Wanderlei Lima de Paulo (Brazil), Oswaldo Luiz do Valle Costa (Brazil)

Cluster analysis for regime identification and forecasting with application to the enhanced index tracking problem

Abstract

This paper deals with the asset allocation problem in the presence of regime switching in asset returns. Considering a financial market subject to changes in regime, it is assumed that the expected value and covariance matrix of the returns of the assets can change according to a Markov chain taking values in a finite set. Generally, to apply a portfolio selection model in a switching regime approach it is necessary to estimate the market parameters and determine the number of regimes. In this paper it is proposed a non-parametric procedure to determine the number of regimes and define in which regime the market belongs to along the time, based on analyzing the historical stock return patterns using cluster analysis tools. The proposed methodology is applied to a portfolio optimization problem with enhanced index tracking and switching regime. The results show a satisfactory performance of the model with regime switches when compared to the case without regime switches.

Keywords: switching regime, cluster analysis, enhanced index tracking, Markov process.
JEL Classification: G11, C10, C61.

Introduction

The portfolio optimization problem is widely studied in the finance literature, under different market models assumptions, utility functions, restrictions and time-horizons. The classical and well known mean-variance single-period model, originally proposed by Markowitz (1959), aims at maximizing the expected return of a portfolio under the restriction of a maximal given level of variance (risk) or, equivalently, minimizing the portfolio variance under the restriction of a minimum given expected return. In Li and Ng (2000) a solution to the multi-period mean-variance problem is presented (see other studies about this theme in Leippold et al. (2004) and Zhou and Li (2000), for instance).

Another portfolio optimization problem related to the classical mean-variance problem consists of establishing an optimal allocation so that the portfolio's return replicates the return of a reference index (benchmark). This problem is the so-called index tracking problem and, in this case, the utility tracking error function of the investor is based on the difference between the portfolio's return and the benchmark's return. Problems of this nature are addressed in Roll (1992), Rudolf et al. (1999), Jorion (2003), Stoyanov et al. (2008), Bajeux-Besnainou et al. (2011) and Chen and Kwon (2012). Within the same spirit, the problem known as enhanced index tracking aims at obtaining returns above the reference index (excess return), while minimizing the deviation of the tracking error, that is, the deviation of the difference between the portfolio's return and the benchmark's return. This kind of problem is studied in Wu et al. (2007), Canakgoz and Beasley (2008), Li et al. (2011) and Guastaroba and Speranza (2012).

In particular the portfolio selection problem in the presence of regime switching in asset returns is an important topic in finance. Usually, a market whose parameters are subject to switching regime is characterized by a Markov switching regime framework (see for instance Zhou and Yin, 2003; Yin and Zhou, 2004; Guidolin and Timmermann, 2007; and Bae et al., 2014). In this paper a financial market model under a multivariate Markov regime switching, where the expected value and covariance matrix of the returns can change according to a Markov chain taking values in a finite set is considered.

In order to work in a Markov switching regime approach it is necessary to estimate the market parameters, determine the number of regimes and define in which regime the market belongs to along the time. In this case an appropriate method is to use a Markov switching model, but it can become technically complex and computationally intensive depending of the number of regimes and variables. Based on the historical stock return patterns using cluster analysis tools it is proposed a simple procedure to determine the number of regimes and classify the market regimes at each instant of time. The proposed methodology was applied to a portfolio optimization problem with enhanced index tracking and switching regime (as presented in Costa and Paulo, 2007). The results show a satisfactory performance of the model with regime switches when compared to the case without regime switches.

The remainder of this paper is organized as follows. Section 1 presents the market model with switching regime. In Section 2 is presented the procedure to identify and determine the number of regimes and estimates the parameters of the market model, based on cluster analysis tools. Section 3 presents an

empirical application of the proposed methodology to a portfolio optimization problem with enhanced index tracking and switching regime. The final section presents some final remarks.

1. Market model with switching regime

The proposed work consideres a financial market with n assets which prices are represented by the random vector $S(t)$, where the components of $S(t)$ are described by $S_l(t)$, with l = 1, …, n, such that $S(t)$ = = $(S_l(t)…S_n(t))'$. The price of an asset at the instant t + 1, $S_l(t + 1)$, is defined by the relation $S_l(t + 1)$ = $(1 + R_l(t))S_l(t)$, in which the vector of returns $R(t)$ = $(R_l(t)…R_n(t))'$ is decomposed as:

$$R(t) = \eta_{\theta(t)} + \Sigma_{\theta(t)}^{1/2} w(t), \quad (1)$$

where the variable $\theta(t)$ characterizes the market regime at the instant t and defines how the asset returns are expected to vary from time t to time $t + 1$ (examples of this approach applied in asset allocation problem can be found in Billio and Pelizzon (2000), Taamouti (2012) and Saunders et al. (2013)).

It is assumed that the variable $\theta(t)$ follows a state Markov process taking values in a finite set {1, …, N} with transition probability matrix P given by:

$$P = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,N} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,N} \\ \vdots & & \ddots & \vdots \\ p_{N,1} & p_{N,2} & \cdots & p_{N,N} \end{pmatrix}, \quad (2)$$

where the value $p_{i,j}$ represents the probability of the market, when in regime i, moves to regime j at the next instant of time. Finally, in (1) Σ_i represents the covariance matrix of the returns, η_i the vector of expected returns when the market operation mode is $\theta(t)$ = i, with i = 1, …, N, and $w(t)$ a vector of random variables with a null mean and covariance matrix equal to the identity matrix and independent of the variable $\theta(t)$, written as:

$$\Sigma_i^{1/2} = \begin{pmatrix} \sigma_{i,1,1} & \cdots & \sigma_{i,1,n} \\ \vdots & \ddots & \vdots \\ \sigma_{i,n,1} & \cdots & \sigma_{i,n,n} \end{pmatrix}, \quad \eta_i = \begin{pmatrix} \eta_{i,1} \\ \vdots \\ \eta_{i,n} \end{pmatrix}. \quad (3)$$

Note that to apply the market model considered here we need to determine the number of regimes, estimate the parameters η_i, Σ_i and $p_{i,j}$, for each regime i = 1,…, N, and define in which regime i the market belongs to for each instant t. An appropriate parametric procedure to estimate these parameters is to use a multivariate Markov switching framework (Hamilton, 1989; Hamilton, 1990), from which it is possible to study unobserved common states (regimes) for several different asset returns (Guidolin and Timmermann, 2007; Taamouti, 2012; Guidolinand Hyde, 2007). Usually, an MMS model is constructed with a predefined number of regimes, so that the choice of the number of regimes is important to provide a sufficient detection rate and not generating a model of high complexity (Zhu et al., 2012; Spezia, 2010; Awirothananon and Cheung, 2009; Psaradakis and Spagnolo, 2003). Finally, from a multivariate Markov switching model (MMS) we can define at each instant t in which regime i the system belongs to, making a probabilistic inference about the unobserved regime $\theta(t)$ given observations on $R(t)$.

However, depending on the number of variables and regimes considered, the application of an MMS model can become technically complex, cumbersome and computationally intensive. Then, using cluster analysis tools in the next section it is proposed a simple non-parametric procedure to determine the number of regimes and define in which regime the market belongs to for each instant of time, as well as estimate the market parameters.

2. Cluster analysis framework

This section presents a methodology to identify and determine the number of regimes and estimate the model's parameters η_i, Σ_i and $p_{i,j}$. Based on Chow et al. (1999) a multivariate distance measure to identify common regimes from past observations of a series of daily returns is applied. Applications of this measure to study turbulence in financial markets can be found in Kritzman et al. (2001), Bauer and Molenaar (2004), Kritzman et al. (2001) and Kritzman et al. (2011).

Considering a market with n financial assets, set $d(t)$ the multivariate distance at each instant t, with t = 1,…, T, given by:

$$d(t) = [r(t) - \bar{r}]\Sigma^{-1}[r(t) - \bar{r}]', \quad (4)$$

where $r(t)$ is the vector of returns, $\bar{r}$ is the vector of average returns and Σ is the covariance matrix, written as:

$$r(t) = \begin{pmatrix} r_1(t) \\ \vdots \\ r_n(t) \end{pmatrix}, \ \Sigma = \begin{pmatrix} \sigma_{1,1} & \cdots & \sigma_{1,n} \\ \vdots & \ddots & \vdots \\ \sigma_{n,1} & \cdots & \sigma_{n,n} \end{pmatrix}, \ \bar{r} = \begin{pmatrix} \bar{r}_1 \\ \vdots \\ \bar{r}_n \end{pmatrix},$$

in which the estimates of $\bar{r}$ and Σ can be obtained by using the past observations of the series of daily vector of returns $\{r(t); t = 1, …, T\}$.

The number of regimes N is identified by analyzing the historical patterns of the multivariate distance series $\{d(t); t = 1,…, T\}$. The series $\{d(t); t = 1,…, T\}$. is segregated in k groups (clusters), written as follows:

$$G_1 = \{d_1^1,\ldots,d_{m_1}^1\},\ G_2 = \{d_1^2,\ldots,d_{m_2}^2\},\ldots,$$
$$G_k = \{d_1^k,\ldots,d_{m_k}^k\}, \quad (5)$$

where m_k is the number of elements of the series $\{d(t);\ t = 1,\ldots,\ T\}$ that were assigned to the group G_k (or cluster k), with $m_1 + \ldots + m_k = T$. In this case $G_i \cap G_j = 0$ whenever $i \neq j$ and $\bigcup_{i=1}^{k} G_i = \{d(t); t = 1,\ldots,T\}$. It is considered that each cluster corresponds to a market regime i, so that the number of regimes is equal to the number of groups established, i.e. $N = k$. Then, from (5) the historical series of the vector of asset returns $\{r(t);\ t = 1,\ldots,\ T\}$ can be divided into N groups, from which the vector of expected returns η_i and covariance matrices Σ_i may be estimated. Due to that, the number of elements in each group G_i should be large enough to allow a reasonable precision for the estimation of the expected returns η_i and covariance matrices Σ_i. Bearing this in mind a possible criterion for the choice of the number of regimes N would be to fix a minimal number of elements for the groups and consider only the cluster solutions that satisfy this restriction.

After that equation (5) has been established, the elements of the expected returns vector η_i and covariance matrix Σ_i, associated to a group G_i, with $i = 1,\ldots,\ N$, may be estimated by:

$$\eta_{i,l} = \frac{1}{m_i}\sum_{t=1}^{T} I_i(t) r_l(t), \text{ with } l = 1,\ldots,n, \quad (6)$$

$$\sigma_{i,s,v} = \frac{1}{m_i - 1}\sum_{t=1}^{T} I_i(t)(r_s(t) - \eta_{i,s})(r_v(t) - \eta_{i,v}), \quad (7)$$
$$\text{with } s,v = 1,\ldots,n,$$

in which the indicator function $I_i(t)$ is such that

$$I_i(t) = \begin{cases} 1, & \text{if } d(t) \in G_i \\ 0, & \text{otherwise} \end{cases}.$$

Within the same spirit, the probabilities of transition among states, $p_{i,j}$, can be calculated as the number of times that there is a switch from regime i to regime j divided by the number of times the system was in regime i, written as:

$$p_{i,j} = \frac{\sum_{t=2}^{T} I_{i,j}(t)}{\sum_{t=2}^{T} I_i(t-1)}, \text{ with } i = 1,\ldots,\ N, \quad (8)$$
$$j = 1,\ldots,N,$$

in which:

$$I_{i,j}(t) = \begin{cases} 1, & \text{if } d(t) \in G_j \text{ and } d(t-1) \in G_i \\ 0, & \text{otherwise} \end{cases}.$$

with $t = 2,\ldots,\ T$. Then, applying (6), (7) and (8) it is possible to estimate the vectors of expected returns η_i, the covariance matrices Σ_i and the transition matrix P, related to each regime $i = 1,...,\ N$.

There are two common methods to clustering a set of observation (or items), hierarchical and non-hierarchical method (for more details see Johnson and Wichern, 2007, Chapter 12). Basically, in the hierarchical method the number of clusters is not specified in advance as occur in non-hierarchical method. However, in the non-hierarchical method the observations may be regrouped during the clustering process, which does not occur in the hierarchical method. To improve the final solution (set of clusters), we can use the hierarchical method as exploratory technique to identify a number of clusters and, in the sequel, use this as input to the non-hierarchical method (in this case the methods are complementary).

As the hierarchical method provides several cluster solutions, the appropriate number of clusters can be determined by cutting off the dendrogram at an arbitrary point (sometimes a subjective choice). An identification of the optimal number of clusters can be done by using some stopping rule index as Calinski/Harabasz pseudo-F and Duda/Hart (see Everitt et al., 2011 for more details). Other methods can be seen in Sugar and James (2003), Sun et al. (2004) and Tibshirani et al. (2001), for instance. Finally, the application of the cluster analysis involves the choice of a convenient similarity measure between the variables. In this paper it is used the classical squared Euclidean distance, but in Bastos and Caiado (2012) it is introduced a new distance measure for clustering financial time series based on variance ratio test statistics.

The methodology presented above allows us to determine the number of regimes and estimate the market parameters established in (1). Moreover, from (5) it is possible to establish in which regime i the market belonged to at each past instant t, considering the past observations of the series of multivariate distance $\{d(t);\ t = 1,\ldots,\ T\}$. However, to apply the model (1) for new observations (i.e. observations that were not considered in the sample $\{r(t);\ t = 1,\ldots,\ T\}$, and which will be denoted by $r_0(t)$ it is necessary to establish a criterion for classifying to which regime a new observed vector of returns $r_0(t)$ belongs to.

Following the ideas in Chow et al. (1999), a simple way would be to assume that the vector of returns $r_0(t)$ is described by a normal distribution with vector of averages $\bar{r}$ and covariance matrix Σ. Then the distance $d_0(t)$ (defined as in (4) with $r_0(t)$ instead of $r(t)$) would follow a chi-squared distribution with

n degrees of freedom, χ^2_n. From a level of significance α a threshold distance, $\bar{d}$, could be defined so that one of the two market regimes (i = 1 or i = 2) could be established in case the observed distance $d_0(t)$ is higher than $\bar{d}$, i.e. $d_0(t) > \bar{d}$. This procedure is somewhat arbitrary to define the threshold $\bar{d}$ and would be only suitable for two regimes. In the sequel it is proposed a procedure to classify regimes based on the set of clusters (or groups) defined in (5). Based on linear classification rule (see Rencher, 2002, Chapter 9), consider the s^{th} threshold $\bar{d}_s$, s = 1,…, N – 1, of the multivariate distance series $\{d(t); t = 1,\ldots, T\}$ as follows:

$$\bar{d}_s = \frac{1}{2}(z_s + z_{s+1}), \tag{9}$$

in which z_t is the center of the group G_t, with i = 1,…, N, defined as:

$$z_i = \frac{1}{m_i}\sum_{k=1}^{m_i} d_k^i, \tag{10}$$

where m_i and $d^i_{m_i}$ are as defined in (5). From this set of threshold distances $\{\bar{d}_s; s = 1,\ldots, N-1\}$ the market regimes are classified as follows:

$$\theta(t) = \begin{cases} 1, \text{ if } d_0(t) \le \bar{d}_1 \\ 2, \text{ if } \bar{d}_1 < d_0(t) \le \bar{d}_2 \\ \vdots \\ N, \text{ if } \bar{d}_{N-1} < d_0(t). \end{cases} \tag{11}$$

The procedure proposed in this section can be summarized as follows:

1. From a series of daily vector of returns $\{r(t); t = 1,\ldots, T\}$ determine the series of multivariate distance $\{d(t); t = 1,\ldots, T\}$ as defined in (4).
2. Using cluster analysis segregate the series of multivariate distance $\{d(t); t = 1,\ldots, T$ into k groups (or clusters), as proposed in this section. Set the number of regimes as $N = k$, m_i, G_i and $d^i_{m_i}$ as in (5).
3. From the set of groups $\{G_i; i = 1,\ldots, N\}$ established in Step 2, estimate for each group the vectors of average returns η_i and the covariance matrices Σ_i applying (6) and (7), respectively. Apply (8) to estimate the transition matrix P.
4. Finally, to classify a new observation $r_o(t)$ which regime is unknown apply the criteria (11).

It should be pointed out that the methodology proposed here to identify and determine the number of regimes and classify a new observation which regime is unknown, characterized by (5)-(11), allows to work with more than two regimes and does not require the normality hypothesis on the vector of returns $r(t)$ as considered in Chow et al. (1999).

3. An empirical application

This section presents an application of the proposed methodology to a portfolio optimization problem with enhanced index tracking and switching regime (as presented in Costa and Paulo, 2007). The model assumes that the market regimes switch according to a finite state Markov chain, in which the returns of the assets are described as in (1).

3.1. Enhanced index tracking problem. Consider that the investor may allocate his/her financial resources in only (n − 1) assets, being asset 1 the reference index (benchmark). Let $U_l(t)$ be the wealth value allocated in each asset l, with l = 2, …, n, and $X_U(t)$ (assume for simplicity that $X_U(t) = X(t)$ from now on) the value of the portfolio related to the investments strategy U, with initial value $X(0) = X_0$ and time horizon T. By taking

$U(t) = (U_2(t) \quad \mathrm{U}(t))'$, $\mathrm{U}(t) = (U_3(t) \quad U_n(t))'$, $R_i(t) = (R_{i,1}(t) \quad R_{i,2}(t) \quad \mathrm{R}_i(t))'$ and $\mathrm{R}_i(t) = (R_{i,3}(t) \quad R_{i,n}(t))'$, we have that

$X(t) = U_2(t) + \mathrm{U}(t)'e$ and

$X(t+1) = (1 + R_{\theta(t),2}(t))U_2(t) + (e + \mathrm{R}_{\theta(t)}(t))'\mathrm{U}(t),$

where e is a (n − 2) dimensional vector with all the components equal to 1. Then, it is possible to show that the value of the portfolio is written as $X(t+1) = (1 + R_{\theta(t),2}(t))X(t) + \mathrm{P}_{\theta(t)}(t)'\mathrm{U}(t)$ with $\mathrm{P}_i(t) = \mathrm{R}_i(t) - R_{i,2}(t)e$.

Let $Y(t)$ represents the value of the reference portfolio associated with a benchmark index. It is supposed that its value follows the recursive equation $Y(t+1) = (1 + R_{\theta(t),1}(t))Y(t)$ with $Y(0) = X(0)$. Notice that the reference portfolio's return is given by $R_{i,1}(t) = \eta_{i,1} + \sum_{s=1}^{n}\sigma_{i,1,s}w_s(t)$. The enhanced index tracking problem consists in finding the investments strategy $\mathrm{U} = (\mathrm{U}(0),\ldots,\mathrm{U}(T-1))$ such that minimizes the functional

$$J((X(0) \quad Y(0))', \theta(0), \mathrm{U}) = \sum_{t=0}^{T} E\left(\delta_{\theta(t)}(t)\left\|X(t) - Y(t)\right\|^2 - \xi_{\theta(t)}(t)\left(X(t) - Y(t)\right)\right). \tag{12}$$

subject to

$$X(t+1) = (1 + R_{\theta(t),2}(t))X(t) + \mathrm{P}_{\theta(t)}(t)'\mathrm{U}(t), \quad (13)$$

$$Y(t+1) = (1 + R_{\theta(t),1}(t))Y(t), \quad (14)$$

where $\delta_i(t)$ and $\xi_i(t)$ are positive real numbers. The quadratic term represents the variability of the portfolio's value and the linear term represents the expected gain related to the reference portfolio. The balancing between the linear and quadratic terms is established through the weights $\delta_i(t)$ and $\xi_i(t)$. Thus, a manager could decide on one of the three investment strategies: achieve an average return higher than the reference index (active management), replicate the return of a reference index (index tracking) or track the reference index with a positive return in relation to the reference index (enhanced index tracking), depending on the values assigned to the parameters $\delta_i(t)$ and $\xi_i(t)$.

The solution for the problem (12)-(14), presented in Costa and Paulo (2007), is of a mode-dependent kind, that is, it depends on the regime of the market along the time. Then its application requires to define for each instant t in which of the states i the market belongs to, as well as to estimate the transition probability matrix (2), the covariance matrix and the expected returns described in (3). To achieve this goal we can apply the procedure proposed in this paper, as described in the next section.

3.2. Numerical example. It is considered a portfolio comprised of six stocks negotiated in the Brazilian stock exchange (BOVESPA), named VALE3, PETR3, BBDC3, GGBR3, ELET3 and USIM3, in which the wealth value can be allocated over the time. Then the financial market model consists of 7 assets, l_1, l_2, l_3, l_4, l_5, l_6 and l_7, being the asset l_1 chosen as the benchmark (Ibovespa index/IBOV). For the purpose of this study the historical stocks prices are considered for the period of 08/01/2008-01/31/2009 (a sample daily return with size $T = 116$). The application of the proposed methodology is presented in the following steps.

Firstly, the series of multivariate distance $\{d(t); t = 1,\ldots, 116\}$ as defined in (4) was calculated. Using STATA software, the hierarchical algorithm (with between-groups linkage cluster method and squared Euclidean distance measure) was applied to cluster the series $\{d(t); t = 1,\ldots, 116\}$ and computed the Duda/Hart indices to choose the optimal number of cluster (as shown in Table 1).

Table 1. Cluster solutions for the Duda/Hart index, Je(2)/Je(1)

Number of clusters	Je(2)/Je(1)	Pseudo-t²
1	0.2686	310.41
2	0.2783	67.43
3	0.2655	237.95
4	0.3168	25.88
5	0.1157	91.74

The conventional rule for choosing the number of optimal clusters is to find the point with the largest Je(2)/Je(1) value that corresponds to a low Pseudo-t^2 value, which has a higher value above and below it. Then, from Table 1, the optimal number of cluster should be four ($k = 4$), which sizes would be $m_1 = 39$, $m_2 = 4$, $m_3 = 56$ and $m_4 = 17$. However, note that the cluster with size $m_2 = 4$ is not appropriate to estimate the covariance matrices Σ_i as established in (3). Thus, using the same criterion of choice from Table 1, two groups ($k = 2$) as input to the $k-$means method was selected. Finally, the historical series of the asset returns was segregated into two groups (G_1 and G_2) with size $m_1 = 21$ and $m_2 = 95$, respectively. By taking the number of regimes $N = 2$, two market regimes were considered, one of higher volatility (regime $i = 1$) and another of lower volatility (regime $i = 2$), as defined in Table 2.

Table 2. Definition of market regimes

Regime	Description
$i = 1$	Market under high average volatility
$i = 2$	Market under low average volatility

From (6) and (7) the vectors of average returns η_i and the covariance matrices Σ_i, for each regime $i = 1$ and $i = 2$, are given by:

$$\eta_1 = (0.01233 \quad 0.02079 \quad 0.01057 \quad 0.00977 \quad 0.01292 \quad 0.01414 \quad 0.02423)',$$

$$\eta_2 = (-0.00605 \quad -0.00744 \quad -0.00485 \quad -0.00544 \quad -0.01063 \quad -0.00286 \quad -0.01268)',$$

$$\Sigma_1 = \begin{pmatrix} 0.00479 & 0.00481 & 0.00502 & 0.00459 & 0.00544 & 0.00331 & 0.00528 \\ 0.00481 & 0.00598 & 0.00563 & 0.00470 & 0.00553 & 0.00270 & 0.00561 \\ 0.00502 & 0.00563 & 0.00645 & 0.00504 & 0.00596 & 0.00282 & 0.00578 \\ 0.00459 & 0.00470 & 0.00504 & 0.00585 & 0.00551 & 0.00301 & 0.00477 \\ 0.00544 & 0.00553 & 0.00596 & 0.00551 & 0.00715 & 0.00430 & 0.00618 \\ 0.00331 & 0.00270 & 0.00282 & 0.00301 & 0.00430 & 0.00410 & 0.00367 \\ 0.00528 & 0.00561 & 0.00578 & 0.00477 & 0.00618 & 0.00367 & 0.00733 \end{pmatrix},$$

$$\Sigma_2 = \begin{pmatrix} 0.00120 & 0.00155 & 0.00140 & 0.00096 & 0.00140 & 0.00069 & 0.00128 \\ 0.00155 & 0.00232 & 0.00194 & 0.00116 & 0.00185 & 0.00069 & 0.00163 \\ 0.00140 & 0.00194 & 0.00201 & 0.00103 & 0.00169 & 0.00060 & 0.00139 \\ 0.00096 & 0.00116 & 0.00103 & 0.00122 & 0.00106 & 0.00067 & 0.00109 \\ 0.00140 & 0.00185 & 0.00169 & 0.00106 & 0.00211 & 0.00080 & 0.00155 \\ 0.00069 & 0.00069 & 0.00060 & 0.00067 & 0.00080 & 0.00100 & 0.00085 \\ 0.00128 & 0.00163 & 0.00139 & 0.00109 & 0.00155 & 0.00085 & 0.00211 \end{pmatrix}.$$

Applying (8) the estimated transition matrix (2) is given by:

$$P = \begin{pmatrix} 0.29 & 0.71 \\ 0.16 & 0.84 \end{pmatrix},$$

Being the size of each group $m_1 = 21$ and $m_2 = 95$, from (10) the center of each cluster (or group) is given by $z_1 = 15$ and $z_2 = 5.19$. Applying (9) we have one threshold distance $\bar{d} = 10.19$. Then, from (11) the market regime for a new observation, i.e. a new vector of returns $r(t)$, can be classified as follows:

$$\theta(t) = \begin{cases} 1, & \text{if } d(t) > 10.19 \\ 2, & \text{if } d(t) \le 10.19 \end{cases}.$$

With the purpose of showing the behavior of the enhanced index tracking problem with switching regime, the previously proposed methodology was applied to the model without regime switches (i.e. $\eta_1 = \eta_2 = \eta$ and $\Sigma_1 = \Sigma_2 = \Sigma$), in which the vector of average returns η and the covariance matrices Σ were estimated using the sample daily return with size $T = 116$ and are given by

$$\eta = (-0.00272 \quad -0.00233 \quad -0.00206 \quad -0.00268 \quad -0.00637 \quad 0.00022 \quad -0.00600)',$$

$$\Sigma = \begin{pmatrix} 0.00190 & 0.00221 & 0.00210 & 0.00166 & 0.00219 & 0.00121 & 0.00210 \\ 0.00221 & 0.00310 & 0.00267 & 0.00186 & 0.00261 & 0.00112 & 0.00251 \\ 0.00210 & 0.00267 & 0.00285 & 0.00180 & 0.00252 & 0.00104 & 0.00227 \\ 0.00166 & 0.00186 & 0.00180 & 0.00209 & 0.00191 & 0.00113 & 0.00184 \\ 0.00219 & 0.00261 & 0.00252 & 0.00191 & 0.00310 & 0.00149 & 0.00252 \\ 0.00121 & 0.00112 & 0.00104 & 0.00113 & 0.00149 & 0.00160 & 0.00146 \\ 0.00210 & 0.00251 & 0.00227 & 0.00184 & 0.00252 & 0.00146 & 0.00325 \end{pmatrix}.$$

Note that the optimization problem aims at finding an optimal allocation at each instant t that minimizes the objective function defined in (12), subject to (13) and (14). From (12), the type of investment strategy can be defined by the balance between the linear and quadratic terms that is established through the weights $\delta_i(t)$ and $\xi_i(t)$, respectively. For the purpose of this work, an enhanced index tracking strategy (named here enhanced management) with $\delta_i(t) = 0.3$ and $\xi_i(t) = 0.1$ was compared to an active management strategy with $\delta_i(t) = 0.1$ and $\xi_i(t) = 0.8$. Setting $X(1) = 100$ and $Y(1) = 100$ as initial values to the portfolio of investments (13) and to the reference portfolio (14), the solution presented in Costa and Paulo (2007) for the problem (12)-(14) was implemented using the Matlab software. Figure 1 and 2 (see the Appendix) show the results for the value of the portfolio investment $X(t)$ and the value of the reference portfolio $Y(t)$, for enhanced management and active management strategies (with and without switching regime).

From Figure 1 and Figure 2 we can see that on average the model with regime switches performs better than the model without regime switches. The performance of an investment can be measured using some kind of indicator as Sharpe ratio and Jensen's alpha (see Sharpe, 1994; Jensen, 1968; and Liptona and Kishb, 2010, for example). To measure the performance of the investment portfolio for each model (with and without regime switches) three indicators were used: mean squared error measure, Sortino ratio (Sortino and van der Meer, 1991) end upside potential ratio (Sortino et al., 1999). The mean squared error measure (MSE) was calculated as follows:

$$MSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(X(t) - Y(t))^2},$$

where $X(t)$ is the value of the portfolio and $Y(t)$ is the value of the reference portfolio. The Sortino ratio (SR) was calculated by:

$$SR = \frac{\frac{1}{T}\sum_{t=1}^{T}(X(t)-Y(t))}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\mathrm{I}^{-}(X(t)-Y(t)))^2}},$$

where $\Gamma(x)$ is such that $\Gamma(x) = x$ if $x < 0$ and $\Gamma(x) = 0$ if $x \geq 0$ (in this case the minimum acceptable return (MAR) is replaced by the reference portfolio $Y(t)$). Finally, the upside potential ratio (UPR) was calculated as follows:

$$UPR = \frac{\frac{1}{T}\sum_{t=1}^{T}I^{+}(X(t)-Y(t))}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(I^{-}(X(t)-Y(t)))^2}},$$

where $I^{+}(x)$ is such that $I^{+}(x) = x$ if $x > 0$ and $I^{+}(x) = 0$ if $x \leq 0$. Notice that the indicators SR and UPR measure the average of excess return and the average of return above the benchmark, respectively, divided by the downside risk (or downside volatility). Table 3 shows the results for the application of the three indicators presented above.

Table 3. Performance measures considering the two strategies: enhanced management and active management

	Enhanced management			Active management		
Type of model	SR	UPR	MSE	SR	UPR	MSE
With regime switches	0.57	29.27	1.68	15.48	34.77	4.99
Without regime switches	1.26	15.79	2.67	2.92	31.05	2.86

From Table 3 note that *SR* for the model with regime is less (greater) than *SR* for the model without regime, considering the enhanced (active) management strategy. On the other hand, we can also see that *UP*R for the model with regime is greater than *UP*R for the model without regime in both strategies. Moreover, from the *MSE* indicator we can see that the value of the portfolio is more (less) adherent to the benchmark when the model with regime is used (compared to the model without regime), considering the enhanced (active) management strategy. Therefore, we can conclude that the model with regime switches, developed using the methodology proposed in this article, performed better than the model without regime.

Conclusion

This paper presents a non-parametric procedure based on cluster analysis tools for determining the number of regimes, estimate the parameters, and define in which regime the market belongs to, for financial markets under regime switching. In this case the expected value and covariance matrix of the asset returns can change according to a Markov chain taking values in a finite set. This new approach is a simple alternative to the classical multivariate Markov switching framework (MMS) used to estimate the market parameters and define the regimes along the time. The application of an MMS model can become cumbersome and computationally intensive when there is a large number of market regimes and variables. The proposed methodology was applied to a portfolio optimization problem with enhanced index tracking and switching regime. The results showed a satisfactory performance of the model with regime switches when compared to the case without regime switches.

References

1. Awirothananon, T. and Cheung, W. (2009). On joint determination of the number of states and the number of variables in Markov-switching models: A Monte Carlo study. Communications in Statistics, *Simulation and Computation*, 38, pp. 1757-1788.
2. Bae, G.I., Kim, C.W. and Mulvey, J.M. (2014). Dynamic asset allocation for varied financial markets under regime switching framework, *European Journal of Operational Research*, 234, pp. 450-458.
3. Bajeux-Besnainou, I., Belhaj, R., Maillard, D., and Portait, R. (2011). Portfolio optimization under tracking error and weights constraints, *The Journal of Financial Research*, 34, pp. 295-330.
4. Bastos, J.A. and Caiado, J. (2012). Clustering financial time series with variance ratio statistics, *Quantitative Finance*, pp. 1-13.
5. Bauer, R., H.R. and Molenaar, R. (2004). Asset allocation in stable and unstable times, *The Journal of Investing*, 13, pp. 72-80.
6. Billio, M. and Pelizzon, L. (2000). Value-at-risk: a multivariate switching regime approach, *Journal of Empirical Finance*, 7, 531-554.
7. Canakgoz, N.A. and Beasley, J.E. (2008). Mixed-interger programming approaches for index tracking and enhanced indexation, *European Journal of Operational Research*, 196, pp. 384-399.
8. Chen, C. and Kwon, R.H. (2012). Robust portfolio selection for index tracking, *Computers & Operations Research*, 39, pp. 829-837.
9. Chow, G., Jacquier, E., Kritzman, M., and Lowry, K. (1999). Optimal portfolios in good times and bad, *Financial Analysts Journal*, 55, pp. 65-73.
10. Costa, O.L.V. and Paulo, W.L. (2007). Indefinite quadratic with linear costs optimal control of Markov jump with multiplicative noise systems, *Automatica*, 43, pp. 587-597.
11. Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011). *Cluster Analysis*, Wiley-Interscience.

12. Guastaroba, G. and Speranza, M.G. (2012). Kernel search: An application to the index tracking problem, *European Journal of Operational Research*, 217, pp. 54-68.
13. Guidolin, M. and Hyde, S. (2007). What tames the celtic tiger? Portfolio implications from a multivariate Markov switching model, *Working Paper*.
14. Guidolin, M. and Timmermann, A. (2007). Asset allocation under multivariate regime switching, *Journal of Economic Dynamics and Control*, 31, pp. 3503-3544.
15. Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica*, 57, pp. 357-384.
16. Hamilton, J.D. (1990). Analysis of time series subject to changes in regime, *Journal of Econometrics*, 45, pp. 39-70.
17. Jensen, M. (1968). The performance of mutual funds in the period from, *Journal of Finance*, 23, pp. 389-416.
18. Johnson, R.A. and Wichern, D.W. (2007). Applied Multivariate Statistical Analysis, Prentice-Hall.
19. Jorion, P. (2003). Portfolio optimization with tracking-error constraints, *Financial Analysts Journal*, 59, pp. 70-82.
20. Kritzman, M., Li, Y., Page, S. and Rigobon, R. (2011). Principal components as a measure of systemic risk, *The Journal of Portfolio Management*, 37, pp. 112-126.
21. Kritzman, M., Lowry, K. and Royen, A.V. (2001). Risk, regimes, and overconfidence, The *Journal of Derivatives*, 8, pp. 32-42.
22. Leippold, M., Trojani, F. and Vanini, P. (2004). A geometric approach to multiperiod mean variance optimization of assets and liabilities, *Journal of Economic Dynamics & Control*, 28, pp. 1079-1113.
23. Li, D. and Ng, W.L. (2000). Optimal dynamic portfolio selection: Multiperiod mean-variance formulation, *Mathematical Finance*, 10, pp. 387-406.
24. Li, Q., Sun, L. and Bao, L. (2011). Enhanced index tracking based on multi-objective immune algorithm, *Expert Systems with Applications*, 38, pp. 6101-6106.
25. Liptona, A.F. and Kishb, R.J. (2010). Robust performance measures for high yield bond funds, *The Quarterly Review of Economics and Finance*, 50, pp. 332-340.
26. Markowitz, H. (1959). Portfolio Selection: Efficient Diversification of Investments, John Wiley, New York.
27. Psaradakis, Z. and Spagnolo, N. (2003). On the determination of the number of regimes in markov-switching autoregressive models, *Journal of Time Series Analysis*, 24, pp. 237-252.
28. Rencher, A.C. (2002). Methods of Multivariate Analysis, Wiley-Interscience.
29. Roll, R. (1992). A mean/variance analysis of tracking error, *The Journal of Portfolio Management*, 18, pp. 13-22.
30. Rudolf, M., Wolter, H.J. and Zimmermann, H. (1999). A linear model for tracking error minimization, *The Journal of Banking and Finance*, 23, pp. 85-103.
31. Saunders, D., Seco, L., Vogt, C. and Zagst, R. (2013). A fund of hedge funds under regime switching, *The Journal of Alternative Investments*, 15, pp. 8-23.
32. Sharpe, W. (1994). The sharpe ratio, *Journal of Portfolio Management*, 21, pp. 49-58.
33. Sortino, F. and van der Meer, R. (1991). Downside risk, *The Journal of Portfolio Management*, 17, pp. 27-31.
34. Sortino, F., van der Meer, R. and Plantinga, A. (1999). The dutch triangle, *The Journal of Portfolio Management*, 26, pp. 50-57.
35. Spezia, L. (2010). Bayesian analysis of multivariate gaussian hidden markov models with an unknown number of regimes, *Journal of Time Series Analysis*, 31, pp. 1-11.
36. Stoyanov, S., Rachev, S., Ortobelli, S. and Fabozzi, F. (2008). Relative deviation metrics and the problem of strategy replication, *Journal of Banking and Finance*, 32, pp. 199-206.
37. Sugar, C.A. and James, G.M. (2003). Finding the number of clusters in a data set: An information-theoretic approach, *Journal of the American Statistical Association*, 98, pp. 750-763.
38. Sun, H., Wang, S. and Jiang, Q. (2004). Fcm-based model selection algorithms for determining the number of clusters, *Pattern Recognition*, 37, pp. 2027-2037.
39. Taamouti, A. (2012). Moments of multivariate regime switching with application to risk-return trade-off, *Journal of Empirical Finance*, 19, pp. 292-308.
40. Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic, *Journal of the Royal Statistical Society*, 63, pp. 411-423.
41. Wu, L.C., Chou, S.C., Yang, C.C. and Ong, C.S. (2007). Enhanced index investing based on goal programming, *The Journal of Portfolio Management*, 33, pp. 49-56.
42. Yin, G. and Zhou, X. (2004). Markowitz's mean-variance portfolio selection with regime switching: From discrete-time models to their continuous-time limits, *IEEE Trans. Automatic Control*, 49, pp. 349-360.
43. Zhou, X. and Li, D. (2000). Continuous-time mean-variance portfolio selection: A stochastic LQ framework, *Applied Mathematics & Optimization*, 42, pp. 19-33.
44. Zhou, X. and Yin, G. (2003). Markowitz's mean-variance portfolio selection with regime switching: A continuous-time model, *SIAM Journal on Control and Optimization*, 42, pp. 1466-1482.
45. Zhu, H., He, Z. and Leung, H. (2012). Simultaneous feature and model selection for continuous hidden markov models, *IEEE Signal Processing Letters*, 19, pp. 279-282.

Appendix

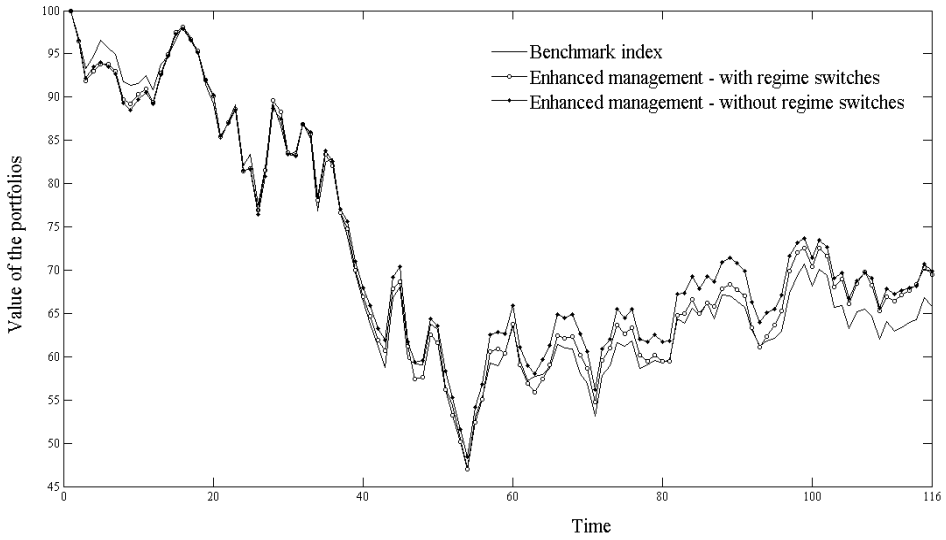

Fig. 1. Values of the portfolio investment *X*(*t*) considering the enhanced management strategy without regime switches and with regime switches

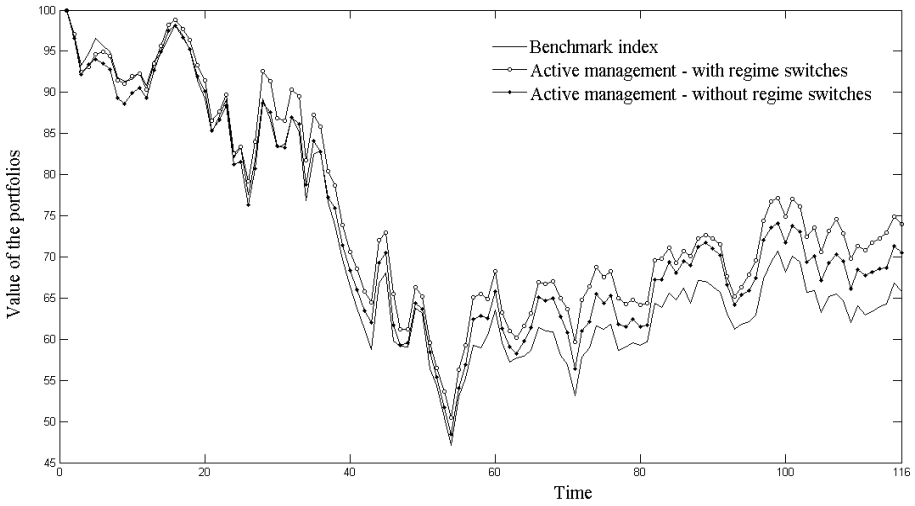

Fig. 2. Values of the portfolio investment X(t) considering the active management strategy without regime switches and with regime switches