Linear and nonlinear models for the analysis of the relationship between stock market prices and macroeconomic and financial factors

Andreia Dionísio, Rui Menezes**,*
*Diana Mendes**, Jacinto Vidigal da Silva**
*University of Évora, Department of Management,
Largo dos Colegiais, 2, 7000 Évora Portugal
andreia@uevora.pt
**ISCTE, Department of Quantitative Methods,
Av. Forcas Armadas, 1649-025 Lisboa, Portugal

Abstract

The main objective of this paper is to assess how mutual information as a measure of global dependence between stock markets and macroeconomic factors can overcome some of the weaknesses of the traditional linear approaches commonly used in this context. One of the advantages of mutual information is that it does not require any prior assumption regarding the specification of a theoretical probability distribution or the specification of the dependence model. This study focuses on the Portuguese stock market where we evaluate the relevance of the macroeconomic and financial variables as determinants of the stock prices behaviour.

Keywords: Nonlinear dependence, macroeconomic and financial factors, mutual information.

1 Introduction

It is quite common to find in the financial literature theories and models based on the efficient market hypothesis, which implies that prediction and forecasting based on historical rates of return or other factors are not possible to perform in practice. This argument has been reinforced by empirical findings that stock prices follow a random walk process. Therefore, an alternative way for studying the relationship between the economic activity represented by macroeconomic factors and the behavior of prices in the stock market lies on the analysis of long-run trends based on monthly observations [Pesaran *et al.* (1995)].

Traditionally, the study of such links has been made on the basis of linear models. However, there are many authors that argue that this type of analysis is in general inconclusive because linear independence is not synonymous of independence, being thus necessary to ascertain the possibility of the existence of nonlinear dependence [Darbellay (1998); Maasoumi *et al.* (2002)].

This paper investigates the relationship between the behavior of certain economic factors and the Portuguese stock market prices by means of linear and nonlinear approaches based upon traditional single equation linear models and global dependence tests (linear and nonlinear) using mutual information and the global correlation coefficient. The main goal is to access dependence in a global way, linear and nonlinear, and independently of any previously assumed model. In this context we use in this paper mutual information in attempting to evaluate the ability of this measure to capture dependence in financial time series.

The paper is organized as follows. Section 2 presents the theoretical framework for accessing the relationship between the behavior of stock markets and various macroeconomic and financial factors. Section 3 presents mutual information as a measure of global dependence, describes the properties of mutual information and the estimation procedure adopted. In Section 4 we describe and justify the data used in our analysis and the results obtained from implementing the methodologies adopted in our study. Both single linear equation models and nonlinear mutual information models were employed in our context as referred to above. The final Section presents some concluding remarks of this study.

2 Background

Asset prices are commonly believed to react sensitively to economic news. Furthermore, daily experience seems to support the view that individual asset prices are influenced by a wide variety of unanticipated events and that some events have more persuasive effects on asset prices than others. In this context, the portfolio theory, based on the diversification effect, focused its attention on the systematic risk. The general conclusion of that theory is that an additional component of long-run return is needed and obtained whenever a particular asset is influenced by systematic economic news and there is no possibility to make extra profit in a diversified portfolio. However, the economic theory, like usually happens in these circumstances, says nothing about the definition of the events with capability to influence asset prices.

There are several studies based in linear models, which results point to the importance of some macroeconomic variables [Chen *et al.* (1986); Pesaran *et al.* (1995); Haugen *et al.* (1996)], business conditions [Fama *et al.* (1989); Fama (1990); Fama *et al.* (1993)] and the real activity [McQueen *et al.* (1993)]. There are alternative approaches that consider the existence of bidirectional relationships between stock returns and macroeconomic variables, revealing in some cases that it is the stock market that leads the real economic activity [see e.g. Fama (1990); Binswanger (2001)].

Most of the models used to study the relationship between the behavior of stock returns and macroeconomic and financial variables were based on linear regression techniques estimated by OLS. In this sense, the possible nonlinear effect was omitted as well as the possible feedback effects. Besides, the estimated coefficients may suffer severe biases since the residuals hardly behave as a white noise. In this sense, the use of nonlinear models to explain in a different way the relationship between the macroeconomic variables and the stock returns may bring about some "fresh air" into this field [e.g. Stuzer (1995); Qi (1999); Maasoumi *et al.* (2002)]

The literature just reviewed allows us to conclude that there exists a potentially important predictability component in stock price movements through the knowledge and exploration of macroeconomic and financial variables, since most of the studies exhibit statistical significance in their relationship.

Globally, we retain an overall impression that there is a set of variables that may affect stock returns and can show some feedback effects. The majority of the studies in this field point to the existence of predictability of stock returns, but the rejection of the efficient market hypothesis based on this evidence was not sufficient for the referred majority of authors.

3 Mutual information

One of the most practical ways to evaluate the (in)dependence between two vectors of random variables $\overrightarrow{X}, \overrightarrow{Y}$ is to consider a measure that assumes the value 0 when there is total independence and 1 when there is total dependence. Let $p_{\overrightarrow{X},\overrightarrow{Y}}(A \times B)$ be the joint probability distribution of $\left(\overrightarrow{X}, \overrightarrow{Y}\right)$ and $p_{\overrightarrow{X}}(A)$, $p_{\overrightarrow{Y}}(B)$ the underlying marginal probability distributions, where A is a subset of the observation space of $\overrightarrow{X}$ and B is a subset of the observation space of $\overrightarrow{Y}$, such that we can evaluate the following expression:

$$\ln \frac{p_{\overrightarrow{X},\overrightarrow{Y}}(A \times B)}{p_{\overrightarrow{X}}(A)\, p_{\overrightarrow{Y}}(B)}. \tag{1}$$

If the two events are independent, then $p_{\overrightarrow{X},\overrightarrow{Y}}(A \times B) = p_{\overrightarrow{X}}(A)\, p_{\overrightarrow{Y}}(B)$, and so equation (1) will take the value zero.

Granger, Maasoumi and Racine (2002) consider that a good measure of dependence should satisfy the following six "ideal" properties:

(a) It must be well defined for both continuous and discrete variables;

(b) It must be normalized to zero if $\overrightarrow{X}$ and $\overrightarrow{Y}$ are independent, and lying between -1 and $+1$, in general;

(c) The absolute value of the measure should be equal to 1 if there is an exact nonlinear relationship between the variables;

(d) It must be similar or simply related to the linear correlation coefficient in the case of a bivariate normal distribution;

(e) It must be metric in the sense that it is a true measure of "distance" and not just a measure of "divergence";

(f) It must be an invariant measure under continuous and strictly increasing transformations.

3.1 Mutual information properties

The concept of mutual information comes from the theory of communication and measures the information of a random variable contained in another random variable. The definition of mutual information goes back to Shannon (1948) and the theory was extended and generalized by Gelfand, Kolmogorov and Yaglom (1956) [in Darbellay (1998a)]. According to Pompe (1998), mutual information is very useful to analyze statistical dependences in scalar or multivariate time series as well as for detecting fundamental periods, detecting optimal time combs for forecasting, modelling and analyzing the (non)stationarity of data. Some of those potentialities have been explored by Granger and Lin (1994) and Darbellay and Wuertz (2000), whose results reveal that mutual information varies in a nonstationary time series framework.

The properties of mutual information appear to confirm its importance as a measure of dependence [Soofi (1997); Darbellay *et al.* (1999), (2000); Darbellay (1998, 1999); Bernhard *et al.* (1999)]. Some of these properties will be presented and explored in this sub-section.

Broadly speaking, there are two ways for estimating mutual information: the first one consists of direct estimation and the second one requires the previous computation of the entropies in order to obtain mutual information.

If p_X, p_Y and $p_{X,Y}$ denote the *pdf* of the random variables X, Y and (X,Y), respectively, then the mutual information can be expressed by the following relation:[1]

$$I\left(X,Y\right)=\int\int p_{X,Y}\left(x,y\right)\log\frac{p_{X,Y}\left(x,y\right)}{p_X\left(x\right)p_Y\left(y\right)}dxdy. \tag{2}$$

[1] The selection of the base of the logarithm is irrelevant, but is convenient to distinguish among results: $\log_2$- entropy measure in bits; $\log_{10}$- entropy measure in dits; $\log_e = \ln$ - entropy measure in nats.

In the case of a continuous distribution, the mutual information assumes non-negative values, so we have $I(X, Y) \geq 0$, where the expression assumes the equality if and only if X and Y are statistically independent [Kullback (1968)]. The mutual information between two random variables X and Y can be considered as a measure of dependence between these variables, or even better, a measure of the statistical correlation between X and Y. However, we can not say that X is causing Y or vice-versa.

The statistic defined in equation (2) satisfies some of the desirable properties of a good measure of dependence described previously, namely *(a)* and, after some transformations, will also satisfy properties *(b)*, *(c)* and *(d)* [Granger *et al.* (2002)].[2]

In order to satisfy properties *(b)* and *(d)* it is convenient to define a measure that can be compared with the linear correlation coefficient. In equation (2), we have $0 \leq I(X, Y) \leq +\infty$, which difficults comparisons for different samples. In this way, we can compare mutual information with covariance, since both are measures of dependence and, for both, comparisons for different samples can be inconclusive.

To obtain a statistic that satisfies property *(d)* without loosing the properties *(a)* to *(c)*, it is convenient to define an equation similar to that in (3). In this context Granger and Lin (1994), Darbellay (1998a) and Soofi (1997), among others, used a standard measure for the mutual information, the global correlation coefficient, defined by:

$$\lambda\left(\overrightarrow{X}, \overrightarrow{Y}\right) = \sqrt{1 - e^{-2I\left(\overrightarrow{X}, \overrightarrow{Y}\right)}}. \tag{3}$$

This measure varies between 0 and 1 being thus directly comparable to the linear correlation coefficient.

The function $\lambda\left(\overrightarrow{X}, \overrightarrow{Y}\right)$ captures the overall dependence, both linear and nonlinear, between $\overrightarrow{X}$ and $\overrightarrow{Y}$. This measure of predictability is based on empirical probability distributions, but it does not depend on any particular model of predictability. In this particular case, the properties mentioned above assume the following form: *(i)* $\lambda\left(\overrightarrow{X}, \overrightarrow{Y}\right) = 0$, if and only if $\overrightarrow{X}$ contains no information on $\overrightarrow{Y}$; *(ii)* $\lambda\left(\overrightarrow{X}, \overrightarrow{Y}\right) = 1$, if exists a perfect relationship between the vectors $\overrightarrow{X}$ and $\overrightarrow{Y}$. This is the limit case of determinism; *(iii)* when modelling the input-output pair $\left(\overrightarrow{X}, \overrightarrow{Y}\right)$ by any model with input $\overrightarrow{X}$ and output $\overrightarrow{U} = f\left(\overrightarrow{X}\right)$, where f is some function of $\overrightarrow{X}$, the predictability of $\overrightarrow{Y}$ by $\overrightarrow{U}$ cannot exceed the predictability of $\overrightarrow{Y}$ by $\overrightarrow{X}$, i.e., $\lambda\left(\overrightarrow{X}, \overrightarrow{Y}\right) \geq \lambda\left(\overrightarrow{U}, \overrightarrow{Y}\right)$.

It is well known that the Gaussian distribution maximize the entropy of Shannon for given first and second moments. This implies that the Shannon entropy of any distribution is bounded upwards by the normal mutual information (NMI), and depends on the covariance matrix [Kraskov *et al.* (2003)].

[2] The demonstration of some theorems about mutual information properties can be found in Kullback, S. (1968). *Information Theory and Statistics,* Dover, New York.

When $d = 2$, that is, for $\left(\overrightarrow{X},\overrightarrow{Y}\right) = (X, Y)$, NMI takes the form [Kullback (1968)]:

$$I(X,Y) = -\frac{1}{2}\log\left(1 - r^2(X,Y)\right). \tag{4}$$

If the empirical distribution is normal, the mutual information can be calculated by equation (4), because the normal distribution is a "linear" distribution in the sense that the linear correlation coefficient in this context captures the overall dependence. In this case, any empirical mutual information must be greater or equal to the normal mutual information [Kraskov *et al.* (2003)].

Intuitively, one would like to have a measure of predictability larger than the measure of linear predictability, i.e. $\lambda \geq r$. Unfortunately, this is not always true [Darbellay (1998)].[3] It is important to note that the difference $(\lambda - r)$ cannot be equated with the nonlinear part of the dependency. Nevertheless, if the distribution is normal, we do have $\lambda\left(\overrightarrow{X},\overrightarrow{Y}\right) = \left|r\left(\overrightarrow{X},\overrightarrow{Y}\right)\right|$, and in $\mathbb{R}^2$ we have $\lambda(X,Y) = |r(X,Y)|$ [Granger *et al.* (1994); Darbellay (1998)].

Maasoumi (1993) shows that the mutual information does not satisfy property *(e)*. In this case, mutual information is just a measure of divergence because it does not satisfy the triangular inequality.

Another important property of the mutual information is additivity, saying that it can be decomposed into hierarchical levels [Shannon (1948); Kraskov *et al.* (2003)], that is $I\left(\overrightarrow{X},\overrightarrow{Y},\overrightarrow{Z}\right) = I\left(\left(\overrightarrow{X},\overrightarrow{Y}\right),\overrightarrow{Z}\right) + I\left(\overrightarrow{X},\overrightarrow{Y}\right)$. It follows that $I\left(\overrightarrow{X},\overrightarrow{Y},\overrightarrow{Z}\right)$ will be always greater or equal to $I\left(\overrightarrow{X},\overrightarrow{Y}\right)$. By the same token, the coefficient of linear determination and the coefficient of linear correlation cannot decrease when one adds more variables to the model.

3.2 The test of independence

Independence is one of the most valuable concepts in econometrics. Thus, according to the properties of mutual information, we can construct an independence test based on the following hypothesis:$H_0 : p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$, $H_1 : p_{X,Y}(x,y) \neq p_X(x)\,p_Y(y)$. If H_0 holds then $I(X,Y) = 0$ and we conclude that the variables are independent. If H_1 holds then $I(X,Y) > 0$ and we reject the null hypothesis of independence. The above hypothesis can be reformulated in the following way:

$$H_0 : I(X,Y) = 0,\ \ H_1 : I(X,Y) > 0.$$

In order to test more accurately for the independence between the variables (or vectors of variables) we will need to compute the critical values of the distribution. There are three different approaches to obtain the critical values of our tests under the null hypothesis: (1) asymptotic approximations for the

[3] A situation that can induce $\lambda < r$ is the small size of the sample. A small size, in this context, is a sample with $n \leq 500$.

null distribution; (2) simulated critical values for the null distribution and (3) permutation-based critical values for the null distribution.

The critical values calculated in this paper for the mutual information are based upon simulated critical values for the null distribution on the basis of a percentile approach (see *Appendix A*). These values were found through a simulation of critical values based upon a white noise for a number of sample sizes. Given that the distribution of mutual information is skewed, we can adopt a percentile approach to obtain critical values.

One of the difficulties for estimating the mutual information from empirical data lies in the fact that the underlying *pdf* is unknown. There are, essentially, three different methods to estimate mutual information: histogram-based estimators; kernel-based estimators; parametric methods. In this paper the method used for estimation of mutual information is the marginal equiquantization (the partition of the space in equiprobable cells).[4]

4 Data and results

The purpose of this study is to evaluate the level of dependence between the Portuguese stock market and a set of economic and financial factors selected according to the relevant literature in this field [see e.g. Chen *et al.* (1986); Asprem (1989); Fama *et al.* (1993); McQueen *et al.* (1993); Pesaran *et al.* (1995); Maasoumi *et al.* (2002)]. The definition and source of the indicators[5] that were selected, as well as the definition of the variables computed on the basis of the selected indicators, are shown in Tables 1 and 2.

Let ER_t be the monthly excess return; $\Delta Lisbor3M_t$ the short-term interest growth rate; $\Delta Swap10_t$ the long-term interest growth rate; ΔDY_t the dividend yield growth rate; ΔEPR_t the earnings price ratio growth rate; ΔIPC_t the CPI growth rate; ΔPIM_t the monthly industrial production growth rate; ΔPIA_t the year on year industrial production growth rate; ΔTD_t the unemployment growth rate and ΔOIL_t the oil price growth rate. The variables were computed in the following way (Table 2).

According to some authors [e.g. Chen *et al.* (1986), Fama (1990), McQueen *et al.* (1993)] we should use the unanticipated changes in the variables, or the respective innovations. Because some of the time series present evidence of significant autocorrelation an seasonality, it was necessary to perform a filtering of these series. The filtered time series were computed in the following way (see Table 3).

We applied linear models to evaluate the relationship between the rate of returns and the macroeconomic and financial variables. The results demonstrate

[4] For a good explanation of this estimation method see Darbellay, G. (1998) and Darbellay, G. and Vadja, I. (1999).

[5] The standard period is 1 month, thus, $E(|t-1)$ denotes the expectation operator at the end of the month $t-1$ conditional on the information set available at the end of the month $t-1$. $X(t)$ denotes the value of the variable X in month t, or the growth rate that prevailed from the end of $t-1$ to the end of t.

Indicator	Symbol	Font and definition
Price index of the portuguese stock market	PI_t	Monthly price index Font: Data base DataStream
Short-term interest rate	$Lisbor3M_t$	Font: Data base Dhatis
Long-term interes rate	$Swap10_t$	Rate of return of a Swap 10 years Font: Data base Dhatis
Dividend yield	DY_t	Dividends/price ratio Font: Data base DataStream
Earnings price ratio	EPR_t	Earnings/price ratio Font: Data base DataStream
Consumer price index	IPC_t	Font: Data base DataStream
Industrial production index	IPI_t	Font: INE
Unemplyment	TD_t	Font: Data base DataStream
Oil prices	OIL_t	Spot oil prices in the USA market Font:http://www.eia.doe.gov/oil_gas/ petroleum/info_glance/prices.html

Table 1: Glossary and definition of indicators. All the indicators are monthly measured and the period in analysis is October 1993 to October 2003. In the statistical analysis, all the indicators show the existence of a unit root according to the Dickey-Fuller test.

that the only significant explanatory variables that are retained by the multivariate model are: ΔDY_t, ΔEPR_t, ΔEPR_{t-1} and $inovIPC_{t-3}$. Not surprisingly, we should note that are precisely the financial variables (ΔDY_t, ΔEPR_t and ΔEPR_{t-1}) that seem to maintain the largest explanatory and predictive power on the *excess return*. In what refers to the macroeconomic variables, only the $inovIPC_{t-3}$ presents statistical significance in a multivariate context, showing a negative correlation with the *excess return*. These results are in accordance with the results reported by other authors, namely *inter alia* Fama (1990), Fama and French (1993), and Maasoumi and Racine (2002).

The analysis of nonlinear dependence is justified because the financial time series may exhibit strong nonlinear components that may be transmitted from one market to the other [see e.g. Hsieh (1991)]. Furthermore, if we only consider the linear relationships or dependencies, we are simultaneously assuming that these relationships are time invariant, which is not usually consistent with the empirical evidence.

In this research work we use mutual information and the global correlation coefficient as a measure of global dependence, where this statistic can be compared with the usual measure of linear correlation. As previously referred, the mutual information has properties that render this measure as an important and widely explored measure and test of independence [e.g. Granger *et al.* (1994); Soofi (1997); Darbellay *et al.* (1999, 2000); Darbellay (1998, 1999); Bernhard *et al.* (1999); Dionísio *et al.* (2003)].

Variable	Definition
ER_t	$\frac{PI_t - PI_{t-1} + D_t}{PI_{t-1}} - Lisbor3M_{t-1}$
$\Delta Lisbor3M_t$	$\ln(Lisbor3M_t) - \ln(Lisbor3M_{t-1})$
$\Delta Swap10_t$	$\ln(Swap10_t) - \ln(Swap10_{t-1})$
ΔDY_t	$\ln(DY_t) - \ln(DY_{t-1})$
ΔEPR_t	$\ln(EPR_t) - \ln(EPR_{t-1})$
ΔIPC_t	$\ln(IPC_t) - \ln(IPC_{t-1})$
ΔPIM_t	$\ln(IPI_t) - \ln(IPI_{t-1})$
ΔPIA_t	$\ln(IPI_t) - \ln(IPI_{t-12})$
ΔTD_t	$\ln(TD_t) - \ln(TD_{t-1})$
ΔOIL_t	$\ln(OIL_t) - \ln(OIL_{t-1})$

Table 2: Glossary and definition of variables refering to the constructed economic factors.

New variable	*Process*
$inovLisbor$	$ARMA(1,0)$ of $\Delta Lisbor3M$
$inovSwap$	$ARMA(1,0)$ of $\Delta Swap10$
$inovIPC$	$ARMA(3,1)$ ofe $\Delta IPCSA$ $\Delta IPCSA$ is the seasonal adjustment of ΔIPC
$inovPIM$	$ARMA(2,0)$ of $\Delta PIMSA$ $\Delta PIMSA$ is the seasonal adjustment of ΔPIM
$inovPIA$	$ARMA(1,1)$ of ΔPIA
$inovTD$	$ARMA(1,1)$ of ΔTD

Table 3: Filtered series to get the unancipated changes in all variables. The seasonal adjustment was realised through the moving average method.

We first computed the mutual information (I), the normal mutual information (NMI), the global correlation coefficient (λ) and the linear correlation coefficient (R) between the excess return in levels and each of the remaining variables measured with lags (see Tables 4 e 5). We should emphasize that mutual information does not establish any causality relationships between the variables under study, it measures the global dependence that may exist between them as a whole. In this way, the mutual information takes into account the bidirectional relationships that can be established between the variables.

According to the results presented in Tables 4 and 5 we can verify that the empirical mutual information (I) is higher in most cases than the normal mutual information (NMI), as well as the global correlation coefficient (λ) is higher than the linear correlation coefficient (R). These differences can reveal the presence of nonlinear dependence for the majority of the pairs of variables under study. The relationships that showed statistical significance were: ER_t and $inovLisbor_t$; ER_t and ΔDY_t; ER_t and ΔEPR_t; ER_t and ΔEPR_{t-1}; ER_t and $inovPIM_{t+2}$; ER_t and $inovPIM_{t-3}$; ER_t and ΔOIL_t; ER_t and ΔOIL_{t-3};

Variable	I $(nats)$	λ	NMI $(nats)$	R
$inovLisbor_t$	0.0413*	0.2816	0.0083	0.1285
$inovLisbor_{t-1}$	0.0175	0.1855	0.0128	0.1591
$inovLisbor_{t-2}$	0.0083	0.1283	0.0060	0.1091
$inovLisbor_{t-3}$	0.0024	0.0692	0.0011	0.0464
$inovSwap_t$	0.0043	0.0925	0.0009	0.0412
$inovSwap_{t-1}$	0.0036	0.0847	0.0030	0.0775
$inovSwap_{t-2}$	0.0195	0.1956	0.0170	0.1830
$inovSwap_{t-3}$	0.0095	0.1372	0.0009	0.0412
ΔDY_t	0.7740**	0.8873	0.2182**	0.5946
ΔDY_{t-1}	0.0103	0.1428	0.0065	0.1136
ΔDY_{t-2}	0.0001	0.0167	0.0006	0.0346
ΔDY_{t-3}	0.0018	0.0599	0.0115	0.1510
ΔEPR_t	0.7108**	0.8710	0.2937**	0.6695
ΔEPR_{t-1}	0.0599*	0.3360	0.0193	0.1944
ΔEPR_{t-2}	0.0001	0.0141	0.0001	0.0100
ΔEPR_{t-3}	0.0083	0.1283	0.0071	0.1187
$inovIPC_t$	0.0010	0.0436	0.0165	0.1800
$inovIPC_{t-1}$	0.0009	0.0424	0.0001	0.0141
$inovIPC_{t-2}$	0.0009	0.0424	0.0009	0.0424
$inovIPC_{t-3}$	0.0262	0.2259	0.0198	0,1970

Table 4: Mutual information (I) in nats, global correlation coefficient (λ), normal mutual information (NMI) and linear correlation coefficient (R) between ER_t and each of the variables (*per si*) for different lags.

Variable	I $(nats)$	λ	NMI $(nats)$	R
$inovPIM_{t+3}$	0.0010	0.0440	0.0005	0.0300
$inovPIM_{t+2}$	0.0342*	0.2571	0.0002	0.0200
$inovPIM_{t+1}$	0.0064	0.1128	0.0000	0.0000
$inovPIM_{t}$	0.0000	0.0000	0.0004	0.0283
$inovPIM_{t-1}$	0.0095	0.1372	0.0025	0.0700
$inovPIM_{t-2}$	0.0006	0.0346	0.0192	0.1942
$inovPIM_{t-3}$	0.0952**	0.4164	0.0000	0.0000
$inovPIA_{t+3}$	0.0014	0.0529	0.0006	0.0346
$inovPIA_{t+2}$	0.0046	0.0957	0.0003	0.0245
$inovPIA_{t+1}$	0.0000	0.0000	0.0036	0.0843
$inovPIA_{t}$	0.0003	0.0245	0.0000	0.0000
$inovPIA_{t-1}$	0.0018	0.0599	0.0023	0.0678
$inovPIA_{t-2}$	0.0030	0.0773	0.0001	0.0100
$inovPIA_{t-3}$	0.0095	0.1372	0.0002	0.0173
$inovTD_{t}$	0.0029	0.0760	0.0273	0.2304
$inovTD_{t-1}$	0.0013	0.0510	0.0002	0.0173
$inovTD_{t-2}$	0.0001	0.0141	0.0001	0.0141
$inovTD_{t-3}$	0.0095	0.1372	0.0224	0.2093
ΔOIL_{t}	0.0361*	0.2639	0.0175	0.1855
ΔOIL_{t-1}	0.0060	0.1092	0.0044	0.0933
ΔOIL_{t-2}	0.0013	0.0510	0.0028	0.0055
ΔOIL_{t-3}	0.0414*	0.2819	0.0001	0.0002

Table 5: Mutual information (I) in nats, global correlation coefficient (λ), normal mutual information (NMI) and linear correlation coefficient (R) between ER_t and each of the variables (*per si*) for different lags.

which seems limiting. The small number of statistically significant global dependences between the variables may be caused by the small samples (about 118 observations) obtained, which can underestimate the value of the mutual information.

The pairs of variables ER_t and ΔDY_t; ER_t and ΔEPR_t and ER_t and ΔEPR_{t-1}, present the highest level of global dependence, which can be an indicator of the presence of nonlinear dependence. The significant differences between λ and R in these cases (and between I and NMI) may be caused by the fact that the variables are not normally distributed and the residuals resulting from estimating the linear regression models presented in Tables **??** e **??** show evidence of autocorrelation and heteroscedasticity. In this context, the simple linear regression analysis may not be sufficient to analyze the level of dependence between the *excess return* and the macroeconomic and financial variables.

If we take into account all the variables that show statistical significance in this preliminary study and calculate the mutual information between them, we obtain the following result:

$$I\begin{pmatrix} ER_t, inovLisbor_t, \Delta DY_t, \Delta EPR_t, \Delta EPR_{t-1}, \\ inovPIM_{t+2}, inovPIM_{t-3}, \Delta OIL_t, \Delta OIL_{t-3} \end{pmatrix} = 1.8517, \tag{5}$$

which means that $\lambda = 0,9876$. The value of the mutual information of equation (5) does not present statistical significance. This fact could be a sign that we should eliminate some variables in order to decrease the degrees of freedom without great impact on the value of mutual information. To this end, we drop each variable individually and in turn, except ER_t, in order to obtain the new values of mutual information. We compute the following models:

$$I\begin{pmatrix} ER_t, inovLisbor_t, \Delta DY_t, \Delta EPR_t, \Delta EPR_{t-1}, \\ inovPIM_{t+2}, inovPIM_{t-3}, \Delta OIL_t \end{pmatrix} = 1.4154^{*} \tag{6}$$

$$I\begin{pmatrix} ER_t, inovLisbor_t, \Delta DY_t, \Delta EPR_t, \Delta EPR_{t-1}, \\ inovPIM_{t+2}, inovPIM_{t-3}, \Delta OIL_{t-3} \end{pmatrix} = 1.5455^{**} \tag{7}$$

$$I\begin{pmatrix} ER_t, inovLisbor_t, \Delta DY_t, \Delta EPR_t, \Delta EPR_{t-1}, \\ inovPIM_{t+2}, \Delta OIL_t, \Delta OIL_{t-3} \end{pmatrix} = 1.3926 \tag{8}$$

$$I\begin{pmatrix} ER_t, inovLisbor_t, \Delta DY_t, \Delta EPR_t, \Delta EPR_{t-1}, \\ inovPIM_{t-3}, \Delta OIL_t, \Delta OIL_{t-3} \end{pmatrix} = 1.5134^{**} \tag{9}$$

$$I\begin{pmatrix} ER_t, inovLisbor_t, \Delta DY_t, \Delta EPR_t, \\ inovPIM_{t+2}, inovPIM_{t-3}, \Delta OIL_t, \Delta OIL_{t-3} \end{pmatrix} = 1.4350^{*} \tag{10}$$

$$I\begin{pmatrix} ER_t, inovLisbor_t, \Delta DY_t, \Delta EPR_{t-1}, \\ inovPIM_{t+2}, inovPIM_{t-3}, \Delta OIL_t, \Delta OIL_{t-3} \end{pmatrix} = 1.2305 \tag{11}$$

$$I\begin{pmatrix} ER_t, inovLisbor_t, \Delta EPR_t, \Delta EPR_{t-1}, \\ inovPIM_{t+2}, inovPIM_{t-3}, \Delta OIL_t, \Delta OIL_{t-3} \end{pmatrix} = 1.3664 \tag{12}$$

$$I\begin{pmatrix} ER_t, \Delta DY_t, \Delta EPR_t, \Delta EPR_{t-1}, \\ inovPIM_{t+2}, inovPIM_{t-3}, \Delta OIL_t, \Delta OIL_{t-3} \end{pmatrix} = 1.4117^{*} \tag{13}$$

The values of mutual information computed in equations (6) to (13) show that when we take away (individually) the variables ΔOIL_{t-3}, ΔOIL_t, $inovPIM_{t+2}$, ΔEPR_{t-1} or $inovLisbor_t$ the mutual information became statistically significant. This fact can be a sign that the information contribution of those variables (which can be interpreted as a sort of marginal mutual information) is not very strong when analyzed jointly with other variables. We should also note that the variables $\Delta DY_t, \Delta EPR_t$ and $inovPIM_{t-3}$ which were statistically significant at 1% in the previous analysis (see Tables 4 and 5) are precisely the variables that show here more informative contribution in a set of variables including ER_t. If we take only the variables $ER_t, \Delta DY_t, \Delta EPR_t$ and $inovPIM_{t-3}$, the value of the mutual information is:

$$I\left(ER_t, \Delta DY_t, \Delta EPR_t, inovPIM_{t-3}\right) = 1.3021,^{**} \tag{14}$$

which is statistically significant and confirms the existence of linear and possibly nonlinear dependence between these variables.

From the present analysis we noticed that the set of macroeconomic and financial variables that are more correlated with the *excess return* is not very different from the one that we found using linear regression analysis. If we apply the same methodology to the variables used in equation (14) (these variables present a level of significance of 1% in the analysis of global dependence displayed in Tables 4 and 5), the mutual information will assume the value presented in equation (15) :

$$I\left(\Delta DY_t, \Delta EPR_t, inovPIM_{t-3}\right) = 0.4311^{**}, \tag{15}$$

thus, the mutual information between ER_t and the set of explanatory variables composed by $\Delta DY_t, \Delta EPR_t, inovPIM_{t-3}$ is:

$$I\left[ER_t, \left(\Delta DY_t, \Delta EPR_t, inovPIM_{t-3}\right)\right] = 0.8710.^{**} \tag{16}$$

The global dependence between ER_t and a vector composed by the variables $\Delta DY_t, \Delta EPR_t$, and $inovPIM_{t-3}$ takes the value of 0, 8710 nats, which corresponds to a global correlation coefficient of $\lambda = 0, 9082$. If we estimate a linear regression model with these variables, namely:

$$ER_t = \alpha + \beta_1 \Delta DY_t + \beta_2 \Delta EPR_t + \beta_3 inovPIM_{t-3} + \varepsilon_t, \tag{17}$$

we would obtain a linear correlation coefficient of $R = 0.7420$, smaller than the correspondent global correlation coefficient. This difference could be generated by the possible presence of nonlinear dependences, which may be a reflex of the leptocurtosis (fat-tails) and skewness of the residuals resulting from the estimation of the equation (17). According to some authors [e.g. Peters (1996)] the presence of fat-tails may be a good sign of the existence of nonlinearities of the variables under study.

In general, we can say that the mutual information and the global correlation coefficient seem to have some advantages relatively to the linear approach, since they have the ability to capture the dependence as a whole (linear and nonlinear). This ability allows for the inclusion of some explanatory variables that do not show a significant explanatory power in linear terms, and incorporate nonlinearities that are important to consider. The results can only be fully explored when it is possible to specify the nonlinear models themselves or the type of nonlinearity that lies behind this dependence. Even so, we believe that it is important to take account of the existence of possible nonlinearities and try to identify them.

The main limitations of the mutual information as a dependence measure between variables are the fact that it does not identify the eventual relationship of causality nor the sign of the implied correlation. Moreover, the mutual information may lose some robustness for small samples and be underestimated in these cases. In this context, we think that the mutual information and the global correlation coefficient could be used as complementary approaches to the traditional linear approach, which leads to a more rich analysis of the phenomenon.

5 Concluding remarks

This paper presents an analysis of the relationship between the Portuguese stock market and a set of macroeconomic and financial factors that were chosen according to the relevant literature in this field. Such relationship was studied using two different approaches, focusing mainly on the short-term component of the market: the single linear equation approach and the global approach that accounts both for linear and nonlinear components. Globally, our results indicate that some explanatory variables appear to have a statistically significant influence on the *excess return* and thus may constitute good proxies for this variable. We can highlight in this context the variables ΔDY_t and ΔEPR_t, which reveals that, for the time period under analysis and the set of variables that were included in our study, the variables that are more related to financial aspects performed better than the macroeconomic variables These results are in line with some of those obtained by Fama and French (1993), according to which the variables related to firms are stronger proxies to the *excess return* of stock prices than the macroeconomic variables.

In the nonlinear approach we explored some of the properties of mutual information (I) and of the global correlation coefficient (λ). The results obtained for these measures are mostly larger than those of the normal mutual information (NMI) and the linear correlation coefficient (R), respectively, which seems to indicate the possibility that there exists a nonlinear dependence between ER_t and the remaining variables. The mutual information does not provide any guidance about the causality that may exist between the variables. Rather it focuses on the dependence between them as a whole, which may constitute an advantage because there is no need to establish *a priori* any structure of dependence.

In our analysis we have seen that the variables ΔDY_t, ΔEPR_t and $inovPIM_{t-3}$ are those that prove to be more deeply related with ER_t. The main differences that we found between the values of the global correlation coefficient (λ) and the corresponding linear correlation coefficient may be caused by the non-normality of the stochastic variables and the fact that the residuals resultant from the estimation of some regressions are not white noise, having undesired evidence of autocorrelation, heteroscedasticity, and non-normality. We should emphasize that the samples used in our study are of small size (about 118 observations), which may lead to an underestimation of the value of mutual information, and weaken the strength of the results that were presented. Taking into account the advantages and the limitations of mutual information as a measure of dependence and test of independence, we believe that such approach can be a useful complement to the measures currently used in the single and multiequation linear approaches, thus promoting a more complete analysis of the phenomenon under study.

References

[1] Asprem, M. (1989). Stock Prices, Asset Portfolios and Macroeconomic Variables in Ten European Countries, *Journal of Banking and Finance*, 13, 589-612.

[2] Bernhard, H. and Darbellay, G. (1999). Performance Analysis of the Mutual Information Function for Nonlinear and Linear Processing, *Acts: IEEE International Conference on Acoustics, Speeche and Signal Processing (USA)*,3, 1297-1300.

[3] Binswanger, M. (2001). Does the Stock Market Still Lead Real Activity? - An Investigation for the -7 Countries, Series A: *Discussion Paper 2001-04*, Solothurn University of Applied Sciences Northwestern, Switzerland.

[4] Chen, N-F., Roll, R. and Ross, S. (1986). Economic Forces and the Stock Market, *Journal of Business*, 59, 3, July, 383-403.

[5] Darbellay, G. (1998). An Adaptative Histogram Estimator for the Mutual Information, *UTIA Research Report n.º 1889*, Acad. Sc., Prague.

[6] Darbellay, G. (1999). An Estimator of the Mutual Information Based on a Criterion for Independence, *Computational Statistics and Data Analysis*, 32, 1-17.

[7] Darbellay, G. and Vadja, I. (1999). Estimation of the Information by an Adaptative Partitioning of the Observation Space, *IEEE Transactions on Information Theory*, 45, May, 1315-1321.

[8] Darbellay, G. and Wuertz, D. (2000). The Entropy as a Tool for Analyzing Statistical Dependence's in Financial Time Series, *Physica A*, 287, 429-439.

[9] Dionísio, A.; Menezes, R. and Mendes, D. (2003) . Teoria da Informação: Ferramenta de Análise da Dependência Estatística en Sucessões Cronológicas Financeiras, in *Gestíon Cientifica Empresarial: Temas de Investigación Actuales*, ed. por J.M. Barreiro, J. A. Diez de Castro, B. Barreiro, R. Ruzo e F. Losada,

[10] Fama, E. (1990). Stock Returns, Expected Returns and Real Activity, *Journal of Finance*, 45, 1089-1108.

[11] Fama, E. and French, K. (1989) . Business Conditions and Expected Returns on Stocks and Bonds, *Journal of Financial Economics*, 25, 23-49.

[12] Fama, E. and French, K. (1993). Common Risk Factors in the Returns on Bonds and Stocks; *Journal of Financial Economics*, 33, 3-56.

[13] Granger, C. and Lin, J. (1994). Using the Mutual Information Coefficiente to Identify Lags in Nonlinear Models, *Journal of Time Series Analysis*, 15, 4, 371-384.

[14] Granger, C, Maasoumi, E. and Racine, J. (2002). A Dependence Metric for Possibly Nonlinear Processes, *UCSD Working Paper*.

[15] Haugen, R. and Baker, N. (1996). Commonality in teh Determinants of Expected Stock Returns, *Journal of Financial Economics*, 41, 401-439.

[16] Hsieh, D. (1991). Chaos and Nonlinear Dynamics: Application to the Financial Markets, *Journal of Finance*, 46, 1839-1877.

[17] Kraskov, A., Stögbauer, H. and Grassberger, P. (2003). Estimating Mutual Information, preprint in http://www.arxiv:cond-mat/0305641.

[18] Kullback, S. (1968). *Information Theory and Statistics*, Dover, New York.Lee, B-S. (1992). Casual Relations Among Stock Returns, Interest Rates, Real Activity, and Inflation, Journal of Finance, 47, 4, 1591-1603.

[19] Maasoumi, E. (1993). A Compendium to Information Theory in Economics and Econometrics, *Econometric Reviews*, 12 (2), 137-181.

[20] Maasoumi, E. and Racine, J. (2002). Entropy and Predictability of Stock Market Returns, *Journal of Econometrics*, 107, 291-312.

[21] McQueen, G. and Roley, V. (1993). Stock Prices, News and Business Conditions, *Review of Financial Studies*, 6, 3, 683-707.

[22] Pesaran, M. and Timmermann, A. (1995). Predictability of Stock Returns: Robustness and Economic Significance, *Journal of Finance*, 50, 1201-1228.

[23] Pompe, B (1998). Ranking Entropy Estimation in NonLinear Time Series Analysis, preprint in *Nonlinear Analysis of Physiological Data;* H. Kantz, J. Kurths and G. Mayer-Kress eds., Springer, Berlin.

[24] Qi, M. (1999). Nonlinear predictability of Stock Returns using Financial and Economic Variables, *Journal of Business and Economic Statistics*, 17, 4, 419-429.

[25] Shannon, C. (1948). A Mathematical Theory of Communication (1 and 2), *Bell Systems Tech*; 27; 379-423 e 623-656.

[26] Soares, J. (1994). *Preços de Acções na Bolsa de Lisboa: Análise, Previsão e Regras de Compra e Venda*, PhD Thesis, IST, Lisbon.

[27] Soofi, E. (1997). Information Theoretic Regression Methods, *Advances in Econometrics - Applying Maximum Entropy to Econometric Problems*; Thomas Fomby and R. Carter Hill eds., Vol. 12.

[28] Stuzer, M. (1995) . A Bayesian Approach to Diagnosis of Asset Pricing Models, *Journal of Econometrics*, 68, 367-397.

6 Appendix A

Critical values tables for testing serial independence through mutual information for $N(0,1)$ data. 5000 replications were computed. *D.F.* are the degrees of freedom for the mutual information, which correspond to the dimension (d) of the analysed vectors.

N=100				*N=200*			
	Percentile				*Percentile*		
DF	*90*	*95*	*99*	*DF*	*90*	*95*	*99*
2	0.0185	0.0323	0.0679	2	0.0092	0.0214	0.0361
3	0.1029	0.1232	0.1933	3	0.0561	0.0701	0.1080
4	0.1059	0.1260	0.1722	4	0.0591	0.0918	0.1318
5	0.2290	0.2580	0.3261	5	0.1049	0.1193	0.1505
6	0.6639	0.7528	0.9663	6	0.5355	0.5956	0.7265
7	0.8996	0.9731	1.1586	7	0.5819	0.6411	0.7802
8	1.3384	1.3839	1.5024	8	0.8378	0.8854	0.9979
9	1.9030	1.9352	2.0142	9	1.2932	1.3267	1.4015
10	2.5266	2.5571	2.6181	10	1.8560	1.8805	1.9258

N=500				*N=1000*			
	Percentile				*Percentile*		
DF	*90*	*95*	*99*	*DF*	*90*	*95*	*99*
2	0.0037	0.0070	0.0144	2	0.0019	0.0041	0.0071
3	0.0222	0.0369	0.0501	3	0.0133	0.0191	0.0311
4	0.0680	0.0788	0.1128	4	0.0340	0.0399	0.0568
5	0.1756	0.2066	0.2712	5	0.0708	0.0865	0.1128
6	0.3084	0.3514	0.4390	6	0.2119	0.2430	0.3046
7	0.4920	0.5391	0.6339	7	0.3635	0.3954	0.4688
8	0.4477	0.4843	0.5659	8	0.4041	0.4414	0.5252
9	0.6661	0.6941	0.7594	9	0.3865	04114	0.4640
10	1.0884	1.1082	1.1483	10	0.6418	0.6585	0.6942

N=2000				*N=2500*			
	Percentile				*Percentile*		
DF	*90*	*95*	*99*	*DF*	*90*	*95*	*99*
2	0.0009	0.0019	0.0033	2	0.0008	0.0015	0.0030
3	0.0061	0.0094	0.0147	3	0.0054	0.0078	0.0129
4	0.0169	0.0203	0.0278	4	0.0134	0.0171	0.0251
5	0.0701	0.0804	0.1030	5	0.0556	0.0648	0.0797
6	0.1370	0.1549	0.1940	6	0.1203	0.1376	0.1738
7	0.2496	0.2733	0.3224	7	0.2181	0.2418	0.2884
8	0.4497	0.4864	0.5508	8	0.3938	0.4217	0.4719
9	0.3036	0.3298	0.3858	9	0.3175	0.3409	0.4024
10	0.3530	0.3669	0.3996	10	0.2931	0.3124	0.3477