Correlations in finance: a statistical approach.

José Manuel López-Alonso, Javier Alda

Optics Department University Complutense of Madrid.
School of Optics. Av. Arcos del Jalón, s/n, 28037 Madrid, Spain
Phone: +34 91 394 68 72, Fax: +34 91 394 68 85
E-mail: jmlopez@opt.ucm.es, j.alda@fis.ucm.es

Abstract

The behaviour of stock markets has been modelled actively during recent years. In some cases the market is modelled as a whole through the time series analysis of some indexes. But the market is made of companies whose time series can be studied independently. In this paper we have paid attention to the characterization of correlations and covariance among different companies in order to extract information about the market. We have used a statistical technique based on the analysis of the covariance matrix between the indexes of companies. When taking into account the sampling uncertainties and high order cumulants of index probability distribution, it is possible to classify automatically trends or clusters of companies in order to identify some independent "submarkets". The method is applied to some finance data sets coming from the Spanish financial market IBEX35.

1. Introduction

Stock market indexes have been studied extensively during recent years. Special attention has been paid to modelization of time series and probability distribution functions for prizes and returns. The selected index evolution is modelled as a stochastic process. Various models has been proposed but heteroscedasticity has been the most accepted one due to the fact that it is suitable for implementing cluster volatility and other features.[1-3]

From other point of view, markets are formed by companies with different types of relations among them. In this sense it could be interesting to develop analytical techniques in order to assess the "market" as a whole. At the same time the "company" point of view has to be preserved. In this paper we propose a method based on a Principal Component Analysis (PCA). The starting point of PCA is the covariance matrix, S, among different companies. It is possible to construct various types of S matrices, depending on the selected parameter under study (real price, returns,..). Then, the correlation structure of the market is introduced by means of covariance matrices. From them, it is possible to build a new set of variables called "Principal Components". The relevant feature of them is that they are uncorrelated. Then, they are natural variables in order to study high correlated ones.

In previous contributions [4-7] we have applied a similar method to the characterization of noise stochastic processes in detector arrays. In this paper we show that a similar approach can be

Noise in Complex Systems and Stochastic Dynamics II, edited by Zoltán Gingl, José M. Sancho, Lutz Schimansky-Geier, Janos Kertesz, Proc. of SPIE Vol. 5471 (SPIE, Bellingham, WA, 2004) · 0277-786X/04/$15 · doi: 10.1117/12.547082

applied to economic data. In section 2 we summarize the principal features of the method. In section 3 we apply it to a selected time series of IBEX35 index reflecting stock prices of 27 Spanish companies. The data are taken as an example and principal conclusions can be extended to other types of markets.

2. Principal Component Analysis.

PCA is a multivariate technique applied to a set of random variables $\{x_i\}_{i=1,..,N}$, being N the number of variables. In our case x_i is the index of the *i*-company. The first step of the procedure is to calculate the covariance matrix **S**, between variables. Principal Components (PC) are linear combinations of the original variables that are uncorrelated and whose variance is arranged in decreasing order.[8] They are calculated through the diagonalization of the covariance matrix, **S**. This produces a set of N eigenvectors, e_α, and eigenvalues, λ_α. PC, Y_α, are calculated as:

$$Y_\alpha = \sum_{i=1}^{N} e_\alpha(i) x_i \quad . \qquad (1)$$

The eigenvalue λ_α represents the variance of PC α. Then, the quantity Ω_α,

$$\Omega_\alpha = \lambda_\alpha \Big/ \sum_{i=1}^{N} \lambda_\alpha \quad , \qquad (2)$$

represents the relative amount of variance explained by PC α. Equation [1] can be seen as a filter process because original variables can be expressed as a function of principal components:

$$x_i = \sum_{\alpha=1}^{N} e_\alpha(i) Y_\alpha \quad . \qquad (3)$$

Then, a new set of filtered variables, x_i^F can be obtained selecting relevant principal components. In reference 4 is described a method to classify PC's into relevant groups or processes. It is based on the probability distribution function of the set of eigenvalues λ_α.[4] Two PCs with the same eigenvalue within uncertainty can be replaced by other two related with the original ones by means of an arbitrary angle rotation through an axis orthogonal to them. This two PCs form a process. The same result can be applied to n consecutive overlapping eigenvalues. The estimation of uncertainty involves fourth-order cumulants of the PC distribution. This procedure has been successfully applied to the classification of random noise in the context of images.[4-7]

Moreover, the method gives us information about the "goodness" of an observation of the variables, $x_i^\beta = \{x_1^\beta, x_2^\beta, ..., x_N^\beta\}_{\beta=1,..,M}$, being M the total number of observations of the variables. PC tend to have a multinormal distribution whose exponential (Mahalanobis distance) follows a chi-square distribution: [6]

$$\chi_\beta^2 = \sum_{\alpha=1}^{N} \frac{Y_\alpha^2(\beta)}{\lambda_\alpha} \rightarrow \chi_{N-1}^2 \quad . \qquad (4)$$

Observations with high χ^2_β could be far away from a threshold in probability. In this case this is a "bad observation": an outlier. Normally, PC departs from pure Gaussian behaviour. Therefore, the chi-square approach for the Mahalanobis distance is only an approximation. Anyway, the probability threshold is usually located at the tail of the distribution and is not very much affected by non-normality in the data set.[6] However, a test has been developed in order to quantify the impact of non-normality in the data (see reference 6 for details).

3. Experimental results.

3.1 Experimental data set.

We have applied the previous method to the Spanish IBEX35 index.[9] We have selected 26 different companies and their daily indexes have been recorded from 2/01/2001 to 21/10/2003 inclusive (704 points). We have analyzed different types of parameters:

- The real prices in time $x_i(t)$ with zero mean.
- The price changes: $Z_i(t) = x_i(t+\delta t) - x_i(t)$.
- The returns in logarithmic form, $S_i(t) = \ln x_i(t+\delta t) - \ln x_i(t)$. We have taken daily data set. Then, due to the high frequency, the change ratio, $R_i(t) = x_i(t+\delta t)/x_i(t)$ is similar to the returns.

The sampling amplitude δt is one day.

3.2 Prices with zero mean.

For the serie of prices of zero mean $x_i(t)$ the results are shown in figures 1-3. In figure 1 is shown the classification of eigenvalues with the variance explained by each PC. There are only 5 non overlapping PC comprising the 97.29% of the total variance. The rest of PC are grouped in a single noise process. Principal Components are shown in figure 2.

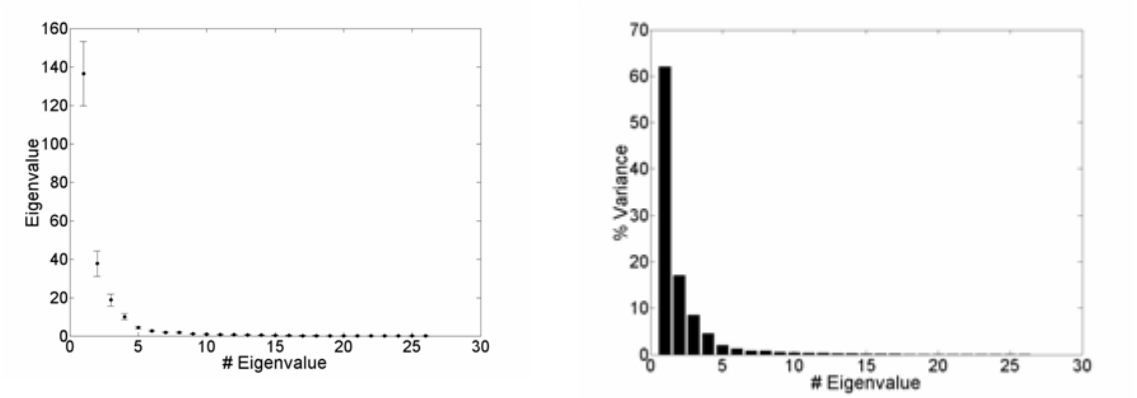

Figure 1: *Classification of principal components into processes (left) and explained variance (right) for the series of prices with zero mean. There is a single noise process comprising around 3% of total variance and five relevant principal components.*

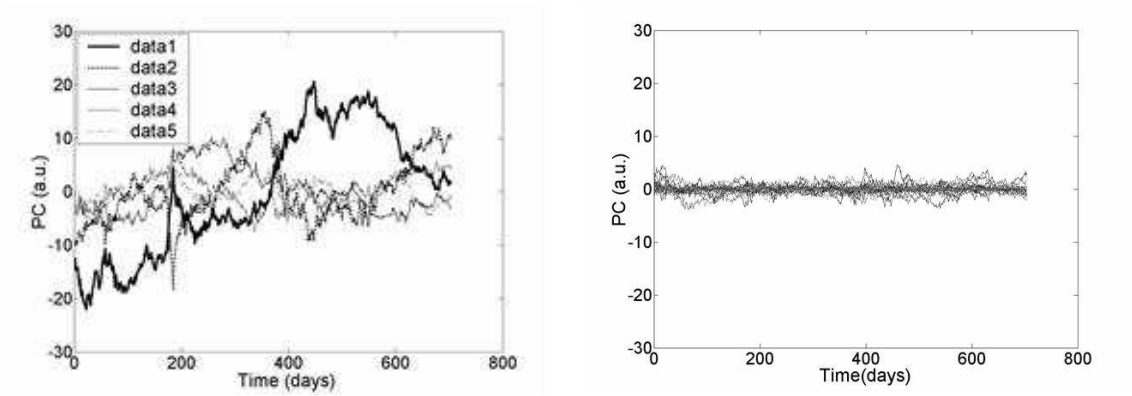

***Figure 2**: View of relevant principal components (left) and those associated to a single noise process (right) for the prices with zero mean.*

It is possible to see how the range of variation of relevant principal components is higher than the range associated with "noise". In figure 3 is calculated the coefficient of non-normality explained in reference 6. The non-normality is condensed over the relevant PC. This confirms the influence of high-order cumulants in the proabability distribution of price indexes. The PC associated in a single noise process behaves "normally". See in figure 3 how the Mahalanobis distance distribution is closer to a chi-square in this case. For both figures, the probability distribution is calculated empirically from the histogram. The solid line is the fitted chi-square distribution.

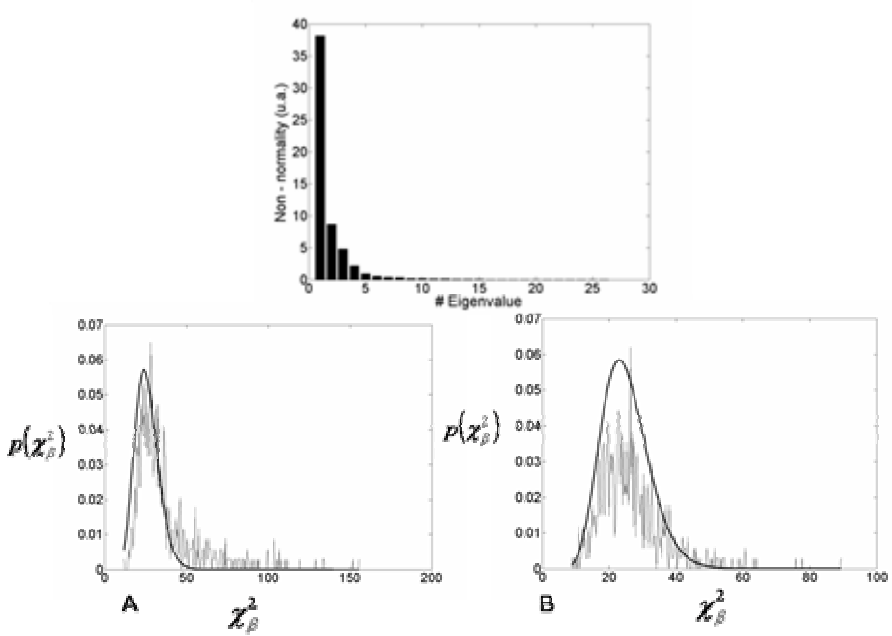

***Figure 3:**Non normality coefficient for the rectified data set of prices with zero mean(up). In the bottom the distribution of Mahalanobis distance for the rectified data with relevant PC (A) and with PC grouped in the "noise"(B).The fitted chi-square distribution is plotted in dotted line.*

3.3 Price changes and returns.

The results of the method for price changes are shown in figure 4. Contrariwise to the previous case, there is only one relevant principal component. The rest are grouped in a single process. This relevant principal component represents 30% of the total variance. The highest non-normality corresponds to this principal component and the distribution of Mahalanobis distance is clearly non chi-square.

Analysis of returns is shown in figure 5. Again, there is only one relevant principal component. Non-normality measure is higher than in the price changes due to the logarithmic non linear transformation. The relevant principal component represents 28.1% of the data set variance. The Mahalanobis distance distribution is again not well fitted by a chi-square distribution. The structure of relevant principal component resembles the relevant one of the price changes (see figure 4D and figure 5D).

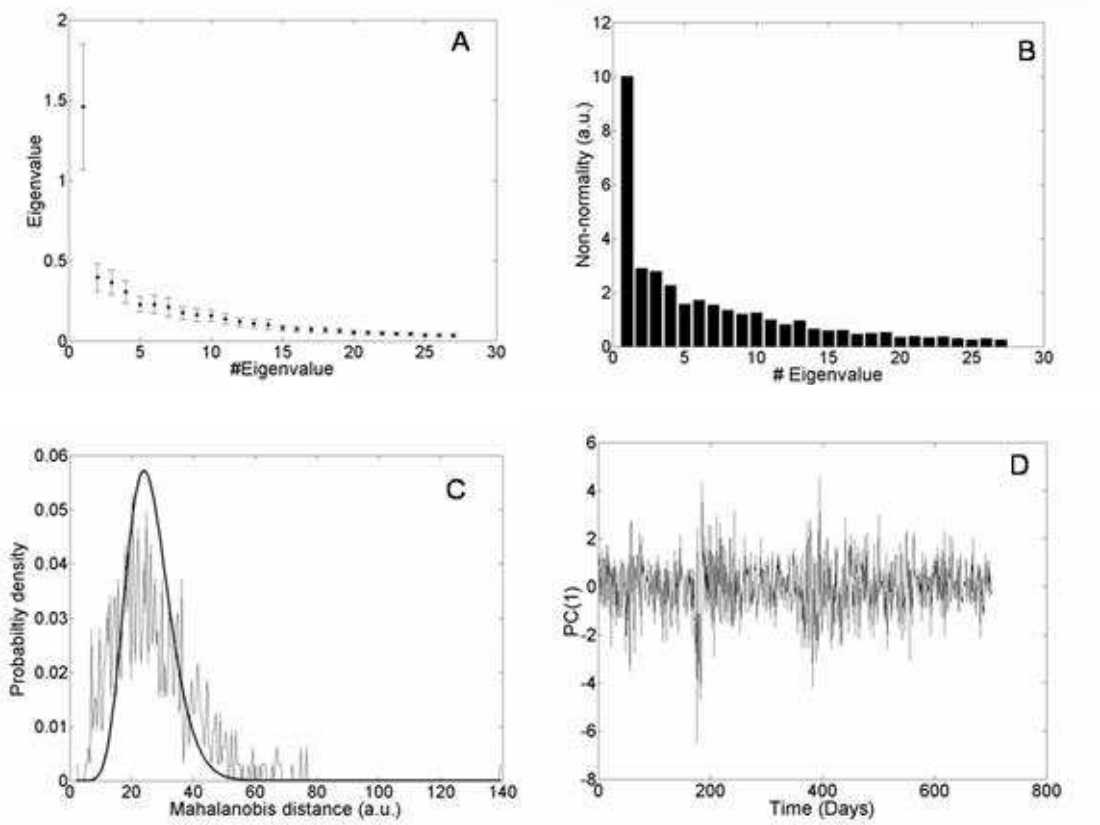

Figure 4: *Results of the method for the price changes. A) Relevant eigenvalues, B) Non-normality for each principal component, C)Probability distribution of Mahalanobis distance, D) View of the relevant principal component (PC(1)).*

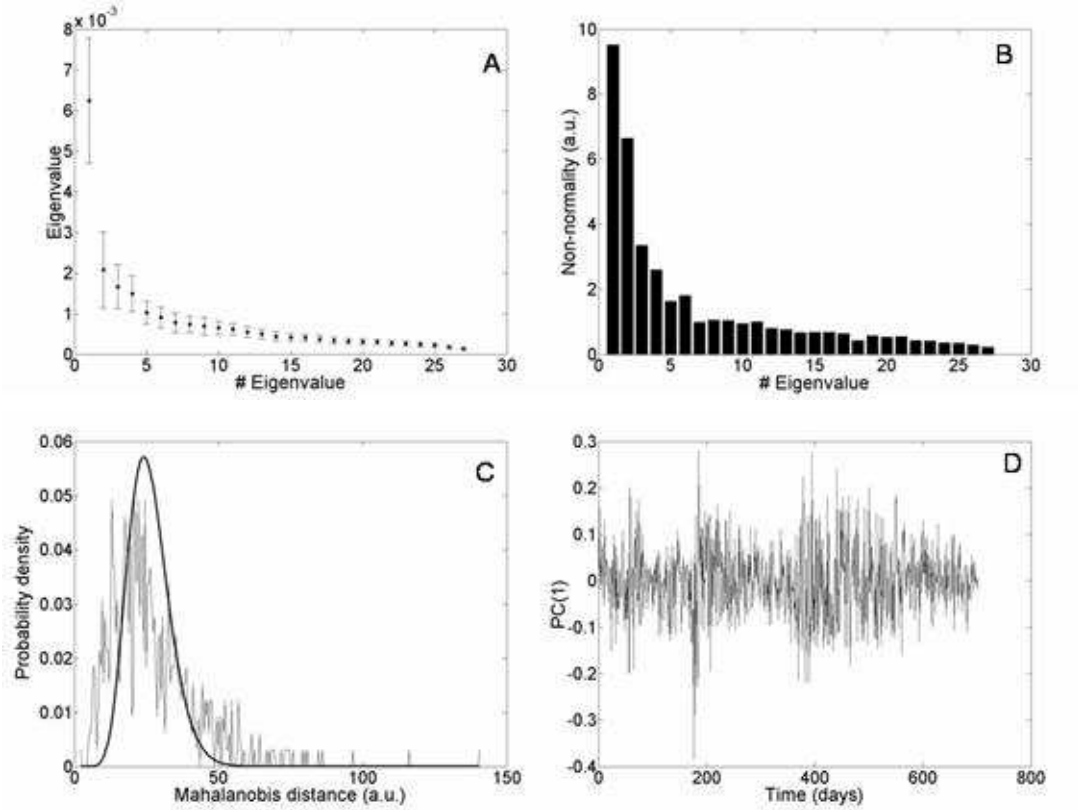

Figure 5: *Results of the method for returns. A) Relevant eigenvalues, B) Non-normality for each principal component, C)Probability distribution of Mahalanobis distance, D) View of the relevant principal component (PC(1)).*

4. Analysis of "trade crashes".

Equation (4) can be used to detect "outliers". It is important to note that Mahalanobis distance is calculated at every moment in time and χ^2_β has contributions of the whole market. In figures 4 and 5 this distribution departs from a pure chi-square. For this reason, instead of using a certain distribution, a cumulative probability function is used for the Mahalanobis distance. In figure 6 are shown these cumulative probabilities for Z(t) and S(t).

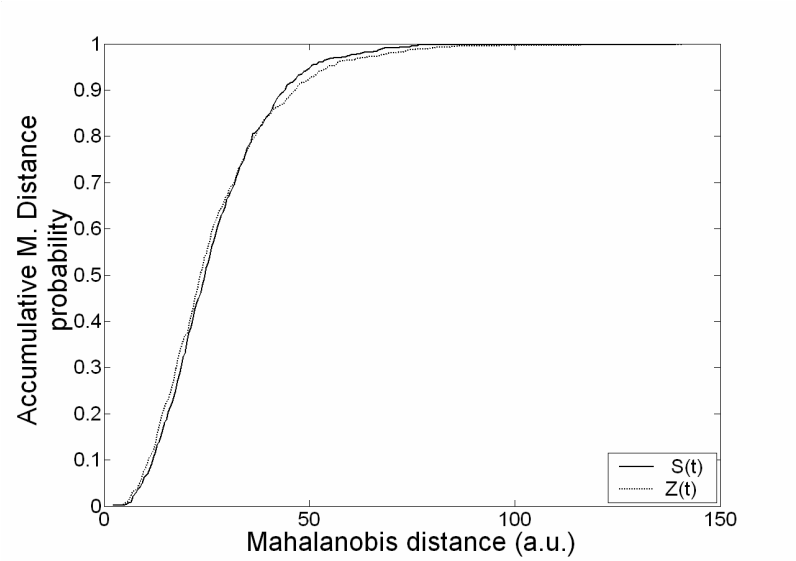

Figure 6: Cumulative probability for the Mahalanobis distance of equation (4), for Z(t) and S(t).

From figure 6 it is possible to choose a threshold in probability for Z(t) and S(t). This threshold marks the maximum probability of occurrence for a possible outlier. Another possibility is to choose the threshold depending on data set[6]. In this sense the threshold is chosen in a way that the probability to obtain only one outlier was negligible. The points above threshold are, in both cases, relevant outliers. These points denotes "market crashes" on Z(t) or S(t). Through threshold selection it is possible to even "quantify" the crash with a probability of occurrence. After that, it is possible to study carefully the data series around those points in order to extract information about the crash.

Figure 7 shows the results of outliers classification for two different thresholds in Z(t) and S(t). In the first one, the probability of occurrence of outliers is less than 1% (Top plots in figure 7). In the bottom of figure 7 the threshold is chosen in a way that now the probability of appearing one outlier in the whole time range is less than 1%. It is possible to see how both types of thresholds gives about the same result for Z(t). For S(t) there is a little change. In all cases the cumulative probability functions used are derived from experiment (see figure 6).

For the highest threshold there is only a crash in both Z(t) and S(t). The time corresponds to the period 31 August 2001 to 10-13 September 2001. Two important things happened in that period in the IBEX. The stock prices of Telecommunications companies (Telefónica SA, Terra, TPI…) decreased very fast during that period following others ".com" companies around the world. Besides, in the same period 11-S terrorist attack happened in New York.

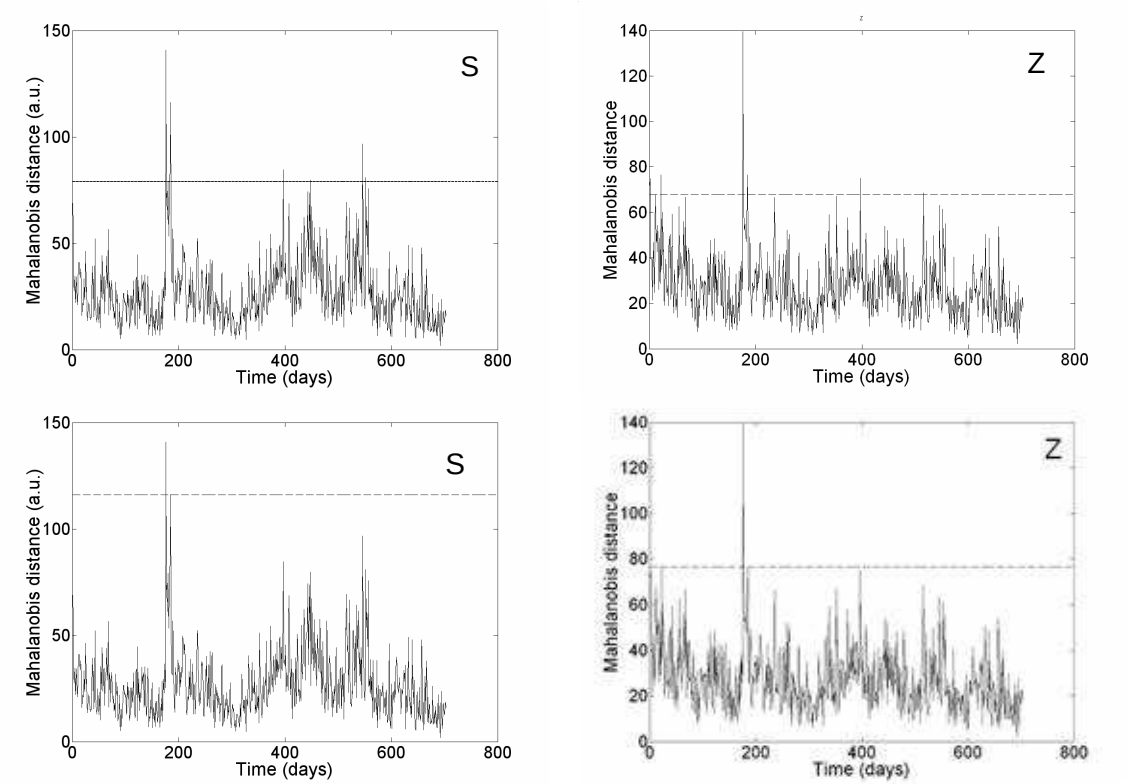

Figure 7: Values in time of Mahalanobis distance for the series of S(t) and Z(t). Probability thresholds for outliers less than 1% (Top) and probability of occurrence of just one single outlier less than 1% (Botton).

The other threshold classifies crashes with of occurrence probability less than 1%. Again, in Z(t) and S(t) appears the previously period of time. Other crisis appears around (maximum points): 22 July 2002, 1 October 2002 and a final region over 28 February 2003 to 20 March 2003. The 1 October 2002 Dow Jones index decreases at the lowest September values till 1997[10]. 22 July 2002 has been recorder too[11] and the mentioned period of March[12].

5. Analysis of time series.

For both Z(t) and S(t) there is only a relevant principal component comprising around the 30% of total variance. Both relevant principal components have a similar time variation (see plot D, figures 4 and 5). Return time series has been extensively modelled. The most relevant model is GARCH model. A GARVH model is used here to model the return series y_t as:

$$\begin{aligned} y_t &= C + e_t \\ \sigma_t^2 &= k + garch(1)\sigma_{t-1}^2 + arch(1)e_{t-1}^2 \end{aligned} \quad , \qquad (5)$$

where e_t are innovations coming from a Gaussian distribution, σ_t^2 is the conditional variance and C, *k,garch(1), arch(1)* are the parameters of the model. The fitted parameters are shown in table 1. In figure 8 is shown the calculated conditional variance with black points in crashes. Comparing it with figure 7 is easy to see if the crash corresponds with a sudden increase or decrease of conditional variance. And the moment with high conditional variance that no

corresponds to crashes (first period of figure 8). Then, the majority of the variance of relevant principal component is given by crashes.

C	0.0014 ± 0.002
K	0.00012 ± 7.5e-5
Garch(1)	0.87 ± 0.03
Arch(1)	0.11 ± 0.02

Table 1: Fitted GARCH parameters for the relevant principal component of the retuns.

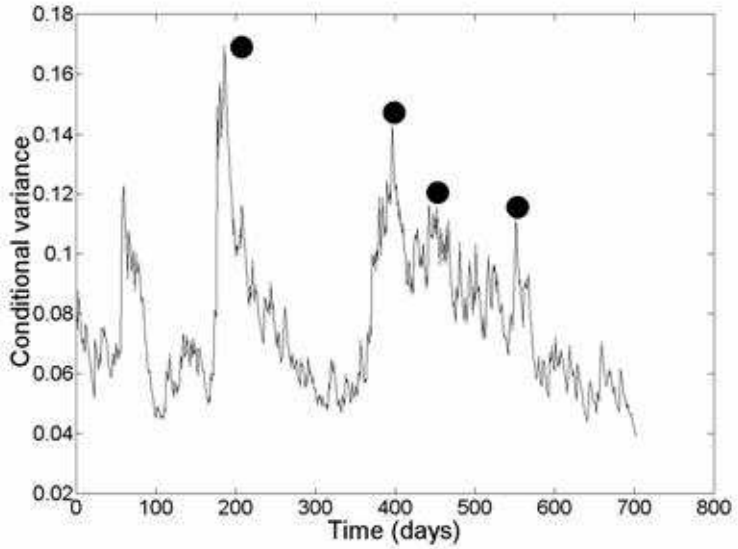

Figure 8: Calculated conditional variance for the return series of relevant principal component. With black circles are marked the crashes classified by the method.

6. Analysis of correlations.

The analysis of correlations is done, again, over the returns series. For this data set there is only a relevant principal component. From the principal component theory and equation (3), the value $e_\alpha^2(i)$ represents the portion of the variance of the company return i, explained by the principal component α. In figure 9 is represented $e_1^2(i)$. Then, the variance of the only relevant principal component is dominated by some companies. One of them is Telefónica SA with strong participation in Telecommunications companies as TPI, Terra, Telefónica Móviles and Sogecable. The rest are two banks (BBVA, related with Telefónica SA too, and BSCH) and a travel Agency: Amadeus. These companies agree well with the description of crashes giving in previous sections (.com companies crisis and terrorist impact). Moreover it is a strong

relation among the analyzed principal component and trade crashes. Actually, these companies have crossed participation of different levels among them.

In figure 10 we plot the correlation matrices for the whole market reconstructed with and without the relevant principal component (equation 3). The mean value of non diagonal correlation for the whole return market is 0.21. Without the relevant principal component this value is -0.0074. Then, relevant principal component is given by an internal correlation among companies in the market and is driven by the companies given in figure 9. Moreover, is also well correlated with the "trades crashes".

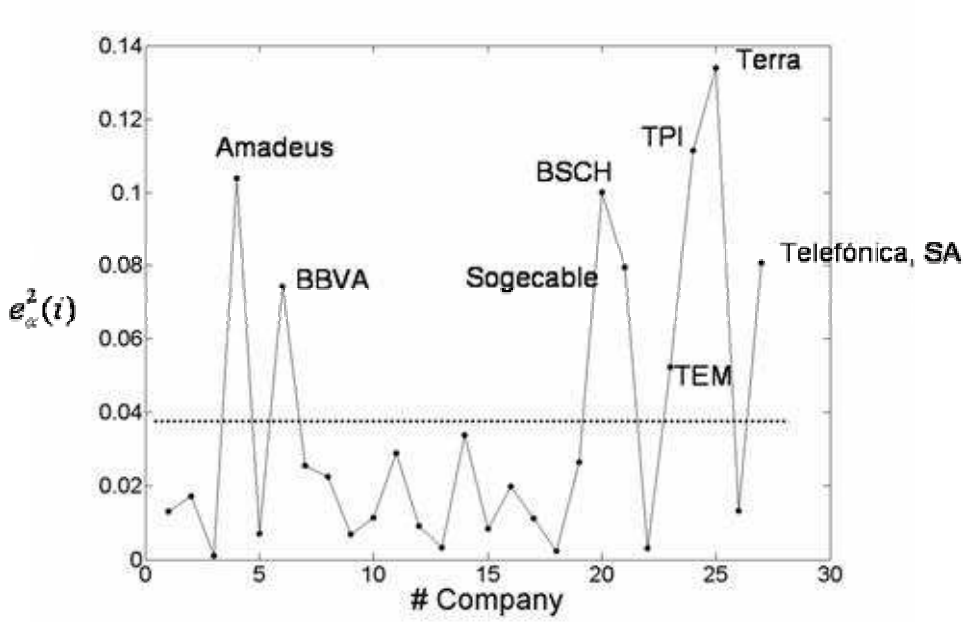

Figure 9: Portion of variance in relevant returns principal component explained by each company.

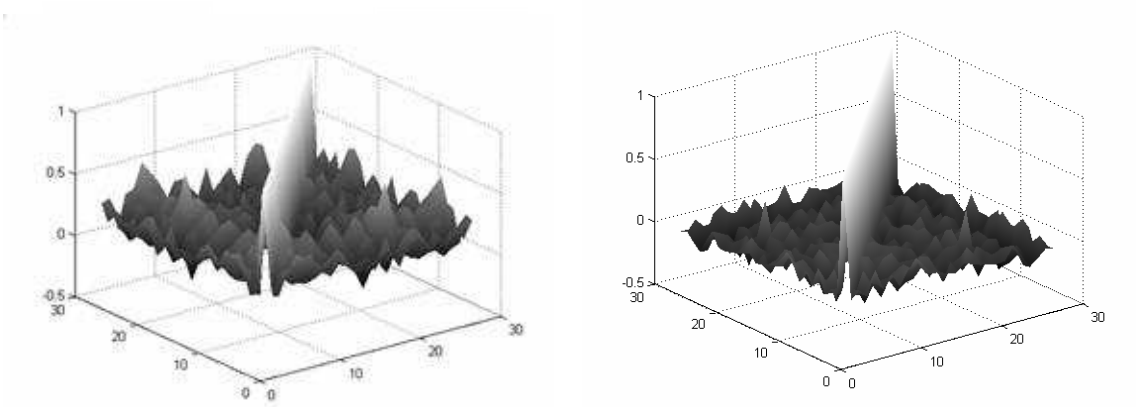

Figure 10: Correlation matrices for the whole return market (left) and without first principal component for returns.

7. Conclusions.

This paper presents a method to characterize correlation and covariance among companies in a trade market. It is based on a Principal Component Analysis and the study of high order cumulants of different index distributions. It permits to select relevant temporal structures and quantify the participation of each company in it. The method permits to classify and quantify trades crashes with different levels of probability. It has been applied to the Spanish IBEX35. In this market a relevant structure containing the 30% of total return variations appears. It is dominated by the trade crashes and companies related with them or being participated by others related with it. This submarket is dominated by Telecommunication companies and banks related with them.

References

1. R.N. Mantenga, H.E. Stanley, *An introduction to econophysics: Correlations and Complexity in Finance, Cambrige* University Press, 1999.
2. R. Engle, "Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of U.K. Inflation," *Econometrica* 50 (1982): 987-1008.
3. B. Mandelbrot,*Fractals and Scaling in Finance: Discontinuity, Concentration, Risk.*New York: Springer, 1997.
4. J. M. López-Alonso, J. Alda, "Principal components characterization of noise for infrared images", Applied Optics, 41, 320-331, (2002).
5. J. M. López-Alonso, J. Alda, "Operational parametrization of the 1/f noise of a sequence of frames by means of the principal components analysis in focal plane arrays", Optical Engineering, **42**,1915-1922, (2003).
6. J. M. López-Alonso, J. Alda, "Bad Pixel identification by means of the principal components analysis", Opt. Eng., 41, 2152-2157, (2002).
7. J.M. López-Alonso, J. Alda, " Automatic classification of Noise for Infrared Images into Processes by means of the Principal Component Analysis" Infrared and Passive Millimeter-wave Imaging Systems: Design, Analysis, Modeling and Testing, R. Appleby, G.C. Holst, D.A. Wikner, Editors, Proc. SPIE Vol. 4719, 95-106, (2002).
8. D. F. Morrison, *Multivariate Statistical Methods*, 3rd ed.McGraw-Hill, Singapore, 1990, Chap. 8.
9. Oficial site Madrid Trade Market, www.bolsademadrid.es
10. Bolletin Skill Digital , http://www.skilldigital.com/boletin/Noticia.asp?IdNoticia=871.
11. Bolletin Skill Digital, http://www.skilldigital.com//boletin/Noticia.asp?IdNoticia=796.
12. Bolletin Skill Digital http://www.skilldigital.com//boletin/Noticia.asp?IdNoticia=1144.