Revista Finanzas y Política Económica, Vol. 13, N.° 2, julio-diciembre, 2021, pp. 513-543
https://doi.org/10.14718/revfinanzpolitecon.v13.n2.2021.9

Rogelio Ladrón de Guevara Cortés[*]
Salvador Torra Porras[**]
Enric Monte Moreno[***]

Recibido: *18 de noviembre de 2020*
Evaluación: *31 de marzo de 2021*
Aprobado: *30 de julio de 2021*

Artículo de investigación

Statistical and computational techniques for extraction of underlying systematic risk factors: a comparative study in the Mexican Stock Exchange

Abstract.

This paper compares the dimension reduction or feature extraction techniques, e.g., Principal Component Analysis, Factor Analysis, Independent Component Analysis, and Neural Networks Principal Component Analysis, which are used as techniques for extracting the underlying systematic risk factors driving the returns on equities of the Mexican Stock Exchange, under a statistical approach to the Arbitrage Pricing Theory. This research is carried out according to two different perspectives. First, an evaluation from a theoretical and matrix scope is done, making parallelism among their particular mixing and demixing processes, as well as the attributes of the factors extracted by each method. Secondly, an empirical study to measure the level of accuracy in the reconstruction of the original variables is accomplished. In general, the results of this research point to Neural Networks Principal Component Analysis as the best technique from both theoretical and empirical standpoints.

Keywords: Neural Networks Principal Component Analysis, Independent Component Analysis, Factor Analysis, Principal Component Analysis, Mexican Stock Exchange.

JEL Classification: G12, G15, C45

[*] Ph.D. in Business Studies. Professor-Researcher. Institute of Research and Graduate Studies in Administrative Sciences (IIESCA). Universidad Veracruzana. Mexico. e-mail: roladron@uv.mx
https://orcid.org/0000-0001-9365-2080

[**] Ph.D. in Economic Sciences. Professor-Researcher. Department of Econometrics, Statistics, and Applied Economy. Faculty of Economics and Business. University of Barcelona. Spain. e-mail: storra@ub.edu.
https://orcid.org/0000-0002-8786-8800

[***] Ph.D. in Digital Signal Processing. Professor-Researcher. Department of Signal Theory and Communications. Barcelona School of Telecommunications Engineering. Polytechnic University of Catalonia. Spain. E-mail: enric.monte@upc.edu
https://orcid.org/0000-XXXX-XXXX-XXXX

Técnicas estadísticas y computacionales para extraer factores de riesgo sistemático subyacentes: un estudio comparativo en la Bolsa Mexicana de Valores

Resumen

Este artículo compara las técnicas de reducción de dimensionalidad o de extracción de características: Análisis de Componentes Principales, Análisis Factorial, Análisis de Componentes Independientes y Análisis de Componentes Principales basado en Redes Neuronales, las cuales son usadas para extraer los factores de riesgo sistemático subyacentes que generan los rendimientos de las acciones de la Bolsa Mexicana de Valores, bajo un enfoque estadístico de la Teoría de Valoración por Arbitraje. Llevamos a cabo nuestra investigación de acuerdo a dos diferentes perspectivas. Primero, las evaluamos desde una perspectiva teórica y matricial, haciendo un paralelismo entre los particulares procesos de mezcla y separación de cada método. En segundo lugar, efectuamos un estudio empírico con el fin de medir el nivel de precisión en la reconstrucción de las variables originales.

Palabras clave: Análisis de Componentes Principales basado en Redes Neuronales, Análisis de Componentes Independientes, Análisis Factorial, Análisis de Componentes Principales, Bolsa Mexicana de Valores.

INTRODUCTION

Classic multivariate dimensional reduction techniques have been widely used in different fields of science to extract the underlying factors from large sets of data or to build synthetic indicators that range from natural to social sciences, such as Physics, Chemistry, Biology, Medicine, Astronomy, Psychology, Education, Management, Marketing, etc. There is a large amount of literature that focuses on the application of mainly, Principal Component Analysis and Factor Analysis, in different fields of knowledge.[1]

In previous research, three different dimension reduction or feature extraction techniques for extracting the underlying systematic risk factors driving the returns on equities in a statistical approach to the Arbitrage Pricing Theory (Ross, 1976) have been presented. This approach assumes *a priori* neither the number nor the nature of either the systematic risk factors or the sensitivity to them; therefore, in this paper, the estimation of both of them is performed by using extraction and econometric techniques in two different stages. The efforts to extract underlying factors with better characteristics led us to advance from classical multivariate techniques, such as Principal Component Analysis (PCA) and Factor Analysis (FA), to more advanced and sophisticated techniques -usually applied in fields like Engineering, Telecommunications, Astronomy, Biochemistry, Bioinformatics, Artificial Intelligence and Robotics- such as Independent Component Analysis (ICA) and Neural Networks Principal Component Analysis (NNPCA).

Although the main objective of each technique is similar -to reduce the dimension or to extract the main features from a set of variables-, they are different in nature, assumptions, principles, and internal algorithms; this makes it difficult and impracticable to compare their results, i.e., the matrices used in the processes of extraction and generation, and the underlying factors extracted. To solve this problem, the main objective of this paper is to propose a methodology to compare the four techniques, based on the degree of accuracy in the reconstruction of the observed variables using the underlying systematic factors extracted using each technique.

[1] For the sake of saving space, a review of literature on the application of multivariate techniques in different fields of knowledge is out of the scope of this paper. However, interested reader can consult Ladrón de Guevara-Cortés & Torra-Porras (2014), Ladrón de Guevara-Cortés, Torra-Porras & Monte-Moreno (2018) and Ladrón de Guevara-Cortés, Torra-Porras & Monte-Moreno (2019), where a deeper review of literature focused in each one of the techniques compared in this paper is done.

To accomplish this objective, first, a theoretical and matrix comparison among techniques is proposed, where their parallelism among their particular mixing and demixing processes is remarked. Then, this paper provides empirical evidence of their reconstruction accuracy using a set of measures usually implemented for forecastings, such as the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE), The Root Mean Square Error (RMSE), the Theil's U statistic (U-Theil), the Confusion Matrix (CM), the Confusion Rate (CR), The chi-squared contrast of independence (χ^2), and the Pesaran & Timmermann's directional accuracy statistic (DA).

Comparative studies of all four techniques in literature are scarce, so the main contribution of this paper is precisely to fill this gap in the financial literature, providing a theoretical and empirical comparative study among PCA, FA, ICA, and NNPCA in the field of finance, by way of matrix parallelism among the four techniques and the analysis of the reconstruction accuracy of the observed returns on equities, using the different components or factors extracted through each technique.

In addition, this study contributes to providing evidence in two particular relevant contexts. On the one hand, the financial market studied, that in this case represents an important emergent Latin-American equity market -the Mexican Stock Exchange-, whose studies related to these kinds of techniques are uncommon as well. On the other hand, the time analyzed corresponds to the period previous to the last recognized financial bubble: the subprime crisis. In the current actual situation where is very likely that another financial and economic crisis strikes, derived from the effects of the COVID-19 pandemic, it is considered necessary to test the performance of these techniques in a similar period to show explanatory insights into these kinds of situations.

Although these types of statistical and computational techniques have both explanatory and forecasting attributes, the aim and scope of this paper are focused only on their explanatory power. The forecasting properties are out of the scope of this paper since they are considered in other additional researches. Likewise, the test of these techniques during crisis and post-crisis periods is being studied in other extensions of this research.

The main results of this research reveal, on the one hand, that from a theoretical perspective NNPCA seems to offer the most suitable attributes for the underlying factors in a statistical context of the Arbitrage Pricing Theory. On the other hand, from an empirical scope, although there is no clear supremacy of any of the four techniques, evidence points to NNPCA as the one with the best performance in the reconstruction of the returns on equities.

The structure of this paper is as follows: section 2 makes a review of literature focused on comparative studies of these techniques; section 3 proposes the matrix parallelism among the used techniques, explaining the attributes of the factors extracted with each one of them; section 4 describes the methodology carried out in the study, and section 5 shows the results of the empirical comparative study. Finally, section 6 draws some conclusions and section 7 presents the references.

REVIEW OF THE LITERATURE.

To the best of our knowledge, only Scholz (2006a) uses and compares three of the aforementioned techniques in a single study, i.e. PCA, ICA, and NNPCA, carried over to molecular data in biochemistry to extract biologically meaningful components. The author explains the benefits and drawbacks of each type of analysis to understand biological issues, concluding that, depending on the characteristics of the data and the purpose of the research, one specific technique is more suitable than the others. For the sake of saving space and considering that one of the main contributions of this paper is the application of these four techniques in the financial context, comparative studies in fields different from Finance and Economics are out of the scope of this research. Nevertheless, the interested reader can easily find references of comparative studies between some of these techniques in the literature of Natural and Social Sciences.

Some recent and relevant researches focused on comparative studies of these types of techniques are, for example, Corominas, Garrido-Baserba, Villez, Olsson, Cortés & Poch (2018), where the authors do a description of the state-of-the art computer-based techniques for data analysis in the context of wastewater treatment plants; and Ayesha, Hanif & Talib (2020) where they present the state-of-the-art dimensionality reduction techniques for high dimensional data and their suitability for different types of application areas such as biomedicine, audio, image and video, genetics, signal-processing, pattern-recognition, etc. Comparative studies in Economics and Finance are not very frequent in literature, and they have dealt with only two of these techniques in the same review. Some relevant references in these fields are the following.[2]

Regarding PCA and FA, Ince & Trafalis (2007) use the components and factors extracted through PCA and FA as the input variables for two different forecasting models to compare their performance for stock price prediction on the NASDAQ. They found that the factors extracted through FA performed better than the components extracted through PCA. More recently, Ibraimova (2019) compares the performance of PCA and FA as dimensionality reduction techniques where the factors extracted by each technique were used as inputs in a model that tried to predict financial distress in companies through machine learning. The best results in predictions were those that used the PCA factors extraction.

Concerning ICA, Bellini & Salinelli (2003) find that the immunization strategies to the US Treasury spot rates curve movements based on ICA perform better than those based on PCA. Lesch, Caille, & Lowe (1999) apply PCA and ICA to perform feature extraction from currency exchange data of the British Pound against the US Dollar, showing that both techniques are capable of detecting deterministic structure in the data, but independent components are much closer in their morphology to the signals.

Back & Weigend (1997) apply ICA and PCA on the Tokyo Stock Exchange, showing that while the reconstruction of the observed stock prices derived from the independent components extracted is outstanding, the reproduction resulting from the principal components is not. Yip & Xu (2000) carry ICA and PCA over to stocks from the S&P 500, finding that ICA gives a better indication of the underlying structure of the US stock market, in terms of the linear relationship between the components extracted through both techniques and some predefined macroeconomic factors.

Rojas & Moody (2001) compare ICA and PCA by investigating the term structure and the interactions between the returns of iShares MSCI Index Funds and the returns of the S&P Depositary Receipts Index; they demonstrate that ICA has more influence on the average mutual information. Lizieri, Satchell & Zhang (2007) compare the ICA and PCA components' capability of capturing the kurtosis in Real Estate Investment Trusts (REIT) in the USA, therefore proving that ICA overcomes PCA. Nevertheless, Wei, Jin & Jin (2005) uncover that, although both techniques produce

[2] Although in the following papers the authors made a theoretical comparison of the techniques utilized, this paper will focus on the comparison of their empirical results. For detailed information about both the theoretical and the empirical comparisons made in those works, the interested reader can consult the sources.

similar results, PCA outperforms ICA in the reconstruction of mutual funds in the Chinese financial market.

On the other hand, Coli, Di Nisio & Ippoliti (2005), in an application of ICA and PCA to a stocks portfolio of the Milan Stock Exchange, uncover that, although the principal components present a minimum reprojection error when they are used to reconstruct the data, the independent components make it easier to distinguish of the trend and the cyclical components. More recently, Sayah (2016), compares PCA-GARCH versus ICA-GARCH models in the context of Basel's III sensitivity based approach for the interest rate risk in the trading book of Banks, finding that in general, the ICA model produced more restrictive results in the Value at Risk (VaR) computation.

Regarding NNPCA, Weigang, Rodrigues, Lihua & Yukuhiro (2007) compare NNPCA and PCA in terms of their dimensional reduction capability, to extract the main feature explaining the trends of withdrawals from an employment time guarantee fund, thereby showing that NNPCA is more suitable than PCA for dimension reduction in this context.[3] On the other hand, Liu & Wang (2011) integrated ICA and PCA with Neural Networks to predict the Chinese Stock Market finding suitable results. In addition, interesting surveys focused on applications of related intelligent computational techniques in financial markets applications can be found in Cavalcante, Brasileiro, Souza, Nobrega & Oliveira (2016), and in machine learning techniques applied to financial market prediction in Miranda, Amorin & Kimura (2019).

On the other hand, there is a working precedent that has undertaken a systematic exploration of dimensionality reduction methods such as Anowar, Sadaoui & Selim (2021). However, the current work aims to find basic factors that can explain the underlying risk factors in the Mexican market. This is why a selection of dimensionality reduction techniques has been made that allows a local approximation to explain the risk. In particular, not all dimensionality reduction techniques allow for explanations. For example, following Scikit-Learn (2021), Multi-dimensional Scaling (MDS) is based on finding a 2D distribution that maintains the metric relationships of the original space, but loses the proportionality relationship when explaining risk and the dynamic aspect that is of interest in this article.

[3] Neither other techniques to produce non-linear components nor other methods to obtain non-linear principal component analysis (NLPCA) different than NNPCA are in the scope of this research. Nevertheless, the interested reader can find in the literature some works where techniques such as the quantum-inspired evolutionary algorithm (QIA) to extract non-linear principal components, or kernel principal component analysis (KPCA) and curvilinear component analysis (CCA) are compared with some of the techniques used in this study.

Finally, techniques such as ISOMAP and Local Linear Embedding (LEE), are based on locally generating the manifold using the observed training points. These are techniques related to differential geometry, which generate a graph that allows surfaces to be modelled in a higher-dimensional space, without crossings or mixtures of regions in different parts of the surface. Although this type of technique may be of interest for understanding neighborhood relationships in the trajectories of values, it does not allow expressing risk with a linear relationship. Something similar happens with the t-distributed stochastic neighbor embedding (t-SNE), in this case, it is an embedding, not a projection, so a mesh is found based on a similarity measure based on proximity of probability distributions utilizing the Kullbach Leiber measure. Although this is a method that allows a reliable low-dimensional representation of the mesh that relates the entry points, it does not allow us to capture the dynamic risk relationships through a matrix, which is the objective of our work.

MATRIX PARALLELISM AMONG PCA, FA, ICA, AND NNPCA.

The four techniques used in this study, PCA, FA, ICA, and NNPCA,[4] can be classified as latent variable analysis, dimension reduction, or feature extraction techniques, whose main objective is to obtain some new underlying synthetic variables - from a set of observed data - capable of reproducing the behavior of the original data, in this context, returns on equities. Strictly speaking, a latent variable analysis technique tries to infer some unobservable artificial variables from a set of observable ones by using some mathematical models. On the other hand, the objective of a dimension reduction technique is only to reduce the dimensionality of the problem by selecting a fewer number of new artificial variables created by the combination of the original ones, via some mathematical or geometric transformation of the observed variables. Finally, a feature extraction technique seeks that the new variables extracted represent the main or most relevant and meaningful components or factors resulting from specific combinations of the observed ones.

Nevertheless, the purpose of this paper is to obtain a set of factors -hidden in the observed variables- to explain, in the best manner, why the returns on equities in the sample behave as they do. Consequently, any of the three approaches to classify

[4] The explanation of each class of analysis is out of the scope of this research since they have been discussed in former studies; this paper will focus only on the comparison among the four techniques. Nevertheless, the interested reader can find a general explanation of PCA, FA, ICA, and NNPCA in Ladrón de Guevara-Cortés & Torra-Porras (2014) and Ladrón de Guevara-Cortés, Torra-Porras & Monte-Moreno (2018, 2019).

Table 1.

Matrix parallelism among techniques to extract the underlying factors of systematic risk.

	Extraction Process	Generation Process	Attributes of the extracted components or factors.
Principal Component Analysis (PCA)	$Z = XA$	$Z = XA'$	1) Linearly uncorrelated components. 2) Linearly mixed.
Factor Analysis (FA)	$F = XC$ (Bartlett's model) $C = PQ$ $P = \Psi-1\Lambda$ $Q = (\Lambda'\Psi-1\Lambda)-1$	$X = 1\mu + FA'$	1) Linearly uncorrelated common factors. 2) Linearly mixed.
Independent Component Analysis (ICA)	$S = WX$	$X = AS$	1) Statistically independent components. 2) Linearly mixed.
Neural Networks Principal Component Analysis. (NNPCA)	$Z = W_2 g(W_1 X)$	$X = W_4 g(W_3 Z)$	1) Nonlinearly uncorrelated components. 2) Nonlinearly mixed.

Notes:

In PCA:
Z = Matrix of principal components.
X = Matrix of data.
A = Matrix of loadings.

In FA:
F = Matrix of common factors.
X = Matrix of data.
Λ = Matrix of loadings.
ψ = Matrix of specific variances or matrix of specificities or uniqueness.
μ = Vector of means.

In ICA:
S = Matrix of independent components or original sources.
X = Matrix of data.
W = Demixing matrix.
A = Mixing matrix.

In NNPCA:
Z = Matrix of nonlinear principal components.
X = Matrix of data.
W_1 = Matrix of weights from the first layer to the second layer.
W_2 = Matrix of weights from the second layer to the third layer.
W_3 = Matrix of weights from the third layer to the fourth layer.
W_4 = Matrix of weights from the fourth layer to the fifth layer.
g = Transferring nonlinear function.

Source: Author's elaboration.

these techniques fits properly as a method for extracting the main factors explaining the behavior of the returns on the equities of the sample. The four classes of analysis include two different processes, the extraction of the underlying factors process and the generation of the original variables process. Table 1 presents matrix parallelism among the extraction and generation processes employed in each technique and the main attributes of their extracted components or factors.

The same kind of analogy can be made to include PCA and FA in the former parallelism as well, taking the matrices of weights in the extraction process ($\mathbf{A}$ and $\mathbf{C}$), the matrices of the extracted components or factors ($\mathbf{Z}$ and $\mathbf{F}$), and the factor loading matrices in the generation process ($\mathbf{A}'$ and $\boldsymbol{\Lambda}'$), respectively. It is important to remark that, although there is matrix parallelism among the elements of these techniques, in this context, the direct comparison of their values is not homogeneous among all of them, e.g., the generation processes in PCA, FA, and ICA include only a linear mixing of the original data matrix and the demixing matrices; however, in NNPCA the process includes a non-linear combination of two matrices of weights and the original data matrix; thus, this technique does not have a single demixing matrix which, when multiplied directly by the data matrix, might produce the extracted factors. A similar situation occurs with the generation process, so it is necessary to use other methods to compare the four techniques, such as the reconstruction accuracy of the observed variables.

On the other hand, strictly speaking, the FA should not be compared directly with the rest of these techniques since the FA includes an independent term corresponding to the specific factors ($\mathbf{U}$), which is not considered in the rest of them.[5] The FA should be compared with the equivalent versions of the other techniques that consider an independent term in the model as well, e.g., the Noisy ICA (N-ICA) or Independent Factor Analysis (IFA) and the Non-linear Factor Analysis (NLFA). Nevertheless, PCA and FA have always been compared and in some cases even confused, since PCA is considered as a method of estimation within the FA, which is incorrect; thus, FA results were included in this review, too. The next step in further research will be to compare FA with the equivalent versions of the independent and non-linear models.

[5] The complete factor analysis model specification includes the matrix of specific factors $uX = 1\mu + F\Lambda' + U$, however, this paper cannot use this matrix in the generation process because it represents the error in the reconstruction of the original variables, which will be known after the reproduction of the variables by: $U = X -(1\mu + F\Lambda')$.

Finally, in the financial context, the most important differences among the four techniques are perhaps the attributes of the components or factors extracted, because they imply a progression from only linearly uncorrelated components in PCA to linearly uncorrelated common factors in FA, then to statistically independent components in ICA, and lastly to non-linearly uncorrelated components in NNPCA. From a theoretical standpoint, the former statement would imply the uncovering of a more realistic latent systematic risk factor structure, as one advance to more sophisticated techniques. This nature of the components or factors extracted through each technique is given mainly for the following conditions: First, while the orthogonal components extracted by using PCA explain the total amount of variance in the observed variables, the orthogonal factors produced by FA explain only the amount of variance explained by common factors, i.e., the covariance among the variables.

Nevertheless, both PCA and FA consider only the second-moment absence of linear correlation; on the other hand, ICA considers higher moment absences of linear correlation, which produce not only linearly uncorrelated components but also statistically independent ones. Finally, while the three former techniques only consider a linear mixing in the extraction and generation processes, NNPCA includes a nonlinear transformation in both processes, which generates not only linearly uncorrelated components but also non-linearly uncorrelated ones.

METHODOLOGY

The data.

The data used in the empirical study correspond to stocks of the Price and Quotation Index (IPC) of the Mexican Stock Exchange (BMV); Table 2 presents the list of the entire sample used in this study. Both the period analyzed and the shares selected respond to the following criteria: 1) the sample used in the cited former studies that allow us to make this comparative study of the results produced by each of the four techniques used in them, 2) the interest in a worldwide recognized pre-crisis period where stock prices were out of the effect of the subprime crisis formation, and 3) the availability of data among the diverse information sources consulted.

In this context, the basic aim was to build a homogeneous and sufficiently broad database, capable of being processed with the feature extraction techniques used in this study. Four different databases to test different expressions and periodicities of the returns on equities were built. On the one hand, two databases are expressed in returns, and the other two, in returns in excesses of the riskless interest rate. On the other hand, two of them have weekly periodicity and the other two a daily one. The weekly databases range from July 7, 2000, to January 27, 2006, and include 20 stocks and 291 observations; whereas the daily databases, from July 3, 2000, to January 27, 2006, contain 22 assets and 1491 quotations.

Extraction of underlying factors and reconstruction of the observed returns.

According to the models in Table 1, the first step was the extraction of the underlying factors by using Matlab® scripts,[6] obtaining also the matrices of weights for the extraction process or demixing matrices and the matrices of loadings of the generation process or mixing matrices. For the estimation of the models, this paper used the following specifications: in PCA, the classic linear version; in FA, the Maximum Likelihood method (MLFA); in ICA, the ICASSO software based on the FastICA algorithm; and in NNPCA, a hierarchical auto-associative neural network or autoencoder.[7] Secondly, the observed variables employing the extracted factors and the mixing matrices were reconstructed. This paper includes the experiments for the four techniques, the four databases, and a test window ranging from two to nine extracted factors.[8]

Measures of reconstruction accuracy.

To obtain a more objective measure of the accuracy of the reconstruction using the systematic risk factors obtained with each technique, some statistics widely employed to evaluate the accuracy of forecasting models in economy and finance were used,

[6] The PCA and FA scripts were elaborated using the functions included in the software; ICA scripts were adapted from Himberg & Hyvärinen (2005); and NNPCA, from Scholz (2006b).

[7] For details about the estimation models, see Ladrón de Guevara-Cortés & Torra-Porras (2014) and Ladrón de Guevara-Cortés, Torra-Porras & Monte-Moreno (2018, 2019).

[8] Since there is not a definite widespread criterion to define the best number of components to extract in all the techniques, nine different criteria usually accepted in PCA and FA literature were used. These criteria were: the eigenvalues arithmetic mean, the percentage of explained variance, the exclusion of the components or factors explaining a small amount of variance, the scree plot, the unretained eigenvalue contrast (Q statistic), the likelihood ratio contrast, Akaike's information criterion (AIC), the Bayesian information criterion (BIC), and the maximum number of components feasible to estimate in each technique. The comparable window across the four techniques indicated the results of the former criteria ranged from two to nine factors.

Table 2.

Stocks used in the study.

No.	Ticker	Name of the Company	Industrial Sector
1	ALFAA	Grupo Alfa	Holding
2	ARA*	Consorcio Ara	Construction: Housing
3	BIMBOA	Grupo Bimbo	Food processing
4	CEMEXCP (1)	Cemex	Cement
5	CIEB	Corporación Interamericana de Entretenimiento	Holding
6	COMERUBC	Controladora Comercial Mexicana	Commerce: retailing and wholesale
7	CONTAL*	Grupo Continental	Food and beverage processing
8	ELEKTRA*	Grupo Elektra	Commercial firms
9	FEMSAUBD	Fomento Económico Mexicano	Beer and beverage
10	GCARSOA1	Grupo Carso	Holding
11	GEOB	Corporación GEO	Construction: Housing
12	GFINBURO	Grupo Financiero Inbursa	Financial services
13	GFNORTEO	Grupo Financiero Banorte	Financial services
14	GMODELOC	Grupo Modelo	Food, tobacco and beverages
15	KIMBERA (1)	Kimberly-Clark de México	Cellulose and paper
16	PE&OLES*	Industrias Peñoles	Ferrous minerals
17	SORIANAB	Organización Soriana	Commerce: retailing and wholesale
18	TELECOA1	Carso Global Telecom	Communications
19	TELMEXL	Teléfonos de México	Communications
20	TLEVICPO	Grupo Televisa	Communications
21	TVAZTCPO	TV Azteca	Communications
22	WALMEXV	Wal-Mart de México	Commerce: retailing and wholesale
Stocks not included in the weekly databases responding to information availability.			

Source: Author's elaboration.

which in this context will represent measures of reconstruction accuracy. These measures, taken from Pérez & Torra (2001) and Diebold & López (1996), are the following: mean absolute error (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE), Theil's U statistic (U-Theil), confusion matrix (CM), confusion rate (CR), chi-squared contrast of independence, and Pesaran & Timmermann's directional accuracy statistic (DA).

The first four are measures of reconstruction accuracy, which represent different expressions to compute the error in the reconstruction of the observed returns; these are their mathematical formulations:

$$MAE = \frac{1}{H}\sum_{h=1}^{H}\left|r_{h-\hat{r}_h}\right| \qquad [1]$$

$$MAPE = \frac{1}{H}\sum_{h=1}^{H}\left|r_h - \hat{r}_h\right|/r_h \times 100 \qquad [2]$$

$$RMSE = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(r_h - \hat{r}_h)^2} \qquad [3]$$

$$U - Theil = RMSE \Big/ \left[\sqrt{\frac{1}{H}\sum_{h=1}^{H} r_h^2} + \sqrt{\frac{1}{H}\sum_{h=1}^{H} \hat{r}_h^2}\right] \qquad [4]$$

Where H denotes the total number of observations; h = 1, ..., H; r_h are the observed returns and $\hat{r}_h$, the reconstructed returns.

The confusion matrix is a contingency table necessary to compute the contrasts for evaluating the direction-of-change reconstruction measures, namely, confusion rate and chi-squared contrast; it is constructed in this manner:

	rh_reconstructed		
		≥ 0	< 0
rh_real	≥ 0	n00	n01
	< 0	n10	n11

[5]

Where n_{ij} indicates the absolute frequency of occurrence of each condition.

The confusion rate shows the percentage of incorrect reconstructions and is calculated by:

$$CR = (n_{01} + n_{10})/H \qquad [6]$$

The chi-squared ($\hat{x}^2$) contrast assumes a null hypothesis of independence between the signs of the reconstruction and their real values; therefore, the rejection of the null hypothesis and the high values of the statistic imply a good performance based on the direction-of-change reconstruction; its formulation is as follows [9]:

$$\hat{\chi}^2 = \sum_{i=0}^{1} \sum_{j=0}^{1} [n_{ij} - n_{i.}n_{.j}H]^2 / [n_{i.}n_{.j}/H] \qquad [7]$$

Where $n_{i.}$ and $n_{.j}$ are the marginal frequencies.

Finally, the DA statistic is another directional accuracy reconstruction measure, with distribution N (0,1), which poses a null hypothesis of independence between the observed and the reconstructed values; its interpretation is similar to the former contrast and is built as follows:

$$DA = [\mathrm{var}(SR) - \mathrm{var}(SRI)]{-0.5}(SR - SRI) \qquad [8]$$

$$SR = H^{-1} \sum_{h=1}^{H} I_i[y_h \cdot \hat{y}_h > 0] \qquad [9]$$

$$SRI = p\hat{p} + (1 - p)(1 - \hat{p}) \qquad [10]$$

$$p = H^{-1} \sum_{h=1}^{H} I_i\,[y_h > 0] \qquad [11]$$

$$\hat{p} = H^{-1} \sum_{h=1}^{H} I_i\,[\hat{y}_h > 0] \qquad [12]$$

$$\mathrm{var}(SR) = H^{-1}[SRI(1 - SRI)] \qquad [13]$$

$$\mathrm{var}(SRI) = H^{-2}[H(2\hat{p} - 1)^2 p(1 - p) + (2p - 1)^2 \hat{p}(1 - p) + 4p\hat{p}(1 - p)(1 - \hat{p})] \qquad [14]$$

Where *SR* denotes the success ratio; *SRI*, the success ratio in the case of independence between the observed and reconstructed values under the null hypothesis, and *I* is an indicative function denoting the occurrence of the condition imposed inside the square brackets.[10]

[9] The degrees of freedom for this contrast are calculated by: $v = (r{-}1)(k{-}1)$, where v denotes the degrees of freedom; r, the number of rows of the confusion matrix; and k, the number of columns.

[10] If the condition is fulfilled, the indicator takes the value of 1.

RESULTS

Graphic analysis.

The results obtained in the reconstruction of the observed returns using the four techniques individually were outstanding at first sight for all of them, making it difficult to determine which one was the best. Figures 1 and 2 present the observed *versus* the reconstructed returns produced by the four techniques, from the first eight stocks of the database of weekly returns when nine factors were extracted.[11] The line plots include all the observations, showing that in general, all the techniques reproduce the real values successfully for the entire period; nevertheless, if a zoom of stem plots is done for the first 50 observations, it can be distinguished that FA and ICA present greater errors in the reconstruction.

Derived from the visual plot analysis it can be detected that, given the number of factors extracted, the four techniques fail to reproduce the highest and lowest peaks in the observations, but, if the number of factors extracted is increased from all the techniques, this problem disappears.[12] In addition, it can be observed that in some cases the best reconstruction of each asset is not produced by the same technique, i.e., while some stocks are reconstructed better by one technique, other shares are better reproduced through another method. All the former results are similar for the entire case of the experiments.

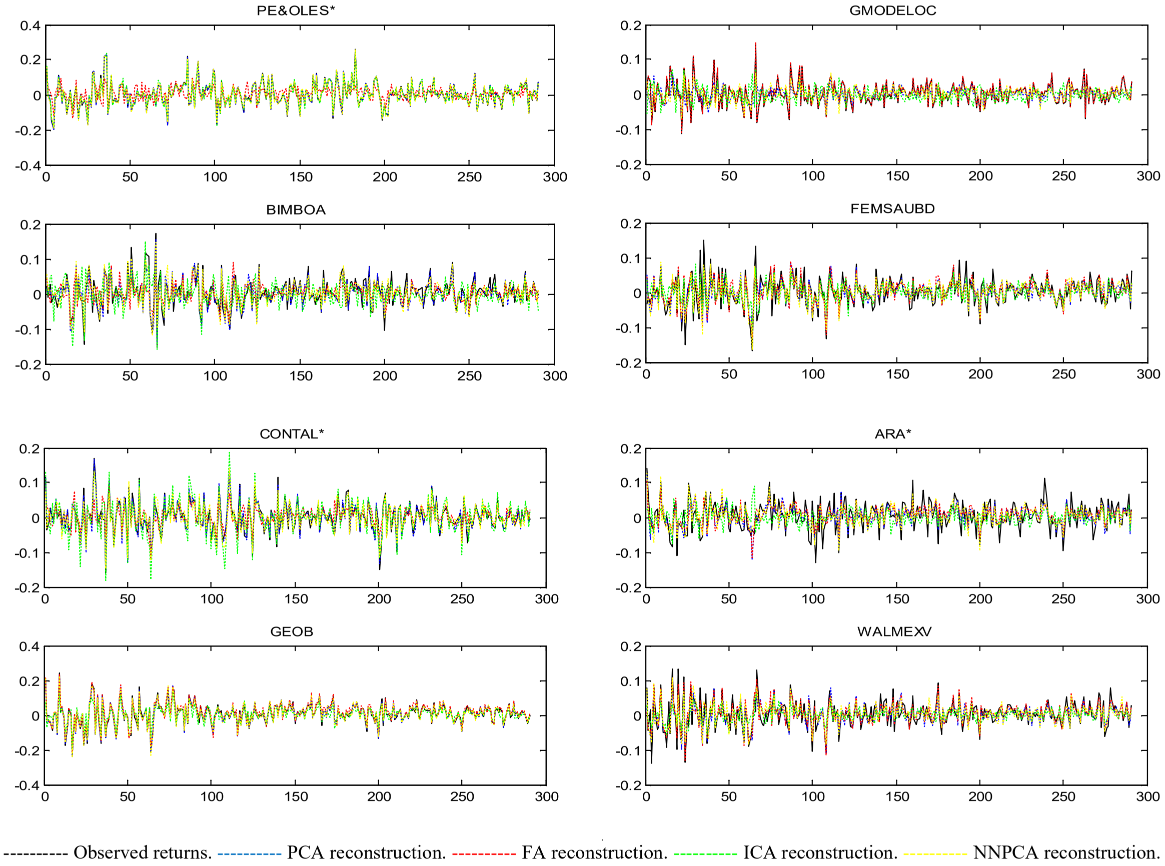
Measures of reconstruction analysis.

All the foregoing measures of accuracy for each stock were computed as well as the arithmetic mean, median, and standard deviation for the MAE, MAPE, RMSE, U-Theil, and CR as proposed synthetic global measures to evaluate the errors in reconstruction for all the assets. In addition, this paper also analyses the results of the directional accuracy statistics χ^2 and DA individually for each stock to test the

[11] For the sake of saving space in this paper, only the results for the database of weekly returns when nine factors were extracted will be presented explicitly; nevertheless, the results and conclusions reported include the entire cases. The rest of the plots are available upon request.

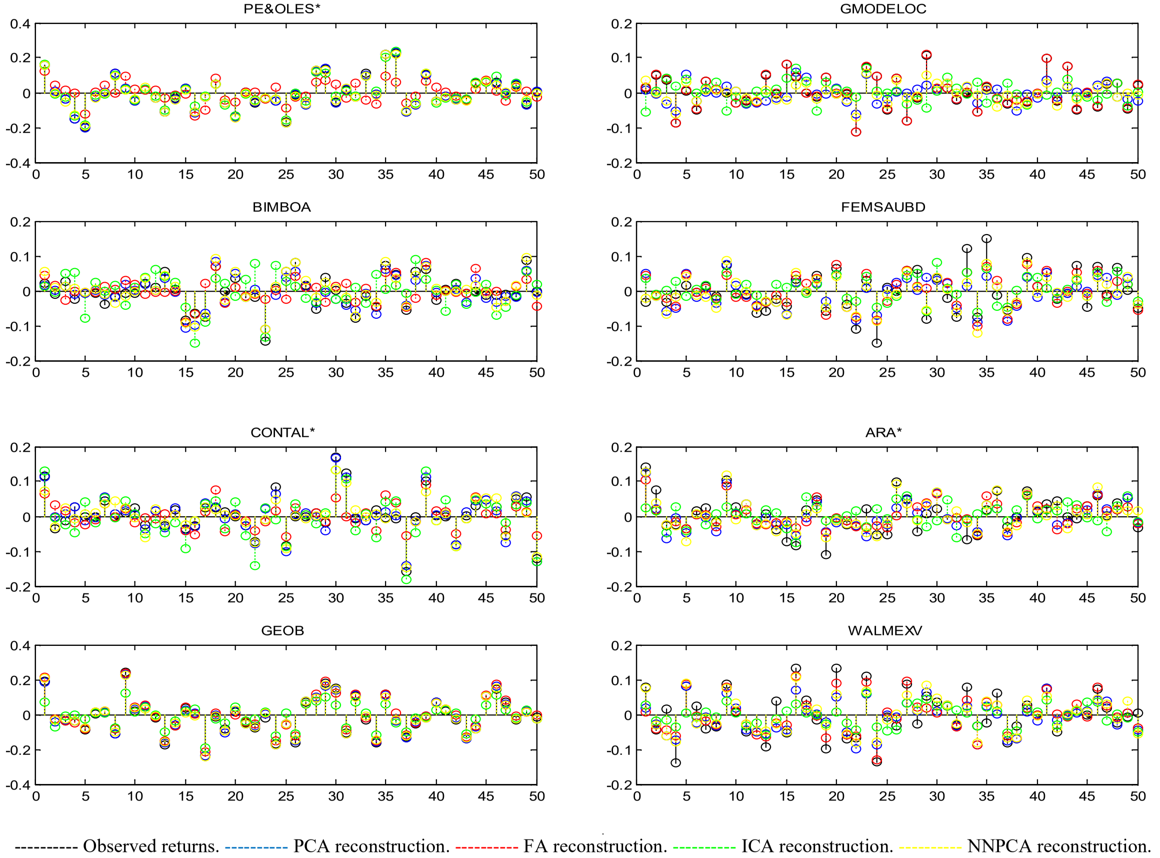
[12] The results of those additional experiments are not presented in this study, those experiments were done only to test that the reproduction capacity of the techniques, considering all the factors feasible to compute in each one of them.

Figure 1. Observed vs. reconstructed returns. Database of weekly returns. Nine underlying factors were extracted. Line plots.



Source: Author's elaboration.

Figure 2. Observed vs. reconstructed returns. Database of weekly returns. Nine underlying factors were extracted. Stem plots.



Source: Author's elaboration.

null hypothesis of independence in the reconstruction process. Therefore, all these calculations for the four extraction techniques, the four databases, and the entire testing window are replicated.

Tables 3 to 6 present the results of the foregoing experiments applied on the database of weekly returns, when nine factors were extracted, for PCA, FA, ICA, and NNPCA, respectively. First of all, it is important to remark that the results for all the techniques are outstanding and reflect a high-quality reconstruction of the returns; however, in trying to find the best of these methods the following distinctions are important. In general, regarding the measures of reconstruction accuracy MAE, MAPE, RMSE, and U-Theil, the smaller errors in the reconstruction - in terms of their arithmetic mean - points to PCA and NNPCA as the best ones. Strictly speaking, PCA scored better results in all the foregoing measures except the U-Theil statistic, but the difference between both techniques in the computed error is really small. However, NNPCA presents a smaller standard deviation of the former statistics, which means less sensitivity to the variations of mean values of the proposed synthetic measures.

In addition, considering that the observed variables are not normally distributed and that the median is a more suitable synthetic measure of the reconstruction accuracy, in this case, NNPCA beat PCA in all the foregoing measures except the MAPE. Regarding the CR, the results are similar; PCA obtained the lowest percentage of incorrect reconstruction in terms of mean, and NNPCA in terms of the median. Concerning the directional accuracy contrasts χ^2 and DA for each stock, the findings of this research show that in almost all the cases the null hypothesis of independence at 5% level of statistical significance is rejected in both tests; therefore, an association can be established between the signs of the predictions and the real values of the returns.[13]

In summary, in almost all cases of the study, the results point to NNPCA as the best technique for the reconstruction in terms of the mean when a smaller number of factors is retained; and to PCA, when a larger number of them are extracted, which leads us to think that NNPCA performs better than the other techniques as

[13] The null hypothesis of independence of the X^2 and DA contrast is rejected in almost all cases; nevertheless, for some specific stocks, the null hypothesis could not be rejected. The effect of these few cases does not significantly affect the overall results and conclusions derived from these statistics.

a dimensional reduction or feature extraction technique. In terms of the median, NNPCA surpasses the rest of the techniques in almost all the cases; besides, in the daily databases, NNPCA shows clearer supremacy over the other techniques in almost all the measures in terms of mean, median, and standard deviation. Nevertheless, this is not a rule, and for some databases, a particular number of factors, and specific measures of accuracy, the results point to other techniques as the best ones.

Additionally, to analyze the performance of each technique in the individual reproduction of the observed variables, the results of the MAE, MAPE, RMSE, U-Theil, and CR obtained in PCA were taken as benchmarks. Then, this set of benchmarks were confronted with the results from the same measures obtained with the rest of the techniques by subtracting the former from the latter. Tables 7 to 9 present said results. The findings of this research reveal that, in terms of the individual reconstruction of the observed returns, in a comparison between FA vs. PCA, ICA vs. PCA, and NNPCA vs. PCA, 50% of the stocks performed equally across these techniques; FA only surpass PCA in 20% of the reproductions –in almost all the measures–, ICA in about 5% - 10% and NNPCA in around 10% - 30%. The former results were similar in the totality of the cases and samples in the study.

Table 3.

Measures of reconstruction accuracy.

Database of weekly returns. Nine underlying factors extracted by Principal Component Analysis.

	PE&OLES*	BIMBOA	GMODELOC	FEMSAUBD	CONTAL*	GEOB	ARA*	WALMEXV	SORIANAB	COMERUBC	ELEKTRA*	TELMEXL	TELECOA1
MAE	0.0024	0.0123	0.0199	0.0220	0.0076	0.0068	0.0225	0.0193	0.0204	0.0219	0.0112	0.0141	0.0157
MAPE	24.3863	148.8622	148.6442	229.4007	83.0108	47.0603	117.8427	163.8752	131.1541	151.0114	82.6411	120.0802	130.2533
RMSE	0.0030	0.0156	0.0269	0.0298	0.0098	0.0086	0.0292	0.0256	0.0261	0.0295	0.0143	0.0185	0.0200
U-Theil	0.0226	0.1921	0.5407	0.4124	0.1136	0.0686	0.4216	0.3628	0.3303	0.3697	0.1276	0.3026	0.2371
CM	145 3	129 28	102 52	125 30	139 15	165 7	132 36	125 29	125 29	120 32	140 14	125 23	134 19
	1 142	23 111	46 91	43 93	14 123	8 111	35 88	41 96	31 106	35 104	18 119	35 108	24 114
CR	0.0137	0.1753	0.3368	0.2509	0.0997	0.0515	0.2440	0.2405	0.2062	0.2302	0.1100	0.1993	0.1478
χ^2	275.2710	122.4291	30.9380	71.3557	186.2886	232.1900	72.9331	77.6900	99.9074	84.3181	176.7072	105.7230	144.0490
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DA	-0.2926	-2.2143	-5.4731	-2.7942	-1.4488	1.5618	-2.7724	-2.7778	-2.6451	-3.3658	-0.8320	-2.5200	-1.4214
p-value	0.3849	0.0134	0.0000	0.0026	0.0737	0.9408	0.0028	0.0027	0.0041	0.0004	0.2027	0.0059	0.0776

	TLEVICPO	TVAZTCPO	GFNORTEO	GFINBURO	GCARSOA1	ALFAA	CIEB	MEAN	MEDIAN	STD. DEV.
MAE	0.0181	0.0178	0.0238	0.0206	0.0194	0.0032	0.0187	0.015890	0.018362	0.006539
MAPE	159.9367	89.1008	113.6115	172.8017	156.5530	31.3949	217.5848	125.960296	130.703660	54.938655
RMSE	0.0240	0.0234	0.0304	0.0263	0.0255	0.0041	0.0241	0.020741	0.024069	0.008668
U-Theil	0.2720	0.2339	0.4020	0.3455	0.3152	0.0329	0.2542	0.267868	0.287273	0.140738
CM	142 16	130 20	139 29	130 28	134 32	152 3	131 22			
	21 112	20 121	33 90	36 97	31 94	5 131	26 112			
CR	0.1271	0.1375	0.2131	0.2199	0.2165	0.0275	0.1649	0.170619	0.187285	0.082030
χ^2	160.8616	152.8821	91.8317	89.8322	90.8314	259.7725	130.2436			
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000			
DA	-0.6675	-2.1683	-1.3742	-2.4386	-1.8620	0.6866	-2.0664			
p-value	0.2522	0.0151	0.0847	0.0074	0.0313	0.7538	0.0194			

Notes: MAE: Mean absolute error. MAPE: Mean absolute percentage error. RMSE: Root mean square error. U-*Theil*: Theil's U statistic. CM: Confusion matrix. CR: Confusion rate. χ^2: Chi-squared independence contrast statistic. DA: Pesaran & Timmerman's directional accuracy statistic. Marked cells represent the best results for each statistic across the four techniques.

Source: Author's elaboration.

Table 4.

Measures of reconstruction accuracy.

Database of weekly returns. Nine underlying factors extracted by Factor Analysis.

	PE&OLES*	BIMBOA	GMODELOC	FEMSAUBD	CONTAL*	GEOB	ARA*	WALMEXV	SORIANAB	COMERUBC	ELEKTRA*	TELMEXL	TELECOA1
MAE	0.04365	0.0242	0.001860577	0.0184815	0.025769	0.013129	0.02513	0.0152253	0.0210067	0.00226127	0.025111	0.012472	0.0038492
MAPE	236.3302	211.1015	26.88911528	183.94481	194.3088	111.2997	138.092	136.07734	146.49039	23.494415	146.7221	109.6418	30.343345
RMSE	0.056267	0.032712	0.00186107	0.024807	0.03448	0.016063	0.03201	0.0203274	0.0268838	0.00226227	0.031976	0.017011	0.0050743
U-Theil	0.535157	0.475224	0.028879393	0.3220112	0.4855	0.126707	0.47565	0.2722695	0.3428755	0.02487502	0.305703	0.273518	0.0572872
CM	106 42	108 49	154 0	135 20	123 31	166 6	145 23	135 19	130 24	152 0	134 20	132 16	150 3
	67 76	46 88	6 131	35 101	49 88	27 92	48 75	40 97	33 104	8 131	47 90	29 114	6 132
CR	0.37457	0.32646	0.020618557	0.1890034	0.274914	0.113402	0.24399	0.2027491	0.1958763	0.02749141	0.230241	0.154639	0.0309278
χ2	18.50807	34.46063	267.8208942	112.2848	58.34573	171.6181	71.0842	103.37939	107.0963	260.539209	85.66236	139.7267	256.09701
p-value	1.69E-05	4.35E-09	0	0	2.2E-14	0	0	0	0	0	0	0	0
DA	-5.026474	-4.816167	0.58447958	-1.738866	-3.20851	1.469671	-1.5632	-1.537311	-2.418094	0.15940352	-1.70586	-1.811926	0.474646
p-value	2.5E-07	7.32E-07	0.72055115	0.0410292	0.000667	0.929175	0.05901	0.0621086	0.007801	0.56332452	0.044017	0.034999	0.6824803

	TLEVICPO	TVAZTCPO	GFNORTEO	GFINBURO	GCARSOA1	ALFAA	CIEB	MEAN	MEDIAN	STD. DEV.
MAE	0.0165638	0.00034505	0.0238034	0.0222714	0.01895164	0.03045	0.010008	0.017727	0.018717	0.010814
MAPE	152.26727	2.18567611	118.31492	167.65046	162.561709	234.033	98.90365	131.5326	142.2914	68.61165
RMSE	0.0218104	0.00037606	0.0303112	0.0290221	0.02389145	0.04082	0.01309	0.023053	0.024349	0.014183
U-Theil	0.24323	0.00357042	0.3922878	0.3920611	0.28960269	0.37273	0.131864	0.27755	0.297653	0.166213
CM	145 13	145 5	155 13	130 28	141 25	134 21	135 18			
	24 109	0 141	55 68	51 82	33 92	54 82	16 122			
CR	0.1271478	0.01718213	0.233677	0.2714777	0.19931271	0.25773	0.116838	0.180412	0.197595	0.103289
χ2	161.21623	271.666438	79.917042	59.279508	101.643086	69.2233	170.719			
p-value	0	0	0	1.366E-14	0	1.1E-16	0			
DA	-0.361109	-0.1767725	-0.4997719	-2.5831351	-1.1655387	-1.8399	-1.58885			
p-value	0.3590089	0.42984355	0.3086179	0.0048953	0.1219005	0.03289	0.056047			

Notes: MAE: Mean absolute error. MAPE: Mean absolute percentage error. RMSE: Root mean square error. U-Theil: Theil's U statistic. CM: Confusion matrix. CR: Confusion rate. χ2: Chi-squared independence contrast statistic. DA: Pesaran & Timmerman's directional accuracy statistic. Marked cells represent the best results for each statistic across the four techniques.

Source: Author's elaboration.

Table 5.

Measures of reconstruction accuracy.

Database of weekly returns. Nine underlying factors extracted by Independent Component Analysis.

	PE&OLES*	BIMBOA	GMODELOC	FEMSAUBD	CONTAL*	GEOB	ARA*	WALMEXV	SORIANAB	COMERUBC	ELEKTRA*	TELMEXL
MAE	0.0084	0.0293	0.0264	0.0220	0.0195	0.0242	0.0296	0.0248	0.0282	0.0270	0.0214	0.0142
MAPE	64.7501	295.9328	244.9408	160.3897	203.1245	136.9776	151.3835	139.9100	155.4740	167.3655	108.3166	128.1428
RMSE	0.0108	0.0378	0.0340	0.0289	0.0260	0.0319	0.0382	0.0325	0.0368	0.0359	0.0274	0.0183
U-Theil	0.0820	0.4886	0.6544	0.4200	0.2746	0.3037	0.6185	0.5559	0.5680	0.5181	0.2827	0.2869
CM	139 9	95 62	68 86	114 41	126 28	149 23	109 59	112 42	97 57	107 45	124 30	125 23
	10 133	43 91	50 87	39 97	27 110	27 92	39 84	55 82	46 91	38 101	19 118	35 108
CR	0.0653	0.3608	0.4674	0.2749	0.1890	0.1718	0.3368	0.3333	0.3540	0.2852	0.1684	0.1993
χ2	219.9453	23.4194	1.7644	58.4982	112.1732	120.3060	31.2677	31.4726	25.0921	53.8406	128.8689	105.7230
p-value	0.0000	0.0000	0.1841	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DA	-0.8740	-6.1286	-9.2642	-3.8614	-3.0952	-0.4115	-5.1348	-4.3296	-6.4027	-4.6385	-2.5123	-2.6381
p-value	0.1911	0.0000	0.0000	0.0001	0.0010	0.3403	0.0000	0.0000	0.0000	0.0000	0.0060	0.0042

	TELECOA1	TLEVICPO	TVAZTCPO	GFNORTEO	GFINBURO	GCARSOA1	ALFAA	CIEB	MEAN	MEDIAN	STD. DEV.
MAE	0.0178	0.0205	0.0121	0.0253	0.0233	0.0212	0.0122	0.0178	0.0212516	0.021704	0.006041
MAPE	113.1419	150.8336	66.1874	122.0431	214.7183	134.3615	90.5096	165.4260	150.69645	145.3718	56.51906
RMSE	0.0224	0.0285	0.0155	0.0317	0.0306	0.0283	0.0161	0.0234	0.0277472	0.028662	0.007873
U-Theil	0.2917	0.3354	0.1475	0.4200	0.3222	0.3856	0.1404	0.2584	0.3677257	0.328806	0.161044
CM	127 26	140 18	134 16	137 31	133 25	128 38	138 17	132 21			
	29 109	30 103	10 131	37 86	41 92	29 96	8 128	30 108			
CR	0.1890	0.1649	0.0893	0.2337	0.2268	0.2302	0.0859	0.1753	0.2300687	0.213058	0.103204
χ2	112.1184	129.6967	196.6535	78.2381	85.4944	83.4104	200.3393	122.4488			
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000			
DA	-2.2336	-1.2033	-1.8223	-2.0039	-2.1520	-2.3332	-1.0017	-2.0753			
p-value	0.0128	0.1144	0.0342	0.0225	0.0157	0.0098	0.1582	0.0190			

Notes: MAE: Mean absolute error. MAPE: Mean absolute percentage error. RMSE: Root mean square error. U-Theil: Theil's U statistic. CM: Confusion matrix. CR: Confusion rate χ2: Chi-squared independence contrast statistic. DA: Pesaran & Timmerman's directional accuracy statistic. Marked cells represent the best results for each statistic across the four techniques.

Source: Author's elaboration.

Table 6.

Measures of reconstruction accuracy.

Database of weekly returns. Nine underlying factors extracted by Neural Networks Principal Component Analysis.

	PE&OLES*	BIMBOA	GMODELOC	FEMSAUBD	CONTAL*	GEOB	ARA*	WALMEXV	SORIANAB	COMERUBC	ELEKTRA*	TELMEXL
MAE	0.0052	0.0151	0.0190	0.0195	0.0142	0.0090	0.0222	0.0197	0.0204	0.0176	0.0133	0.0139
MAPE	51.1458	160.0609	161.1749	206.1618	145.5141	59.4703	124.7358	176.8132	133.1683	132.5406	94.8276	120.0370
RMSE	0.0069	0.0190	0.0252	0.0261	0.0187	0.0113	0.0280	0.0252	0.0256	0.0226	0.0172	0.0184
U-Theil	0.0514	0.2356	0.4886	0.3475	0.2214	0.0904	0.3960	0.3558	0.3235	0.2632	0.1546	0.2984
CM	137 11	130 27	120 34	128 27	124 30	164 8	140 28	131 23	121 33	117 35	142 12	124 24
	8 135	30 104	48 89	45 91	18 119	10 109	36 87	45 92	33 104	25 114	17 120	34 109
CR	0.0653	0.1959	0.2818	0.2474	0.1649	0.0619	0.2199	0.2337	0.2268	0.2062	0.0997	0.1993
χ2	220.0596	106.6076	54.6440	73.6059	131.7443	221.1705	86.8420	82.7119	86.3830	101.1126	186.2767	105.5344
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DA	-1.2342	-2.0357	-3.4890	-2.2399	-3.4755	1.5446	-1.8161	-2.1455	-3.2318	-3.4614	-0.6019	-2.6371
p-value	0.1086	0.0209	0.0002	0.0125	0.0003	0.9388	0.0347	0.0160	0.0006	0.0003	0.2736	0.0042
	TELECOA1	TLEVICPO	TVAZTCPO	GFNORTEO	GFINBURO	GCARSOA1	ALFAA	CIEB	MEAN	MEDIAN	STD. DEV.	
MAE	0.0165	0.0178	0.0174	0.0180	0.0210	0.0209	0.0087	0.0138	0.01616	0.017492	0.00454	
MAPE	138.6046	149.9463	84.2911	110.6404	175.8470	192.1690	88.1679	132.3092	131.8813	132.8544	41.84477	
RMSE	0.0205	0.0238	0.0232	0.0225	0.0265	0.0273	0.0112	0.0175	0.020835	0.022573	0.005842	
U-Theil	0.2433	0.2689	0.2296	0.2700	0.3476	0.3433	0.0916	0.1784	0.25996	0.266042	0.109921	
CM	129 24	137 21	132 18	147 21	124 34	138 28	143 12	133 20				
	25 113	22 111	17 124	22 101	38 95	37 88	15 121	22 116				
CR	0.1684	0.1478	0.1203	0.1478	0.2474	0.2234	0.0928	0.1443	0.174742	0.182131	0.063749	
χ2	127.6172	143.4507	167.7673	141.3357	72.8895	85.2414	192.6076	146.8721				
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000				
DA	-1.9858	-1.2245	-2.0516	-0.4957	-3.1148	-1.5406	-0.4987	-1.7104				
p-value	0.0235	0.1104	0.0201	0.3101	0.0009	0.0617	0.3090	0.0436				

Notes: MAE: Mean absolute error. MAPE: Mean absolute percentage error. RMSE: Root mean square error. U-Theil: Theil's U statistic. CM: Confusion matrix. CR: Confusion rate χ2: Chi-squared independence contrast statistic. DA: Pesaran & Timmerman's directional accuracy statistic. Marked cells represent the best results for each statistic across the four techniques.

Source: Own elaboration.

Table 7.

Factor Analysis (FA) vs. Principal Component Analysis (PCA).

Measures of reconstruction accuracy obtained in FA minus measures of reconstruction accuracy obtained in PCA.

	PE&OLES*	BIMBOA	GMODELOC	FEMSAUBD	CONTAL*	GEOB	ARA*	WALMEXV	SORIANAB	COMERUBC	ELEKTRA*	TELMEXL	TELECOA1
MAE	0.041267	0.000000	0.011866	0.000000	-0.018012	0.000000	-0.003515	0.000000	0.018217	0.000000	0.006300	0.000000	0.002600
MAPE	211.943910	0.000000	62.239320	0.000000	-121.755131	0.000000	-45.455934	0.000000	111.298008	0.000000	64.239433	0.000000	20.249655
RMSE	0.053226	0.000000	0.017069	0.000000	-0.025029	0.000000	-0.005036	0.000000	0.024655	0.000000	0.007417	0.000000	0.002817
U-Theil	0.512601	0.000000	0.283107	0.000000	-0.511798	0.000000	-0.090409	0.000000	0.371920	0.000000	0.058101	0.000000	0.054066
CR	0.182131	0.000000	-0.147766	0.000000	-0.106529	0.000000	-0.096220	0.000000	-0.068729	0.000000	0.075601	0.000000	-0.226804
								FA > PCA		FA = PCA		FA < PCA	
	TLEVICPO	TVAZTCPO	GFNORTEO	GFINBURO	GCARSOA1	ALFAA	CIEB	Num.	%	Num.	%	Num.	%
MAE	0.000000	-0.004110	0.000000	0.000630	0.000000	-0.019683	0.000000	6	30%	10	50%	4	20%
MAPE	0.000000	-27.797823	0.000000	15.336333	0.000000	-127.517004	0.000000	6	30%	10	50%	4	20%
RMSE	0.000000	-0.005238	0.000000	0.000815	0.000000	-0.027252	0.000000	6	30%	10	50%	4	20%
U-Theil	0.000000	-0.090554	0.000000	0.012559	0.000000	-0.344814	0.000000	6	30%	10	50%	4	20%
CR	0.000000	-0.006873	0.000000	0.065292	0.000000	-0.030928	0.000000	3	15%	10	50%	7	35%

Notes: FA > PCA: Cases where FA reproduces worse than PCA. i.e., FA's error in reproduction is greater than PCA's one.

FA = PCA: Cases where FA reproduce just the same as PCA. i.e., FA's error in reproduction is equal to PCA's one.

FA < PCA: Cases where FA reproduce better than PCA. i.e., FA's error in reproduction is less than PCA's one.

Source: Author's elaboration.

Table 8.

Independent Component Analysis (ICA) vs. Principal Component Analysis (PCA).

Measures of reconstruction accuracy obtained in ICA minus measures of reconstruction accuracy obtained in PCA.

	PE&OLES*	BIMBOA	GMODELOC	FEMSAUBD	CONTAL*	GEOB	ARA*	WALMEXV	SORIANAB	COMERUBC	ELEKTRA*	TELMEXL	TELECOA1
MAE	0.005984	0.000000	0.016928	0.000000	0.006492	0.000000	-0.000021	0.000000	0.011908	0.000000	0.017351	0.000000	0.007094
MAPE	40.363779	0.000000	147.070553	0.000000	96.296550	0.000000	-69.011048	0.000000	120.113756	0.000000	89.917366	0.000000	33.540749
RMSE	0.007710	0.000000	0.022171	0.000000	0.007074	0.000000	-0.000986	0.000000	0.016158	0.000000	0.023273	0.000000	0.008961
U-Theil	0.059434	0.000000	0.296461	0.000000	0.113678	0.000000	0.007623	0.000000	0.160974	0.000000	0.235103	0.000000	0.196886
CR	0.051546	0.000000	0.185567	0.000000	0.130584	0.000000	0.024055	0.000000	0.089347	0.000000	0.120275	0.000000	0.092784
								ICA > PCA		ICA = PCA		ICA< PCA	
	TLEVICPO	TVAZTCPO	GFNORTEO	GFINBURO	GCARSOA1	ALFAA	CIEB	Num.	%	Num.	%	Num.	%
MAE	0.000000	0.005476	0.000000	0.007847	0.000000	0.005008	0.000000	9	45%	10	50%	1	5%
MAPE	0.000000	-23.965160	0.000000	24.319920	0.000000	16.354031	0.000000	8	40%	10	50%	2	10%
RMSE	0.000000	0.006965	0.000000	0.010777	0.000000	0.006396	0.000000	9	45%	10	50%	1	5%
U-Theil	0.000000	0.193085	0.000000	0.237705	0.000000	0.148364	0.000000	10	50%	10	50%	0	0%
CR	0.000000	0.092784	0.000000	0.147766	0.000000	0.054983	0.000000	10	50%	10	50%	0	0%

Notes: ICA > PCA: Cases where ICA reproduce worse than PCA. i.e., ICA's error in reproduction is greater than PCA's one.

ICA = PCA: Cases where ICA reproduce just the same as PCA. i.e., ICA's error in reproduction is equal to PCA's one.

ICA < PCA: Cases where ICA reproduce better than PCA. i.e., ICA's error in reproduction is less than PCA's one.

Source: Author's elaboration.

Table 9.

Neural Networks Principal Component Analysis (NNPCA) vs. Principal Component Analysis (PCA).

Measures of reconstruction accuracy obtained in NNPCA minus measures of reconstruction accuracy obtained in PCA.

	PE&OLES*	BIMBOA	GMODELOC	FEMSAUBD	CONTAL*	GEOB	ARA*	WALMEXV	SORIANAB	COMERUBC	ELEKTRA*	TELMEXL	TELECOA1
MAE	0.002834	0.000000	0.002735	0.000000	-0.000915	0.000000	-0.002446	0.000000	0.006667	0.000000	0.002218	0.000000	-0.000298
MAPE	26.759518	0.000000	11.198693	0.000000	12.530643	0.000000	-23.238905	0.000000	62.503355	0.000000	12.410033	0.000000	6.893027
RMSE	0.003870	0.000000	0.003358	0.000000	-0.001718	0.000000	-0.003784	0.000000	0.008836	0.000000	0.002694	0.000000	-0.001172
U-Theil	0.028865	0.000000	0.043488	0.000000	-0.052073	0.000000	-0.064887	0.000000	0.107867	0.000000	0.021756	0.000000	-0.025570
CR	0.051546	0.000000	0.020619	0.000000	-0.054983	0.000000	-0.003436	0.000000	0.065292	0.000000	0.010309	0.000000	-0.024055
								NNPCA > PCA		NNPCA = PCA		NNPCA < PCA	
	TLEVICPO	TVAZTCPO	GFNORTEO	GFINBURO	GCARSOA1	ALFAA	CIEB	Num.	%	Num.	%	Num.	%
MAE	0.000000	0.000394	0.000000	-0.000022	0.000000	-0.004385	0.000000	5	25%	10	50%	5	25%
MAPE	0.000000	12.938022	0.000000	2.014203	0.000000	-18.470846	0.000000	8	40%	10	50%	2	10%
RMSE	0.000000	-0.000346	0.000000	-0.000487	0.000000	-0.006875	0.000000	4	20%	10	50%	6	30%
U-Theil	0.000000	-0.007039	0.000000	-0.006815	0.000000	-0.106481	0.000000	4	20%	10	50%	6	30%
CR	0.000000	-0.006873	0.000000	0.020619	0.000000	-0.024055	0.000000	5	25%	10	50%	5	25%

Notes: NNPCA > PCA: Cases where NNPCA reproduce worse than PCA. i.e., NNPCA's error in reproduction is greater than PCA's one.

NNPCA = PCA: Cases where NNPCA reproduce just the same as PCA. i.e., NNPCA's error in reproduction is equal to PCA's one.

NNPCA < PCA: Cases where NNPCA reproduce better than PCA. i.e., NNPCA's error in reproduction is less than PCA's one.

Source: Own elaboration.

CONCLUSIONS

From the theoretical standpoint, NNPCA constitutes the best technique, since the underlying factors extracted present better attributes; they are nonlinearly uncorrelated, warranting not only linearly uncorrelated systematic risk factors for the Arbitrage Pricing Theory (APT) model but also nonlinearly uncorrelated ones.

However, the findings in the empirical study do not demonstrate clear supremacy of one technique over the others since all the techniques successfully reproduce the observed returns; nevertheless, broadly speaking and based on its theoretical supremacy and the evidence uncovered, NNPCA can be pointed out as the best technique to reconstruct the observed returns on equities of the sample.

MLFA was the technique with the worst performance in the reconstruction; although its results were good enough, the other techniques simply performed better. However, the clarification stated in the section on matrix parallelism about the direct comparison of FA with the other kinds of analysis used in this study must not be forgotten. A future step in the research will be to compare FA to its equivalent versions for the independent and non-linear models.

According to the attributes of the components or factors produced by each technique, it can be expected that the results in the reconstruction should be better as one moves from basic techniques such as PCA and FA to advanced methods like ICA and NNPCA. However, in general, the ICA reconstruction was worse than the PCA in terms of the first four measures of reconstruction accuracy in almost all cases. Further research will be necessary to find out the reasons for these results.

Additionally, it can be concluded that the four techniques performed a successful reconstruction of the observed returns; nevertheless, the supremacy of one of them over the others is very sensitive to the number of components or factors retained, the expression of the model, and the specific asset analyzed. Consequently, it might be stated that the selection of one technique or the other will depend mainly on the number of dimensions to retain and the specific stock object of study; nevertheless, further research concerning this issue will be necessary.

Finally, some natural expansions of this work would be the search for some other measures to evaluate the accuracy of the reproduction – both in univariate and in multivariate terms – and some other methodologies to compare the results of the four techniques; a deeper study regarding the univariate and multivariate statistics and the morphology of the components and factors extracted; and the interpretation of the underlying factors of systematic risk, namely, the risk attribution process. Likewise, after having tested these techniques in a pre-crisis period free of almost any prices distortion originated by natural speculative movements during a crisis period, other extensions of this research would be the testing the accuracy of the reproduction of the observed returns produced by the multifactor generative models of returns generated by each technique in both a crisis and a post-crisis period.

ACKNOWLEDGEMENTS

The authors acknowledge Cristina Ubago at Gaesco, Spain, for the financial data provided as well as to the referees of this paper for their valuable comments.

DECLARATION OF INTEREST CONFLICTS

There is not any conflict of interest in the elaboration of this paper and all the ethical principles generally accepted in the scientific community have been observed.

REFERENCES

1. Anowar, F., Sadaoui, S., & Selim, B. (2021). A conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, *40*(5), p.p. 1000378-. https://doi.org/10.1016/j.cosrev.2021.100378

2. Ayesha, S., Hanif, M. K., Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, 59 (July 2020), p.p. 44-58. https://doi.org/10.1016/j.inffus.2020.01.005

3. Back, A. & Weigend, A. (1997). A first application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems*, 8 (4), p.p. 473-484. https://doi.org/10.1142/S0129065797000458

4. Bellini, F. & Salinelli, E. (2003). Independent Component Analysis and Immunization: An exploratory study. *International Journal of Theoretical and Applied Finance*, 6(7), p.p. 721-738. https://doi.org/10.1142/S0219024903002201

5. Cavalcante, R.C., Brasileiro, R.C., Souza, L.F., Nobrega, J.P., Oliveira, A.L.I. (2016). Computational Intelligence and Financial Markets: A Survey and Future Directions. *Expert Systems with Applications,* 55 (15 August 2016), p.p. 194-211. https://doi.org/10.1016/j.eswa.2016.02.006

6. Coli, M., Di Nisio, R., & Ippoliti, L. (2005). Exploratory analysis of financial time series using independent component analysis. In: *Proceedings of the 27th international conference on information technology interfaces*, p.p. 169-174. Zagreb: IEEE. https://doi.org/10.1109/ITI.2005.1491117

7. Corominas, Ll., Garrido-Baserba, M., Villez, K., Olson, G., Cortés, U., & Poch, M. (2018). Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. *Environmental Modelling & Software*, 106 (Agosto 2018), p.p. 89-103. https://doi.org/10.1016/j.envsoft.2017.11.023

8. Diebold, F.X. & Lopez, J.A. (1996). Forecast evaluation and combination. In: G.S. Madala & C.R. Rao (eds.), *Handbook of statistics, Vol.14. Statistical Methods in Finance*, p.p. 241-268. Amsterdam: Elsevier. https://doi.org/10.3386/t0192

9. Himberg, J. & Hyvärinen, A. (2005). Icasso: software for investigating the reliability of ICA estimates by clustering and visualization. Retrieved from at: http://www.cis.hut.fi/projects/ica/icasso/about+download.shtml [2 February 2009].

10. Ibraimova, M. (2019). *Predicting Financial Distress Through Machine Learning (Publication No. 139967)* [Unpublished Master's Thesis]. Universitat Politécnica de Catalunya. Retrieved from: http://hdl.handle.net/2117/131355

11. Ince, H. & Trafalis, T. B. (2007). Kernel principal component analysis and support vector machines for stock price prediction. *IIE Transactions 39(6):* p.p. 629-637. https://doi.org/10.1109/IJCNN.2004.1380933

12. Ladrón de Guevara-Cortés, R., Torra-Porras, S. & Monte-Moreno, E. (2019). Neural Networks Principal Component Analysis for estimating the generative multifactor model of returns under a statistical approach to the Arbitrage Pricing Theory. Evidence from the Mexican Stock Exchange. *Computación y Sistemas*, *23*(2), p.p. 281-298. http://dx.doi.org/10.13053/CyS-23-2-3193

13. Ladrón de Guevara-Cortés, R., Torra-Porras, S. & Monte-Moreno, E. (2018). Extraction of the underlying structure of systematic risk from Non-Gaussian multivariate financial time series using Independent Component Analysis. Evidence from the Mexican Stock Exchange. *Computación y Sistemas*, *22*(4), p.p. 1049-1064 http://dx.doi.org/10.13053/CyS-22-4-3083

14. Ladrón de Guevara Cortés, R., & Torra Porras, S. (2014). Estimation of the underlying structure of systematic risk using Principal Component Analysis and Factor Analysis. *Contaduría y Administración, 59*(3), p.p. 197-234. http://dx.doi.org/10.1016/S0186-1042(14)71270-7

15. Lesch, R., Caille, Y., & Lowe, D. (1999). Component analysis in financial time series. In: *Proceedings of the 1999 Conference on Computational intelligence for financial engineering*, p.p. 183-190. New York: IEEE/IAFE. http://dx.doi.org/10.1109/CIFER.1999.771118

16. Lui, H. & Wan, J. (2011). Integrating Independent Component Analysis and Principal Component Analysis with Neural Network to Predict Chinese Stock

Market. *Mathematical Problems in Engineering*, 2011, p.p. 1-15. https://doi.org/10.1155/2011/382659

17. Lizieri, C., Satchell, S. Satchell & Zhang, Q. (2007). The underlying return-generating factors for REIT returns: An application of independent component analysis. *Real Estate Economics, 35*(4): p.p. 569-598. https://doi.org/10.1111/j.1540-6229.2007.00201.x

18. Miranda-Henrique, B., Amorin-Sobreiro, V., Kimura, H. (2019). *Experts Systems with Applications*, *124* (15 jun 2019), p.p. 226-251. https://doi.org/10.1016/j.eswa.2019.01.012

19. Pérez, J.V. & Torra, S. (2001). Diversas formas de dependencia no lineal y contrastes de selección de modelos en la predicción de los rendimientos del Ibex35. *Estudios sobre la Economía Española 94* (marzo, 2001), p.p. 1-42. Retrieved from: http://documentos.fedea.net/pubs/eee/eee94.pdf

20. Rojas, S., & Moody, J. (2001). Cross-sectional analysis of the returns of iShares MSCI index funds using Independent Component Analysis. *CSE610 Internal Report*, Oregon Graduate Institute of Science and Technology. Retrieved from: http://www.geocities.ws/rr_sergio/Projects/cse610_report.pdf

21. Ross, S.A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13*(3): p.p. 341-360. https://doi.org/10.1016/0022-0531(76)90046-6

22. Sayah, M. (2016). Analyzing and Comparing Basel III Sensitivity Based Approach for the Interest Rate Risk in the Trading Book. *Applied Finance and Accounting*, *2*(1), p.p. 101-118. https://doi.org/10.11114/afa.v2i1.1300

23. Scholz, M. (2006a). *Approaches to analyzing and interpret biological profile data.* [Unpublished Ph.D. Dissertation]. Postdam University. Retrieved from: https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/696/file/scholz_diss.pdf

24. Scholz, M. (2006b). Nonlinear PCA toolbox for Matlab®. Retrieved from: http://www.nlpca.org/matlab. [8 September 2008].

25. Scikit-Learn (2021, July 12). Manifold Learning. https://scikit-learn.org/stable/modules/manifold.html#

26. Wei, Z., Jin, L. & Jin, Y. (2005). Independent Component Analysis. *Working Paper.* Department of Statistics. Stanford University.

27. Weigang, L., Rodrigues, A. Lihua, S. & Yukuhiro, R. (2007). Nonlinear Principal Component Analysis for withdrawal from the employment time guarantee fund. In: S. Chen, P. Wang & T. Kuo (eds.), *Computational Intelligence in Economics and Finance. Vol. II*, p.p. 75-92. Berlin: Springer-Verlag. https://doi.org/10.1007/978-3-540-72821-4_4

28. Yip, F. & Xu, L. (2000). An application of independent component analysis in the arbitrage pricing theory. In: S. Amari et al. (eds.) *Proceedings of the International Joint Conference on Neural Networks*, p.p. 279-284. Los Alamitos: IEEE. https://doi.org/10.1109/IJCNN.2000.861471