UDC: 658.147; 339.72
JEL Classification: C10, E44, G15, O57
Keywords: principal components analysis; returns; interaction

World Equity Markets: A New Approach for Segmentation

José Dias CURTO – José Castro PINTO – João Eduardo FERNANDES*

1. Introduction

The continuous process of economic integration among world economies has a positive effect on the relationships of the various financial markets which are characterized by a high degree of returns correlation. The main interest of this paper, given the increasing globalization of equity markets, is to empirically demonstrate that, *using Principal Components Analysis (hereinafter referred to as "PCA")*, the returns of some markets are particularly associated with groups of nations and, simultaneously, try to explain the returns of each one based on the returns of the associated groups. *To the best of our knowledge, and despite the large number of PCA applications in finance, no empirical studies exist where this multivariate statistical technique is applied to segment world equity markets into related geographic areas / integrated economies. We conclude that PCA is very useful for this purpose.*

Therefore, the focus of this study is on the relationships among the returns of the most important national equity indices (25 countries in the sample) from the six major continents (Europe, North and South America, Africa, Asia and Australia) for the period from January 1995 to August 2003. To understand the interdependencies among the stock indices, and according to the expected level of returns correlation, we need not analyze all of the 25 indices individually but only on a reduced number of dimensions extracted from PCA. The application of this statistical analysis allows us not only to verify some linear associations between equity markets (which can also be achieved by correlation analysis), but also to determine the magnitude of the impact of the other equity markets on the returns of a particular one.

PCA is a multivariate statistical technique particularly suitable for analyzing the patterns of complex and multidimensional relationships that transforms a large number of related observable variables into a smaller set of new non-directly-observable composite dimensions (named by components) which can be used to represent relationships among those variables. New dimensions can be estimated as linear combinations of the entire set of observed variables and the correlations among the variables can be attributed to such shared components.

Although it is possible that all the variables contribute to the components,

* All authors: ISCTE Business School (Lisbon, Department of Quantitative Methods). Corresponding author: dias.curto@iscte.pt.

it is important that only a subset of variables characterizes each component, as indicated by their large coefficients in the linear combination. The identification of such underlying components that replace the original set of variables, greatly simplifies the description and the understanding of complex phenomena such as the interaction of world stock markets.

As the reduction of dimensionality is one of the objectives of PCA, the parsimony is also important for this statistical technique. It is important to explain the relationships among sets of observed variables using just a few components with a minimum loss of information in the original set. For example, if the number of components equals the number of variables, no simplification or summarization occurs and PCA has no value added. Components should also be meaningful, i.e. a good PCA solution should be both simple and interpretable.

PCA is particularly useful for investors' decision making. First, it is simpler to analyze a reduced number of dimensions instead of all 25 international stock indices. PCA meets this objective providing a reduced number of principal components that represent the wide set of international stock markets.

Second, if risk avoidance is very important for investors, the international diversification decisions can be thoroughly simplified using PCA. As the covariance and the correlation matrices are n-dimensional spaces, where n is the number of markets, it is very difficult for investors to have very strong convictions about each of the 300 covariances/correlations between all of the stock-market indices. They must somehow simplify their beliefs about the covariation of the 25 indices and PCA summarizes the covariances/correlations matrices in the largest few principal components. The compression of investors' beliefs into a much smaller number of dimensions is more likely to approximate their evaluation of a portfolio's risk. Moreover, the significant reduction in computation time makes PCA a very useful tool for risk management in large portfolios (Alexander, 2001).

In conclusion, the quest for the simplification of a complex reality led us to consider the application of PCA in order to find equity-market segments which have significant influence on the return of a given equity market. PCA, as we will demonstrate in the next section, has been commonly used in some empirical studies in the financial field. However, as far as we know, it has never been used for this purpose.

The outline of the paper is as follows. In the next section we discuss some applications of PCA in empirical finance. Section 3 provides an empirical application including some theoretical discussion about PCA and the final section summarizes our concluding remarks.

2. Data Reduction Methods in Finance

Data reduction methods, namely PCA, have been widely used in empirical finance. The main application fields are stock prices and stock returns, options implied volatilities, futures maturities, interest rates, exchange rates and conditional volatility.

Feeney and Hester (1964) apply PCA to both stock prices and rates of return to construct three alternative indices to a widely quoted stock-market index, the Dow Jones Industrial Average (DJIA). These alternative indices are linear combinations of prices and returns (trending and not trending adjusted) of the 30 DJIA stocks. As each one of the principal components resulting from PCA is a linear combination of prices or returns, the three indices proposed by Feeney and Hester (1964) are estimated by extracting characteristic roots from the stock prices and returns covariance matrices or more familiarly by the method of principal components. As there is little point in investors constructing an index if the market information is represented by all 30 extracted characteristic roots (which is exactly the same number of stocks) the authors only use the information of the vectors associated with the largest two characteristic roots to construct each one of the three indices.

PCA also supports some kinds of multivariate GARCH (Generalized Autoregressive Conditional Heteroskedasticity) models. As financial volatilities move together over time across assets and markets, it is widely accepted that returns volatility is driven by a small number of common dimensions. Factor ARCH models (Engle – Ng – Rothschild, 1990) are based on PCA to extract the principal components that represent the common variation on returns volatility.

The ISMA Centre (where Carol Alexander is Chair in Risk Management) has also contributed to popularizing the PCA data reduction method as a general procedure in finance. Despite the large number of papers where PCA is applied resulting from the research this centre, we will concentrate on the applications of chapter six of Alexander (2001). The first one relates to US and UK government and corporate bond yields where two principal components are used to represent a system of 20 yields, explaining nearly 96 % of the total variation. Thus, PCA allows a substantial reduction in dimensionality: from 20 yields to two principal components. PCA is also applied to the term structures of future prices and gives and important contribution to modeling volatility smiles and skews.

PCA is also proposed by Loretan (1997) as an easy-to-implement method to reduce the effective dimensionality of stress scenarios for market risk in financial instruments such as exchange rates, stock prices and interest rates from nine different countries: Belgium, Canada, France, Germany, Japan, the Netherlands, Switzerland, the United Kingdom and the United States. He found that stock prices and exchange-rate returns series are more highly correlated than short-term interest rates. This suggests that dimensiona-lity reduction may apply for certain groups of series, but not for others.

Lam and Ang (2006) applied PCA to analyze the relationship between globalization and stock-market returns. They found that global factors offer four times more explanatory power than domestic factors for developed-market stock returns. On the other hand, domestic factors are as important as global ones for the emerging economies. They apply PCA to summarize the information that is extracted from the large set of macroeconomic variables included in the study. The principal components are used as explanatory variables in the regression model where the dependent variable is the stock-market returns of both developed and emerging economies.

TABLE 1 World Equity Indices

North/Latin America		Europe / Africa / Middle East		Asia / Pacific	
Index	**Country**	**Index**	**Country**	**Index**	**Country**
DJIA	USA	FTSE100	UK	NIKKEY225	Japan
SPTSX	Canada	CAC40	France	HSI	Hong-Kong
MEXBOL	México	DAX	Germany	AS51	Australia
IBOVESPA	Brazil	IBEX35	Spain	SESALL	Singapore
MERVAL	Argentina	MIB30	Italy	SENSEX	India
IPSA	Chile	PSI20	Portugal	KOSPI200	South Korea
IBVC	Venezuela	WIG	Poland		
IGBVL	Peru	BUX	Hungary		
		XU100	Turkey		
		EFGIEFG	Egypt		
		SASEIDX	Saudi Arabia		

As the dominant principal components are orthogonal to each other, PCA minimizes the multicollinearity problem in the regression model.

According to this brief exposition, and to the best of our knowledge, no empirical studies exist where PCA is applied to segment world equity markets into related geographic areas/integrated economies. Therefore, the stock-market segmentation is the most important contribution of this paper.

3. Data, Methodology and Empirical Results

The data consist of weekly end-of-session quotes for 25 series of the most important general national equity indices. The period analyzed ranges from January 1995 to August 2003 (the data were extracted from the Bloomberg database). *Table 1* presents the names, regions and country names for each index.

As the data input to PCA must be stationary and the prices are generally non-stationary, they have to be transformed, commonly into returns, before PCA is applied (Feeney – Hester, 1964), (Alexander, 2001). The returns also need to be standardized with zero mean and unit standard deviation before the application of this statistical technique. Otherwise the first principal component will be dominated by the input variable with the greatest volatility.

The continuously compounded percentage rates of return (not adjusted for dividends as the volatile component of a stock's return is generally attributable to stock-price appreciation and depreciation) are calculated by taking the first differences of the logarithm of series:

$$r_{jt} = 100 \cdot [\ln(P_{jt}) - \ln(P_{jt-1})] \tag{1}$$

where P_{jt} is the week closing value for the stock index j at time t.

The sample of returns includes 450 observations after an initial observation is lost due to the differencing process. *Table 2* summarizes the basic statistical properties of returns. For most of the world stock indices in the sample, the returns' empirical distributions appear to be somewhat

TABLE 2 Statistical Properties of Returns

Countries	Mean	Median	Maxi-mum	Mini-mum	Std Dev.	Skew-ness	Kurtosis	JB	*P*-value
Argentina	0.100	0.284	23.767	–19.588	5.449	0.032	4.837	63.320	0.000
Australia	0.125	0.187	4.999	–5.260	1.668	–0.131	3.208	2.097	0.350
Brazil	0.298	0.499	21.747	–25.057	5.388	–0.542	5.280	119.539	0.000
Canada	0.130	0.265	9.311	–11.760	2.391	–0.590	6.157	213.002	0.000
Egypt	0.032	–0.123	26.610	–15.939	3.647	0.909	10.804	1203.971	0.000
Chile	0.073	0.128	9.958	–11.305	3.001	–0.230	4.588	51.278	0.000
France	0.126	0.131	11.034	–12.126	3.127	–0.090	3.630	8.039	0.018
Germany	0.122	0.320	12.887	–14.079	3.470	–0.276	4.515	48.782	0.000
Hong–Kong	0.075	0.100	13.917	–19.921	3.851	–0.424	5.814	161.929	0.000
Hungary	0.397	0.435	14.736	–33.016	4.395	–1.068	11.581	1465.952	0.000
India	0.020	0.054	12.079	–13.353	3.661	0.007	4.370	35.203	0.000
Italy	0.123	0.111	19.297	–13.921	3.339	0.207	6.544	238.721	0.000
Japan	–0.142	–0.120	11.047	–11.292	3.050	0.027	3.857	13.821	0.001
Mexico	0.267	0.348	17.503	–17.716	4.059	–0.131	5.163	89.014	0.000
Peru	0.078	–0.083	17.248	–11.206	2.974	0.592	8.090	512.049	0.000
Portugal	0.082	0.107	15.565	–16.564	2.795	–0.336	8.547	585.485	0.000
Poland	0.224	0.057	13.904	–19.244	4.118	–0.169	4.828	64.755	0.000
Singapore	–0.041	–0.207	12.119	–18.689	3.051	–0.353	7.054	317.452	0.000
Saudi Arabia	0.259	0.165	9.755	–6.254	1.947	0.464	5.980	182.691	0.000
South Korea	–0.025	–0.211	15.246	–17.014	5.159	–0.018	3.683	8.759	0.013
Spain	0.191	0.310	13.586	–11.633	3.075	–0.126	4.571	47.495	0.000
Turkey	0.844	0.821	25.781	–30.367	7.060	–0.137	4.790	61.476	0.000
UK	0.071	0.267	10.069	–8.864	2.341	–0.183	4.328	35.593	0.000
USA	0.196	0.374	8.090	–15.385	2.485	–0.731	6.620	285.830	0.000
Venezuela	0.538	0.190	26.598	–24.773	4.811	0.659	8.334	566.097	0.000

asymmetric, as reflected by the negative and the positive estimates of skewness. Except for Australia, all the series returns also have heavy tails and show strong departure from normality (skewness and kurtosis coefficients are all statistically different from those of the standard normal distribution which are 0 and 3, respectively). The Jarque-Bera (JB) test also rejects the null hypothesis of normality at the 5% level of significance (except for Australia).

The excess of kurtosis and the non-normality are stylized facts of financial returns (Mandelbrot, 1963), (Fama, 1965).

3.1 The Appropriateness of PCA

The appropriateness of PCA is commonly evaluated in terms of the variables measurement scale, the sample size and the correlations among observed variables.

The variables for PCA are generally assumed to be of metric (quantitative) measurement, including interval and ratio scales. These types of scale provide the highest level of measurement precision, permitting nearly all the mathematical operations to be performed.

Regarding the sample size, the general agreement states that the number of observations must exceed the number of variables. However, there is no consensus about the number of observations to be included in the sample. As a general rule it is desirable that the minimum of observations must be five times the number of variables to be analyzed, and a more acceptable range would be a ten-to-one ratio. Consequently, if the main purpose of a study is to find out what components underlie a group of variables, it is essential that the sample should be sufficiently large to enable this to be done reliably. As we have 450 observations this requirement is clearly fulfilled.

As mentioned before, one goal of PCA is to identify a small number of components that can be used to represent linear relationships among sets of observed variables. Therefore, one of the basic assumptions of PCA is that variables which share common components must be strongly correlated and variables are highly collinear when there are only a few important sources of information in the data that are common to many variables.

There are some indicators traditionally used to draw conclusions about the appropriateness of PCA according to the strength of the linear relationship among observed variables.

First, the strength of linear relationships can be represented by the correlation coefficient between pairs of variables. If correlations are small, it is unlikely that they share common components and obviously PCA is not appropriate: if visual inspection reveals no substantial number of correlations greater than 0.3, PCA probably is not appropriate (Hair et al., 1998). The correlation coefficient is computed as:

$$r_{X_i X_j} = \frac{s_{ij}}{s_i s_j} \tag{2}$$

where s_{ij} is the covariance between ith and jth variables; s_i and s_j are the standard deviations of the ith and jth variables, respectively.

In our study, the correlation matrix is presented in *Table 3*. Except for three countries (Venezuela, Saudi Arabia and Egypt) all the other correlations are positive and statistically significant at the 5% significance level. However, even for those countries, correlations are still significant for many cases (the estimated values that exceed the 5% level are italicized). This means that all the countries have a large correlation with at least one of the other countries in the set, suggesting that they may constitute one or more components. According to the strong linear relationship of returns, PCA seems to be appropriate.

The Kaiser-Meyer-Olkin (KMO) is another measure to quantify the degree of intercorrelations among the observed variables. It compares the observed correlation coefficients with the partial correlation coefficients. It is computed as:

$$KMO = \frac{\sum_{i=1}^{k} \sum_{\substack{j=1 \\ i \neq j}}^{k} r_{ij}^2}{\sum_{i=1}^{k} \sum_{\substack{j=1 \\ i \neq j}}^{k} r_{ij}^2 + \sum_{i=1}^{k} \sum_{\substack{j=1 \\ i \neq j}}^{k} p_{ij}^2} \tag{3}$$

where r_{ij} and p_{ij} are the simple correlation coefficient and the partial correlation coefficient between the ith and jth variables, respectively.

TABLE 3 Correlation Matrix

	CAN	MEX	BRA	UK	FRA	GER	SPA	ITA	POR	JAP	H-K	AUS	SIN	IND	SKO	ARG	CHI	VEN	PER	POL	EGY	SAR	HUN	TUR
USA	0.627	0.499	0.409	0.675	0.648	0.656	0.584	0.548	0.336	0.280	0.428	0.503	0.406	0.193	0.265	0.351	0.425	0.150	0.273	0.287	*0.039*	0.127	0.354	0.173
CAN		0.460	0.371	0.560	0.611	0.594	0.549	0.488	0.464	0.334	0.425	0.499	0.404	0.232	0.326	0.319	0.326	0.162	0.283	0.345	*0.055*	0.136	0.399	0.199
MEX			0.520	0.472	0.467	0.452	0.511	0.416	0.404	0.286	0.399	0.448	0.356	0.215	0.248	0.538	0.433	0.265	0.325	0.286	*0.107*	*0.086*	0.369	0.233
BRA				0.373	0.413	0.426	0.446	0.349	0.391	0.279	0.302	0.397	0.279	0.172	0.234	0.487	0.557	0.291	0.428	0.205	*0.059*	*0.092*	0.356	0.271
UK					0.779	0.744	0.700	0.646	0.483	0.328	0.538	0.499	0.429	0.184	0.345	0.360	0.356	0.150	0.256	0.298	*0.026*	*0.099*	0.427	0.235
FRA						0.830	0.771	0.752	0.580	0.362	0.475	0.481	0.395	0.211	0.323	0.374	0.368	0.133	0.280	0.344	*0.055*	0.119	0.425	0.253
GER							0.763	0.723	0.570	0.346	0.495	0.503	0.416	0.236	0.331	0.350	0.346	0.145	0.264	0.366	*0.061*	0.163	0.470	0.262
SPA								0.724	0.621	0.323	0.453	0.528	0.393	0.236	0.344	0.434	0.403	0.188	0.347	0.325	*0.010*	*0.072*	0.487	0.269
ITA									0.512	0.285	0.375	0.470	0.323	0.226	0.272	0.364	0.322	0.128	0.259	*0.070*	*0.070*	*0.097*	0.436	0.246
POR										0.218	0.353	0.351	0.293	0.259	0.237	0.296	0.271	*0.078*	0.233	0.370	*0.045*	*0.064*	0.456	0.238
JAP											0.367	0.371	0.355	0.183	0.316	0.205	0.188	*0.101*	0.222	0.267	*0.015*	*0.057*	0.243	0.156
H-K												0.482	0.653	0.225	0.439	0.280	0.276	0.162	0.236	0.276	*0.017*	*0.032*	0.344	0.146
AUS													0.414	0.261	0.362	0.289	0.351	0.150	0.298	0.365	*0.046*	0.115	0.386	0.240
SIN														0.272	0.402	0.302	0.249	0.160	0.244	0.297	*0.058*	0.085	0.262	0.154
IND															0.279	0.121	0.187	0.154	0.216	0.217	0.129	*0.045*	0.247	0.100
SKO																0.161	0.233	0.101	0.146	0.285	0 .090	0.097	0.302	0.171
ARG																	0.391	0.243	0.390	0.199	*0.034*	*0.046*	0.297	0.189
CHI																		0.261	0.447	0.267	0.084	0.136	0.381	0.180
VEM																			0.194	0.120	0.094	*0.071*	0.239	0.162
PER																				0.211	*0.057*	*0.042*	0.316	0.191
POL																					0.147	*0.051*	0.534	0.257
EGY																						0.173	*0.086*	*0.081*
SAR																							0.128	*0.072*
HUN																								0.379

TABLE 4

KMO value	Classification
0.90–1.00	Marvelous
0.80–0.90	Meritorious
0.70–0.80	Middling
0.60–0.70	Mediocre
0.50–0.60	Miserable
0.00–0.50	Unacceptable

TABLE 5 KMO and Bartlett's Test

Kaiser-Meyer-Olkin measure of sampling adequacy		0.934
Bartlett's test of sphericity	Approx. chi-square	5349
	Degrees of freedom	300
	Significance	0.000

If the sum of squared partial correlation coefficients between all pairs of variables is small when compared to the sum of squared correlation coefficients, the KMO index is close to 1 and PCA is *marvelous* (Kaiser, 1974). If the KMO index is relatively small, PCA is not appropriate, since correlations between pairs of variables cannot be explained by other variables. Kaiser (1974) proposed the following classifications according to the values of the KMO index (*Table 4*).

As a typical instrument to evaluate the appropriateness of PCA, the Bartlett's test of sphericity was also performed in our survey.

The Bartlett's test of sphericity tests the null hypothesis that the population correlation matrix is an identity matrix and, therefore, the non-zero correlations in the sample correlation matrix must be due to sampling error. If the null is not rejected, variables are uncorrelated and PCA is inappropriate.

$$BS = -\left[n - 1 - \frac{1}{6}(2p + 5)\right] \ln|\mathbf{R}| \quad \text{or} \quad BS =- \left[n - 1 -\frac{1}{6}— (2p + 5)\right]\sum_{i=1}^{p}\ln(\lambda_i)$$

$$BS \overset{a}{\cap} \chi^2_{\left[\frac{1}{2}p(p-1)\right]} \tag{4}$$

where $\mathbf{R}$ is the correlation matrix, n is the number of observations, p is the number of variables and λ_I is one of the eigenvalues of the correlation matrix. *Table 5* shows the results of the KMO and the Bartlett's test regarding our sample.

As we can observe, Table 5 presents a marvelous result for the KMO statistic (between 0.90 and 1.0) and the null hypothesis on the Bartlett's test is clearly reject; it is unlikely that the population correlation matrix of returns should be an identity matrix. Thus, if PCA is conducted, the components extracted will account for a large amount of variance of the observed variables.

Finally, the partial correlation coefficients were also computed in order to measure the strength of the relationship among variables. The negative

TABLE 6 The Anti-image Correlation Matrix

	USA	CAN	MEX	BRA	UK	FRA	GER	SPA	ITAL	POR	JAP	H-K	AUS	SIN	IND	SKO	ARG	CHI	VEN	PER	POL	HUN	TUR	EGY	SAR
USA	0.930	–0.410	–0.100	–0.140	–0.200	–0.100	–0.080	0.050	–0.010	0.150	0.080	–0.020	–0.060	–0.050	–0.040	0.100	–0.010	–0.120	–0.020	0.070	–0.010	–0.030	0.000	0.070	–0.030
CAN		0.910	–0.020	0.140	0.060	–0.160	–0.080	–0.010	0.050	–0.080	–0.090	0.100	–0.170	–0.140	–0.040	–0.030	–0.080	0.070	0.020	–0.070	–0.010	–0.080	0.100	–0.090	–0.070
MEX			0.930	–0.170	–0.010	0.030	0.050	–0.110	0.010	–0.040	–0.080	–0.050	–0.050	–0.020	–0.060	0.080	–0.360	–0.030	–0.050	0.150	–0.070	0.030	–0.010	–0.010	–0.070
BRA				0.910	0.060	–0.040	–0.100	0.060	0.040	–0.110	–0.110	0.090	–0.150	–0.030	0.010	–0.060	–0.260	–0.290	–0.030	–0.190	0.140	0.010	–0.120	0.000	0.010
UK					0.950	–0.200	–0.120	–0.170	0.040	–0.050	–0.050	–0.120	–0.020	–0.060	0.110	–0.170	0.030	–0.050	0.020	0.000	0.070	–0.060	–0.060	0.020	0.050
FRA						0.930	–0.290	–0.180	–0.250	–0.150	–0.130	–0.080	0.150	0.090	0.020	–0.030	–0.020	–0.060	0.060	–0.010	–0.050	0.170	–0.020	–0.030	–0.010
GER							0.950	–0.240	–0.100	–0.080	0.000	–0.140	0.010	0.070	0.020	0.060	0.070	0.050	–0.040	0.030	–0.110	–0.060	0.000	0.060	–0.100
SPA								0.930	–0.230	–0.210	0.020	0.080	–0.170	–0.030	0.030	–0.030	–0.040	–0.080	–0.110	–0.040	0.150	–0.130	0.060	–0.030	0.120
ITAL									0.950	–0.020	0.020	0.100	–0.130	0.000	–0.030	–0.010	–0.060	0.000	0.060	0.000	0.040	–0.100	–0.020	0.020	–0.050
POR										0.920	0.090	0.000	0.040	–0.080	–0.160	0.070	–0.020	0.120	0.190	0.050	–0.120	–0.140	–0.050	–0.010	0.010
JAP											0.930	–0.010	–0.080	–0.060	0.060	–0.020	–0.050	0.110	0.010	–0.060	–0.080	0.000	–0.080	0.090	0.060
H–K												0.850	–0.240	–0.530	0.000	–0.130	–0.070	0.020	0.020	–0.040	0.110	–0.090	0.080	0.140	0.010
AUS													0.920	0.100	–0.010	–0.010	0.140	–0.050	–0.030	–0.030	–0.190	0.080	–0.100	0.010	0.000
SIN														0.840	–0.030	–0.080	–0.060	0.000	–0.100	0.020	–0.140	0.150	0.020	–0.100	0.030
IND															0.790	–0.120	0.110	–0.100	–0.110	–0.080	0.000	–0.060	0.000	–0.050	0.080
SKO																0.900	0.000	0.010	–0.040	0.090	–0.090	–0.070	0.010	–0.030	–0.040
ARG																	0.910	–0.070	–0.160	–0.200	0.030	–0.070	0.000	0.050	0.060
CHI																		0.930	–0.040	–0.200	–0.070	–0.130	0.070	–0.020	–0.040
VEN																			0.900	–0.010	0.050	–0.110	–0.060	–0.060	–0.010
PER																				0.910	–0.020	–0.060	–0.030	–0.060	0.050
POL																					0.840	–0.390	–0.070	– 0.080	0.150
HUN																						0.900	–0.190	–0.010	–0.170
TUR																							0.900	–0.070	–0.090
EGY																								0.560	–0.170
SAR																									0.550

value of this measure is called the anti-image correlation and the matrix with these correlations is shown in *Table 6*. The measures of sampling adequacy (hereinafter referred to as "MSA") for each variable are printed on the diagonal of that matrix and reasonably large values are needed for a good PCA. For the *i*th variable MSA is:

$$MSA_i = \frac{\sum_{j \neq i} r_{ij}^2}{\sum_{j \neq i} r_{ij}^2 + \sum_{j \neq i} a_{ij}^2} \tag{5}$$

From the analysis of this matrix we conclude that as the proportion of large partial correlation coefficients is small and the values for MSA are greater than 0.80 except for a few cases, it means that returns are strongly correlated and PCA is appropriate.

As all the previous results provided evidence that PCA is feasible, the next step is to extract the components.

3.2 Extracting an Initial Solution

As mentioned before, components can be estimated as linear combinations of the variables:

$$C_j = \sum_{i=1}^{k} w_{ji} X_i = w_{j1} X_1 + w_{j2} X_2 + \ldots + w_{jk} X_k \tag{6}$$

where w_i are known as component score coefficients and k is the number of variables.

PCA is based on an eigenvalue and eigenvector analysis of $\mathbf{V} = \mathbf{X}'\mathbf{X}/T$, the $(k \times k)$ symmetric matrix of correlations between the normalized variables in **X**. Each principal component (hereinafter referred to as "PC") is a linear combination of these columns, where the weights are chosen in such a way that:

1. The first PC explains the greatest amount of the total variation in **X**; the second component explains the greatest amount of the remaining variation, and so on.
2. To be orthogonal to the other components, the next component must be derived from the proportion of the variance remaining after the first component has been extracted. Thus the second component may be defined as the linear combination of variables that accounts for the most residual variance after the effect of the first component has been removed from the data. Subsequent components are defined similarly, until all the variance in the data is exhausted. The PCs are uncorrelated with each other.

It is shown that this can be achieved by choosing the weights from the set of eigenvectors of the correlation matrix.

In PCA the components are estimated as linear combinations of the observable variables. While it is possible that all variables contribute to each component, hopefully only a subset of variables characterizes it, as indicated by their large coefficients in the linear equation. Therefore, the general model for the *i*th standardized variable can be written as:

$$X_i = A_{i1}C_1 + A_{i2}C_2 + \ldots + A_{ik}C_k + U_i \tag{7}$$

where the *C's* are the common components, the *U* is the unique component and the *A's* are the coefficients used to combine the *k* extracted components. The unique components are assumed to be uncorrelated with the common components.

To decide how many components are needed to represent the data, it is helpful to examine the percentage of total variance explained by each. The total variance is the sum of the variance of each variable. For simplicity, all variables and components are expressed in standardized form, with mean 0 and a standard deviation of 1. Since the variance of each variable is 1, the total variance is equal to the number of observed variables in the analysis.

There are several criteria for determining the number of components we should retain. The choice of criterion may depend on the average size of communalities and the number of variables and observations. The Kaiser criterion has been recommended for situations where the number of variables is less than 30 and the average communality is greater than 0.7 or when the number of subjects is greater than 250 and the mean communality is near or greater than 0.6 (Stevens, 1992). Based on this criterion, only the components having eigenvalues greater than 1 are considered significant; all the others are disregarded. The rationale for this criterion is that any individual component should account for the variance of at least a single variable.

As our sample meets the Kaiser requirement, the PCs were extracted based on this criterion.

In our study, the first components obtained from PCA are shown in *Table 7*.

As mentioned above, it is possible to compute as many components as variables in the original set. Therefore, 25 components were extracted (the number of original variables). The eigenvalue associated with the first component is 9.177. Since this is greater than 1.0 (Kaiser criterion), it explains more variance than a single variable. The variance explained by this component is 36.709 % (9.177/25) of the total variance of the observed variables. The next four components also have an eigenvalue greater than 1.0 and therefore explain more variance than a single variable. The remaining components have eigenvalues less than 1.0 and therefore explain less variance than a single variable.

Hence in this study, five components have been extracted to represent the 25 indices returns. The cumulative percentage of variance explained by the first five components is 58.6 %; this means that 58.6 % of the common variance shared by the 25 stock indices' returns can be accounted for by the 5 components.

Table 8 displays the coefficients that relate the returns to the five components. In Portugal, for example, the standardized returns (r_{POR}) can be expressed as:

$$r_{POR} = 0{,}6503C_1 - 0{,}1357C_2 - 0{,}1133C_3 + 0{,}2384C_4 - 0{,}2202C_5 \tag{8}$$

where C_j represent each one of the five extracted principal components.

TABLE 7 Components Extraction

	Initial eingenvalues		
Component	Total	% of var.	Cum. %
1	9.177	36.709	36.709
2	1.667	6.667	43.376
3	1.425	5.700	49.076
4	1.270	5.079	54.154
5	1.108	4.434	58.588
6	0.917	3.668	62.257
7	0.859	3.435	65.692
8	0.814	3.255	68.947
9	0.759	3.035	71.982
10	0.731	2.924	74.906
11	0.699	2.797	77.703
12	0.661	2.644	80.347
13	0.657	2.627	82.974
14	0.569	2.278	85.251
15	0.525	2.100	87.352
16	0.518	2.070	89.422
17	0.435	1.741	91.163
18	0.394	1.574	92.737
19	0.369	1.477	94.214
20	0.330	1.319	95.533
21	0.294	1.175	96.708
22	0.247	0.987	97.696
23	0.234	0.935	98.631
24	0.192	0.769	99.400
25	0.150	0.600	100.000

TABLE 8 Component Matrix

Country	Component				
	1	2	3	4	5
USA	0.738	–0.155	–0.158	–0.105	0.219
CAN	0.721	–0.142	–0.003	–0.011	0.102
MEX	0.676	0.254	–0.126	–0.140	0.071
BRA	0.619	0.434	–0.211	–0.140	0.011
UK	0.797	–0.292	–0.134	–0.033	0.081
FRA	0.832	–0.308	–0.200	0.078	0.055
GER	0.831	–0.295	–0.145	0.113	0.066
SPA	0.830	–0.177	–0.212	0.047	–0.063
ITA	0.748	–0.259	–0.220	0.155	0.007
POR	0.650	–0.136	–0.113	0.238	–0.220
JAP	0.479	–0.040	0.319	–0.214	–0.063
H–K	0.645	–0.143	0.352	–0.341	0.014
AUS	0.685	–0.027	0.166	–0.090	0.011
SIN	0.588	–0.059	0.429	–0.336	0.080
IND	0.362	0.127	0.403	0.074	–0.054
SKO	0.485	–0.053	0.521	–0.079	0.007
ARG	0.549	0.365	–0.270	–0.224	–0.002
CHI	0.566	0.455	–0.150	–0.078	0.091
VEN	0.289	0.490	0.017	–0.034	0.080
PER	0.471	0.465	–0.092	–0.141	–0.101
POL	0.507	0.077	0.308	0.359	–0.316
HUN	0.631	0.169	0.095	0.370	–0.293
TUR	0.373	0.216	0.032	0.376	–0.320
EGY	0.106	0.235	0.259	0.462	0.444
SAR	0.163	0.094	0.090	0.352	0.674

TABLE 9 Communalities

Country	Initial	Extraction
USA	1.000	0.652
CAN	1.000	0.551
MEX	1.000	0.562
BRA	1.000	0.635
UK	1.000	0.747
FRA	1.000	0.835
GER	1.000	0.816
SPA	1.000	0.772
ITA	1.000	0.699
POR	1.000	0.560
JAP	1.000	0.382
H-K	1.000	0.676
AUS	1.000	0.505
SIN	1.000	0.653
IND	1.000	0.318
SKO	1.000	0.516
ARG	1.000	0.558
CHI	1.000	0.564
VEN	1.000	0.331
PER	1.000	0.476
POL	1.000	0.586
HUN	1.000	0.659
TUR	1.000	0.431
EGY	1.000	0.544
SAR	1.000	0.622

The relationship between returns and components is expressed by the coefficients of this matrix, called *Component Loadings* as they indicate how much weight is assigned to each component. The countries have been listed in terms of the size of their loadings on the component to which they are most closely related. Components with large coefficients (in absolute value) for a variable are closely related to that variable. For example, component 1 is the component with the largest loading for the standardized returns in Portugal.

Since the estimated components are uncorrelated, the component loadings represent the unique contribution of each component and can also be interpreted as the correlations among them and the original variables. In order to evaluate the five-component model, we can compute the proportion of the variance of returns explained by the model. As the components are orthogonal, the total proportion of variance explained by the model is simply the sum of the variance proportions explained by each component. The proportion of variance in one original variable that is accounted for by the common components is called the *communality* of the variable and can be computed as the factor loadings sum of squares. The communalities of returns are shown in *Table 9*. For example, in USA case, the value obtained is given by:

$$(0.738)^2 + (-0.155)^2 + (-0.158)^2 + (0.105)^2 + (0.219)^2 = 0.652$$

The better represented countries are France, Germany and Spain with more than 75 % of the respective variances explained, and the worst represented countries are India, Venezuela and Japan with a communality of less than 40 %.

3.3 Rotation

Sometimes most of the variables will load on the same component, making its interpretation ambiguous. Ideally, the analyst would like to find that each variable loads high on one component and approximately zero on all the others. In general, the component pattern can be clarified by "rotating" the components in F-dimensional space, since rotation focuses on transforming the components to make them more interpretable. For example, in our survey South Korea and Peru present similar loadings respectively to the $C1$, $C3$ and $C1$, $C2$ components (see *Table 10*).

In the rotation process, the reference axes of the components are turned about the origin until some other position has been reached. The ultimate effect of rotating the component matrix is to redistribute the variance from earlier components to later ones to achieve a simple, theoretically more meaningful component pattern.

Rotation does not affect the goodness of fit of the PCA solution. That is, the communalities and the percentage of variance explained do not change.

The simplest case of rotation is an orthogonal rotation in which the axes are maintained at 90 degrees. A variety of algorithms can be used for orthogonal rotation. The most commonly used method is the varimax method,

TABLE 10 Component Matrix

	Component				
	1	2	3	4	5
FRA	0.831	–0.308	–0.200	0.078	0.055
GER	0.831	–0.295	–0.144	0.113	0.066
SPA	0.830	–0.177	–0.212	0.046	–0.063
UK	0.797	–0.292	–0.134	–0.033	0.081
ITA	0.748	–0.259	–0.220	0.155	0.006
USA	0.738	–0.155	–0.158	–0.105	0.219
CAN	0.721	–0.142	–0.003	–0.011	0.102
AUS	0.685	–0.027	0.165	–0.090	0.011
MEX	0.676	0.254	–0.126	–0.140	0.071
POR	0.650	–0.136	–0.113	0.238	–0.220
H–K	0.645	–0.143	0.352	–0.341	0.014
HUN	0.631	0.169	0.095	0.370	–0.293
BRA	0.619	0.434	–0.211	–0.140	0.011
SIN	0.588	–0.059	0.429	–0.336	0.080
CHI	0.566	0.455	–0.150	–0.077	0.091
ARG	0.549	0.365	–0.270	–0.224	–0.002
POL	0.507	0.077	0.308	0.358	–0.316
JAP	0.479	–0.040	0.319	–0.214	–0.063
PER	0.471	0.465	–0.092	–0.141	–0.101
VEN	0.289	0.490	0.017	–0.034	0.080
SKO	0.485	–0.053	0.521	–0.079	0.007
IND	0.362	0.127	0.403	0.074	–0.054
EGY	0.106	0.235	0.259	0.462	0.444
TUR	0.373	0.216	0.032	0.376	–0.320
SA	0.163	0.094	0.090	0.352	0.674

TABLE 11 Rotated Component Matrix

	Component				
	1	2	3	4	5
FRA	0.862	0.158	0.201	0.158	0.031
GER	0.837	0.142	0.230	0.190	0.074
UK	0.795	0.170	0.283	0.075	0.009
ITA	0.786	0.142	0.111	0.220	0.033
SPA	0.780	0.267	0.188	0.232	–0.056
USA	0.700	0.289	0.258	–0.050	0.096
CAN	0.611	0.213	0.331	0.117	0.092
POR	0.584	0.128	0.102	0.434	–0.057
BRA	0.296	0.719	0.117	0.127	0.006
CHI	0.238	0.682	0.117	0.123	0.118
ARG	0.296	0.675	0.081	0.029	–0.081
PER	0.108	0.638	0.142	0.183	–0.057
MEX	0.405	0.580	0.228	0.090	0.042
VEN	–0.048	0.526	0.108	0.121	0.161
SIN	0.258	0.186	0.742	–0.013	0.012
H–K	0.366	0.164	0.714	0.005	–0.073
SKO	0.158	0.040	0.663	0.195	0.108
JAP	0.208	0.142	0.552	0.101	–0.065
AUS	0.435	0.256	0.469	0.170	0.038
IND	0.030	0.116	0.434	0.309	0.141
HUN	0.328	0.258	0.174	0.671	0.063
POL	0.193	0.070	0.303	0.669	0.064
TUR	0.139	0.198	0.005	0.610	0.022
SA	0.153	0.056	0.005	–0.081	0.768
EGY	–0.052	0.048	0.052	0.183	0.709

which attempts to minimize the number of variables with high loadings on just one component.

Thus, in this study the varimax method was applied. As can be observed, the above mentioned interpretation problems regarding South Korea and Peru are now much more clear.

From analysis of *Table 11* it is clear that each component is associated with a group of country indices that have strong economic, political and geographical relationships among them. The first component is associated with the Western developed economies of North America (USA and Canada) and of the European Union (France, Germany, UK, Italy, Spain and Portugal); the second component is associated with the Central and South American countries (Brazil, Chile, Argentina, Peru, Mexico, and Venezuela); the third with the Asian and Pacific countries (Singapore, Hong Kong, South Korea, Japan, Australia, India); the fourth with the Eastern European countries (Hungary, Poland and Turkey). Finally the fifth component is associated with the North African and Middle Eastern countries (Egypt and Saudi Arabia).

Moreover, the application of PCA also allows us to determine the magnitude impact of the other equity markets on the returns of a particular one. For example, observing the component loadings associated with Portugal, one can say that its returns depend firstly on the western equity markets (0.584) and second on the Eastern European countries (0.434).

TABLE 12 Component Score Coefficient Matrix

Country	Component				
	1	2	3	4	5
USA	0.171	0.029	–0.006	–0.195	0.072
CAN	0.114	–0.025	0.046	–0.056	0.056
UK	0.008	0.215	–0.006	–0.079	0.004
FRA	–0.036	0.313	–0.068	–0.039	–0.033
GER	0.201	–0.062	–0.008	–0.094	–0.007
SPA	0.233	–0.081	–0.079	–0.032	0.009
ITA	0.217	–0.094	–0.059	–0.009	0.042
POR	0.179	–0.015	–0.085	0.030	–0.075
JAP	0.217	–0.077	–0.123	0.035	0.008
H–K	0.119	–0.076	–0.107	0.232	–0.086
AUS	–0.066	–0.016	0.271	–0.018	–0.083
SIN	–0.030	–0.032	0.346	–0.129	–0.086
IND	0.009	0.007	0.156	–0.008	–0.001
SKO	–0.076	–0.008	0.381	–0.141	–0.016
MEX	–0.129	–0.025	0.210	0.160	0.078
BRA	–0.101	–0.092	0.343	0.046	0.056
ARG	–0.011	0.308	–0.073	–0.097	–0.095
CHI	–0.050	0.298	–0.056	–0.039	0.063
VEM	–0.130	0.256	0.008	0.012	0.104
PER	–0.105	0.291	–0.012	0.039	–0.089
POL	–0.084	–0.101	0.069	0.437	–0.009
HUN	–0.037	–0.009	–0.053	0.412	–0.012
TUR	–0.057	0.012	–0.106	0.418	–0.036
EGY	–0.054	–0.019	–0.003	0.078	0.580
SAR	0.063	–0.011	–0.046	–0.148	0.653

In the Portuguese case, for example, the predicted values for standardized returns are given by:

$$\hat{P}_t = 0.584C_1 + 0.128C_2 + 0.102C_3 + 0.434C_4 - 0.057C_5$$

To conclude this analysis, the Component Score Coefficient Matrix is presented (*Table 12*). This matrix can be used to calculate a score for each observation of the original series, for a given component. By this we can use the five-score component series instead of the original twenty-five index series to study the evolution of the world's equity indices.

4. Conclusions

In this study PCA was applied in a quite different context from the traditional approaches in the financial field. The obtained results showed that the application of PCA, in order to determine subsets of equity markets, is an effective tool.

With the help of this multivariate statistical technique we were able to identify five groups of nations whose equity indices are closely related. Although the total variance explained by these five components is not very high (about 60 %), the interpretation of each one of these components is very clear. Each component is mainly associated with a group of countries which have highly integrated economies, or at least a common cultural/geographical background. This means that, although the globalization of markets can be considered an uncontroversial fact, these empirical results show that different subsets of equity markets have a different influence on the returns of a specific country.

For example, the Portuguese stock-market returns are mainly and positively influenced by the most efficient capital markets represented by the first extracted component (USA, Canada and European Union countries). Using the component matrix (Tables 10 and 11) we also conclude that the Eastern European countries are the second most important influence for the Portuguese returns. As PCA is more than a correlation result, it is also possible to quantify how these markets influence the Portuguese returns using the component matrix coefficients.

The present investigation also shows that the component scores can be used to study the evolution of the world's equity indices, instead of having to work with a large number of series. *Each principal component is a linear combination of all 25 stock market indices (columns of Table 12) and the standardized return of each country is a linear combination of the five principal components extracted (Table 11 presents the coefficients associated with each one of the components). Thus, we can predict the returns of each stock market using just five explanatory dimensions instead of the other 24 stock markets.* In short, this report aimed to show the power of PCA to summarize a complex dataset, making it a useful tool to understand complex relationships.

Finally, just a few words related to the application of PCA model. Investors can summarize the correlations among the 25 stock-market indices in the largest few principal components that result from PCA. This data reduction can simplify the investors' evaluation of a portfolio's risk. *This means that investors may focus their attention only on those markets which have significant impact on their portfolio, instead of dispersing their attention on markets whose impact is residual.*

REFERENCES

ALEXANDER, C. (2001): *Market models a guide to financial markets*. John Wiley & Sons, 2001.

BRYANT, E. H. – ATCHLEY, W. R. (1975): *Multivariate statistical methods: within-groups covariation.* Dowden, Hutchinson and Ross, Inc., Stroudsburg (Pa.), 1975.

BRYMAN, A. – CRAMER, D. (1997): *Quantitative data analysis with SPSS for windows*. Routledge, London, 1997.

CATTELL, R. B. (1966): The meaning and strategic use of factor analysis. In: R. B. Cattell (Ed.): *Handbook of Multivariate Experimental Psychology*. Chicago, Rand McNally, 1966.

DUNTEMAN, G. H. (1989): Principal components analysis. *Sage University Paper*.

ENGLE, R. – NG, V. – ROSTHSCHILD, M. (1990): Asset pricing with a Factor-ARCH covariance structure: empirical estimates for treasury bills. *Journal of Econometrics*, vol. 45, 1990, pp. 213–37.

FAMA, E. (1965): The behaviour of stock prices. *Journal of Business*, vol. XXXVIII, January 1965, no. 1, pp. 34–105.

FEENEY, G. J. – HESTER, D. D. (1964): Stock market indices: a principal components analysis. *Cowles Foundation for Research in Economics, Discussion paper*, no. 175.

GORSUCH, R. L. (1983): *Factor Analysis*. Lawrence Erlbaum Associates, 1983.

HAIR, J. – TATHAM, R. – ANDERSON, R. – BLACK, W. (1998): *Multivariate data analysis*. 5th Edition. Princeton University Press, New Jersey, 1998.

HOTELLING, H. (1933): Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, vol. 24, 1933, pp. 417ů41 and 498–520.

JACKSON, J. E. (1991): *A User's Guide to Principal Components*. John Wiley & Sons, 1991.

LAM, S. S. – ANG, W. W. (2006): Globalization and stock market returns. *Global Economy Journal*, vol. 6, 2006, no. 1.

LORETAN, M. (1997): *Generating market risk scenarios using principal components analysis: methodological and practical considerations*. Federal Reserve Board, 1997.

KAISER, H. F. (1974): An index of factor simplicity. *Psychometrika*, vol. 39, 1974, pp. 31–36.

MANDELBROT, B. (1963): The variation of certain speculative prices. The *Journal of Business*, vol. 36, 1963, pp. 394–419.

PEARSON, K. (1901): On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, Ser. B, vol. 2, 1901, pp. 559–72.

STEVENS, J. (1992): *Applied multivariate statistics for the social sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

SUMMARY

JEL Classification: C10, E44, G15, O57
Keywords: interaction; principal components analysis; returns

World Equity Markets: A New Approach for Segmentation

José Dias CURTO – ISCTE Business School, Lisbon (corresponding author, dias.curto@iscte.pt)
José Castro PINTO – ISCTE Business School, Lisbon
João Eduardo FERNANDES – ISCTE Business School, Lisbon

This paper is an assessment of international equity-market integration and uses an innovative approach to segment equity markets into related geographic areas. Our focus is on the relationships among the returns of the dominant national equity indexes by continent. To understand how these indexes have evolved, we will concentrate on a reduced number of dimensions extracted from principal components analysis. We will demonstrate that each one of these components is particularly associated with certain groups of nations and less associated with others.