Latent class models for financial data analysis: Some statistical developments

Luca De Angelis*

Dipartimento di Scienze Statistiche, University of Bologna, Italy

November, 2010

* All correspondence to: Luca De Angelis, Dipartimento di Scienze Statistiche, Alma Mater Studiorum Universita' di Bologna, Via delle Belle Arti, 41, 40126, Bologna. Italy.

E-mail: l.deangelis@unibo.it; Phone: +39-051-2094628; Fax: +39-051-232153.

Latent class models for financial data analysis: Some statistical developments

Abstract

I exploit the potential of latent class models for proposing an innovative framework for financial data analysis. By stressing the latent nature of the most important financial variables, expected return and risk, I am able to introduce a new methodological dimension in the analysis of financial phenomena. In my proposal, (i) I provide innovative measures of expected return and risk, (ii) I suggest a financial data classification consistent with the latent risk-return profile, and (iii) I propose a set of statistical methods for detecting and testing the number of groups of the new data classification. The results lead to an improvement in both risk measurement theory and practice and, if compared to traditional methods, allow for new insights into the analysis of financial data. Finally, I illustrate the potentiality of my proposal by investigating the European stock market and detailing the steps for the appropriate choice of a financial portfolio.

Keywords: Latent variable; Latent class model; Financial data analysis; Risk-return profile; Portfolio choice.

1. Introduction

Statistical methods for latent variables have a longstanding tradition in both theoretical and empirical researches and cover a wide range of academic and operational fields. Notwithstanding the relevant progresses made in the last years, the usefulness of latent variables in financial studies is still largely unexplored. In this paper I propose to analyze the most important financial variables by stressing their latent nature and I suggest to resort to the statistical methodology developed for the analysis of latent variable in order to introduce an innovative assessment of financial phenomena.

In traditional financial studies, stocks are analyzed on the basis of two dimensions: the risk and the expected return. In this framework, the contribution of statistical methodology can be relevant, since both risk and expected return are variables which are not directly observable and, therefore, can be measured by means of the numerous statistical methods developed for latent variables. More specifically, the stock's risk - expected return profile can be seen as a latent variable underlying the stock's performance and the observed return values can be used as the indicator variables which enable the development of a measurement procedure. Thus, it is crucial to define a methodological process which is able both to assess the latent nature of the risk and the expected return and to guide me to discriminate stocks under their risk-return profile.

To achieve this purpose, I propose to analyze the financial data by exploiting the potential of the latent class (LC) analysis. This methodology developed by Lazarsfeld and Henry (1968) for sociological researches is an extremely powerful tool in order to obtain a straightforward classification of the observations (Magidson and Vermunt, 2001). LC analysis classifies multivariate data following a model-based approach rather than a procedure based on measures of distances and similarities such as cluster analysis, multidimensional scaling, or correspondence analysis (Banfield and Raftery, 1993). Furthermore, with respect to other classification procedures, within LC framework, it is possible to test the suitability of the fitted model through different statistical indicators.

Latent class models are widely used in social sciences research, e.g. psychology (Bouwmeester et al., 2004), sociology (Moors and Vermunt, 2007), medicine and psychiatry (Leask et al., 2009), marketing (Bijmolt et al., 2004), or archaeometry (Moustaki and Papageorgiou, 2005) for its flexibility and classification potentiality. However, to my

knowlegde, the application of this statistical methodology to financial data analysis is at a very preliminary level.

The first aim of the paper is to obtain groups of stocks characterized by homogenous risk-return profiles. My proposal is to identify these groups with the latent classes, achieved within LC analysis. In this way I am also able to suggest a solution to a second open issue in financial data analysis: my proposal allows me to use statistical indicators and tests developed for determining the number of latent classes in order to define the number of groups of stocks with different financial features. This approach introduces a methodological dimension in the classification procedure of financial data which are usually grouped without involving any scientific rule.

Furthermore, in order to show the potentiality of my proposal, I address one of the most widespread cases of financial data analysis, the choice of a portfolio. As a matter of fact, the stock's classification achieved by LC analysis leads to an improvement in the risk measurement methods and, therefore, in the diversification processes which can be employed to define a portfolio.

In Section 2, I introduce the LC model I suggest for financial data analysis and I illustrate the latent nature of financial variables. I complete the methodological section with the description of the classification procedure I use, in comparison to the existing proposals. In Section 3, I present an empirical analysis of the European stock market using LC analysis and illustrate the implications that my proposal has on the definition of a financial portfolio. In Section 4, I conclude.

2. Latent class methods for financial data classification

The LC analysis is usually performed using some observed indicators which express the manifest variables included in the model in order to obtain inference about the latent variable of interest and the subsequent classification of the units into the latent classes.

Denoting by $\mathbf{y}'_h = (y_{1h}, y_{2h}, ...y_{ih}, ..., y_{ph})$ the vector of the p manifest variables for the hth sample unit, for $h = 1, 2, \ldots, n$, the LC model is specified as

$$f(\mathbf{y}_h) = \sum_{x=1}^{K} \eta_x f(\mathbf{y}_h \mid x). \quad (1)$$

Therefore, in LC analysis is assumed that observed data are modelled as drawn from a factor space consisting of *K* latent classes. For each class *x* there is an associated proportion, η_x, which is referred to as the prior probability and it is assumed that $\sum_{x=1}^{K} \eta_x = 1$.

In this framework, all the relationship among the observed variables is explained by the latent variable *x*. In other words, the manifest variables are assumed to be independent conditional on the latent classes. This is known as the local independence assumption and implies that

$$f(\mathbf{y}_h) = \sum_{x=1}^{K} \eta_x \prod_{i=1}^{p} f(y_{ih} \mid x). \quad (2)$$

The conditional distribution, $f(y_{ih} \mid x)$, indicates the probability of assuming a particular value for the *i*th manifest variable given that the unit is classified into latent class *x*.

These conditional distributions are assumed to be Gaussians with (conditional) mean $\mu_x(i)$ and variance $\sigma_x^2(i)$ for the *i*th manifest variable y_i in class *x*:

$$f(y_{ih} \mid x) = \frac{1}{\sqrt{2\pi\sigma_x^2(i)}} \exp\left\{-\frac{1}{2}\frac{(y_{ih} - \mu_x(i))}{\sigma_x^2(i)}\right\}.$$

The normality assumption for continuous manifest variables in model-based clustering is well established since it is assumed that Equation (1) is not necessarily "true", but rather that Gaussian distribution is treated as a cluster shape prototype, given that many distributions can be approximated closely by a Gaussian mixture (Coretto and Hennig, 2010). A further development of the analysis of financial variables is to assume an alternative conditional

distribution, e.g. the *t*-distribution as proposed by McLachlan and Peel (2000) for mixture models.

The conditional distribution, $h(x \mid \mathbf{y}_h)$, which denotes the probability of a unit belonging to latent state x given $\mathbf{y}$ is referred to as the posterior probability and can be written as

$$h(x \mid \mathbf{y}_h) = \frac{\eta_x f(\mathbf{y}_h \mid x)}{f(\mathbf{y}_h)}.$$

Posterior probabilities are used to perform the classification of the units into the K latent classes, see Section 2.2.

The log-likelihood function for a random sample of size n is given by

$$LL = \sum_{h=1}^{n} \log f(\mathbf{y}_h) = \sum_{h=1}^{n} \log \left[\sum_{x=1}^{K} \eta_x \prod_{i=1}^{p} f(y_{ih} \mid x) \right] \tag{3}$$

for $x = 1, \ldots, K$. The log-likelihood in Equation (3) can be maximized by exploiting the iterative procedure of the EM algorithm (Dempster et al., 1977) under the constraint $\sum_{x=1}^{K} \eta_x = 1$.

The EM algorithm provides the maximum likelihood estimations of the LC model parameters following the iterative procedure reported below:

Step 1. Choose initial values for the posterior probabilities $h(x \mid \mathbf{y}_h)$.

Step 2. Achieve a first approximation for the model parameters using the following equations:

Estimated prior probabilities:

$$\hat{\eta}_x = \sum_{h=1}^{n} \hat{h}(x \mid \mathbf{y}_h) / n$$

Estimated conditional means:

$$\hat{\mu}_x(i) = \sum_{h=1}^{n} y_{ih} \hat{h}(x \mid \mathbf{y}_h) / (n \hat{\eta}_x) \tag{4}$$

Estimated conditional variances:

$$\hat{\sigma}_x^2(i) = \sum_{h=1}^{n} (y_{ih} - \hat{\mu}_h(i))^2 \hat{h}(x \mid \mathbf{y}_h) / \sum_{h=1}^{n} \hat{h}(x \mid \mathbf{y}_h)$$

Step 3. Achieve a new estimate for the posterior probability:

$$\hat{h}(x \mid \mathbf{y}_h) = \frac{\hat{\eta}_x \hat{f}(\mathbf{y}_h \mid x)}{\hat{f}(\mathbf{y}_h)}. \tag{5}$$

Step 4. Return to Step 2 and continue until convergence is attained.

In order to avoid the problem of likelihood convergence into a local maximum which occurs frequently in LC models, especially when the number of variables is large and the sample size is small (Wedel and DeSarbo, 1995), many different starting values are used. For more background and details on the EM algorithm, see McLachlan and Krishnan (1997).

Vermunt and Magidson (2005) suggest to switch to the Newton-Raphson algorithm once the EM procedure is close to the final solution. By using both the algorithms, the convergence is achieved faster.

2.1. Latent nature of financial variables

In financial studies and, in particular, in the framework of standard portfolio theory (Markowitz, 1952), the analyses are performed on the basis of two variables which cannot be directly measured: the risk and the expected return. An accurate and precise measurement of these variables requires special attention and a considerable methodological effort since no empirical corresponding quantities exist. The expected return and the risk are thus two latent

variables which underlie and characterize the financial phenomena. As consequence, a set of observed variables, e.g. the mean and some variability measures of past stock's returns, is usually employed in order to achieve an approximation of the risk and expected return variables.

More rigorously, the unobservable expected return variable E(r) is usually approximated by

$$\mathrm{E}(r) \approx \bar{r}$$

where $\bar{r}$ indicates the mean of the past stock's returns. The unobservable risk, denoted with V, can be obtained as a function of the standard deviation σ and some percentiles τ_ν of the observed stock's return distribution:

$$V = f(\sigma, \tau_\nu).$$

The variable τ_ν, where usually $\nu = 0.01$ and/or $\nu = 0.05$, denotes the extreme values of the stock's return distribution which correspond to the potential biggest losses and play a fundamental role in investment decision-making process.

The latent variables E(r) and V lead to the stock's risk – return profile, which is also a latent variable and represents the main interest in this study because it summarizes the latent characteristics of the financial variables.

In this paper, my aim is to analyze the latent nature of these variables by exploiting the potential of the latent class models for measuring and making easily interpretable the stock's risk - return profile. With the purpose of specifying a LC model able to accurately analyze financial data and to guide in the selection of a financial portfolio, I suggest to consider seven (manifest) variables which allows me to achieve a measurement of the latent expected return and risk variables.

First, I resort to the stock's mean return, $\bar{r}$, in order to approximate the expected return. Second, I use the standard deviation, σ, the first and the fifth percentiles ($\tau_{0.01}$ and $\tau_{0.05}$), of the stock's monthly return distribution as proxies of the risk. Third, as a further approximation of the latent expected return variable, I include the ninetieth percentile, $\tau_{0.90}$, as a proxy of the stock's potential to create wealth.

Finally, I propose a further generalization in order to account for the heteroskedasticity which characterizes financial variables. To achieve this purpose, I take advantage of the great

flexibility of LC models, which simply allow to evaluate the effects of high variability (i.e. volatility) periods by including the stock's performance during turmoil phases among the manifest variables. In particular, I first endogenously detect the periods characterized by high variability and, as second step, I compute the mean, $\bar{r}_C$, and the standard deviation, σ_C, of the stock's return associated with these turmoil periods.

My feeling is that LC models represent an extremely appealing solution in order to deal with both the latent nature of financial variables and their heteroskedasticity.

2.2. Classification of financial data

Classification of financial data is a relevant and innovative topic which is experiencing an increasing interest in the statistical literature as testified by the several recent works. Many authors contributed to the classification of financial variables by means of different clustering procedures. Among the others, Pattarin et al. (2004), following a pioneering analysis by Brown and Goetzmann (1997), used principal components and a genetic algorithm for clustering mutual funds styles; Da Costa et al. (2005) proposed a straightforward cluster analysis for classifying stocks according to a risk – return criterion; Caiado and Crato (2007), Otranto (2008), and Lisi and Otranto (2008) based their classification processes on heteroskedasticity models; and Basalto et al. (2007) referred to a clustering procedure based on the Hausdorff distance. Recently, Dias et al. (2010) proposed a mixture latent Markov model which allows for static and dynamic classifications of stock market indexes.

My purpose is to contribute to the existing literature on financial data classification by exploiting the model-based clustering procedure of LC models (Vermunt and Magidson, 2003). This approach allows me to follow a strict methodological process for defining the stock's classification in homogenous groups with respect to their latent risk - return profile.

The first step of the analysis is the definition of the number *K* of latent classes and it represents a relevant improvement because it allows me to introduce a methodological

dimension in financial variable classification. The issues about robustness and reliability of tests and criteria for model selection in LC analysis have been widely discussed (e.g., Nylund et al., 2007). I agree with the concerns about the uncritical use of these indicators but I also believe that they can contribute to the current procedures used in financial data classification, for which the choice of K is somewhat arbitrary.

In the following, I resort to both the Akaike information criterion and the likelihood ratio test for comparing nested LC models. The Akaike information criterion (Akaike, 1974) is expressed as

$$AIC = -2[\max LL] + 2m$$

where LL is the log-likelihood function in Equation (3) and m denotes the number of parameters.

The likelihood ratio test statistic is computed as

$$LRT = -2(\max LL \mid H_0 - \max LL \mid H_1) \tag{6}$$

where H_0 refers to the more restricted model and H_1 to the more general model (e.g., a LC model with $K + 1$ classes). P-values are estimated by parametric bootstrap; replication samples are generated from the probability distribution defined by the maximum likelihood estimates under H_0. The estimated bootstrap p-value is defined as the proportion of bootstrap samples with a larger LRT value than the original sample.

Once I define the number of latent classes, the next step is to allocate the units into the K groups. The classification method, which belongs to the unsupervised learning structure family (Vermunt and Magidson, 2003), is based on the posterior probabilities estimated by the LC model $\hat{h}(x \mid \mathbf{y}_h)$ given by Equation (5). More precisely, units are allocated to the latent class with the highest posterior probability:

$$\arg \max_{x=1,\ldots,K} \hat{h}(x \mid \mathbf{y}_h) .$$

This method of assignment is sometimes referred to as empirical Bayes modal or modal a posteriori estimation (Skrondal and Rabe-Hesketh, 2004).

The estimated posterior probabilities are also used to compute the estimated proportion of classification errors (E) which is defined as

$$E = \frac{\sum_{h=1}^{n} \left[1 - \max \hat{h}(x \mid \mathbf{y}_h)\right]}{n}.$$

Obviously, a good classification should have a value of E close to zero.

Another statistic based on the estimated posterior probabilities is the R-squared based on entropy (Magidson and Vermunt, 2005)

$$R^2_{entropy} = 1 - \frac{\frac{1}{n}\sum_{x=1}^{K}\sum_{h=1}^{n} -\hat{h}(x \mid \mathbf{y}_h) \log \hat{h}(x \mid \mathbf{y}_h)}{\sum_{x=1}^{K} -\hat{h}(x) \log \hat{h}(x)} \tag{7}$$

where $\hat{h}(x)$ are the estimated marginal latent probabilities, which are defined as

$$\hat{h}(x) = \frac{\sum_{h=1}^{n} \hat{h}(x \mid \mathbf{y}_h)}{n}.$$

This indicator is particularly useful for evaluating the classification power of the estimated LC model and can be interpreted as a measure for determining how well the latent classes are separated. I use the R-squared measure based on entropy in Equation (7) also as a further indicator for helping me in the model selection procedure.

3. Empirical analysis: The European stock market

Among the many developments of LC analysis which can be illustrated in the field of financial data, I propose an investigation of the European financial market. In particular, I analyze the monthly return distribution from January 1999 to August 2010 of the 50 stocks included in the Dow Jones' EUROSTOXX 50 index, using the seven variables described in Section 2.1 which define the vector of manifest variables $\mathbf{y}_h$ in Equation (1).

The endogenous detection of the periods characterized by high variability is performed according to the extreme negative values of the stock market index returns (observations lower than the fifth percentile of the index monthly return distribution), and including two time-observations before and after that date. Furthermore, the standard deviation value in these periods must be at least 1.5 times higher than the standard deviation computed on the whole data series, otherwise it will not be considered as an high variability phase. Following these criteria, the periods characterized by high variability on which I compute variables $\bar{r}_C$ and σ_C are the following: from July to December 2001, from June 2002 to March 2003, and from December 2007 to May 2009.

3.1. LC model estimation and results

The estimation of the LC models for different values of *K* allows me to define the number of classes which can better explain the relationships existing among the manifest variables. In my proposal, *K* represents the number of groups which characterizes the new financial data classification. The LC models are estimated using jointly the EM and the Newton-Raphson algorithms and 50 different starting values in order to avoid the problem of likelihood convergence into a local maximum.

Table 1 reports the results of the log-likelihood values (*LL*), the number of parameters (*m*), the Akaike information criterion (*AIC*), the classification error (*E*), and the R-squared entropy ($R^2_{entropy}$) from LC model estimation with different number of classes.

The results in Table 1 show that, according to the Akaike information criterion, the best model is the 7-class LC model, thus indicating the presence of seven underlying different groups of stocks. The LC model with seven latent classes provides the highest value for the R-squared entropy ($R^2_{entropy} = 0.9877$) and a low classification error ($E = 0.0048$). Furthermore, the likelihood ratio test with p-value achieved by bootstrap, introduced in Equation (6), confirm the validity of the 7-class LC model specification. The comparison between the model with 7 classes (H_1) and the models with K from 1 to 6 (H_0) provides, in all the cases, p-values below 0.05, thus rejecting the null hypotheses. On the contrary, the likelihood ratio test which compares models with 7 and 8 latent classes leads to a non-rejection of H_0: $K = 7$ against the alternative H_1: $K = 8$ ($LRT = 21.937$ with (bootstrap) p-value = 0.084).

This set of statistical indicators and tests indicates that the 7-class LC model provides the best fit to the data and allows me to identify the number of groups in which classifying the 50 European stocks following a strict methodological process and avoiding any a prioristic assumption[1].

[INSERT TABLE 1 ABOUT HERE]

The results related to the 7-class LC model estimation are shown in Table 2 which illustrates prior probabilities and the conditional means of each manifest variable[2]. The latent

[1] Traditional procedures and subjective beliefs could suggest a lower number of relevant stock's groups, but, since my proposal includes tests for the choice of *K*, I will favour a statistical based method.
[2] It is worth mentioning that the Wald tests are all highly significant, denoting that the conditional means for each manifest variable strongly differ between the classes. Hence, the estimated LC model is able to markedly define the financial features of each group.

classes are numbered according to their sizes, i.e. on the basis of prior probabilities $\hat{\eta}_x$ reported on the first row. Class 1 is the modal group and collects the 21.8% of the stocks, while Class 7 is the smallest, with only the 6.2% of the stocks. The details of prior probabilities indicate the presence of some small groups, e.g. Classes 6 and 7, and some bigger ones, such as Classes 1, 2, and 3 which, if cumulated, cluster almost 60% of the stocks considered in the analysis.

The interpretation of the financial characteristics, i.e. the (latent) risk-return profile, of each latent class can be obtained on the basis of the conditional means, $\hat{\mu}_x(i)$, reported in Table 2. For example, Class 7 is characterized by the highest conditional mean for variables $\bar{r}$ and $\tau_{0.90}$, thus suggesting a high level of expected return. However, the evaluation of the variables related to the risk allows me to define Class 7 as the group with the highest level of risk as well. In particular, this latent class is characterized by the highest conditional mean value for variables σ, $\tau_{0.01}$, $\tau_{0.05}$, and also the lowest mean return during periods of high variability, $\hat{\mu}_7(\bar{r}_C) = -4.536$. Class 2 is characterized by the lowest mean return, $\hat{\mu}_2(\bar{r}) = 0.006$, and strongly underperforms also during high variability periods, $\hat{\mu}_2(\bar{r}_C) = -3.765$. It is also worth noting that, for Class 3, all the manifest variables associated to latent risk take the lowest conditional mean values. However, this group of stocks is characterized by the lowest conditional mean for variable $\tau_{0.90}$, an average value of mean return, $\hat{\mu}_3(\bar{r}) = 0.473$, and the highest mean return during periods of high variability, $\hat{\mu}_3(\bar{r}_C) = -2.021$. Therefore, Class 3 conjugates a moderate expected return with a very low level of risk. Conversely, the stocks classified into Class 6 are particularly volatile, as expressed by the conditional mean values for variables σ and σ_C, and they are affected by large drops in prices: conditional means for variables $\tau_{0.01}$ and $\tau_{0.05}$ are -36.57 and -19.91, respectively. Classes 4 and 5 are characterized by quite similar values of expected return (see the respective values of $\hat{\mu}_x(\bar{r})$ and $\hat{\mu}_x(\tau_{0.90})$ in

Table 2). However, latent class 4 shows a lower level of latent risk than Class 5, according to the other conditional means reported in Table 2. Finally, Class 1 is characterized by a moderately low level of risk and an average expected return.

[INSERT TABLE 2 ABOUT HERE]

The characterization of the risk-return profiles of the seven groups of stocks facilitates a correct financial evaluation. On one hand, a profitable investment should avoid Classes 2 and 6. On the other hand, an attractive portfolio should include stocks classified into Classes 3 and 4, and, for a higher level of risk, also those belonging to latent class 7.

3.2. Implications on financial data analysis: portfolio choice

The results related to the LC model estimation illustrated in Section 3.1 lead, first, to an innovative evaluation of the latent characteristics of each stock and, second, to a new financial data classification based on the latent risk-return profile.

In this section, I propose to further develop the analysis of the latent class model by adding a final step, which allows me to create a financial portfolio characterized by an optimal risk-return profile[3].

In the traditional framework of the standard portfolio theory, the set of optimal portfolios, called efficient frontier, is achieved by minimizing the risk for a given value of mean return and by maximizing the mean return for a given value of risk. Furthermore, it is crucial to evaluate the interrelations among the stocks participating to the portfolio.

In Figure 1 is illustrated, for each latent class reported in Table 2, the efficient frontier obtained by including the stocks assigned to each latent class. The comparison of these efficient frontiers shows how the different groups defined by the LC analysis are well-

[3] This is only one example of the many possible financial data analyses one might develop using this proposal.

separated and heterogeneous one to the other. For this reason, the different latent classes are particularly useful for defining effective investment strategies, i.e. a new efficient frontier characterized by superior performances.

[INSERT FIGURE 1 ABOUT HERE]

On the left side of Figure 1, that is in the lower risk area, it is possible to observe how the best frontiers are achieved by means of Class 3, Class 4, and Class 7, for the lowest, the average, and the highest mean return levels, respectively. Therefore, by jointly using these three latent classes, I achieve a set of optimal portfolios which can be exploited for setting an appealing investment strategy. The efficient frontier derived from the combination of the stocks classified into Classes 3, 4, and 7, depicted in Figure 1, clearly shows the best performance. Furthermore, on the right side of Figure 1, that is in the high risk area, are located the efficient frontiers related to Classes 2 and 5, for lowest mean returns values, and Class 6. Finally, it is worth noting how positions of efficient frontiers in Figure 1 are strictly consistent with the latent class characteristics illustrated in Table 2 and Section 3.1.

5. Conclusions

I propose an innovative framework for financial data analysis. First, I take into consideration the latent nature of the most important variables used for analyzing the financial phenomena and, within this framework, I suggest to base the measurement of both expected return and risk by resorting to the statistical methodology developed for the study of latent variables. In particular, I show how LC models allow me to emphasize the relevance of the statistical methods in financial data analysis and, hence, to introduce a new methodological dimension in the assessment of financial phenomena.

The flexibility of LC models enables me to include into the expected return and risk measurement a wide set of information and to overcome the traditional automatic correspondences between expected return and mean from one side, and between risk and standard deviation from the other side.

Furthermore, I derive an innovative stock's classification into homogenous groups under their latent risk-return profiles which enourmosly facilitates both the understanding of the underlying features characterizing the financial phenomena and the choice of an efficient portfolio on the basis of a rigorous statistical procedure.

The number of groups of the new data classification is determined on the basis of different statistical indicators, thus it is not decided in advance as in cluster analysis methods, e.g. k-means. In this paper, I use a traditional criterion, the AIC, and two more recent indicators for assessing the goodness of fit of the estimated model and guiding me in the model selection issue: the bootstrap likelihood ratio test and an R-squared measure based on entropy.

I also develop an empirical analysis, finding evidence of seven groups characterized by strongly different risk - return profiles in which classifying the 50 stocks included in the Dow Jones EUROSTOXX 50 index.

My proposal is particularly appealing since contributes to the definition of new and enhanced methods for both financial risk measurement and portfolio diversification processes, providing a more methodologically correct measurement of the latent variable risk with respect to the traditional risk assessment procedures.

Finally, I suggest to use this framework in a dynamic approach: the addition of new temporal observations to the data set allows a constant update of the stock's classification and, consequently, the possible evolution of the investment decisions.

References

Akaike, H., 1974. A New Look at the Statistical Model Identification. IEEE Trans. Automat. Control 19, 716-723.

Banfield, J.D., Raftery, A.E., 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics 49, 803-821.

Basalto, N., Bellotti, R., De Carlo, F., Facchi, P., Pantaleo, E., Pascazio, S., 2007. Hausdorff clustering of financial time series. Physica A 379, 635-644.

Bijmolt, T.H., Paas, L.J., Vermunt , J.K., 2004. Country and consumer segmentation: Multi-level latent class analysis of financial product ownership. Int. J. Res. Market. 21, 323-340.

Bouwmeester, S., Sijtsma, K., Vermunt, J.K., 2004. Latent class regression analysis for describing cognitive developmental phenomena: an application to transitive reasoning. Eur. J. Dev. Psychol. 1, 67-86.

Brown, S.J., Goetzmann, W.N., 1997. Mutual fund styles. J. Financ. Econ. 43, 373-399.

Caiado, J., Crato, N., 2007. A GARCH-based method for clustering of financial time series: International stock markets evidence. In: Skiadas, C.H. (Ed.), Recent Advances in Stochastic Modeling and Data Analysis, World Scientific Publishing, NJ, 542–551.

Coretto, P., Hennig, C., 2010. A simulation study to compare robust clustering methods based on mixtures. Adv. Data Anal. Classif. 4, 111-135.

Da Costa, N., Cunha, J., Da Silva, S., 2005. Stock selection based on cluster analysis. Econ. Bull. 13, 1-9.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Stat. Soc. B Stat. Meth. 39, 1-38.

Dias, J.G., Vermunt, J.K., Ramos, S., 2010. Mixture hidden Markov models in finance research. In: Fink, A., Lausen, B., Seidel, W. and Ultsch, A. (Eds.), Advances in data analysis, data handling and business intelligence, 451-459. Springer, Berlin-Heidelberg.

Lazarsfeld, P.F., Henry, N.W., 1968. Latent structure analysis. Houghton Mill, Boston, Mass.

Leask, S.J., Vermunt, J.K., Done, D.J., Crowd, T.J., Blows, M., and Boks, M.P., 2009. Beyond symptom dimensions: Schizophrenia risk factors for patient groups derived by latent class analysis. Schizophr. Res. 115, 346-350.

Lisi, F., Otranto, E., 2008. Clustering mutual funds by return and risk level. Crenos Working Papers, 200813, Centre for North South Economic Research, University of Cagliari and Sassari, Sardinia, Italy.

Magidson, J., Vermunt, J.K., 2001. Latent class factor and cluster models, bi-plots and related graphical displays. Socio. Meth. 31, 223-264.

Markowitz, H., 1952. Portfolio selection. J Finance 8, 77-91.

McLachlan, G., Krishnan, T., 1997. The EM algorithm and extensions. Wiley, New York.

McLachlan, G., Peel, D., 2000. Robust mixture modelling using the *t*-distribution. Stat. Comput. 10(4), 339-348.

Moors, G., Vermunt, J.K., 2007. Heterogeneity in postmaterialist value priorities. Evidence from a latent class discrete choice approach. Eur. Socio. Rev. 23, 631-648.

Moustaki, I., Papageorgiu, I., 2005. Latent class models for mixed variables with applications in archaeometry. Comput. Stat. Data Anal. 48, 659-675.

Nylund, K.L., Asparouhov, T., Muthén, B.O., 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling 14(4), 535-569.

Otranto, E., 2008. Clustering heteroskedastic time series by model-based procedures. Comput. Stat. Data Anal. 52, 4685-4698.

Pattarin, F., Paterlini, S., Minerva, T., 2004. Clustering financial time series: An application to mutual funds style analysis. Comput. Stat. Data Anal. 47, 353-372.

Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Chapman & Hall/CRC, London.

Vermunt, J.K, Magidson, J., 2003. Latent class models for classification. Comput. Stat. Data Anal. 41, 531-537.

Vermunt, J. K., Magidson, J., 2005. Latent GOLD 4.0 User's Guide, Statistical Innovations Inc., Belmont, Mass.

Wedel, M., DeSarbo, W.S., 1995. A mixture likelihood approach for generalized linear models. Journal of Classification 12, 1-35.

Table 1

Results from LC models estimation with different number of classes: log-likelihood, number of parameters, Akaike information criterion, classification error, R-squared entropy

K	LL	m	AIC	E	$R^2_{entropy}$
1	-825.75	14	1679.5	-	-
2	-709.32	29	1476.6	0.0024	0.9840
3	-640.98	44	1370.0	0.0046	0.9848
4	-610.17	59	1338.3	0.0126	0.9717
5	-591.91	74	1331.8	0.0104	0.9782
6	-571.89	89	1321.8	0.0109	0.9787
7	-552.42	104	1312.8	0.0048	0.9877
8	-542.71	119	1323.4	0.0052	0.9876

Table 2

Results related to the 7-class LC model estimation: prior probabilities and conditional means for the manifest variables $\bar{r}$, σ, $\tau_{0.01}$, $\tau_{0.05}$, $\tau_{0.90}$, $\bar{r}_C$ and σ_C.

Profile	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
$\hat{\eta}_x$	0.2184	0.1981	0.1794	0.1212	0.1204	0.1008	0.0616
$\hat{\mu}_x(\bar{r})$	0.538	0.006	0.473	0.974	0.841	0.244	1.244
$\hat{\mu}_x(\sigma)$	7.581	10.209	6.003	9.107	10.612	13.222	16.176
$\hat{\mu}_x(\tau_{0.01})$	-18.46	-25.93	-14.53	-22.02	-25.85	-36.57	-36.95
$\hat{\mu}_x(\tau_{0.05})$	-11.39	-16.34	-10.05	-13.27	-16.96	-19.91	-24.02
$\hat{\mu}_x(\tau_{0.90})$	9.33	10.67	7.35	11.23	12.64	12.58	19.42
$\hat{\mu}_x(\bar{r}_C)$	-2.222	-3.765	-2.021	-2.637	-3.090	-3.024	-4.536
$\hat{\mu}_x(\sigma_C)$	9.625	13.345	6.872	12.250	15.570	20.176	18.388

Fig. 1. Efficient frontiers for the seven latent classes estimated by the LC model and a combination of Classes 3, 4, and 7