Analysis of stock market indices through multidimensional scaling

J. Tenreiro Machado, Fernando B. Duarte, Gonçalo Monteiro Duarte

A B S T R A C T

We propose a graphical method to visualize possible time-varying correlations between fif- teen stock market values. The method is useful for observing stable or emerging clusters of stock markets with similar behaviour. The graphs, originated from applying multidimen- sional scaling techniques (MDS), may also guide the construction of multivariate econo- metric models.

Keywords: Multidimensional scaling, Stock market daily values, Time-varying correlation, Econophysics

1. Introduction

Economical indexes measure the performance of segments of the stock market and are normally used to benchmark the performance of stock portfolios. This paper proposes a descriptive method which analyses possible correlations/similarities in international stock markets. Its results are expected to guide the design of statistical models aiming to test hypotheses of interest. Ultimately, the method can even lead to the postulation of new hypotheses. The study of the correlation of inter- national stock markets may have different motivations. Economic motivations to identify the main factors which affect the behaviour of stock markets across different exchanges and countries. Statistical motivations to visualize correlations in order to suggest some potentially plausible parameter relations and restrictions.The understanding of such correlations would be helpful to the design good portfolios [16,18].

Bearing these ideas in mind the outline of our paper is as follows. In Section 2 we give the fundamentals of the multidimensional scaling (MDS) technique, which is the core of our method, and we discuss the details that are relevant for our specific application. In Section 3 we apply our method for daily data on fifteen stock markets, including major American, Asian/Pacific, and European stock markets. In Section 4 we conclude the paper with some final remarks and potential topics for further research.

2. Multidimensional scaling

Generally speaking MDS techniques develop spatial representations of psychological stimuli or other complex objects about which people make judgements (e.g., preference, relatedness), that is they represent each object as a point in a m-dimensional space. What distinguishes MDS from other similar techniques (e.g., factor analysis, cluster analysis) is that in MDS there are no preconceptions about which factors might drive each dimension. Therefore, the only data needed is a measure for the similarity between each possible pair of objects under study. The result is the transformation of the data into similarity measures which can be represented by Euclidean distances in a space of unknown dimensions [4]. The greater

the similarity of two objects, the closer they are in the *m*-dimensional space. After having the distances between all the objects, the MDS techniques analyse how well they can be fitted by spaces of different dimensions. The analysis is normally made by gradually increasing the number of dimensions until the quality of fit (measured for example by the correlation between the data and the distance) is little improved with the addition of a new dimension. In practice a good result is nor- mally reached well before the number of dimensions theoretically needed to a perfectly fit is reached [8,14,19,24].

In the MDS method a small distance between two points corresponds to a high correlation between two stock markets and a large distance corresponds to low or even negative correlation [17]. A correlation of one should lead to zero distance between the points representing perfectly correlated stock markets. MDS tries to estimate the distances for all pairs of stock markets to match the correlations as close as possible. MDS may thus be seen as an exploratory technique without any dis- tributional assumptions on the data. The distances between the points in the MDS maps are generally not difficult to inter- pret and thus may be used to formulate more specific models or hypotheses. Also, the distance between two points should be interpreted as being the distance conditional on all the other distances. One possibility to obtain such an approximate solu- tion is given by minimizing the stress function. The obtained representation of points is not unique in the sense that any rotation or translation of the points retains the distances [5]. To formalize MDS, we need some notation. Let *n* be the number of different objects and let the dissimilarity for objects *i* and *j* be given by d_{ij}. The coordinates are gathered in an *n* x *p* matrix X, where *p* is the dimensionality of the solution to be specified in advance by the user. Thus, row *ir* from X gives the coordinates for object *i* on dimension *r*. Let d_{ij} be the Euclidean distance between rows *i* and *j* of X defined as

$$d_{ij} = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2} \qquad (1)$$

that is, the length of the shortest line connecting points *i* and *j* on dimension *r*. The objective of MDS is to find a matrix X such that d_{ij} matches d_{ij} as closely as possible. This objective can be formulated in a variety of ways but here we use the raw-Stress r^2,

$$\sigma^2 = \sum_{i=2}^{n} \sum_{j=1}^{i-1} w_{ij} (\delta_{ij} - d_{ij})^2 \qquad (2)$$

i

proposed by Kruskal [13], who was the first one to propose a formal measure for doing MDS, where w_{ij} is a user defined weight that must be nonnegative. This measure is also referred to as the least-squares MDS model. Note that due to the symmetry of the dissimilarities and the distances, the summation only involves the pairs *i*, *j* where *i* > *j*. For example, many MDS programs implicitly choose $w_{ij} = 0$ for dissimilarities that are missing. The minimization of r^2 is a complex problem. Therefore, MDS programs use iterative numerical algorithms to find a matrix X for which r^2 is a minimum. In addition to the raw Stress measure there exist other measures for doing Stress. One of them is normalized raw Stress, which is simply raw Stress divided by the sum of squared dissimilarities. The advantage of this measure over raw Stress is that its value is independent of the scale and the number of dissimilarities. The second measure is Kruskal's Stress-1 which is equal to the square root of raw Stress divided by the sum of squared distances. A third measure is Kruskal's Stress-2, which is similar to Stress-1 except that the denominator is based on the variance of the distances instead of the sum of squares. Another measure that seems reasonably popular is called S-Stress and it measures the sum of squared errors between squared distances and squared dissimilarities.

In order to assess the quality of the MDS solution we can study the differences between the MDS solution and the data. One convenient way to do this is by inspecting the so-called Shepard diagram [21]. A Shepard diagram shows both the transformation and the error. Let p_{ij} denotes the proximity between objects *i* and *j*. Then, a Shepard diagram plots simultaneously the pairs (p_{ij}, d_{ij}) and (p_{ij}, d_{ij}). By connecting the (p_{ij}, d_{ij}) points a line is obtained representing the relationship between the proximities and the disparities. The vertical distances between the (p_{ij}, d_{ij}) points and (p_{ij}, d_{ij}) are equal to $d_{ij} - d_{ij}$, that is, they give the errors of representation for each pair of objects. Hence, the Shepard diagram can be used to inspect both the resid- uals of the MDS solution and the transformation.

3. Analysis of stocks markets

In this section we study numerically the fifteen selected stock markets, including six American markets, six European markets and three Asian/Pacific markets.

Our data consist of the *h* daily close values of *s* = 15 stock markets from January 2, 2000, up to December 31, 2009, to be denoted as $x_i(t)$, $1 \leq t \leq h$, $i = 1, \ldots, s$. The stock markets are listed in Table 1.

The data are obtained from data provided by Yahoo Finance web site [12], and they measure indexes in local currencies.

Fig. 1 depicts the time evolution, of daily, closing price of the fifteen stock markets versus year with the well-know noisy and ''chaotic-like'' characteristics.

Assuming that financial index prices are random variables one of the most important analyses' parameter of the financial indexes it is the *volatility* [3]. Volatility measures variability or dispersion about a central tendency. Normally is defined as the deviation from their mean. The historical volatility is the volatility of a series of index prices where we look back over the historical price. The historical volatility estimate, for each index *i*, is given by

Table 1
Fifteen stock markets and value of volatility values.

i	Stock market index	Abbreviation	Country	Volatility (W_P)
1	All Ordinaries	AORD	Australia	0.5118
2	EURONEXT BEL-20	BFX	Belgium	0.6827
3	Cotation Assiste en Continu	CAC	France	0.7922
4	Deutscher Aktien Index	DAX	German	0.8412
5	Dow Jones Industrial	DJI	USA	0.6599
6	Footsie	FTSE	UK	0.6729
7	Iberia Index	IBEX	Spain	0.7511
8	Bolsa Mexicana de Valores	MXX	Mexico	0.7737
9	NASDAQ	NDX	USA	1.1129
10	New York Stock Exchange	NYA	USA	0.6883
11	Standard & Poor's	SP500	USA	0.7024
12	Shanghai Stock Exchange	SSEC	China	0.8415
13	Swiss Market Index	SSMI	Swiss	0.6590
14	Straits Times Index	STI	Singapore	0.6761
15	Toronto Stock Exchange	TSX	Canada	0.6553

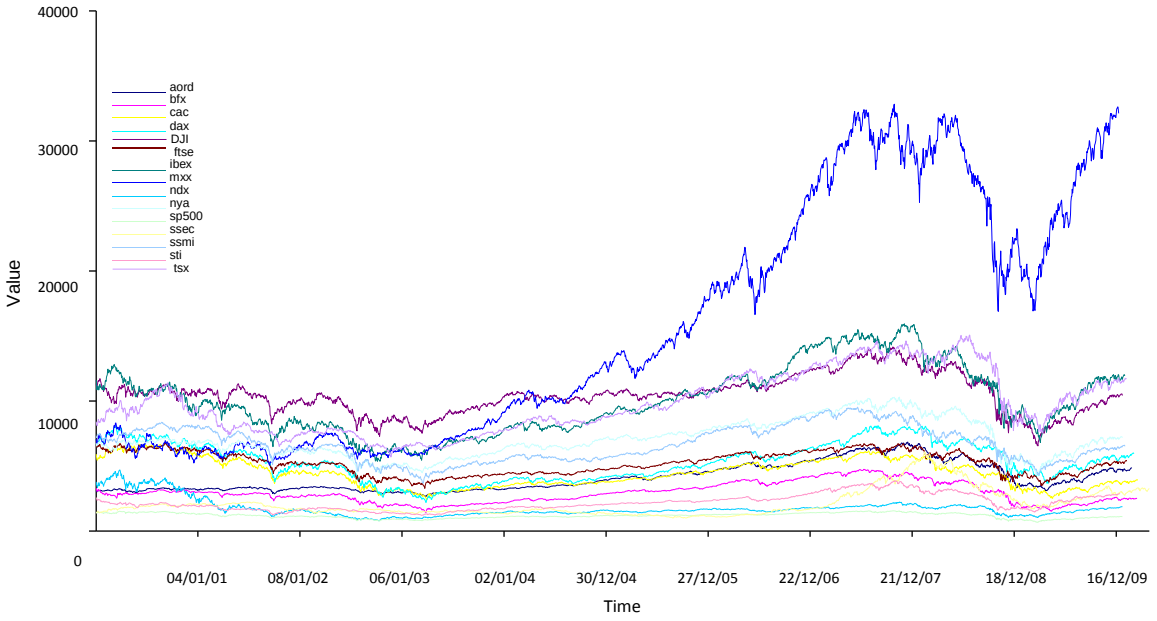

Fig. 1. Time series for the fifteen indexes from January 2000, up to December 2009.

$$\Psi = \sqrt{\frac{1}{h-1}\sum_{j=1}^{h}(u_j - \bar{u})^2}, \qquad (3)$$

where $u(t) = \ln(x(t)) - \ln(x(t-1))$, $1 < t \leq h$ and $\bar{u}$ is the arithmetic average of the u_j.

The parameter W gives the estimated volatility per interval of observation. To enable have the volatilities for different period lengths, usually scale this estimate with a factor h, according the period length, which is the number of intervals in the period length.

$$\Psi_p = \Psi \cdot \sqrt{h}.$$

Since our interval data observation is daily and our period is one decade we use $h = 2510$. The parameter volatility values are shown in the Table 1.

In the sequel, this section is organized in two subsections, the first adopts an analysis based on the correlation of the time evolution and the second adopts a metrics based on histogram distances.

3.1. MDS analysis based on time correlation

In this subsection, we apply the MDS method described in Section 2 to the time correlation of the selected stock markets.

For the fifteen markets, we consider the time correlations between the daily close values. We first compute the correlations among the fifteen stock markets obtained a $s \times s$ matrix and then apply MDS. In this representation, points represent the stock markets.

In order to reveal possible relationships between the market stocks index the MDS technique is used. In this perspective several MDS criteria are tested. The Sammon criterion revealed good results and is adopted in this work [1,9,15]. For this purpose we calculate $s \times s$ matrix M based on a correlation coefficient c_{ij}, that provides a measurement of the similarity be- tween two indexes and is defined in Eq. (4). In matrix M each cell represents the time correlation between a pair of indexes:

$$c_{ij} = \left(\frac{\frac{1}{h}\sum_{t=1}^{h} x_i(t) \cdot x_j(t)}{\sqrt{\frac{1}{h}\sum_{t=1}^{h} x_i^2(t) \cdot \frac{1}{h}\sum_{t=1}^{h} x_j^2(t)}} \right)^2, \qquad (4)$$

$i, j = 1, \ldots, s$. Fig. 2 shows the 3D locus of each index positioning in the perspective of expression (4). Fig. 4 depicts the stress and the Shepard plots for the MDS. The stress plot, as function of the dimension of the representation space, revealing that a three dimensional space describe a with reasonable accuracy the ''map'' of the fifteen signal indexes. Moreover, the Shepard plot shows that a good distribution of points around the 45 degree line is obtained.

For comparison it was decided to confront MDS with an alternative visualization method. For that purpose are adopted dendograms using the same information matrices of the MDS case. The dendogram of Fig. 3 shows the hierarchical clustering of the fifteen indexes with matrix M based on the time correlation. To generate the dendograms we selected the MultiDend- ograms hierarchical clustering package, configured for the ''Unweighed Average'' clustering method [2,11]. Several other methods (Single Linkage, Complete Linkage, Weighted Average, Unweighed Centroid, Weighted Centroid, Join Between- Within) were tested leading to dendograms qualitatively of the same type.

We observe that the two visualisation techniques give similar conclusions. For *pros* the MDS we have a more intuitive mapping and for *cons* we have the requirement of a 3-dimensional chart.

There are empirical conclusions one can draw from the graphs in Fig. 2. The indexes seem to be organized according to their characteristics on the three dimensional MDS suggesting that we may group the fifteen indexes into three clusters:

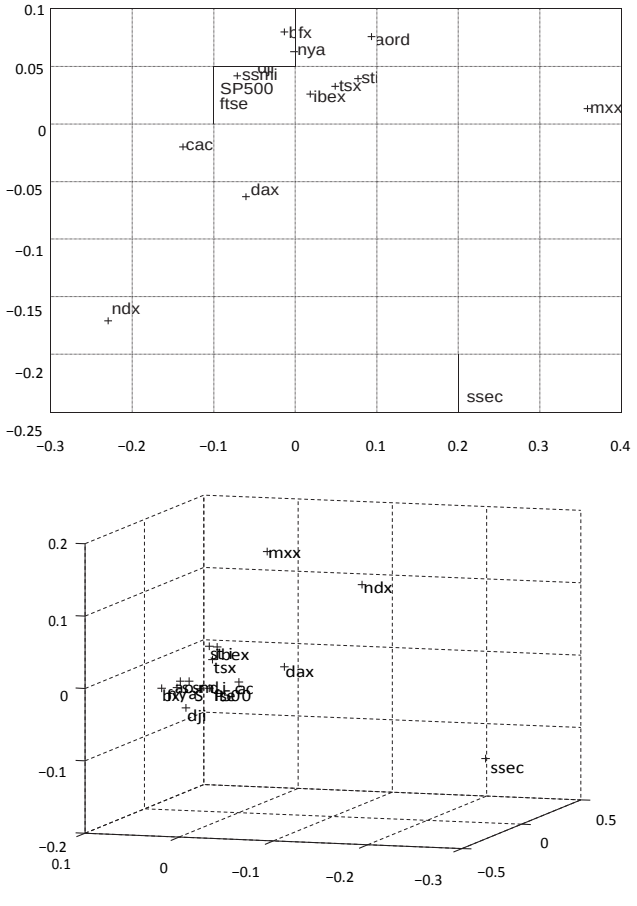

Fig. 2. Two (top) and three (bottom) dimensional MDS graphs for the fifteen indexes using time correlation.

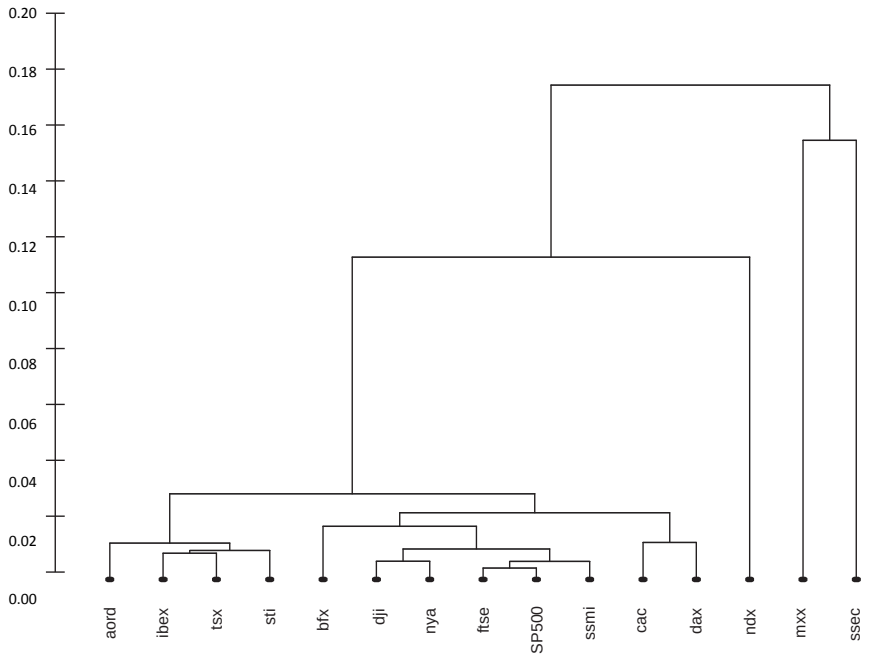

Fig. 3. Dendogram for the fifteen indexes using time correlation.

i. Cluster A, on the left: AORD, BFX, CAC, DJI, FTSE, IBEX, NYA, SP500, SSMI, STI and TSX;
ii. Cluster B, on the right: DAX, NDX and SSEC;
iii. Cluster C, on the top: MXX.

Let explore each:

The cluster A groups the majority of the indexes and may be considered to represent the norm. The indexes grouped on Cluster B are the ones with the highest volatilities (i.e. greater than 0.80). The emergence of Cluster B may suggest that inves- tors have a different behaviour in volatile markets. In fact the standard financial theory shows that there is a negative rela- tion between volatility and expected return [6,10]. Therefore, some investors worry and have an extra level of concern as they watch the value of their portfolios move more violently and it may originate irrational responses or at least decisions which are not aligned with the normal practices.

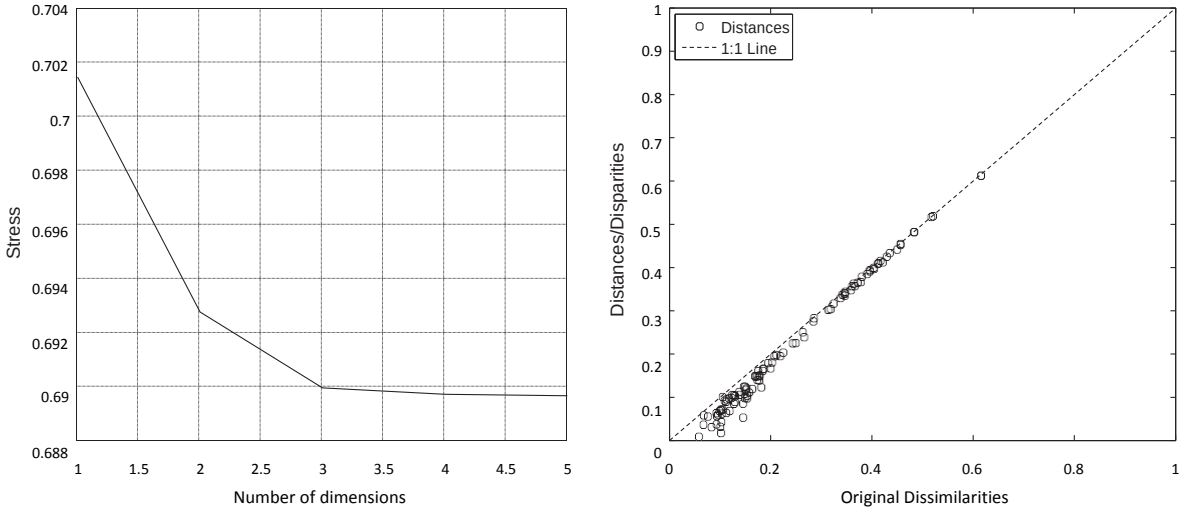

Fig. 4. Stress (left) and Shepard (right) plots of 3D MDS representation of the fifteen indexes *vs* number of dimension using time correlation.

3.2. *MDS analysis based on histogram*

For each of the fifteen indexes we draw the corresponding histogram of relative frequency and we calculate statistical descriptive parameters like the arithmetic mean ($\mathbf{l}_i$), the standard deviation ($\mathbf{r}_i$) and the Pearson's Kurtosis coefficient c_i. The values of the statistical descriptive parameters are listed in Table 2.

For all the fifteen indexes we calculate the ''histogram's distances'' [7,20,22,23], d^a and d^b using the equations:

$$d_{ij}^{a} = \sqrt{\frac{(\mu_i - \mu_j)^2}{\mu_i^2 + \mu_j^2} + \frac{(\sigma_i - \sigma_j)^2}{\sigma_i^2 + \sigma_j^2}}, \tag{5}$$

$$d_{ij}^{b} = \sqrt{\frac{(\mu_i - \mu_j)^2}{\mu_i^2 + \mu_j^2} + \frac{(\sigma_i - \sigma_j)^2}{\sigma_i^2 + \sigma_j^2} + \frac{(\gamma_i - \gamma_j)^2}{\gamma_i^2 + \gamma_j^2}}, \tag{6}$$

where $i, j = 1, \ldots, s$.

Figs. 5 and 8 show the 2D locus of each index positioning in the perspective of the expressions (5) and (6), respectively demonstrating differences between the corresponding MDS plots.

Figs. 7 and 10 depict the stress as function of the dimension of the representation space based on d^a and d^b distances, revealing that a two dimensional space describe with reasonable accuracy the ''map'' of the fifteen signal indexes unlike that seen in the MDS based on time correlation. Moreover, the resulting Shepard plot shows that a good distribution of points around the 45 degree line is obtained for the two distances.

The dendograms of Figs. 6 and 9 show the hierarchical clustering of the fifteen indexes, using the ''Unweighed Average'' clustering method with histogram's distance d^a and d^b based matrices, respectively.

Table 2
Statistical descriptive parameters.

i	$\mathbf{l}_i$	$\mathbf{r}_i$	c_i
1	4082.52	1074.17	−0.51
2	2956.11	788.96	−0.67
3	4475.70	1071.94	−1.04
4	5324.72	1440.27	−1.00
5	10472.98	1454.40	−0.11
6	5248.86	871.57	−1.24
7	10042.82	2583.65	−0.72
8	15372.58	9168.40	−1.27
9	1753.80	701.35	4.05
10	7034.10	1404.76	−0.58
11	1187.55	198.44	−0.85
12	2079.76	1031.71	2.96
13	6689.10	1337.82	−0.95
14	2179.07	615.19	−0.22
15	9789.64	2360.58	−0.95

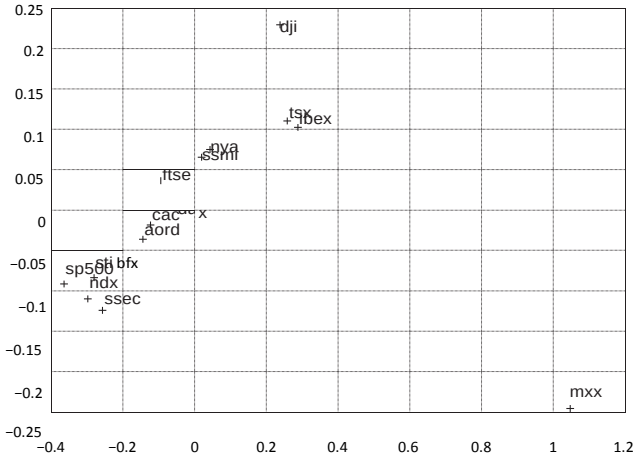

Fig. 5. Two dimensional MDS graphs for the fifteen indexes using histogram's distance d^a.

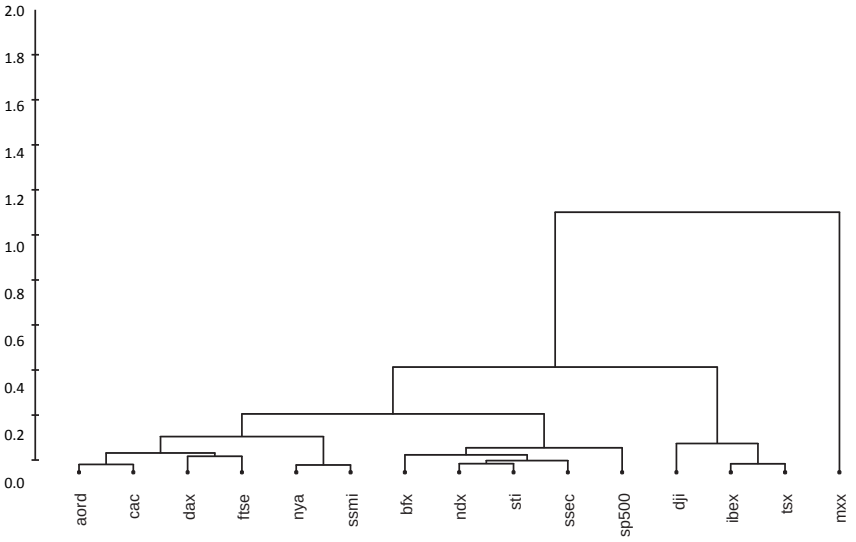

Fig. 6. Dendogram for the fifteen indexes using histogram's distance d_a.

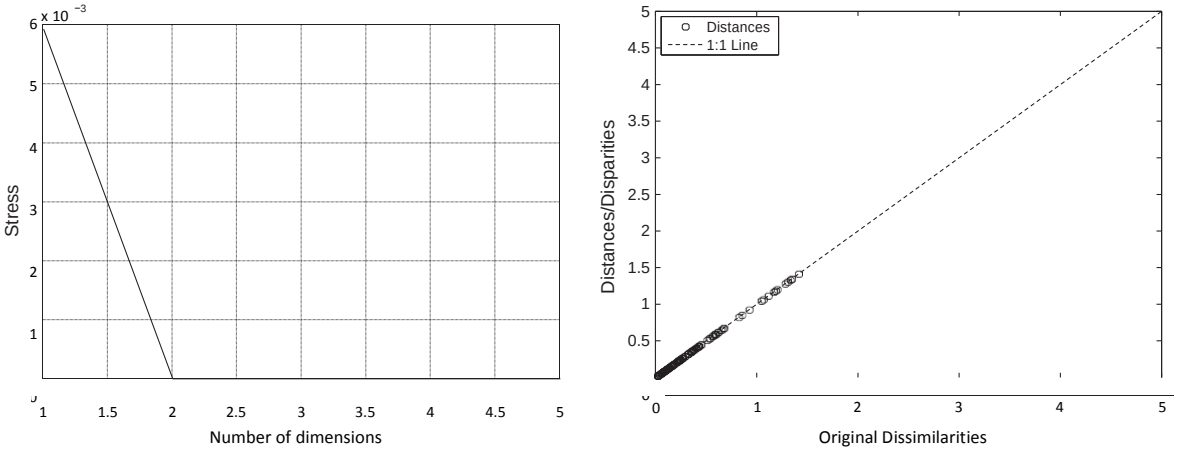

Fig. 7. Stress (left) and Shepard (right) plots of 2D MDS representation for all indexes *vs* number of dimension, using histogram's distance d^a.

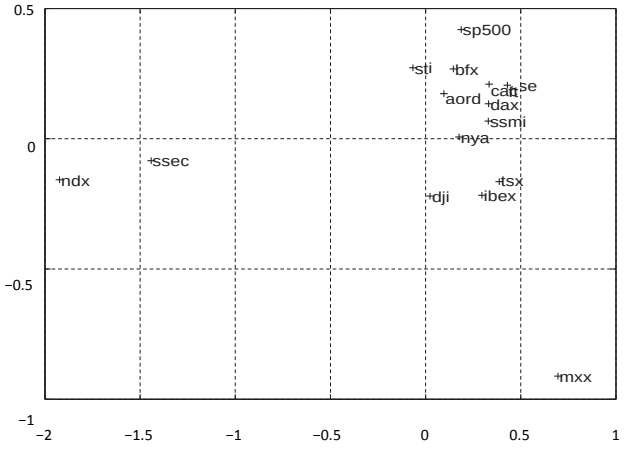

Fig. 8. Two dimensional MDS graphs for the fifteen indexes using histogram's distance d^b.

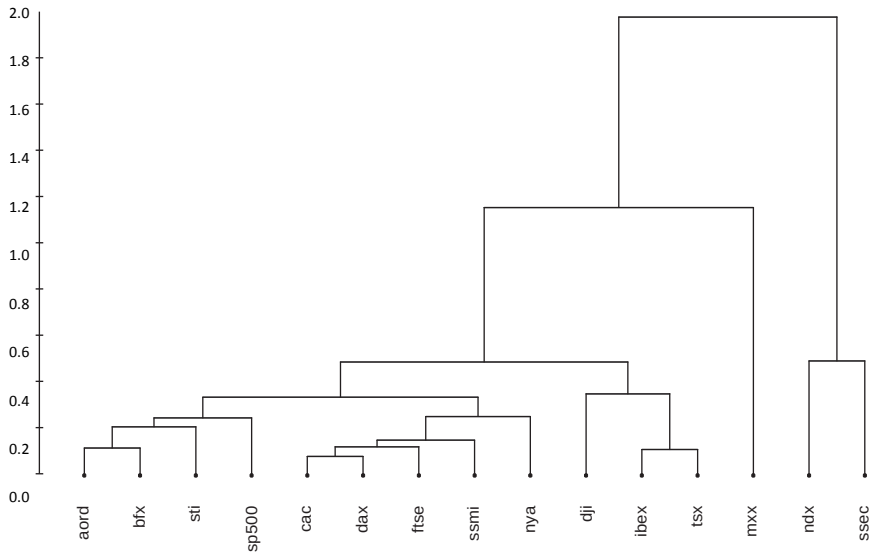

Fig. 9. Dendogram for the fifteen indexes using histogram's distance d_b.

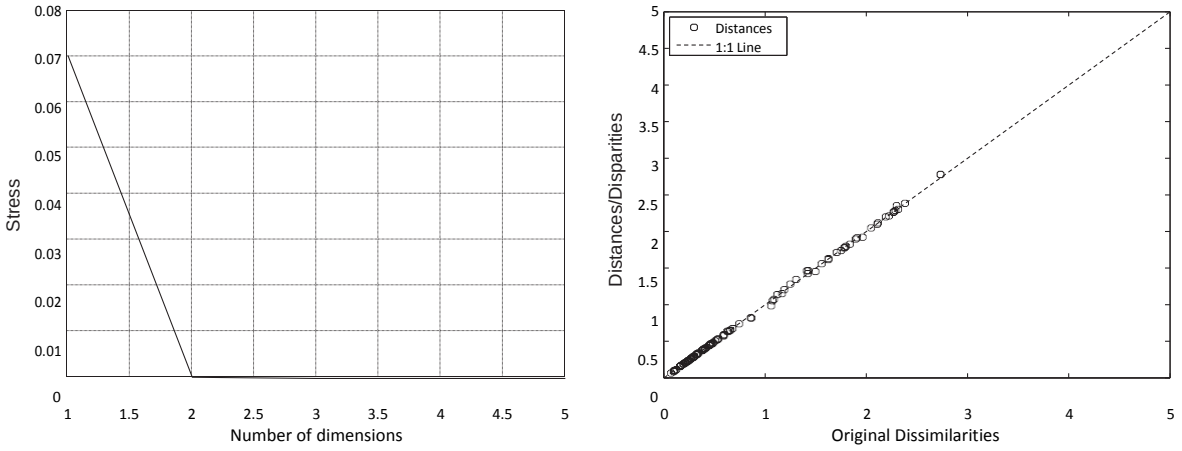

Fig. 10. Stress (left) and Shepard (right) plots of 2D MDS representation for all indexes *vs* number of dimension, using histogram's distance d^b.

We observe again that the base information is identical, but that MDS is more intuitive. Furthermore, in these two cases, 2-dimensional chart are sufficient.

Curiously in the chart corresponding to the MDS based on correlation (Fig. 2) we can see an V shape with the NDX index at the vertex, and the BFX and AORD at the corners. The MXX and the SSEC indexes are out of the angle form. However, in the chart corresponding to the MDS based on the histogram distances (Figs. 5 and 8) such shape cannot be found. Instead d^a leads to a long ''S'' curve having the DJI and the SSEC indexes as extremes emerges can be observed in Fig. 5. On the other hand, d^b produces the map of Fig. 8 where the SSEC, NDX and MXX are far apart from the rest of the points similarly to what occurs in the map of Fig. 2.

It is interesting to note that in all cases the MXX index behaves differently from the other (i.e., is not part of the shapes and regularities formed). Perhaps this may explained by the fact that Mexico was less affected by the *dot.com* crisis in the beginning of the period under study, since then it was strongly emerging from its own *Mexican Peso Crisis*.

4. Conclusion

In this paper, we proposed simple graphical tools to visualize time-varying correlations between stock market behaviour. We illustrated our MDS-based method daily close values of fifteen stock markets. There are several issues relevant for further research. A first issue concerns applying our method to alternative data sets, with perhaps different sampling frequencies or returns and absolute returns, to see how informative the method can be in other cases. A second issue concerns taking the

graphical evidence seriously and incorporating it in an econometric time series model to see if it can improve empirical specification strategies.

References

[1] Ahrens B. Distance in spatial interpolation of daily rain gauge data. Hydrol Earth Syst Sc 2006;10:197–208.

[2] Benabdeslem K, Bennani Y. Dendogram-based svm for multi-class classification. J Comput Inf Technol, GIT 2006;14(4):283–9. doi:10.2498/ cit.2006.04.03.

[3] Black F, Scholes M. The pricing of options and corporate liabilities. J Polit Econ 1973;81(3):637–54.

[4] Borg I, Groenen P. Modern multidimensional scaling:theory and applications. New York: Springer; 2005.

[5] Buja A, Swayne D, Littman M, Dean N, Hofmann H, Chen L. Data visualization with multidimensional scaling. J Comput Graph Stat 2008;17(85):444–72.[6] Campbell JY. Stock returns and the term structure. J Financ Econ 1987(18):373–400.

[7] Chaa SH, Srihari S. On measuring the distance between histograms. Pattern Recogn 2002(35):1355–70. [8] Cox T, Cox M. Multidimensional scaling. New York: Chapman & HallCrc; 2001.

[9] Duarte FB, Machado JT, Duarte GM. Dynamics of the dow jones and the nasdaq stock indexes. Nonlinear Dyn 2010;61(4):691–705. [10] Fama EF, Schwert GW. Asset returns and inflation. J Financ Econ 1977(5):115–46.

[11] Fernández A, Gómez S. Solving non-uniqueness in agglomerative hierarchical clustering using multidendograms. J Classif 2008;1(25):43–65. doi:10.1007/s00357-008-9004-x.

[12] <http://finance.yahoo.com> (2010).

[13] Kruskal J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964;29(1):1–27. <http://dx.doi.org/10.1007/BF02289565>.

[14] Kruskal J, Wish M. Multidimensional scaling. Newbury Park, CA: Sage Publications, Inc; 1978.

[15] Machado JT, Duarte GM, Duarte FB. Identifying economic periods and crisis with the multidimensional scaling. Nonlinear Dyn 2010. doi:10.1007/s11071-010-9823-2. ISSN: 0924-090x.

[16] Nigmatullin R. Universal distribution function for the strongly-correlated fluctuations: general way for description of different random sequences. Commun Nonlinear Sci Numer Simulat 2010;15(3):637–47.

[17] Nirenberg S, Latham PE. Decoding neuronal spike trains: how important are correlations? Proc Nat Acad Sci 2003;100(12):7348–53.

[18] Plerou V, Gopikrishnan P, Rosenow B, Amaral L, Stanley H. Econophysics: financial time series from a statistical physics point of view. Phys A 2000;279:443–56.

[19] Ramsay JO. Some small sample results for maximum likelihood estimation in multidimensional scaling. Psychometrika 1980;45(1):139–44.

[20] Serratosa F, Sanroma G, Sanfeliu A. A new algorithm to compute the distance between multi-dimensional histograms. Lect Notes Comput Sci 2008;4756(1):115–23.

[21] Shepard R. The analysis of proximities:multidimensional scalling with an unknown distance function. Psychometrika I 1962(27):219–46.

[22] Sierra B, Lazkano E, Jauregi E, Irigoien I. Histogram distance-based bayesian network structure learning: a supervised classification specific approach. Decis Support Syst 2009;48(1):180–90. <http://dx.doi.org/10.1016/j.dss.2009.07.010>.

[23] Werman M, Peleg S, Rosenfeld A. A distance metric for multidimensional histograms. Comput Vision Graph Image Process 1985(32):328–36.[24] Woelfel J, Barnett GA. Multidimensional scaling in Riemann space. Qual Quant 1982;16(6):469–91.