REVISTA MEXICANA DE ECONOMÍA Y FINANZAS
Nueva Época
REMEF
(THE MEXICAN JOURNAL OF ECONOMICS AND FINANCE)

Revista Mexicana de Economía y Finanzas
Nueva Época
Volumen 15 Número 1, Enero - Marzo 2020, pp. 1-16
DOI: https://doi.org/10.21919/remef.v15i1.446
(Recibido: 2/julio/2019, aceptado: 8/octubre/2019)

imef
fundación de investigación

Hierarchical PCA and Applications to Portfolio Management

Marco Avellaneda[1]

Courant Institute of Mathematical Sciences, NYU, USA

Abstract

It is widely known that the common risk-factors derived from PCA beyond the first eigenportfolio are generally difficult to interpret and thus to use in practical portfolio management. We explore an alternative approach (HPCA) which makes strong use of the partition of the market into sectors. We show that this approach leads to no loss of information with respect to PCA in the case of equities (constituents of the S&P 500) and also that the associated common factors admit simple interpretations. The model can also be used in markets in which the sectors have asynchronous price information, such as single-name credit default swaps, generalizing the works of Cont and Kan (2011) and Ivanov (2016).
JEL Classification: C02, C65, G24
Keywords: returns, blocks, PCA, HPCA, portfolio

PCA jerárquico y aplicaciones a la gestión de cartera

Resumen

Es ampliamente conocido que los factores de riesgo comunes derivados del PCA más allá de la primera eigenportafolio son generalmente difíciles de interpretar y, por lo tanto, de utilizar en la gestión práctica de la cartera. Exploramos un enfoque alternativo (HPCA) que hace un fuerte uso de la partición del mercado en sectores. Demostramos que este enfoque no conduce a la pérdida de información con respecto al PCA en el caso de la renta variable (constituidos por el S&P 500) y también que los factores comunes asociados admiten interpretaciones simples. El modelo también se puede utilizar en mercados en los que los sectores tienen información asincrónica de precios, como single-name swaps de incumplimiento de crédito, generalizando las obras de Cont y Kan (2011) e Ivanov (2016).
Clasificación JEL: C02, C65, G24
Palabras clave: rendimiento, bloques, PCA, HPCA, portafolio

[1]251 Mercer Street, New York, NY 10012
*No declared funding source for research development

1 Introduction

Principal Components Analysis (PCA) and random matrix theory (RMT) have become widespread tools for data analysis. PCA (Joliffe (2002) [6]) provides a mathematical and objective approach to extract economic information from the correlation matrices of asset returns. In this approach, the analyst extracts common risk factors from the eigenvectors and eigenvalues of the correlation matrix.

The first eigenvector of the correlation of stock returns corresponds to the solution of the variational problem

$$V^{(1)} = argmax \left\{ V^t R V; ||V|| = 1 \right\}. \tag{1}$$

Here, R is the correlation matrix of daily returns and $||.||$ is the Euclidean norm in R^n, n being the total number of assets. Equation 1 shows that the principal eigenvector is represents the direction (line) which "captures the most variance" as described by the correlation matrix. The first eigenvector satisfies

$$RV^{(1)} = \lambda^{(1)} V^{(1)} \tag{2}$$

PCA also finds recursively additional (orthogonal) directions beyond $V^{(1)}$ which capture the most variance. The other eigenvectors and eigenvalues are computed in the same way as Eq. (1)) with the maximization to the sub-space orthogonal to the space spanned by the ones computed previously, i.e,

$$V^{(k)} = argmax \left\{ V^t R V; ||V|| = 1, V^{(k)t} V^{(l)} = 0,\ 1 \leq l < k \right\}. \tag{3}$$

The eigenvalues satisfy $\lambda^{(1)} > \lambda^{(2)} \geq ... \geq \lambda^{(n)}$. Assume that the data corresponds to the daily returns of a group of stocks. The Karhunen-Loeve representation of the standardized returns is

$$X_j = \sum_{k=1}^{n} \sqrt{\lambda^{(k)}} V_j^{(k)} F^{(k)} \tag{4}$$

where

$$F^{(k)} = \frac{1}{\sqrt{\lambda^{(k)}}} \sum_{i=1}^{n} V_i^{(k)} X_i. \tag{5}$$

By construction, $F^{(k)}$ are uncorrelated and have variance 1. Since these random variables are linear combinations of the daily standardized returns of the assets, we call them (standardized) "eigenportfolio (EP) returns", with the caveat that the actual portfolio "weights" are obtained by dividing each entry of the eigenvector by the volatility of the asset (Avellaneda and Lee 2008, 2010) [1].[2]

PCA is a framework for learning about the common factors which affect the returns of a given group of assets. The first eigenportfolio, associated with the r.v. $F^{(1)}$, is a common risk factor which explains the maximum variability. We can write a one-factor model for each asset, namely

$$X_j = \beta_j F^{(1)} + \epsilon_j \tag{6}$$

where β_j is the regression coefficient of the standardized return on the first EP. The "residuals" ϵ_j in equation 6 are uncorrelated with $F^{(1)}$, which is nice. However, they are generally correlated for different stocks.

[2]We consider correlations instead of covariances because it mathematically simpler to work in dimensionless units, i.e. to reduce to the case when all the volatilities are equal to one.

The regression coefficients satisfy

$$\beta_j = \sqrt{\lambda^{(1)}}\, V_j^{(1)}, \quad j = 1, ..., n. \tag{7}$$

In the case of economic data, which is noisy, the consensus is to disregard EPs which correspond to low eigenvalues. In a celebrated paper, Laloux *et al* (2000) [7] proposed to use random matrix theory (RMT) to establish a cutoff in the number of EPs use to model the standardized returns, namely

$$X_j = \sum_{k=1}^{m} \beta_j^{(k)}\, F^{(k)} \; + \epsilon_j \tag{8}$$

where $\beta_j^{(k)}$ are "factor loadings" and (with a slight abuse of notation) ϵ_j are residuals obtained after "defactoring" relatively to the m eigenportfolios. The number m is a cutoff which is to be determined from the context.

According to [7], the eigenvalues of a pure noise matrix follow the Marcenko-Pastur distribution and have a spectrum which, for large matrices,is asymptotically bounded from above by $\lambda^{+,MP} = (1+\sqrt{n/T})^2$, where T is the number of observations. Asymptotics should hold in the limit $n/T \to \gamma$ (a constant) as n and T both tend to infinity. The way to use RMT to calculate the cutoff is to construct the correlation matrix $R_{i,j}^{(m)} = Corr(\epsilon_i, \epsilon_j)$ for m large enough and verify that its top eigenvalue is of the order of $\lambda^{+,MP}$. One can also compare the empirical distribution of eigenvalues with the Marcenko-Pastur probability distribution.

PCA aided by RMT is an elegant approach to analyzing correlation matrices of financial data and can also be applied to may areas of science. The main strength of the method is that it can detect common risk factors based on a matrix of asset returns, without any additional information. In other works, PCA "lets the data speak for itself". Generally speaking, PCA explains the most variability with the smallest number of factors. Most studies tend to justify the PCA approach by recognizing that it produces some factors which have *ex-post* economic interpretations, such as equating $EP^{(1)}$ with the Sharpe Market Portfolio (Boyle 2017) [2], or attempt to interpret higher-order EPs in terms of industry sectors [1]. In the case of fixed-income, the EPs are often identified with "parallel shifts", or with long-term vs short-term oscillations of the yield curve (Litterman and Scheinkman, 1991) [8].

2 The identification problem

One of the frequent criticisms of PCA in Finance is that the common risk factors generated by higher-order eigenportfolios – aside from the first eigenportfolio – are difficult to interpret and appear to be unstable across time. We call this the *identification problem*. Because of it, many portfolio managers favor traditional factor models such as Barra; see Shkolnik et. al. (2016) [9] for alternative approaches to model financial correlations.

The identification problem in PCA reflects the uncertainty, or unreliability, of cross-asset correlations. From a practical point of view, as the size of trading universe increases, the correlations of assets which are not economically related (a tech stock with an energy stock, or with a foreign stock) are difficult to quantify and may be noisy. This could be due to several reasons: the lack of "explanation" for the relation between the stocks, or perhaps that their prices are not sampled simultaneously (e.g. if they are end-of-day prices in different time-zones) or that the number of observations is not large compared to the number of assets considered. For example, empirical correlations of price changes of out-of-the money options with different underlying assets may not be as reliable or significant as the data would suggest.

To mitigate the identification problem, we should seek a factor model which can recognize the economic

nature or function of the asset as well as the statistical properties of returns. This lead us to the model described hereafter.

3 Hierarchical PCA

The hierarchical PCA (HPCA) applies to markets which can be partitioned into several sectors or asset-classes. Consider first an abstract market, in which the empirical data matrix of asset returns, with dimensions $T \times n$, can be partitioned into "blocks of columns" labeled $k = 1, 2, ..., b$. These blocks have dimensions $T \times n_k$ with $k = 1, 2, ..., b$. Each block represents data sampled from a sector. For simplicity, we assume that the indices of the securities are organized so that blocks which are adjacent to one another in the matrix and do not overlap. We have a few concrete situations in mind:

- The blocks represent data of industry sectors for equities in the same economy (e.g. sectors associated with the 500 or so stocks in the S&P 500 index). In this case, the columns of a block correspond to the historical standardized returns of the stocks in the sector observed over T consecutive dates.

- Each block represents a stock or index and all of the derivatives written on it. In this case, the columns in a block represent the returns of the stock and the changes of the implied volatilities of options with different strikes and tenors written on the stock (Dobi 2015 [4]).

- In the context of credit derivatives, the data represents changes in credit spreads for CDS. The blocks correspond to CDS referencing the same obligor (issuer) but with different tenors (Cont and Kan (2011) [3], Ivanov (2017) [5]).

Define the function $I(j) = k$ if asset j is in block k. According to Eq. (4) we can write, for each asset in the "big universe",

$$X_j = \beta_j \, F^{(1,I(j))} + \epsilon_j, \tag{9}$$

where β_j is the regression coefficient of the returns of asset j on the first factor of block $I(j)$ and ϵ_j is the residual.

We shall make the following assumption ("HPCA assumption"):

$$\boxed{\text{If } I(i) \neq I(j), \text{ then } Corr(\epsilon_i, \epsilon_j) = 0.} \tag{10}$$

The assumption states that residuals are uncorrelated if their assets belong to different sectors. Equation (9) defines the asset statistics within each block exactly, and the model is completed by specifying the joint statistics of the factors $F^{(1,k)}$, $k = 1, 2, ..., b$. The HPCA assumption says nothing new regarding intra-block correlations, which are set equal to the empirical correlations between asset returns within the same sector or block. Of course, the intra-block correlations could be further denoised using RMT if necessary ([4]).

Using the HPCA assumption Eq. (10), the proposed model has the modified correlation matrix for asset returns:

$$\begin{aligned} \tilde{R}_{ij} &= R_{ij} \text{ if } I(i) = I(j) \\ &= \beta_i \, \beta_j \, \overline{\rho}^{I(i)I(j)} \text{ if } I(i) \neq I(j) \end{aligned} \tag{11}$$

where $\overline{\rho}^{k,k'} = Corr(F^{(1,k)}, F^{(1,k')})$.

Proposition 1 *Eq. (11) corresponds to a symmetric non-negative matrix with* $\tilde{R}_{ii} = 1$ *for all* i*. In particular, it corresponds to the correlation matrix of a system of standardized random variables.*

Proof. To check non-negative definiteness, note that for all $\theta \in R^n$ we have

$$\theta^t \ \tilde{R}\theta = \sum_{k=1}^{b} \sum_{I(i)=I(j)=k} \theta_i\theta_j(R_{ij} - \beta_i\beta_j) + \sum_{k,k'=1}^{b} \Big(\sum_{I(i)=k} \theta_i\beta_i \Big) \Big(\sum_{I(j)=k'} \theta_j\beta_j \Big) \ \overline{\rho}^{k,k'}. \tag{12}$$

For any k, the matrix $R_{ij} - \beta_i\beta_j$ restricted to sector k is identical to the sector correlation, except for the fact that the eigenvalue corresponding to $V^{(1,k)}$ is set to zero. In particular, it is non-negative definite. Moreover, the matrix $\overline{\rho}^{k,k'}$ is also a correlation matrix, so it is non-negative definite. Since both summands are non-negative it follows that $\theta^t \ \tilde{R}\theta \geq 0$ for all $\theta \in R^n$.

A concrete implementation of the data model is achieved as follows: let $\psi_1, ..., \psi_b$ be Gaussian random variables with mean zero and covariance matrix $\overline{\rho}$, and let ζ_{ik} , $i : I(i) = b, \ k = 1, ..., b$ be i.i.d. standardized Gaussian random variables which are independent of the ψ's. The data model is

$$X_i = \beta_i \ \psi^{I(i)} \ + \sum_{\{j:j\geq 2, \ I(j)=I(i)\}} \gamma_{ij} \ \zeta_{j\, I(i)} \tag{13}$$

The random variables need not be necessarily Gaussian: they can be multivariate Student-t, or they can be transforms of arbitrary distributions connected by a Gaussian or t-Copula; see for instance [5].

The multivariate distribution associated with HPCA presents an alternative model to the classical PCA (Eq. (8)). It has a tree structure: in the equity example discussed below, the top vertex corresponds to the "market"; there are 11 branches corresponding to industry sectors, and each of the 11 vertices has branches corresponding to the stocks in each sector.

Hierarchical models with more than two layers arise naturally. For instance, HPCA can be used to model "world portfolios", in which the first layer consists of countries or regions, the second to industry sector indices in each country, and the third layer could describe the individual securities in each region/sector.

For another useful example, consider a stock market in which stocks belong to different industry sectors, and then, include columns associated with equity options returns. In this case, the tree has three layers because we can associate to each stock an additional sub-group: the block consisting of the returns of implied volatilities (on a constant delta/time-to-maturity grid) and the stock returns. Now the root corresponds to the full market, the first layer corresponds to industry sectors, the second layer corresponds to stocks and the third layer represents an individual name with all the associated option-implied volatilites.

A similar approach works for credit derivatives. In this case, the returns of the CDS with different tenors referencing each obligor constitute a block associated with an obligor. These blocks can be grouped by industry sectors or, alternatively, blocks could be generated according to membership in a credit index (CCX.IG, CDX.HY, CDX.HV), or both; [5].

In summary, if financial data can be grouped into blocks or sectors with clear economic interpretation, with multiple instruments associated with each block, we can generate a data model with tree-like structure from the HPCA assumption in Eq. (10). This approach combines information available for each asset (sector, sub-sector, reference obligor, option underlying asset) with the explanatory power of PCA. For simplicity, we will consider the analysis of a two-layer HPCA. Adding more layers is mathematically straightforward.

4 Spectral analysis

The HPCA assumption Eq. (10)gives rise to explicitly computable eigenvalues and eigenvectors for the matrix $\tilde{R}$ defined in Eq. (11).

Proposition 2.

1. *For each sector $k = 1, ..., b$, let $\lambda^{(1,k)} > \lambda^{(2,k)} \geq ... \geq \lambda^{(n_k,k)}$ denote the n_k eigenvectors of the sector correlation matrix, ordered from largest to smallest, and let $V^{(i,k)}$ be the corresponding eigenvectors. Define the n-dimensional vectors*

$$\begin{aligned} W_j^{(i,k)} &= V_j^{(i,k)} \quad if \;\; I(j) = k \\ &= 0 \quad\;\; if \;\; I(j) \neq k, \end{aligned} \tag{14}$$

which correspond to the embedding of the sector-level eigenvectors, $V^{(i,k)} \in R^{n_k}$, into the large space R^n. The vectors $W^{(i,k)}$, $i = 1, ..., n_k$, $k = 1, ..., b$ form an orthogonal basis of R^n.

2. *The subspace Ω of R^n generated by the vectors $W^{(1,k)}$, $k = 1, ..., b$, viz.*

$$\Omega = \Big\{ \sum_{k=1}^{b} \alpha_k W^{(1,k)} : (\alpha_1, ..., \alpha_b) \in R^b \Big\}, \tag{15}$$

is invariant under the action of $\tilde{R}$ viewed as an operator from R^n to R^n.

3. *Consider the $b \times b$ matrix*

$$M^{k,k'} := \sqrt{\lambda^{(1,k)}} \sqrt{\lambda^{(1,k')}} \overline{\rho}^{k,k'}. \tag{16}$$

Let $\mu^{(1)}, ..., \mu^{(b)}$ denote the eigenvalues of M, ranked in decreasing order, and let $(\alpha^{(k)} = (\alpha_1^{(k)}, ..., \alpha_b^{k)})$ $k = 1, ..., b$ represent the corresponding normalized eigenvectors (defined up to sign). The vectors

$$\tilde{W}^{(1,k)} = \sum_{p=1}^{b} \alpha_p^{(k)} \, W^{(l,p)} \tag{17}$$

are eigenvectors of $\tilde{R}$, with corresponding eigenvalues $\mu^{(k)}$, for $k = 1, ..., b$.

4. *For each sector k and each j, $2 \leq j \leq n_k$, the vector $W^{(j,k)}$ is an eigenvector of $\tilde{R}$, with eigenvalue $\lambda^{(j,k)}$.*

This proposition completely characterizes the eigenvalues and eigenvectors of the HPCA correlation matrix relating them to the eigenvalues and eigenvectors of sector PCAs.[3] Thus, the HPCA assumption eliminates the identification problem for common factors: "eigenportfolios" have concrete meanings attached to the information about the correlations of sectors. In the examples to follow, we shall compare HPCA with PCA and show that the former is an excellent substitute for the full empirical correlation matrices when we model multivariate financial data.

[3]The proof of Proposition 2 is straightforward: one just has to observe that $\beta_i = \sqrt{\alpha^{I(i)}} V_i^{(1,I(i))}$ and calculate explicitly the action of $\tilde{R}$ on each of the vectors $W^{(j,k)}$.

5 Application: S&P 500 constituents

We consider data for $n = 434$ equities which are constituents of the S&P500 index. The data ranges from February 22, 2012 to February 16, 2018. We consider the correlation matrix of standardized stock returns, and define the sectors as General Industry Classification groups (GICs), so $b = 11$; see Table 1.

Cuadro 1. GIC sectors and number of companies in each sector.

GIC (k)	Description	Number of companies(n_k)
1	Consumer Discretionary	73
2	Consumer Staples	56
3	Energy	27
4	Financials	59
5	Health Care	51
6	Industrials	57
7	Information Technology	58
8	Materials	23
9	Real Estate	27
10	Telecommunication Services	3
11	Utilities	28

5.1 Eigenvalues

We considered the full empirical correlation matrix[4] and the HPCA correlation matrix $\tilde{R}$ ("HPCA matrix"). The spectrum of the HPCA matrix is very similar than the one of the empirical correlation matrix R, with the difference that the latter eigenvalues at the top of the spectrum are slightly larger the eigenvalues of the HPCA matrix. This is due to the fact that PCA explains more variance with fewer common factors (see Figure (5.1)). On the other hand, the sum of eigenvalues is equal to $n = 434$ in both cases, which means that for high enough rank, the higher-order eigenvalues of HPCA are larger than those of PCA. The lowest eigenvalues of R are infinitesimal, and the latter matrix is degenerate. At the bottom of the spectrum (not shown here) the HPCA spectrum has much higher eigenvalues (separated from zero) than PCA, since they are bounded from below by the lowest eigenvalue from all the sectors. Thus, the HPCA matrix is better conditioned than the full empirical matrix.

[4]In the sequel we refer to the full empirical correlation matrix as the "PCA matrix", for short.

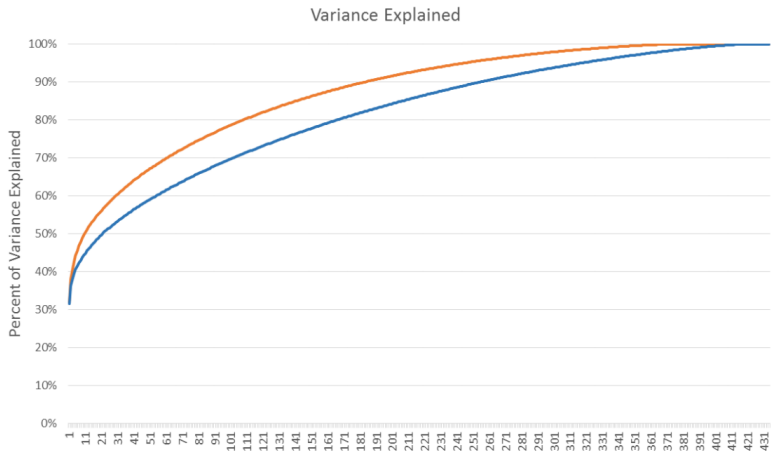

Figure 1. X=axis: rank (k) of the eigenvalues, sorted in decreasing order. Y-axis: sum of the first k eigenvalues divided by $n = 434$. The PCA curve rises faster than HPCA, due to the nature of the PCA algorithm.

Cuadro 2. Top 25 eigenvalues of PCA and HPCA, sorted in decreasing order.

PCA	HPCA	Eigenportfolio		PCA	HPCA	Eigenportfolio
138.87	137.19	Multi-sector		2.79	2.18	Industrials
26.84	20.70	Multi-sector		2.52	2.15	Consumer Disc.
11.88	8.18	Multi-sector		2.46	2.14	Healthcare
7.70	5.91	Multi-sector		2.36	2.09	Inf. Technology
6.87	4.93	Multi-sector		2.32	2.03	Multi-sector
5.75	3.69	Multi-sector		2.24	1.94	Technology
5.16	3.38	Consumer Disc.		2.20	1.93	Industrials
4.70	2.88	Multi-sector		2.18	1.92	Energy
3.90	2.80	Financials		2.13	1.80	Consumer Disc.
3.61	2.68	Multi-sector		2.06	1.59	Inf. Technology
3.48	2.67	Healthcare		2.01	1.57	Industrials
3.02	2.53	Cons. Cyclical		1.96	1.57	Healthcare
2.87	2.25	Healthcare				

The column "Eigenportfolio" gives an interpretation of the corresponding HPCA eigenportfolio. "Multi-sector" corresponds to a $\mu^{(k)}$-eigenvalue and eigenvector, which are combinations of the *first* eigenportfolios for each of the 11 sectors (space Ω). The other eigenvalues/eigenvectors correspond to higher-order eigenvalues/eigenvectors for individual GIC sectors. Notice that, after sorting, some of the GIC eigenportfolios are more important in terms of explaining variability than multi-sector portfolios.

5.2 Eigenvectors

We turn to empirical analysis of the eigenvectors of the HPCA and the empirical correlation matrices, *i.e.* to the issue of identification problem for PCA/HPCA. The first eigenvectors for HPCA and PCA are plotted in Figures (5.2) and (5.2). Since the first eigenvector of M has positive entries and the first eigenvectors of sector correlations also have positive entries due to the positive correlations of stocks ([1],[2] ; EV1 loadings are positive for both PCA and PCA. Figure (5.2) superimposes both eigenvectors. The ordering of the X-axis is alphabetical in each sector and sectors are grouped displayed in increasing order of GIC according to Table (5). The two eigenvectors are practically indistinguishable in the sense that their average difference is of order 1.0×10^{-5} and the standard deviation (centered RMS distance) is 5.3×10^{-3}. The RMS error is one

order of magnitude smaller than the average size of each entry in the eigenvectors which is approximately equal to 4.7×10^{-2}, in both cases.

This identifies the first eigenportfolio of the market as a "portfolio of first eigenportfolios" of different sectors (GICs). The difference in explanatory power between the two eigenvectors is the difference between the corresponding eigenvalues, divided by the number of stocks, namely $(138.87 - 137.19)/434 = 0.39\%$, which is negligible in this context. In particular, this suggests that using the first HPCA eigenportfolio as a proxy for the market portfolio gives rise to a better description of the market portfolio and an easier way to allocate to each stock. For instance, the first EV could be proxied by a capitalization-weighted sector ETF.[5].

For eigenvectors 2 through 5 Figures (5.2) through (5.2), we find that the PCA eigenvectors correspond to "noisy versions" of the corresponding HPCA eigenvectors. The latter are essentially long-short sector eigenportfolios. The discrepancy increases when we consider higher-order eigenvalues, beyond 5. Eigenvectors #6 aren't similar as shown in Figure (5.2). The PCA eigenvector contains both positive and negative signs within the Consumer Discretionary sector. Eigenvector 7 in HPCA is the first which is concentrated in a single sector, which is Consumer Discretionary (Fig. (5.2). The remaining eigenvectors up to rank 10 are displayed in Figures (5.2) to (5.2).

The main conclusions are: (a) most of the top eigenvalues and corresponding eigenvectors are related to the inter-sector correlation $\overline{\rho}$. This provides an interpretation for these eigenportfolios, or common risk factors, as "portfolios of long-only sector portfolios". (b) The remaining eigenvectors may be quite different. The HPCA defines the factors into "sector-sector" and "long-short intra-sector". PCA eigenvectors, in contrast, become increasingly difficult to interpret as simple sector-sector interactions or intra-sector interactions.

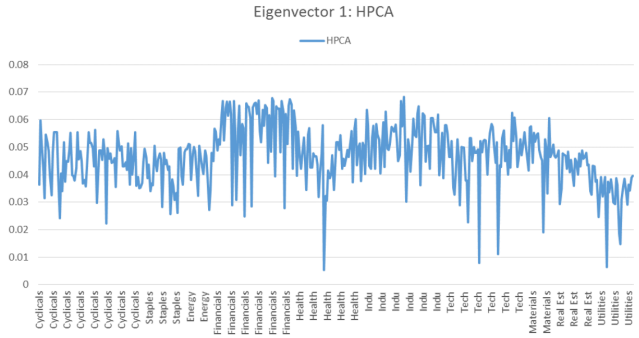

Figure 2. First eigenvector of HPCA. Variance explained= 30%.

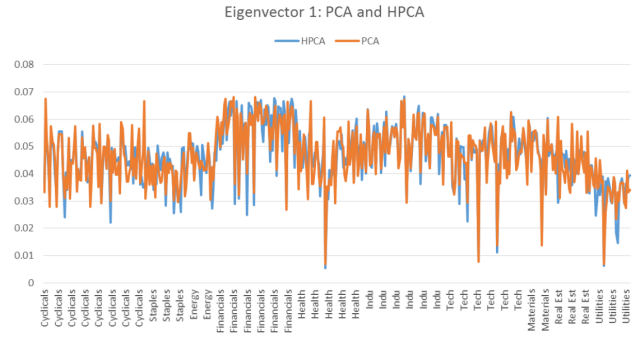

Figure 3. Comparison of the first eigenvectors of HPCA and PCA, which have approximately the same explanatory value. Their Euclidean distance (RMS error) is 5.5×10^{-3}, which is an order of magnitude smaller than the average entry size.

[5]A careful analysis of this idea, including out-of-sample tracking error analysis, will be done in a separate publication.

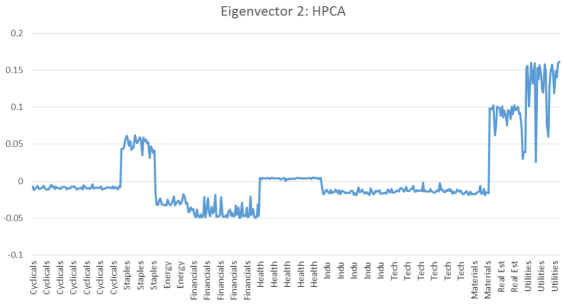

Figure 4. Second eigenvector of HPCA. The variance explained is 4.7% for HPCA and 6.1% for PCA.

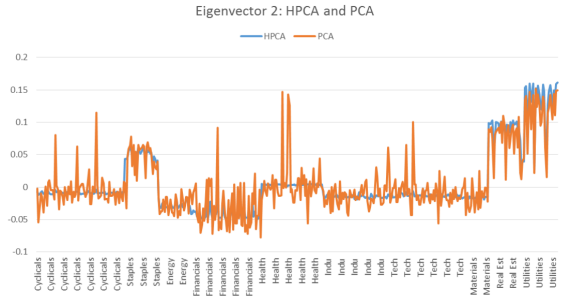

Figure 5. Comparison of the second eigenvectors. The PCA eigenvector is essentially a noisy version of the HPCA eigenvector.

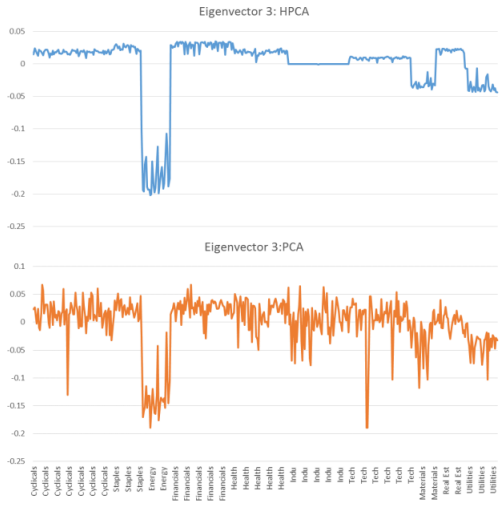

Figure 6. The third eigenvectors of HPCA: one can observe again that PCA EV3 is a noisy version of HPCA EV3.

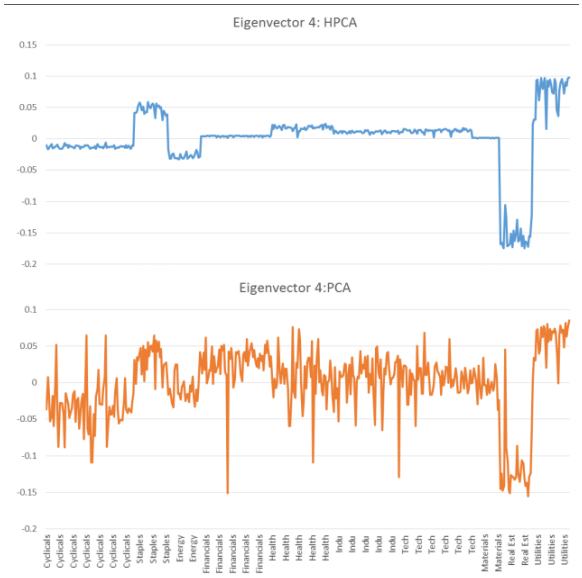

Figure 7. The fourth eigenvectors. Notice the similar loadings for sectors.

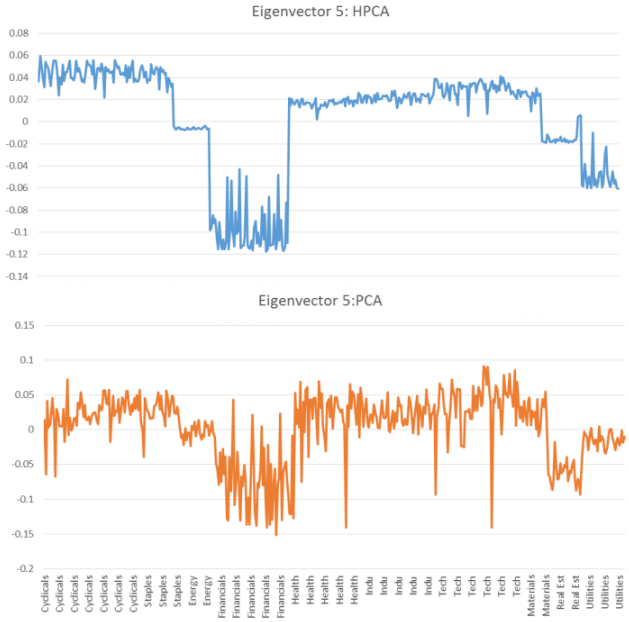

Figure 8. The fifth eigenvectors. Notice the similar loadings for sectors.

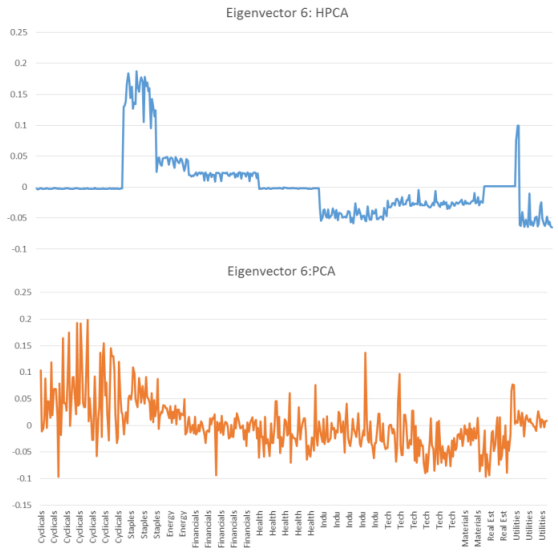

Figure 9. The sixth eigenvectors. In this case, PCA presents a different shape and is not "localized" on any sector. The leftmost part of the PCA eigenvector corresponds to Consumer Discretionary.

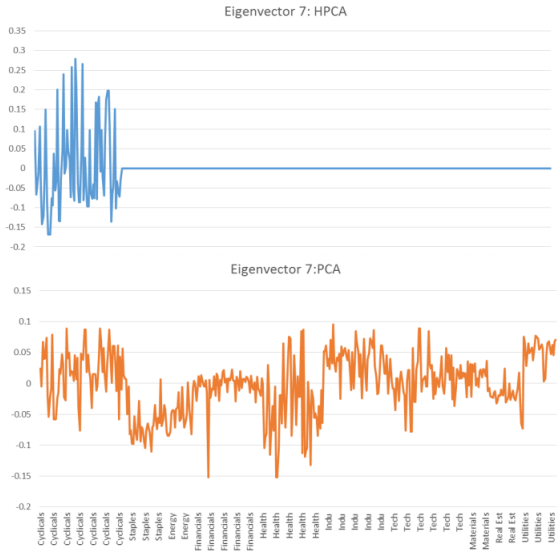

Figure 10. The seventh eigenvectors. The HPCA is essentially an eigenvector localized on the Consumer Discretionary sector (the second eigenvector of this sector. The PCA eigenvector is completely delocalized.

Figure 11. Eight eigenvectors.

Figure 12. Ninth eigenvectors. The HPCA eigenvector is localized in the Financials sector.

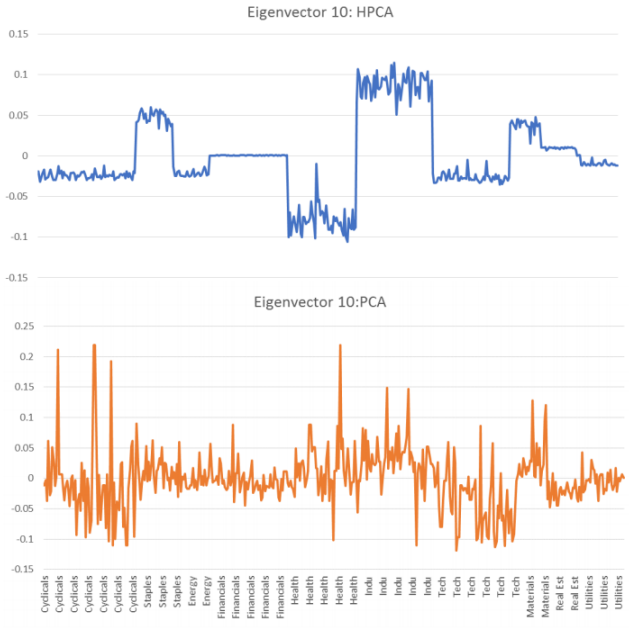

Figure 13. Tenth eigenvectors.

6 Analysis of residuals via RMT & Conclusion

To further evaluate the HPCA, we considered both models (HPCA,PCA) with a cutoff $m = 30$, and compared the multivariate statistics. We expect that after removing $m \approx 30$ eigenvectors, the correlations of the residuals (both intra- and inter- sector) should be small.

Empirically, the top eigenvectors of the correlation matrices of residuals are approximately 6.8 (HPCA) and 7.7 (PCA), which correspond to an approximate average correlation of $7.3/434 = 1.7\%$. We compared the histograms of the eigenvalues for the corresponding correlation matrices and found that they are very near each other. We also compared the histograms with a discretization of the Marcenko-Pastur distribution (mimicking the comparable histogram for the large-matrix limit), suggesting that the residuals behave like a random matrix in both models; see Fig. (6). The majority of the lines, in both cases, are below the Marcenko-Pastur cutoff $\lambda^+ = 2.36$, as postulated by RMT, and have comparable sizes to the MP distribution. There are, nevertheless, some lines above the MP threshold in both models (which are essentially equal), but they decreasing in magnitude as λ increases,and could perhaps be interpreted as finite-size fluctuations.

This calculation suggests that using the full empirical correlation matrix is not more informative than using the HPCA model, which uses only the sector correlation matrices, and in which intra-sector correlations are derived from the correlations of the EV1 for different sectors. Clearly, the HPCA provides a simpler description of common risk factors than PCA. The HPCA is therefore a viable alternative to PCA in the analysis of multivariate data in Finance, which should be of interest for asset-allocation and portfolio risk-management.

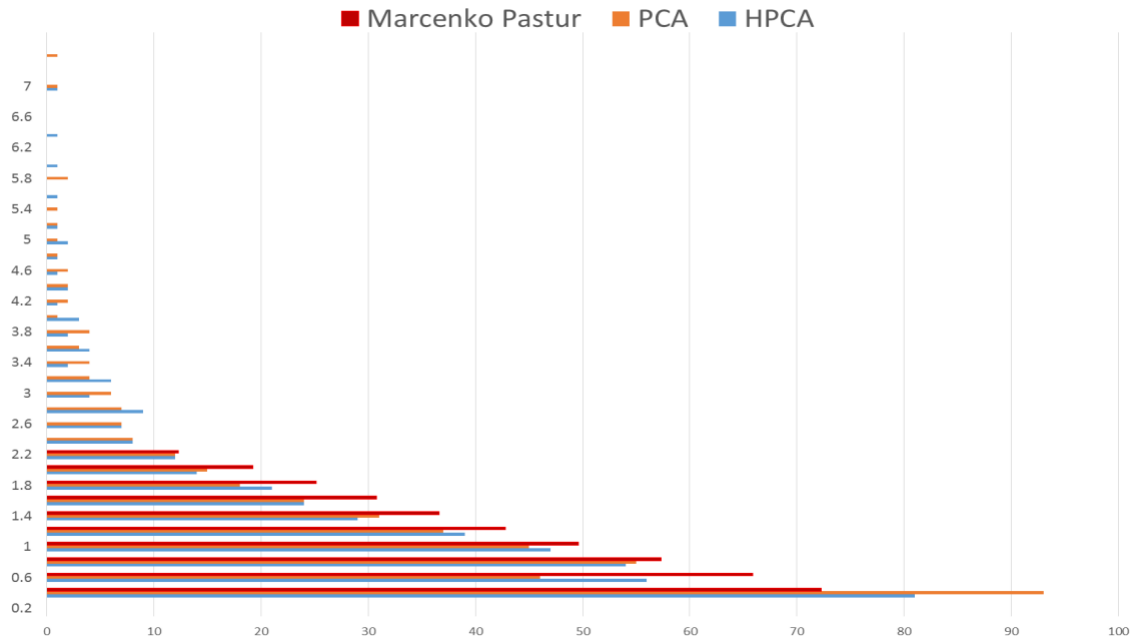

Figure 14. Histograms of residuals for HPCA (blue) and PCA (orange) after removing the first 3 eigenportfolios. For reference we display the "histogram" of the Marcenko-Pastur (MP) for corresponding to the same ratio of rows to columns (1508×434). The histograms of HPCA and PCA are comparable. Both are localized below the critical MP level of 2.36, with a smooth "leakage" as expected due to finite-size effects.

References

[1] Avellaneda, M. and Lee, JH, Statistical arbitrage in the US equities market, Quantitative Finance, 2010, vol. 10, issue 7, 761-782

[2] Boyle, Phelim P., Positive Weights on the Efficient Frontier (October 9, 2012). Available at SSRN: https://ssrn.com/abstract=2159445 or http://dx.doi.org/10.2139/ssrn.2159445

[3] Cont, R and Kan, Y.H., Statistical Modeling of Credit Default Swap Portfolios (April 1, 2011). Available at SSRN: https://ssrn.com/abstract=1771862 or http://dx.doi.org/10.2139/ssrn.1771862

[4] Dobi, Doris, Modeling Volatility Risk in Equity Options A Cross-Sectional Approach, Scholars' Press, (June 2018), NYU Ph.D. Dissertation, August 2014.

[5] Ivanov, S, Initial margin estimations for credit default swap portfolios, RISK Journal of Financial Market Infrastructures, July 2017

[6] Jollife, I.T., Principal Compoment Analysis, 2nd edition, Springer, New York, 2002.

[7] Laloux, L., Cizeau, P., Potters, M. and Boucheaud, J.-P., Random matrix Theory and Financial Correlations, Mathematical Methods in Applied Sciences, 2000.

[8] Litterman, R., and Scheinkman, J., Common factors affecting bond returns, The Journal of Fixed Income, 1991

[9] Shkolnik, A.D., Goldberg L., Bohn, J.R., Identifying Broad and Narrow Financial Risk Factors with Convex Optimization (August 20, 2016). Available at SSRN: https://ssrn.com/abstract=2800237 or http://dx.doi.org/10.2139/ssrn.2800237