RELATÓRIOS COPPEAD

430

Abril 2017

DETERMINANTS OF STOCK MARKET CLASSIFICATIONS

Beatriz Vaz de Melo Mendes
Ramon A.C. Martins

Relatórios COPPEAD é uma publicação do Instituto COPPEAD de Administração da Universidade Federal do Rio de Janeiro (UFRJ)

Editora
Leticia Casotti

Editoração
Lucilia Silva

Ficha Catalográfica
Cláudia de Gois dos Santos

M538d Mendes, Beatriz Vaz de Melo.

Determinants of stock market classifications. / Beatriz Vaz de Melo Mendes, Ramon A. C. Martins. – Rio de Janeiro: UFRJ/COPPEAD, 2017.

15 p.; 27 cm. – (Relatórios COPPEAD; 430)

ISSN 1518-3335
ISBN 978-85-7508-117-4

1. Finanças. 2. Cópulas (Estatística matemática). I. Título. II. Série.

CDD: 332

Determinants of stock market classifications

Beatriz V. M. Mendes[ab1] **and** **Ramon A. C. Martins**[a]

[a]COPPEAD Graduate School of Business - UFRJ, Rio de Janeiro, Brazil
[b] Institute of Mathematics - UFRJ, Rio de Janeiro, Brazil

Abstract

In this paper we use discriminant analysis to describe and predict market classifications. Potential discriminators are derived from relevant characteristics of market indices, in particular from the returns' volatility. Using a training data set, an initial screening on the predictors is carried out by the Kruskal-Wallis test and a model based simple rule is constructed. The statistical significance of research results is confirmed by the high ratio of correct classifications (96.6%) along with formal statistical tests. Using a validation data set this rule is applied to allocate markets to one of the previously defined groups: Developed, Emerging, or Frontier. The easy to implement quantitative approach for classifying markets was able to anticipate market reclassifications, and erroneous classifications were found to be exactly those markets reclassified in the following classification review.

Key-words: Discriminant analysis, Developed, Emerging, and Frontier markets, GARCH models.

1 Introduction

Stock market classifications are important for financial institutions such as banks and investment funds, as well as for financial agents including institutional and private investors, financial regulators, and so on. Any risk manager needs to understand the environment under which he is operating and how the market is perceived by the financial world and, certainly, the market classification as "Developed", "Emerging", or "Frontier", contributes to this perception.

Market classifications are provided by various sources such as the IMF, MSCI, Dow Jones, among others. All of them have similar discriminating variables such as qualitative measures on the country economic development, the size and liquidity of equity markets, and market accessibility for foreign investors.

However, these classifications are not frequently updated, contrasting with the speed at which new projects are created. Thus, investment opportunities may be missed. Note that according to the MSCI, the *misclassification of a market within a global index may significantly increase the cost, tracking error and overall risk of a portfolio tracking the index.* It would be interesting to have a model based discriminating rule based on quantitative variables that could be easily (re)computed at any time. Multivariate statistical techniques

[1]Corresponding author. Coppead/UFRJ, PO Box 68514, Rio de Janeiro, RJ 21941-972, Brazil.

do exist to provide empirical evidence for the abovementioned classifications. Among them we select the discriminant analysis.

Discriminant analysis is a statistical technique which classifies an element into one out of several pre-defined groups dependent upon the observed values of selected predictors or discriminators. It derives a linear combination of the variables which best discriminates among the groups. The technique derives a meaningful predictive rule and has the advantage of considering a large profile of variables common to the elements in the groups.

Discriminant analysis has been used in the financial theory and practice. For example, to analyze financial institutions bankruptcy or financial distress, or to model credit rating. Altman (1968) was the first to apply discriminant analysis to analyze corporate bankruptcy. Taffler (1982) used discriminant analysis for the identification of British companies at risk of failures. Koh and Killough (1990) constructed a classification model to help external auditors to assess the going-concern status of his clients. Shirata (2012) used discriminant analysis to predict Japanese corporate bankruptcy. Zhu and Li (2010) proposed an effective indicator system and established the credit evaluation models of China's listed companies using financial data.

In this paper we use discriminant analysis to classify and predict market development status as Developed, Emerging, or Frontier. A related article is Alrgibi et al. (2010) , where the authors focus on the identification of the variables that would help an analyst to classify a market either as Developed or Emerging. Our paper, however, differs from Alrgibi et al. (2010) in many aspects. First, we include the third market category of Frontier markets. Second, we differ on how markets are previously classified. Alrgibi et al. (2010) obtain the groups' definition from an indicator of stock market development, the annual per capita Gross Domestic Product (GDP), whereas we use information provided by the Dow Jones. Third, the choice of discriminating variables. They use indicators of corporate depth, market activity, and size of market, and we focus on quantitative variables extracted from econometric models. Finally, we provide a more comprehensive study, considering two samples, one for estimation (training sample) and another one for validation.

The data analyzed are from forty stock markets and the quantitative variables are the relevant characteristics of indices' returns. Strong emphasis is set on volatility features. Noting that volatility by itself is an important issue, our guess was that some measures derived from the volatility would carry information with discriminating power. Some authors have used indeed volatility as a measure of stock market development, for instance, Levine (1997). According to Demirguc-Kunt and Levine (1996), high volatility in market returns is not necessarily a sign of underdevelopment, being, actually, an indicator of development.

Carroll and Collins (2012) compare volatility features of a stock index from a small economy, the Irish Stock Exchange Quotient (ISEQ), and the largest developed economy, the Dow Jones Industrial Average. They found differences in persistence measures, and question whether or not these differences could be generalized and used to compare small and large economies. Here we extend this research to three market classifications.

We found that discriminant analysis is very efficient when applied to market classifications, with results which are in line with those obtained through qualitative data. This means that the variables selected carry enough stock market information useful for classifying. In addition, the technique was able to anticipate changes in the classifications provided by the Dow Jones. The advantage of having a quantitative methodology for market classifications is that the analysis can be reproduced or updated at any time, since variables can be easily obtained.

In Section 2 we briefly review the basics on Fisher's discriminant analysis and present the discriminating variables. In Section 3 we carry on the empirical analysis. An initial sample of forty indices from the three market classifications is used to compute the predictors and establish the function(s) which best discriminate between countries in three mutually exclusive groups: Developed, Emerging, and Frontier. Using a secondary sample we validate the results and show that the discriminant technique is a reliable model for predicting market development status. Section 4 concludes and presents suggestions for further research.

2 Methodology

2.1 Discriminant analysis

Discriminant analysis is a statistical technique that allocates elements to g previously defined groups using p independent variables, the discriminators or predictors. It provides a prediction rule which allows someone to decide to which group a (new) element (in our study, a stock market) belongs to, according to its degree of similarity with other elements. One advantage is the reduction of the analyst's space dimensionality.

More formally, consider g populations $\pi_i, i = 1, ..., g$, to be classified based on p independent random variables $\mathbf{X}' = [X_1, X_2, ..., X_p]$. Let $\boldsymbol{\mu}_i = E(\mathbf{X}|\pi_i) \in \mathbb{R}^p$ represent the expected values of the variables given group i, $\overline{\boldsymbol{\mu}} = \frac{1}{g}\sum_{i=1}^{g} \boldsymbol{\mu}_i \in \mathbb{R}^p$ be the mean vector of combined populations, $\mathbf{B}_{\boldsymbol{\mu}} = \sum_{i=1}^{g}(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})'$ represent the between groups sum of cross-products, and $\boldsymbol{\Sigma}_i = \mathrm{cov}(\mathbf{X}|\pi_i)$ be the $p \times p$ covariance given group i. No assumptions are made about the populations' joint distributions. The method only requires that all $\boldsymbol{\Sigma}_i$ are equal and of full rank. Let $\boldsymbol{\Sigma}$ represent this common covariance matrix.

The idea is to transform the p-variate vector $\mathbf{X}$ to the univariate variable Y such that the Y's derived from different groups are separated as much as possible. This can be done through the linear combination

$$Y = \mathbf{a}'\mathbf{X} \tag{1}$$

for some $\mathbf{a} \in \mathbb{R}^p$. Then, the expected value for group i is $\mu_{iY} = \mathbf{a}'E(\mathbf{X}|\pi_i) = \mathbf{a}'\boldsymbol{\mu}_i$, the overall mean is $\overline{\mu}_Y = \mathbf{a}'\overline{\boldsymbol{\mu}}$, and the common variance of Y is $\sigma_Y^2 = \mathbf{a}'\mathrm{cov}(\mathbf{X})\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$.

Consider the ratio of the sum of squared distances from groups to overall mean of Y, to the variance of Y is

$$\frac{\sum_{i=1}^{g}(\mu_{iY} - \overline{\mu}_Y)^2}{\sigma_Y^2} = \frac{\mathbf{a}'\mathbf{B}_{\boldsymbol{\mu}}\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}. \tag{2}$$

The ratio (2) measures the variability between the groups in the Y-space relative to the common variability within groups. The discriminant analysis looks for the value $\hat{\mathbf{a}}$ maximizing (2). It means that a market will be included into the group which is the closest in terms of distance of units from the centroid of the group.

The solution $\hat{\mathbf{a}}$ must be obtained using training samples of size n_i from groups π_i, $i = 1, 2, ..., g$. Let $\mathbf{x}_i$ denote the $n_i \times p$ data matrix from population π_i, and $\mathbf{x}'_{ij}$ its jth row. Let $\overline{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$, and consider the sample covariances $\mathbf{S}_i = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)'$, $i = 1, 2, ..., g$, $j = 1, 2, ..., n_i$, and the overall mean estimate $\overline{\mathbf{x}} = \frac{1}{g}\sum_{i=1}^{g}\overline{\mathbf{x}}_i$, the $p \times 1$ vector average taken over all of the sample observations in the training set.

Let $\mathbf{B} = \sum_{i=1}^{g}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})'$ be the sample between groups matrix estimate of $\mathbf{B}_{\boldsymbol{\mu}}$, and let $\mathbf{W}/(n_1 + n_2 + ... + n_g - g)$ be the estimate of $\boldsymbol{\Sigma}$, where $\mathbf{W} = \sum_{i=1}^{g}(n_i - 1)\mathbf{S}_i$. The $\hat{\mathbf{a}}$ which maximizes $\frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}$ is given by the eigenvectors $\hat{\mathbf{e}}_i$ of $\mathbf{W}^{-1}\mathbf{B}$.

Let $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_s > 0$ be the $s \leq min(g-1, p)$ non-zero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ and, $\hat{\mathbf{e}}_1, ..., \hat{\mathbf{e}}_s$, the corresponding normalized eigenvectors. The linear combination $\hat{\mathbf{e}}'_1\mathbf{x}$ is the first sample discriminant. The second sample discriminant is $\hat{\mathbf{e}}'_2\mathbf{x}$, and so on, a total of s sample discriminants, corresponding to s discriminant variables Y_k, $k = 1, 2, \cdots, s$. Note that for $g = 3$ the maximum number of discriminants is 2. The technique while reducing the dimensionality promotes the maximum possible separation of groups.

The discriminants provide the basis for the classification rule for the (new) members. For each population π_i the vector $\mathbf{Y} = [Y_1, \cdots, Y_s]'$ has mean vector $\boldsymbol{\mu}_{iY} = [\mu_{iY_1}, \cdots, \mu_{iY_s}]'$ where $\boldsymbol{\mu}_{iY} = \mathbf{a}'_i\boldsymbol{\mu}_i$, and covariance matrix $\mathbf{I}$. Therefore the deviation of $\mathbf{Y}$ with respect to the population i center is measured by $(\mathbf{Y} - \boldsymbol{\mu}_{iY})'(\mathbf{Y} - \boldsymbol{\mu}_{iY}) = \sum_{j=1}^{s}(Y_j - \mu_{iY_j})^2$.

Let $\mathbf{y}$ be an observation of $\mathbf{Y}$ The classification rule allocates $\mathbf{y}$ to group π_k if the squared distance between $\mathbf{y}$ and $\boldsymbol{\mu}_{kY}$ is smaller than this distance for all other $\boldsymbol{\mu}_{iY}$, $i \neq k$. That is $\sum_{j=1}^{s}(y_j - \mu_{kY_j})^2 \leq \sum_{j=1}^{s}[\mathbf{a}'_j(\mathbf{x} - \boldsymbol{\mu}_i)]^2 \quad$ for all $i \neq k$.

Consider a (possible new) element with predictors $\mathbf{x}$ and suppose that just two discriminants are needed. The technique allocates $\mathbf{x}$ as a member of π_k whenever

$$\sum_{j=1}^{2}[\hat{\mathbf{e}}'_j(\mathbf{x} - \overline{\mathbf{x}}_k)]^2 \leq \sum_{j=1}^{2}[\hat{\mathbf{e}}'_j(\mathbf{x} - \overline{\mathbf{x}}_i)]^2 \quad \text{for all } i \neq k. \tag{3}$$

For details on multivariate techniques see Johnson and Wichern (1992).

2.2 Searching for relevant discriminators

The first task is then the identification of the variables able to discriminate and classify the stock markets. We examine the distribution of series of index returns searching for features carrying relevant information on corresponding markets. They will form the $\mathbf{X}$ variables in model (1) and classification rule (3). Let r_t, $t = 1, 2, \cdots, T$ represent the indices log-returns.

The potential predictors considered include estimates of some features of the unconditional returns distribution and estimates of parameters of econometric models. They are chosen due to their relevance reported in the literature. We start by drawing some basic statistics from the underlying unconditional distribution of the returns.

The sample mean

$$\overline{r} = \frac{1}{T}\sum_{t=1}^{T} r_t, \tag{4}$$

the sample variance

$$S^2 = \frac{1}{T-1}\sum_{t=1}^{T}(r_t - \overline{r})^2, \tag{5}$$

the sample skewness coefficient

$$\widehat{A} = \frac{1}{(T-1)S^3}\sum_{t=1}^{T}(r_t - \overline{r})^3, \tag{6}$$

its absolute value

$$|\widehat{A}| = \frac{1}{(T-1)S^3}\sum_{t=1}^{T}(r_t - \overline{r})^3, \tag{7}$$

and the sample kurtosis coefficient

$$\widehat{K} = \frac{1}{(T-1)S^4}\sum_{t=1}^{T}(r_t - \overline{r})^4, \tag{8}$$

are respectively the empirical estimates of the unconditional mean $\mu = E[r_t]$, the population variance $\sigma^2 = E[(r_t - \mu)^2]$, the population skewness $A = \frac{E[(r_t-\mu)^3]}{\sigma^3}$ and its absolute value, and population kurtosis $K = \frac{E[(r_t-\mu)^4]}{\sigma^4}$.

Measuring correlation between markets is also an important issue. Some authors, for instance Bekaert and Harvey (1995), Todorov and Bidarkota (2011) and Gupta et al. (2011), recognize that less developed markets show weak integration with most important financial markets. Let X and Y represent the log-returns of two market indices. The next estimator is

$$\widehat{\rho} = (\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y}))/S_X S_Y, \tag{9}$$

the empirical estimate of the Pearson linear correlation coefficient between X and Y, $\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$.

To measure monotone correlation we use the Spearman ρ_S and the Kendall's τ coefficients. Let r_{x_i} and r_{y_i} represent the ranks of observations x_i and y_i from X and Y. The empirical estimate of ρ_S is

$$\widehat{\rho_s} = 1 - \frac{6}{T(T^2-1)}\sum_{i=1}^{T} d_i^2 \tag{10}$$

where $d_i = r_{x_i} - r_{y_i}$, the difference between the ranks of x_i and y_i.

Two pairs of observations (x_i, y_i) and (x_j, y_j) are concordant if $x_i > x_j$ and $y_i > y_j$, or if $x_i < x_j$ and $y_i < y_j$. They are discordant if either $x_i > x_j$ and $y_i < y_j$, or $x_i < x_j$ and $y_i > y_j$. In the case $x_i = x_j$ or $y_i = y_j$, classifications do not apply. Let T_c and T_d be respectively the number of concordant and discordant pairs. The estimate of τ is

$$\widehat{\tau} = \frac{T_c - T_d}{T(T-1)/2}. \tag{11}$$

Now we consider discriminators derived from econometric models. The autoregressive moving average ARMA(p, q) process is usually applied to model the conditional mean $\mu_t = E[r_t|I_{t-1}]$, where I_t represents the set of all information available up to time t, and may be specified as

$$r_t - a_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} - \sum_{j=1}^{q} \theta_j a_{t-j} \tag{12}$$

where the error a_t follows a white noise process.

To model to conditional variance $\sigma_t^2 = var(r_t|I_{t-1})$ we apply the Generalized Autoregressive Conditionally Heteroskedastic GARCH(m, s) model, a generalization of the ARCH model proposed in the seminal work of **?** (see Bollerslev (1986), Engle et al. (1987), Nelson (1991), Glosten et al. (1993)). Under this model the conditional variance follows

$$\sigma_t^2 = \omega + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \tag{13}$$

where $a_t = \sigma_t \varepsilon_t$, $\varepsilon_t \sim$ F(0, 1), and where $\omega \geq 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ and $\sum_{i=1}^{m} \alpha_i + \sum_{j=1}^{s} \beta_j < 1$ to guarantee a positive finite unconditional variance.

The GARCH volatility equation does not distinguish between positive and negative returns. Many studies have empirically shown the asymmetric impact of returns on volatility, the *bad news* effect, with large negative returns increasing volatility (for instance, Fischer (1976)).

Some GARCH extensions deal with this phenomenon, including the EGARCH model of Nelson (1991). The log of the conditional variance of the simple but powerful EGARCH$(1, 1)$ model is given by

$$\log(\sigma_t^2) = \omega_E + \gamma a_{t-1} + \alpha_{1E}(|a_{t-1}| - E|a_{t-1}|) + \beta_{1E} \log(\sigma_{t-1}^2). \tag{14}$$

The specification $\log(\sigma_t^2)$ being less restrictive facilitates estimation.

All abovementioned volatility models capture only short memory. Very often the autocorrelation function of squared returns shows a hyperbolic decay rate, characteristic of long memory. The Fractionally Integrated GARCH, FIGARCH(m,d,s), captures the effect of long memory through the parameter d. Its volatility equation may be written as

$$\sigma_t^2 = \omega[1 - \beta(L)]^{-1} + \{1 - [1 - \beta(L)]^{-1} \phi(L)(1 - L)^d\} a_t^2 \tag{15}$$

where

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - ... - \phi_v L^v$$

$$\beta(L) = 1 - \beta_1 L - \beta_2 L^2 - ... - \beta_s L^s$$

and where $v = max(m, s)$, $\phi_i = \alpha_i + \beta_i$ and $0 < d < 1$. As in the GARCH specification, the polynomials $\phi(L)$ e $\beta(L)$ capture the short memory in the volatility.

Models are estimated by maximum likelihood assuming that a_t follows either a Normal or a t-student distribution with υ degrees of freedom, and that $p = q = s = m = 1$. Best model definition, including the choice of residuals distribution and definition of orders p, q, m and s, is indicated by the Akaike criterion (Akaike, 1973). All fits are carefully checked with respect to the assumptions made. All maximum likelihood estimates

$$\hat{\delta}_{MLE} = (\widehat{\phi_0}, \widehat{\phi_1}, \widehat{\theta}_1, \widehat{\omega}, \widehat{\alpha_1}, \widehat{\beta}_1, \widehat{\alpha_{1E}}, \widehat{\beta_{1E}}, \widehat{\gamma}, \widehat{d}, \widehat{\upsilon}) \tag{16}$$

of the parameters $\delta = (\phi_0, \phi_1, \theta_1, \omega, \alpha_1, \beta_1, \alpha_{1E}, \beta_{1E}, \gamma, d, \upsilon)$ will be tested as discriminators X_i.

Further quantities X_i are also derived from the conditional models. Consider the variance return (rv_t), defined in Engle and Patton (2001) as the proportional change in conditional variance σ_t^2, that is, $rv_t = 100 * \ln\left(\frac{\sigma_t^2}{\sigma_{t-1}^2}\right)$. The volatility of the volatility (VoV) is defined as the standard deviation of the random component of the variance return

$$\text{VoV} = \sqrt{var(rv_t)}. \tag{17}$$

Other X_is are related to the so called stylized facts on return volatility. Probably the most important and evident one is the volatility clustering. Large prices changes usually follow observed large changes (same for small changes), the persistence phenomenon affecting predictions. The GARCH(1,1) and EGARCH(1,1) models have persistence respectively given by $\alpha_1 + \beta_1$ (denoted by P) and β_{1E}.

Another persistence measure is the half-life of volatility. Let $\sigma^2_{t+k|t}$ denote, conditionally to information at time t, the expected value of the k periods ahead variance of returns, that is, $\sigma^2_{t+k|t} = E[(r_{t+k} - \mu_{t+k})^2 | I_t]$, where μ_{t+k} denotes the expected return. Given that at time $t+1$ the $\sigma^2_{t+1|t}$ has moved away from its reversion level, the unconditional variance σ^2, the 'half-life' of volatility is defined as the time κ taken for the volatility to move halfway back towards σ^2. That is

$$\kappa \ : \ \left|\sigma^2_{t+\kappa|t} - \sigma^2\right| = \frac{1}{2}\left|\sigma^2_{t+1|t} - \sigma^2\right|. \tag{18}$$

According to Carroll and Collins (2012), the half-life of volatility of the GARCH(1,1) and the EGARCH(1,1) models are respectively given by

$$\kappa = \frac{\ln[(\alpha_1 + \beta_1)/2]}{\ln(\alpha_1 + \beta_1)} \tag{19}$$

and

$$\kappa_E = \frac{\ln(\beta_{1E}/2)}{\ln(\beta_{1E})}. \tag{20}$$

The maximum likelihood estimators of (17), (19), and (20), as well as the persistence estimates are also considered as potential predictors X_i. They are computed based on the GARCH(1,1) and EGARCH(1,1) fits and denoted as $\widehat{VoV}$, $\widehat{VoV_E}$, $\widehat{\kappa}$, $\widehat{\kappa_E}$, and $\widehat{P}$. We compute a total of twenty four variables.

3 Classifying the markets

3.1 Data

Historical data were downloaded from the Bloomberg terminal. An initial sample, used for estimation of the classification rule, was composed of daily closing prices of thirty stock market indices equally divided into the three market classifications: Developed, Emerging, and Frontier. A validation sample was composed of extra ten series from the same categories. All market related information covered the thirteen year period from 01/01/2001 through 05/02/2014. The corresponding classifications were obtained from the Dow Jones Indexes Country Classification System. The classifications are determined based on *a rules-based methodology that incorporates objective data, and practically guided subjectivity that allows for committee input and market feedback. The review process begins with analysts examination of countries based on three broad categories of metrics for each country: Market & Regulatory Structure, Trading Environment, and Operational Efficiency. These categories reflect the market characteristics that are often considered by investors when determining the relative level of development and ease of investment.*

The Developed markets, hereafter denoted as Group 1, are Germany, Belgium, United States, France, Hong Kong, United Kingdom, Israel, Japan, Luxembourg and Suitzerland. Group 2 is composed of the Emerging markets South Africa, Brazil, Chile, China, Colombia, Philippines, India, Mexico, Russia and Turkey. Argentina, Bahrain, Croatia, United Arab Emirates, Jamaica, Latvia, Namibia, Nigeria, Kenya and Tunisia compose Group 3, the Frontier markets. Tables containing, for all groups, the values of the twenty-four discriminating variables defined in Section 2 are available to the reader upon request.

Before the estimation of the discriminant rule we investigate the ability of the selected variables to discriminate the groups. We carry out the Kruskall-Wallis (KW) test, which verifies if any two samples come from the same distribution, see Kruskal and Wallis (1952). The KW robust statistic is based on ranks and basically determine whether or not there are statistical significant differences between the variables means for the pre-defined groups. We observed that test results changed dramatically when the Namibia index (and in some extent also the Argentina index) were excluded from the analysis. Therefore, in what follows we exclude Namibia from the analysis. Table 1 provides the statistical significance of the KW test applied to the predictors.

The average return is marked higher for groups 2 and 3, with three exceptions, China, Bahrain and Croatia. The KW test confirms and indicate that the mean return from developed markets is different from the other two markets, in line with results in Bekaert and Harvey (1995). According to the KW test, the unconditional variance, the asymmetry, and the kurtosis coefficients are not promising classifiers, whereas all three correlation coefficients seem to be helpful for discriminating Group 3. The correlation coefficients were computed pairing each index with the Dow Jones Industrial Average.

Table 1 reveals that all variables drawn from the econometric models (except $\widehat{\theta_1}$ and $\hat{d}$) reject the null for at least two pairs of categories. Three statistics show up as those carrying greater discriminatory power and, if we assume the 10% significance level, we can also include the half-life of volatility $\widehat{\kappa_E}$. The first one, $\widehat{\phi_0}$, is related to the unconditional mean. The other two are the autoregressive estimate $\widehat{\alpha_{1E}}$, and the leverage term estimate $\hat{\gamma}$, from the EGARCH model. Results for the estimates $\hat{\nu}$ and $\hat{\omega}$ come from the GARCH(1,1) fit based on the t-student distribution.

Table 1: The 1% (***), 5% (**), and 10% (*) statistical significance of the Kruskal Wallis test applied to the discriminating variables. The symbol $\surd$ means acceptance of the null hypothesis.

i & j	Null Hypotheses: $F_i = F_j$; i, j =D,E,F							
	$F(\bar{r})$	$F(S^2)$	$F(\widehat{A})$	$F(\lvert\widehat{A}\rvert)$	$F(\widehat{K})$	$F(\widehat{\rho})$	$F(\widehat{\rho_S})$	$F(\widehat{\tau})$
D & E	***	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
D & F	***	$\surd$	$\surd$	*	**	***	***	***
E & F	$\surd$	*	$\surd$	*	$\surd$	**	**	**
	$F(\widehat{\phi_0})$	$F(\widehat{\phi_1})$	$F(\widehat{\theta_1})$	$F(\widehat{\omega})$	$F(\widehat{\alpha_1})$	$F(\widehat{\beta_1})$	$F(\widehat{\kappa})$	$F(\widehat{\kappa_E})$
D & E	***	$\surd$	**	***	$\surd$	$\surd$	***	*
D & F	**	***	$\surd$	***	***	***	***	***
E & F	**	**	$\surd$	$\surd$	***	***	$\surd$	*
	$F(\widehat{\nu})$	$F(\widehat{\alpha_{1E}})$	$F(\widehat{\beta_{1E}})$	$F(\widehat{\gamma})$	$F(\widehat{P})$	$F(\widehat{d})$	$F(\widehat{VoV})$	$F(\widehat{VoV_E})$
D & E	$\surd$	***	$\surd$	**	***	$\surd$	$\surd$	$\surd$
D & F	***	***	***	***	***	$\surd$	***	***
E & F	***	***	*	***	$\surd$	$\surd$	***	**

Notation in table: D, E, and F represent the Developed, Emerging, and Frontier markets. $F(T)$ represents the distribution of the statistic T.

The $\widehat{\alpha_1}$ and $\widehat{\beta_1}$ GARCH estimates are very efficient when distinguishing the frontier markets, but the moving average estimate looses relevance when the *bad news* term is included in the volatility equation. On the other hand, persistence from the GARCH(1,1) fit ($\widehat{P}$) seems to be powerful when one of the groups tested is Group 1. Figure 1 illustrates and show the conditional volatility estimated for two indices, the Hong Kong index at the left hand side, and the Kenya index at the right hand side. Most developed and emerging indices show the pattern observed for the HSI index, with greater persistence in volatility when compared to indices from frontier markets.

The KW statistic also rejects the nulls of tests involving the estimates of the volatility of the volatility from both the GARCH and EGARCH fits from the frontier markets. Note that indices from developed and emerging markets usually show larger persistence an therefore smaller volatility of volatility.

It is possible that some of the measurements will have a high degree of correlation with each other. This means that we may find a relatively small number of selected measurements

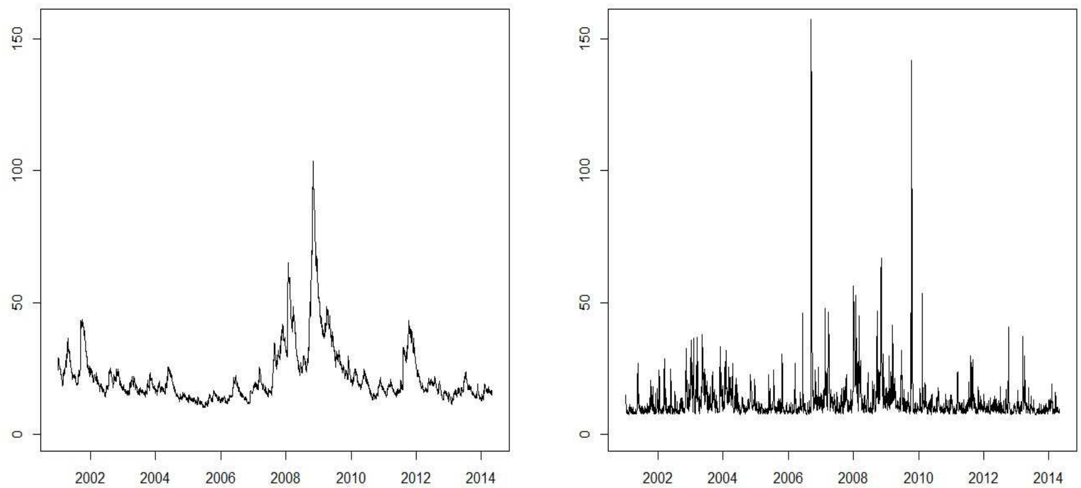

Figure 1: *The conditional volatility estimated for two indices, the Hong Kong index at the left hand side, and the Kenya index at the right hand side.*

from the original list of 24 variables carrying the same amount of information to build up a prediction model. To check that we computed the linear correlation coefficients for all pairs of variables with some interesting findings.

The mean return $\bar{r}$ is not correlated with all but one variable, $\widehat{\phi_0}$, for which the correlation is equal to 0.68. The volatility of volatility $\widehat{VoV_E}$ is highly negatively correlated with $\widehat{\kappa_E}$ and $\widehat{\beta_{1E}}$ (-0.81 and -0.93, respectively), also an expected result since the greater the persistence of the volatility, the longer its reversion time, and the smaller the VoV. The $\widehat{VoV}$ and the $\widehat{VoV_E}$ are positively related (0.80) thus, if one of them is in the classification rule, the other one should not be. The leverage term $\widehat{\gamma}$ shows a strong positive correlation with all three correlation coefficients. Thus either $\widehat{\gamma}$ is a predictor in the rule or one of the correlations is a classifier since the correlations among $\widehat{\rho}$, $\widehat{\rho_S}$, and $\widehat{\tau}$ are aproximately one. Finally, both pairs $\widehat{\alpha_1}$ & $\widehat{\alpha_{1E}}$, and $\widehat{\beta_1}$ & $\widehat{\beta_{1E}}$ show strong positive correlation around 96%.

3.2 Discriminant Analysis

After constructing the variable profile we apply the discriminant analysis to identify the linear combinations of the predictors that can successfully differentiate the market categories, helping to provide authenticity to the terms Developed, Emerging, and Frontier.

We use the `R` and the `SPlus` packages to perform a step-wise discriminant analysis. All details required by the statistical technique are implemented in the functions used, including several tests such as the Box and the Bartlett tests of homogeneity of covariances among groups, the only assumption made by the statistical model. The algorithm takes into consideration the correlations commented above, perform tests for the equality of means, the Wilk's lambda, Pillai trace, and Hotelling-Lawley trace tests, among others. In spite of our findings on the correlated variables and the results from the KW test, all variables are passed as inputs to the computer program. However, since we 24 regressors but only 29 data points (10, 10, and 9 in each group), we restrict the number of X_is composing the linear combinations to be at most 5.

The best discriminant rule maximizes the ratio of the between-group sum of squares to the within-group sum of squares, and also minimizes the apparent error rate, the proportion of misclassifications within the training sample. The search found 4 winning solutions, all extremely accurate with a rate of 96.55% (28/29) correct classifications in the training sample, the only market missclassified being Israel. They are the following classification rules (C), all constructed without the Namibia index.

$$
\begin{aligned}
&C1\\
y_1 &= -8.469\overline{r} - 36.714\widehat{\gamma} - 0.017\widehat{VoV_E}\\
y_2 &= -42.432\overline{r} + 4.480\widehat{\gamma} + 0.008\widehat{VoV_E}\\
\\
&C2\\
y_1 &= -12.403\overline{r} - 33.483\widehat{\gamma} + 19.085\widehat{\beta_{1E}}\\
y_2 &= -42.062\overline{r} - 7.122\widehat{\gamma} + 7.559\widehat{\beta_{1E}}\\
\\
&C3\\
y_1 &= -9.656\overline{r} - 33.732\widehat{\gamma} - 12.404\widehat{\alpha_1}\\
y_2 &= -41.856\overline{r} + 3.212\widehat{\gamma} + 6.037\widehat{\alpha_1}\\
\\
&C4\\
y_1 &= -10.724\overline{r} - 33.399\widehat{\gamma} + 8.872\widehat{\beta_1}\\
y_2 &= -42.048\overline{r} + 4.508\widehat{\gamma} - 3.813\widehat{\beta_1}
\end{aligned}
$$

The sample mean is present in all four rules which are composed by only three variables. We immediately note that the results of the KW test and the information on the variables' degrees of collinearity are very pertinent. Recall that $\widehat{\gamma}$, as well as $\widehat{\alpha_{1E}}$, were indicated by the KW test, rejecting the null for all three pairs of groups. The leverage estimate $\widehat{\gamma}$ is indeed a predictor in all four classifiers, but $\widehat{\alpha_{1E}}$ is present in only one rule. However, $\widehat{\alpha_1}$ compose classifier C3, and $\widehat{\alpha_1}$ is highly positively correlated $\widehat{\alpha_{1E}}$.

The third variable in the rules comes from a conditional model. Rules 1 and 2 require only the EGARCH fit and might be preferred, whereas rules 3 and 4 require both the GARCH and EGARCH fits. The estimate $\widehat{\beta_{1E}}$, which measures persistence of the EGARCH(1,1) model, was not seen as a strong discriminator according to the KW test, but it is highly positively correlated with $\widehat{\beta_1}$ which is present in C4. For the sake of experimentation (and curiosity) we tried substituting the variable $\bar{r}$ by $\widehat{\phi_0}$ since they are mathematically related, empirically highly positively correlated, and indicated by the KW test. However, this solution was not as good as the ones reported, with four erroneous classifications out of 29. All this shows that the multivariate technique takes into consideration all aspects involving the relationship of the variables. It is interesting to note that the correlations between variables in the rules are all smaller than 0.30 in absolute value, being half of them negative.

Figure 2 shows at the left side hand the positions of indices from the training sample in

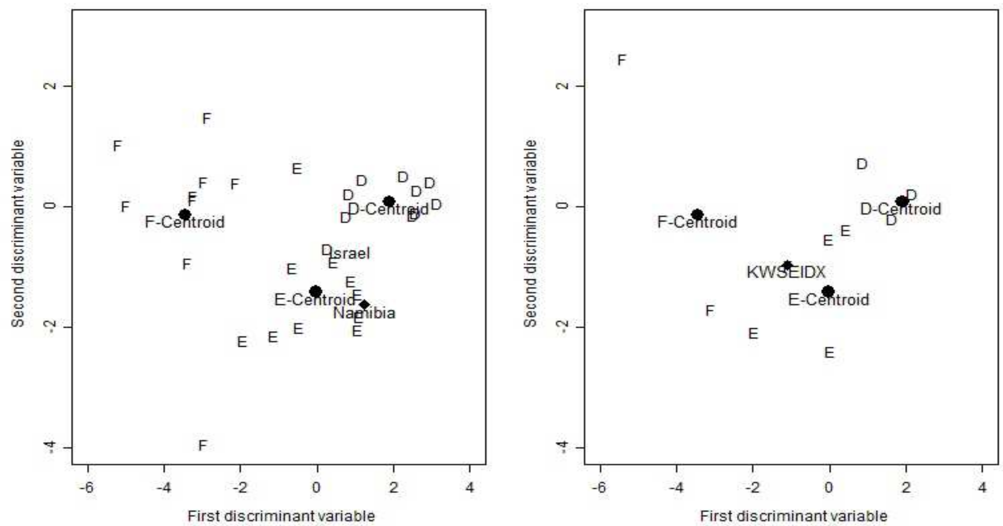

Figure 2: *Positions of indices in the classification plane defined by rule C1. At the left hand side the training sample, and at the right hand side the validation sample.*

the space of classifier C1. The TA-100 index of Israel was the one missclassified as emergent by all 4 rules. Actually, this market was re-classified as developed in 2009. Thus during 8 out of the 13 years covered by our data it was indeed an emerging market. It is interesting to observe how the Namibia index stands out of the remaining frontier markets. All winning rules missclassified Namibia as an emerging market. Note that in the composition of the FTB-098 index there are 24 out of 32 firms from South Africa, Australia, and Canada.

The classification rules are now tested using a validation sample containing 10 indices. Three are developed markets: Australia (ASX200), Spain (IBEX35) and Ireland (ISEQ). Four are emerging: Egypt (EGX30), Hungarian (BUX), Indonesia (JCI) and Taiwan (TAIEX), and three are frontier indices: Kuwait (KWSEIDX), Malta (MALTEX) and Sri Lanka (COLOMBO). All four rules applied to the validation sample resulted in only one out of ten incorrect classification, the Kuwait index as emerging. The positions of the validation data in the classification space is shown at the right hand side of Figure 2, with index KWSEIDX located close to the Group 2 centroid. This may indicate that this market is under revision and is about to be raised to the condition of emergent.

The extremely high rates of correct classification in-sample and out-of-sample by the four wining rules suggest that besides the index returns sample means, other freely continuously available variables from (E)GARCH volatility models covering different aspects of the returns indices, are important tools helping taking decisions in the process of classifying economies.

After almost twenty years of research on market development status, it is observed that developed and emerging markets show many similarities. Therefore, some characteristics which have early distinguished these markets such as the correlations between volatility measurements (Camilleri and Galea, 2009) have now raised and had lost relevance. Another example is the large first lag returns autocorrelations, nowadays apply to the distinction of emerging and frontier markets. Another possibility not considered in this paper would be the market stratifications by economic size and/or geographical position. This would take the research back to qualitative data, something we were avoiding.

4 Discussions

Considering how influential could be for local and foreign investors the knowledge of a country development status, and the speed at which new ideas travel around the integrated world, in this study we proposed to classify markets using discriminant analysis. Results obtained using quantitative variables are in line to those obtained through qualitative data, and the technique was able to anticipate changes in the classifications with a high degree of accuracy. A simple classification rule was constructed with the econometric models based variables having the highest ability to distinguish the Developed, Emerging, and Frontier markets. Differently from the qualitative variables used by the classification agencies, not always revealed to the public, the discriminators used here can be obtained and updated at any time.

The empirical analysis used forty important stock indices from the three groups. Basic statistics and functions of estimates from GARCH (1,1), EGARCH(1,1), and FIEGARCH(1,1) models were tested as relevant predictors. The Kruskal-Wallis test was applied to verify the statistical significance of differences between groups. The best classifiers producing the maximum separation of the pre-defined groups were composed by only three variables. They were validated using a sample of ten markets which were not part of the training sample.

The small error rate presented by the classifiers indicate that the return series of a market stock index contains much of the total information present in the wide range of economic variables usually adopted by classifier agents, here the Dow Jones. All we need are the mean return and some features of the returns' volatility to construct a rule with an overall accuracy rate of 96.55% in predicting market group. Erroneous classifications were found to be exactly those reclassified in the following market classification review.

Further work will employ other multivariate techniques to validate conventional market classifications. Copulas and pair-copulas are possibilities for group allocation. After finding the best pair-copula fit for each group we compute some copula based measures. The log-likelihood value could be a tool for allocating new members. The log-likelihood value could also be used to identify atypical observations, with a large significant change indicating an influential observation, and any element producing an arbitrary change in parameters' values would indicate model breakdown.

References

E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, XXIV(4):589–609, 1968.

R. Taffler. Forecasting company failure in the uk using discriminant analysis and financial ratio data. *Journal of the Royal Statistical Societyf*, 145(3):342–358, 1982.

H. C. Koh and L. N. Killough. The use of multiple discriminant analysis in the assessment of the going-concern status of an audit client. *Journal of Business Finance & Accounting*, 2(17):179–192, 1990.

C. Y. Shirata. Financial ratios as predictors of bankruptcy in japan: an empirical research. *https://www.researchgate.net/*, 2012.

K. Zhu and J. Li. Studies of discriminant analysis and logistic regression model application in credit risk for china's listed companies. *Management Sc. and Engineering*, 4(4):24–32, 2010.

G. Alrgibi, M. Ariff, and L. Murray. What factors discriminate developed and emerging capital markets? *Applied Economic Letters*, 13(17):1293–1298, 2010.

R. Levine. Financial development and economic growth: views and agenda. *Journal of Economic Literature*, (35):688–726, 1997.

A. Demirguc-Kunt and R. Levine. Stock markets, corporate finance, and economic growth: an overview. *The World Bank Economic Review*, V(10):223–239, 1996.

T. Carroll and J. Collins. Volatility models and the iseq index, 2012.

R. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*, volume 4. Prentice hall Englewood Cliffs, NJ, 1992.

G. Bekaert and C.R. Harvey. Emerging equity market volatility. *Journal of Financial economics*, 43(1):29–77, 1995.

G. Todorov and P. V. Bidarkota. Time-varying risk and risk premiums in frontier markets. *Available at SSRN 1947412*, 2011.

R. Gupta, T. Jithendranathan, and A. Sukumaran. Looking at new markets for international diversification: frontier markets perspective. In *24th Australasian Finance and Banking Conference*, 2011.

T Bollerslev. Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, (31):307–327, 1986.

R. F. Engle, D. M. Lilien, and R. P. Robins. Estimating time varying risk premia in the term structure: the arch-m model. *Econometrica: J. of the Econometric Society*, pages 391–407, 1987.

D. B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, pages 347–370, 1991.

L. R. Glosten, R. Jagannathan, and D. E Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance*, 48(5):1779–1801, 1993.

B. Fischer. Studies of stock price volatility changes. *Proceedings of the 1976 Meetings of the A.S.A., Business and Economic Statistics Section*, pages 177–181, 1976.

H Akaike. Information theory and an extension of the maximum likelihood principle. In *Second Intern. Symposium on Information Theory*, pages 267–281, 1973.

W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

S. J. Camilleri and G.a Galea. The diversification potential offered by emerging markets in recent years. *The FEMA Research Bulletin*, 1(3):21–37, 2009.

COPPEAD
UFRJ

Rua Pascoal Lemme, 355 Ilha do Fundão Rio de Janeiro - RJ CEP:21941-918 Caixa Postal 68514 21941-972
Tel:2598-9898 Fax:2598-9848 e-mail:atendimento@coppead.ufrj.br Home Page:www.coppead.ufrj.br