectors and the fifteenth has its coefficients almost entirely in the Financial sector. See Table 6 for more details on portfolios.

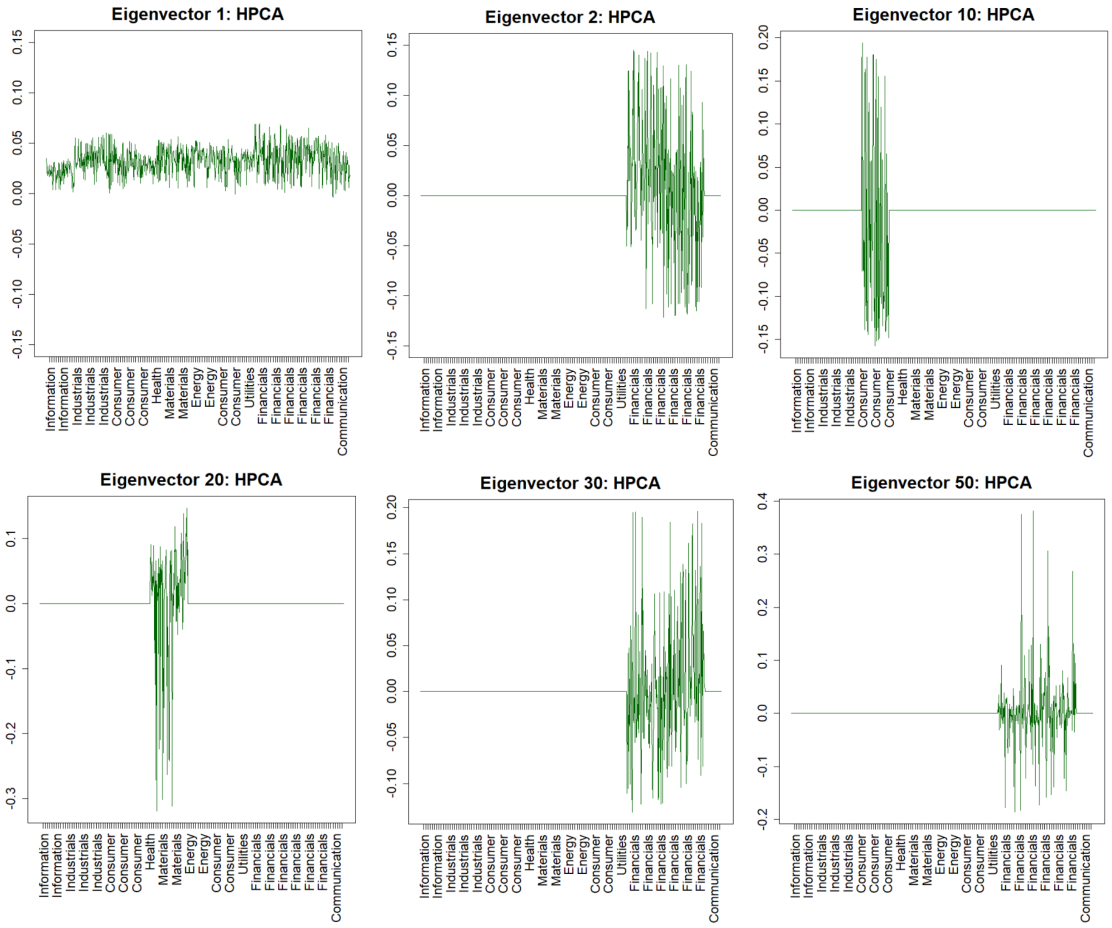

Figure 14: Higher-order eigenvectors of HPCA for Emerging markets clustered by GICS. Higher-order HPCA eigenvectors are localized in one or a few sectors.

4.5.3 Eigenvalue Analysis: clusters based on Countries

Like all the previous cases, the curve of the PCA cumulative explained variance increases faster than the HPCA curve. The behavior here is slightly different that the behavior of the GICS-based HPCA for Emerging markets. Specifically, the different of the curves of the two approaches (PCA and HPCA) here is lower than the counterpart for the GICS-based HPCA case.

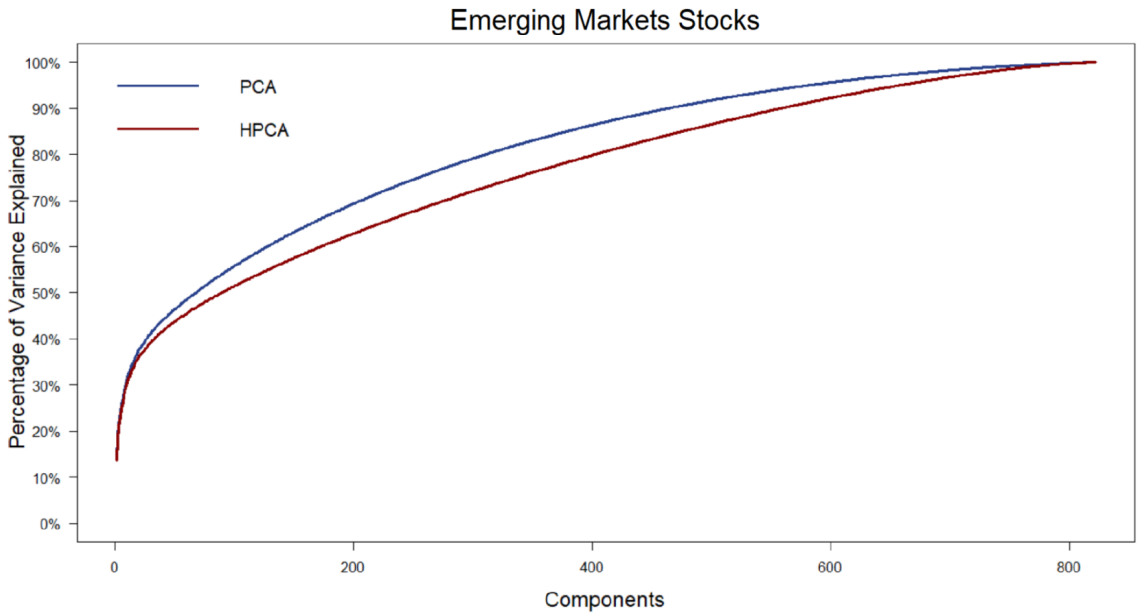

Figure 15: Cumulative variance explained by eigenvalues of PCA and HPCA classified by country.

Table 7 shows that HPCA has slightly lower eigenvalues. Furthermore, it shows that almost all eigenportfolios are built based on multiple countries.

Emerging Stocks					
Eigen	PCA	HPCA	PCA (%)	HPCA (%)	Eigenportfolio
Eigen 1	113.67	111.51	13.83%	13.57%	Multi-country
Eigen 2	39.44	38.20	4.80%	4.65%	Multi-country
Eigen 3	23.77	22.03	2.89%	2.68%	Multi-country
Eigen 4	20.68	19.00	2.52%	2.31%	Multi-country
Eigen 5	14.87	13.93	1.81%	1.69%	Multi-country
Eigen 6	11.27	11.33	1.37%	1.38%	China
Eigen 7	10.81	10.85	1.32%	1.32%	Multi-country
Eigen 8	9.30	9.66	1.13%	1.18%	Multi-country
Eigen 9	8.68	8.22	1.06%	1.00%	Multi-country
Eigen 10	7.14	6.88	0.87%	0.84%	Multi-country
Eigen 11	6.35	5.92	0.77%	0.72%	Multi-country
Eigen 12	6.07	5.66	0.74%	0.69%	China
Eigen 13	5.94	5.42	0.72%	0.66%	Multi-country
Eigen 14	5.37	4.62	0.65%	0.56%	Multi-country
Eigen 15	4.81	4.45	0.59%	0.54%	Multi-country

Table 7: First 15 HPCA and PCA eigenvalues clustered by country for Emerging markets.

4.5.4 Eigenvector Analysis: clusters based on Countries

Figure 16 shows that, unlike the previous GICS-based HPCA case, the distribution of the eigenvectors across the spectrum is less clear here.

As usual, the first eigenvector has all the positive coefficients. The other eigenvectors are less intuitive, although it is noticeable that the twentieth eigenportfolio is concentrated almost entirely in China and the fiftieth eigenportfolio in Korea. Table 7 provides more details.

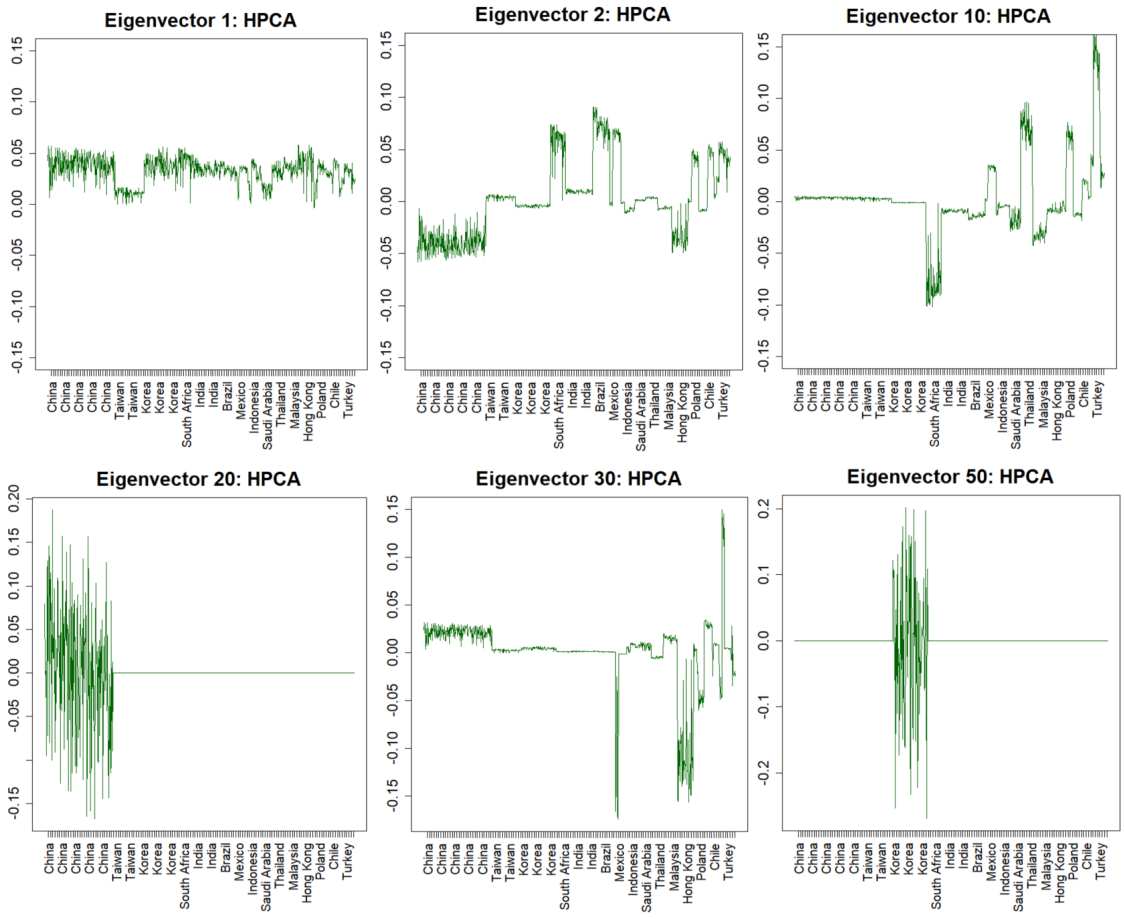

Figure 16: Higher-order eigenvectors of HPCA for Emerging markets clustered by country.

5 HPCA Statistically Generated Clusters

Practitioners usually employ pre-built (static) clusters such as GICS and countries to build factor models. Stocks belonging to the same GICS or country share common factors that capture –to some extend– their joint dynamics.

However, these types of models have some shortcomings. Stock markets and their components change almost continuously, producing substantial changes in their

behavior that are not captured by static clusters. This is not desirable for risk and portfolio management for various reasons.

It goes against the diversification of a (seemingly diversified) strategy. Many investment portfolios base their mandates on diversifying their allocations among sectors, sub-sectors, countries, etc., to avoid high and undesirable idiosyncratic risk. However, there are several factors beyond the sector and/or the country that affect the behavior of portfolio holdings. For example, when interest rates rise sharply, capital-intensive companies are negatively affected and diversification vanishes.

Trading strategies, such as the so-called sector/country rotation may also been affected for the same reasons. Securities that belong to a specific sector/country can change their behavior sharply under the changes of a market regime and the strategy that worked ex-ante may stop working overnight.

To mitigate this problem and account for hidden risk factors, we adopt a purely statistical technique. This is a simple and still powerful tool that dynamically adapts to changes in market conditions over time, which makes it suitable for managing trading portfolios. Also, it is a parsimonious approach since it does not rely on too many parameters. The user only needs to define the number of clusters, which depends on the number of K eigenvectors, without specifying any other parameters or hyper-parameters.

5.1 Description of the Algorithm

Based on matrix diagonalization (PCA), the algorithm constructs new features in the space that retain the behavior of each component based on linear combinations of its main characteristics, leading to statistical clusters of similar-behaved securities. Figure 17 illustrates how the space is divided into different quadrants (clusters) to which each stock belongs.

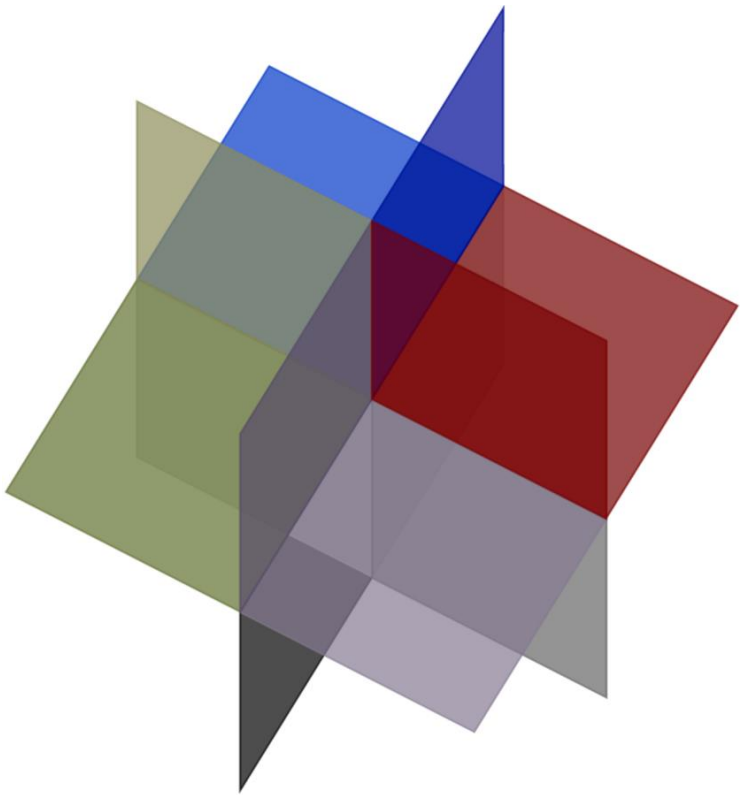

Figure 17: The space is divided into different quadrants (clusters) to which each asset belongs based on the sign of the eigenvectors.

To run the statistical clustering, it is needed to estimate the eigenvectors as in Eq. (3) and define the number of factors K, omitting the first one (Eq. (2)) since it has all the coefficients positive.[15] Then, each security is clustered appropriately according to the sign of the coefficients in each eigenvector.

5.2 HPCA with Statistical Clustering

In this section, we analyze HPCA using the statistical clustering method. We set the number of eigevectors K equal to 4. Therefore, the expected number of clusters is $2^4 = 16$, although not necessarily all the clusters will have components. Some of them can be empty, *ergo* removed.

5.2.1 Eigenvalue Analysis

Clustering using a statistical approach delivers similar patterns on the curve of cumulative explained variance as the static approaches. The PCA curve rises faster than the HPCA curve.

[15]This applies to correlated markets, e.g. stocks. If other (uncorrelated) assets are included, this is not necessary.

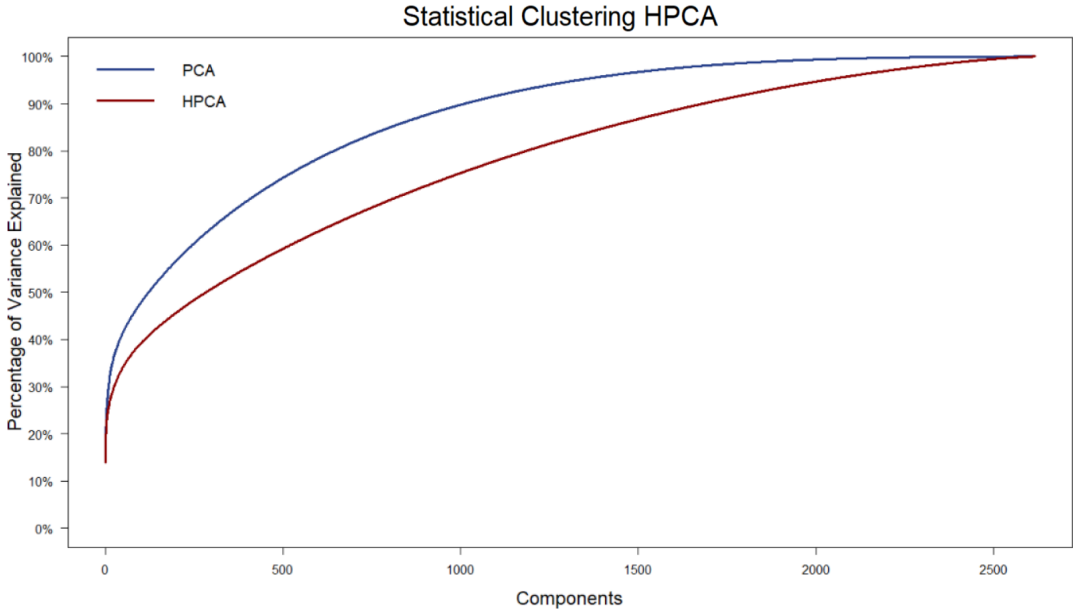

Figure 18: Cumulative variance explained by eigenvalues of PCA and HPCA using statistical clustering.

Table 8 shows that HPCA has lower eigenvalues.

Statistical Clustering				
Eigen	PCA	HPCA	PCA (%)	HPCA (%)
Eigen 1	376.72	371.51	14.41%	14.21%
Eigen 2	149.51	147.68	5.72%	5.65%
Eigen 3	73.52	66.99	2.81%	2.56%
Eigen 4	52.10	42.09	1.99%	1.61%
Eigen 5	32.44	32.17	1.24%	1.23%
Eigen 6	28.17	23.85	1.08%	0.91%
Eigen 7	25.16	23.36	0.96%	0.89%
Eigen 8	23.12	21.52	0.88%	0.82%
Eigen 9	20.73	12.42	0.79%	0.48%
Eigen 10	15.13	9.83	0.58%	0.38%
Eigen 11	14.58	9.64	0.56%	0.37%
Eigen 12	11.67	8.92	0.45%	0.34%
Eigen 13	11.27	8.41	0.43%	0.32%
Eigen 14	10.60	7.92	0.41%	0.30%
Eigen 15	9.71	7.63	0.37%	0.29%

Table 8: First 15 HPCA and PCA eigenvalues using statistical clustering.

5.2.2 Eigenvector Analysis

Higher-order eigenvectors are concentrated in a few clusters. The second eigenvector is not easy to associate since it is significant across the whole spectrum. See the plot below.

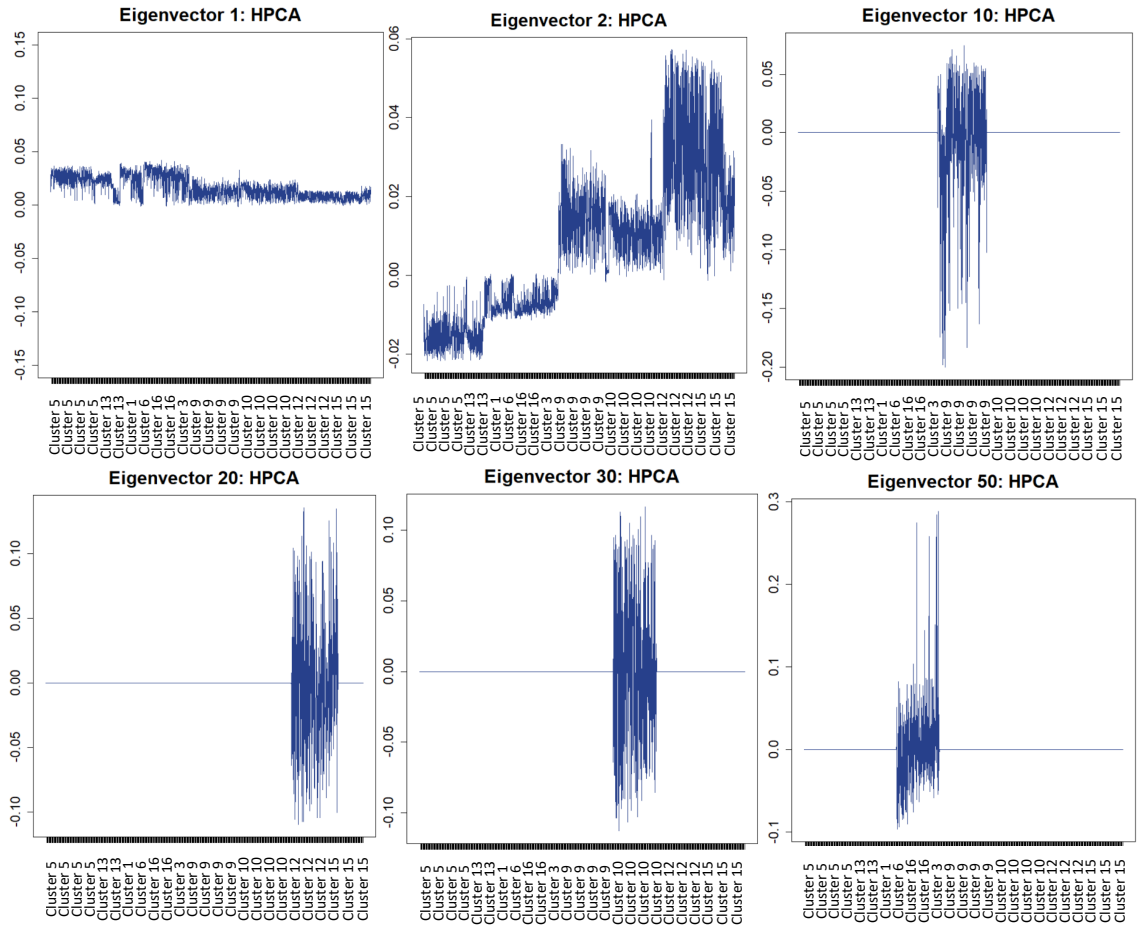

Figure 19: Higher-order eigenvectors of HPCA clustered by the statistical approach. For example, the tenth eigevector is localized in *Cluster 9*. The twentieth is localized mostly in *Cluster 12*. The thirtieth eigenvector is totally concentrated in *Cluster 10* and the fiftieth in *Clusters 6 and 16*. See Table 12 for details on each cluster.

An interesting question at this point is oriented towards the meaning of each group.[16] If the clustering technique works well, one would expect assets that share common factors, such as sectors or countries, to belong to the same cluster. However, this is not necessarily true because there are myriad "hidden" factors that are difficult to identify and remain important drivers of asset returns. In the end, that is the objective of the statistical clustering approach; identify clusters that are not easy to see with common factors such as sectors or countries.

[16]See in the Appendix Table 11 the number of components of each cluster.

Table 12 in the Appendix shows the main sectors and countries to which the components (assets) belong within each cluster.

There are two types of sectors. *Cyclical sectors*: Consumer Discretionary, Financials, Real Estate, Industrials, Information Technology, Materials and Communication, and *defensive sectors*: Consumer Staples, Energy, Health Care and Utilities.

The main insight obtained from this analysis is that in clusters 2, 5, 8, 9, 10, 12, 14, 15 and 16 the three main sectors are cyclical. In the other clusters, at least two out of three are also cyclical. Only cluster 13 has two defensive sectors. As for the countries, the most interesting insights are obtained from clusters 2, 4, 5, 12, 13, 14 and 15, where almost all the companies belong to China or to the US. Therefore, the statistical clustering approach is doing a good job by identifying common factors such as sector (distinguishing between cyclical and defensive sectors) and countries but also it takes into account other non-explicit (statistical) factors. This is evident in the next section, where we apply this procedure to portfolio management.

6 Application to Portfolio Management

We demonstrate how practitioners might apply HPCA to portfolio management. First we show that, after transaction costs, the main eigenportfolio has a similar performance to that of the market, as explained in Subsection 2. Second, we use the statistical clustering approach with HPCA to build statistical factor models and build an investment portfolio.

6.1 First Eigenportfolio ≈ Market Portfolio

We construct a portfolio based on the first eigenvector for the US stock market. The portfolio is rebalanced monthly and the correlation matrix is estimated with a rolling window of approximately 125 days.

As expected, the first eigenportfolio is a good proxy of the market. This confirm the points mentioned before.

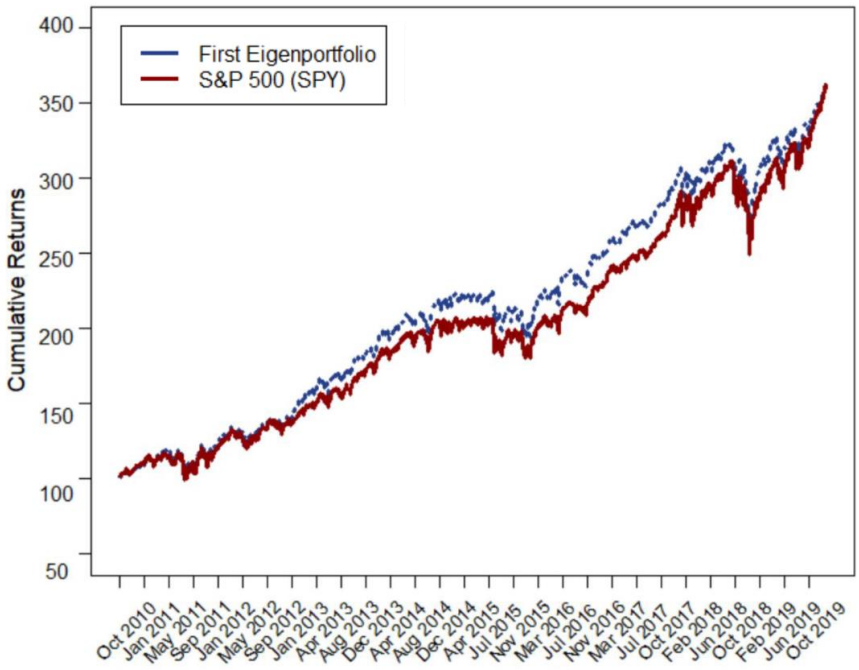

Figure 20: Cumulative returns of the market portfolio, represented via its tradeable ETF (SPY), and the first eigenportfolio, rebalanced monthly.

6.2 Statistical Clustering Factor Model

We blend the statistical clustering approach with HPCA and use the (ordered) eigenvalues and its associated eigenvectors to build a statistical factor model for the covariance matrix and the expected returns. The model correlation matrix using K components of HPCA reads

$$\mathcal{C} = \hat{\mathcal{O}}\hat{\Lambda}\hat{\mathcal{O}}^T + \zeta^2 \tag{8}$$

$$\zeta_j^2 = \sum_{i=K+1}^{N} \lambda^{(i)} (\mathcal{O}_j^{(i)})^2 \tag{9}$$

where ζ^2 is the (uncorrelated) idiosyncratic risk and $\hat{O}$ and $\hat{\Lambda}$ are the modified orthogonal matrix of eigenvectors and the diagonal matrix of eigenvalues, respectively. We define the expected returns as

$$E(r) = \sum_{i=i}^{K} \beta_j^{(K)} F^{(K)} + \epsilon_i \tag{10}$$

where $\beta_j^{(K)}$ are the factor loadings and ϵ_i the residuals.

To select the number K of factors used to reconstruct the correlation matrix and compute the expected returns we used the so-called effective rank (eRank) method. Let the singular value decomposition of the $T \times N$ matrix of standardized log-return

$$R = UDV \tag{11}$$

where U and V are $T \times T$ and $N \times N$ unitary matrices and D is the diagonal matrix with singular values in decreasing order. Then, the associated probability distribution is

$$\mathcal{P}_j = \frac{\sigma_j}{\|\sigma\|_1} \quad \text{for} \quad j = 1, ..., Q \quad \text{for} \quad Q = M \wedge T \tag{12}$$

where $\|.\|_1$ is the $L-1$ norm. Then, the effective rank is defined as

$$eRank(R) = exp\{\mathcal{H}(\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_Q)\} \tag{13}$$

where $\mathcal{H}$ is the Shannon entropy.[17]

Figure 21 shows the cumulative returns of the different strategies analyzed.

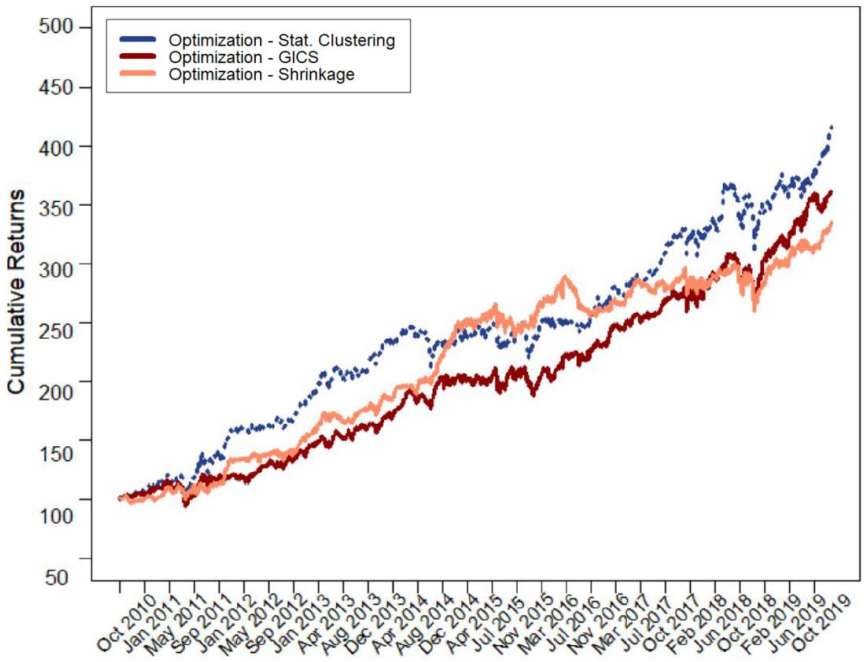

Figure 21: In this plot we show the cumulative returns of three optimized portfolios based on HPCA. The blue line is the strategy performed using the statistical clustering approach, the red line represents the strategy based on the GICS clustering and salmon line represents the classical Sharpe ratio optimization with shrinkage covariance matrix.

Table 9 displays the main performance statistics of the three optimization techniques, the passive S&P 500 Index and the First Eigenportfolio, showing that the optimization approach based on statistical clustering with HPCA outperforms all the other portfolios.

[17] See [Kakushadze and Yu, 2017] for more implementation details.

	CAGR	Std Dev	Sharpe ratio	MaxDD	Calmar ratio
First Eigen	15.36%	14.31%	1.07	21.7%	0.71
Stat. Clustering	16.91%	13.65%	1.24	16.3%	1.04
GICS	15.13%	13.06%	1.16	18.0%	0.84
Shrinkage	14.19%	11.48%	1.24	14.7%	0.97
S&P 500	15.13%	13.99%	1.08	19.9%	0.76

Table 9: Main performance statistics of the proposed strategies. The risk-free rate was set to zero.

Conclusions

Throughout this paper, we have empirically compared the performance of PCA and HPCA. We tested different approaches for more than 2600 stocks, grouping them by market regions: United States, Europe, China, and Emerging stock markets. Main results provide evidence that HPCA works remarkably well for different markets, tackling one of the main shortcomings of classical PCA, the so-called identification problem. The HPCA portfolios, unlike the PCA ones, are easy to interpret and embeds an economic/financial intuition behind. Thus, these portfolios can be used as factors to build models for risk and portfolio management, as well as for different trading strategies such as statistical arbitrage.

Clustering stock returns by GICS and countries provides outstanding results when HPCA is applied. However, these types of clustering approaches have some drawbacks. First, the user needs to specify the clusters, which are not always available for all markets. Second, static groups like these cannot quickly adjust to changes in market conditions. In this matter, we showed that under different market regimes, the correlation between different assets changes sharply and so, the behavior of returns. This, in turn, could affects the covariance matrix estimators and the performance of well-known trading strategies such as factor/cluster rotation. A seemingly well-diversified portfolio can get concentrated in a few components if market conditions change and the user does not account for that timely.

To tackle these issues inherent in static clusters, we propose a purely statistical cluster approach, in which, the user no longer needs to specify the clusters, since they are obtained from the spectral decomposition of the correlation matrix of the stock returns. This approach is more promising to account for changes in the behavior of assets. Furthermore, by definition this tool is parsimonious provided that the user only needs to fix up one parameter, the number of K eigenvectors.[18] To illustrate an application, we show it in the context of portfolio optimization for the US stock market. We provide evidence that using HPCA statistical-based factor models outperform other classical portfolio construction methodologies such as the shrinkage covariance matrix and the HPCA GICS-based factor models.

[18] Although not covered in this paper, the user could use some heuristics to set the K number of eigenvectors.

References

Avellaneda, M. (2019) Hierarchical PCA and Applications to Portfolio Management. *Working Paper – NYU Courant.*

Avellaneda, M. and Lee, J.H. (2010) Statistical arbitrage in the U.S. equity market. *Quantitative Finance* 10(7): 761-782.

Boyle, P. (2014) Positive Weights on the Efficient Frontier. *North American Actuarial Journal* 18(4): 462-477.

Cizeau, P., Potters, M. and Bouchaud, J-P. (2000) Correlation Structure of Extreme Stock Returns. *Quantitative Finance* 1(2): 217-222.

Connor, G. and Korajczyk, R.A. (1988) Risk and Return in an Equilibrium APT: Application of a New Test Methodology. *Journal of Financial Economics* 21(2): 255-289.

Cont, R. and Da Fonseca, J. (2002) Dynamics of implied volatility surfaces. *Quantitative Finance* 2(1): 45-60.

Fabozzi, F.J., Focardi, S.M. and Kolm, P.N. (2010) Quantitative Equity Investing: Techniques and Strategies. *Hoboken, NJ: John Wiley & Sons, Inc.*

Fabozzi, F.J., Kolm, P.N. and Focardi, S.M. (2002) Robust Financial Modeling of the Equity Market: From CAPM to Cointegration. *Hoboken, NJ: John Wiley & Sons, Inc.*

Fabozzi, F.J., Kolm, P.N., Pachamanova, D.A. and Focardi, S.M. (2007) Robust Portfolio Optimization and Management. *Hoboken, NJ: John Wiley & Sons, Inc.*

Fama, E.F. and French, K.R. (1992) The Cross-Section of Expected Stock Returns. *Journal of Finance* 47(2): 427-465.

Fama, E.F. and French, K.R. (1993) Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics* 33(1): 3-56.

Fama, E.F. and French, K.R. (2015) A five-factor asset pricing model. *Journal of Financial Economics* 116(1): 1-22.

Golan, A., Judge, G.G. and Miller, D. (1996) Maximum Entropy Econometrics: Robust Estimation with Limited Data. *Wiley.*

Jollife, I.T. (2002) Principal Compoment Analysis. 2nd edition, *Springer, New York.*

Kakushadze, Z. (2015) Heterotic Risk Models. *Wilmott Magazine* 2015(80): 40-55.

Kakushadze, Z. and Yu, W. (2017) Statistical Risk Models. *Journal of Investment Strategies* 6(2): 1-40.

Laloux, L., Cizeau, P., Potters, M. and Boucheaud, J.-P. (2000) Random matrix Theory and Financial Correlations. *Mathematical Methods in Applied Sciences* 1(2): 217-222.

Ledoit, O. and Wolf, M. (2014) Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis* 139: 360-384.

Litterman, R.B. and Scheinkman, J. (1991) Common Factors Affecting Bond Returns. *Journal of Fixed Income* 1(1): 54-61.

Marčenko, V.A. and Pastur, L.A. (1967) Distribution of Eigenvalues for Some Sets of Random Matrices. *Mathematics of the USSR-Sbornik* 1(4): 457-483.

Meucci, A. (2010) Factors on Demand: Building a Platform for Portfolio Managers, Risk Managers and Traders. *Risk* 23(7): 84-89.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L., Guhr, T. and Stanley, H. (2002) Random matrix approach to cross correlations in financial data. *Physical Review E* 65(6).

Ross S. (1976) The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory* 13(3), 341-360.

Roy, O. and Vetterli, M. (2007) The effective rank: A measure of effective dimensionality. In: *Proceedings – EUSIPCO 2007, 15th European Signal Processing Conference.* Poznań, Poland (September 3-7), pp. 606-610.

Sharpe W.F. (1964) Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *Journal of Finance* 19(3), 425-442.

Torun, M. U., Akansu, A. N. and Avellaneda, M. (2011) Portfolio risk in multiple frequencies. *IEEE Signal Processing Magazine* 28(5), 61-71.

Appendix

Europe	Number	Emerging	Number
United Kingdom	147	China	288
France	89	Korea	114
Germany	77	Taiwan	87
Switzerland	52	India	80
Sweden	44	Brazil	55
Italy	33	Hong Kong	50
Spain	26	South Africa	44
Netherlands	25	Malasya	43
Denmark	21	Thailand	37
Belgium	18	Saudi Arabia	31
Total	**532**		**829**

Table 10: Number of companies by countries in European and Emerging markets.

Cluster	Number	Percentage (%)
Cluster 1	24	0.91%
Cluster 2	9	0.34%
Cluster 3	90	3.41%
Cluster 4	18	0.68%
Cluster 5	368	13.94%
Cluster 6	194	7.35%
Cluster 7	1	0.04%
Cluster 8	20	0.76%
Cluster 9	406	15.38%
Cluster 10	352	13.33%
Cluster 11	1	0.04%
Cluster 12	388	14.70%
Cluster 13	207	7.84%
Cluster 14	82	3.11%
Cluster 15	127	4.81%
Cluster 16	353	13.37%

Table 11: Number of assets that belong to each cluster. Some clusters, such as *Cluster 7* and *Cluster 11*, only have one component.

Cluster	Main Sectors	Sectors (%)	Main Countries	Countries (%)
Cluster 1	Materials; Health Care; Consumer Discretionary	50.00%	United Kingdom; Germany; Italy	62.5%
Cluster 2	Materials; Financials; Industrials	77.78%	China; Germany; Italy	77.78%
Cluster 3	Materials; Financials; Consumer Staples	48.31%	China; India; Korea	34.83%
Cluster 4	Financials; Health Care; Consumer Discretionary	61.11%	China; Brazil; Korea	72.22%
Cluster 5	Information Technology; Industrials; Consumer Discretionary	50.81%	United States; United Kingdom; Ireland	97.55%
Cluster 6	Industrials; Financials; Consumer Staples	42.48%	United Kingdom; Germany; France	37.15%
Cluster 8	Materials; Financials; Information Technology	55.00%	United Kingdom; Germany; France	50.00%
Cluster 9	Materials; Financials; Consumer Discretionary	49.62%	China; Taiwan; Korea	39.00%
Cluster 10	Materials; Financials; Information Technology	47.57%	China; India; Korea	43.33%
Cluster 12	Materials; Industrials; Information Technology	53.35%	China; Taiwan; Hong Kong	96.91%
Cluster 13	Consumer Staples; Real Estate; Utilities	49.10%	United States; Brazil; China	80.01%
Cluster 14	Consumer Discretionary; Communication; Financials	49.41%	China; Korea; Thailand	75.60%
Cluster 15	Consumer Discretionary; Information Technology; Industrials	49.20%	China; Indonesia; Malaysia	96.03%

Cluster 16	Financials; Industrials; Consumer Discretionary	55.24%	United Kingdom; France; Germany	52.11%

Table 12: Main countries and industries belonging to each cluster.