Latent Class Analysis for Portfolio Choice

Michele Costa and Luca De Angelis

Abstract We exploit the potential of latent class analysis in order to propose an innovative framework for financial portfolio development. By stressing the latent nature of the most important financial variables, the expected return and the risk, we are able to introduce a methodological dimension in relevant steps of portfolio analysis. First, we provide a test for the number of possible investment choices. Second, we are able to include in the decision process also the information related to economic and financial environment. Our results lead to an improvement in the risk management methods and, if compared to traditional portfolio strategies, allow us to achieve better investment opportunities.

Key words: Latent variables, Latent class analysis, Statistical analysis of financial data, Financial 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 we propose to address one of the most widespread cases of financial decisions, the choice of a portfolio, by means of the statistical methodology developed for latent variables.

In standard portfolio theory, stocks are selected on the basis of two dimensions: the risk and the expected return. Furthermore, and most importantly, it is crucial to

[1] Michele Costa, Dipartimento di Scienze Statistiche, Università di Bologna; email: michele.costa@unibo.it

Luca De Angelis, Dipartimento di Scienze Statistiche, Università di Bologna; email: l.deangelis@unibo.it

evaluate the interrelations among the assets participating to the portfolio. On the whole, a portfolio is preferred when it maximizes the expected return for a given risk level, or it minimizes the risk for a given level of expected return.

In this framework, the contribution of statistical methodology can be relevant, since the risk and the expected return are variables which are not directly observable and, therefore, they can be analyzed 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 are the indicator variables which allow us to develop 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 help us to discriminate the stocks under their risk-return profile.

To achieve this purpose, we propose to base the decision-making process in portfolio choice 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 (see Magidson and Vermunt, 2001, and Moustaki and Papageorgiou, 2005, among the others).

Our first result is represented by the latent classes, obtained within LC models, which group stocks characterized by homogenous risk-return profiles. The stock's allocation achieved by LC analysis leads to an improvement in the risk management methods and, therefore, in the diversification processes which are employed to define an efficient portfolio. Furthermore, our results, compared to the traditional procedures, lead to better investment opportunities. The inclusion of information related to economic and financial framework increases the reliability of the estimates and their usefulness at both strategic and operative level.

Finally, the use of LC models allows us to introduce a further methodological dimension in the portfolio analysis and value the special relevance of statistical methods in financial decision processes. The use of LC analysis in the framework of financial studies is still at a preliminary stage, but, as suggested by our results, it promises interesting developments.

2 A Latent Class Model for Financial Variables

The LC analysis is usually performed using some observed indicators which express the manifest variables and the covariates included in the model in order to obtain maximum likelihood estimation of the parameters and the subsequent classification of the stocks into the latent classes.

With the purpose of specifying a LC model able to guide in the selection of a financial portfolio, we suggest to consider four indicator variables. First, we resort to the stock's mean return, M, in order to approximate the expected return. Second, we use the standard deviation, S, and the first percentile, P, of the stock's monthly return distribution as approximation of the risk. We also extend the existing literature on the role of LC models in financial field (Lisi and Otranto, 2008) by including also some stock's characteristics, such as the economic sector, C, and the market index membership, I, as covariates. Finally, we propose a further generalization in order to

account for the stock's behaviour during a financial crisis, *T*. To achieve this purpose we take advantage of the great flexibility of LC models, which simply allow to evaluate the effects of high volatility periods by including the stock's performance in turmoil phases among the manifest variables. The inclusion of this latter indicator allows us to evaluate a more sophisticated measurement of the latent risk. Therefore, we compute *T* as the standard deviation of the stock's daily return distribution in periods associated with well-known financial crises and/or big drops in stock markets.

The latent class analysis allows to determine one latent variable *X* characterized by *K* classes which can be interpreted as the new stock's classification into homogenous groups characterized by their own risk-return profile.

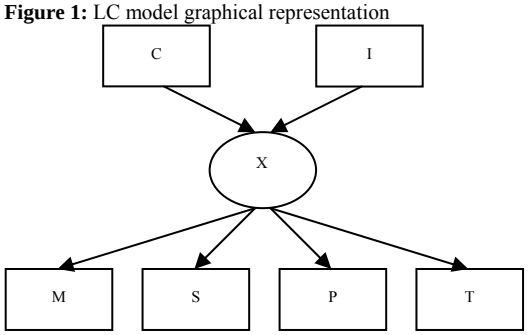
The LC model is specified as

$$f(Z) = \sum_{x=1}^{K} \pi_{X=x|C,I} \prod_{i=1}^{p} g(Z_i \mid X = x)$$

where *f*(*Z*) denotes the observed values of all the manifest variables $Z_i = M, S, P,$ and T (in our case, $p = 4$) and the covariates *C* and *I*, whereas $\pi_{X=x|C,I}$ indicates the probability of belonging to the latent class *x*, for $x = 1, \ldots, K$ (prior probability, given covariates *C* and *I*). The conditional probabilities $g(Z_i \mid X = x)$ indicate the probability of assuming a particular value for one of the four manifest variable given that the stock is classified to latent class *x*. These conditional probabilities are assumed to be normally distributed.

In this framework, all the relationship among the indicator variables *M*, *S*, *P*, and *T* is explained by the latent variable *X* which is influenced by the covariates *C* and *I*. In other words, the manifest variables are assumed to be independent conditional on the latent classes. This is known as the local independence assumption which is the pillar of LC models. As shown in Figure 1, no relationship exists among the four manifest variables and between the covariates and the indicators.

Figure 1: LC model graphical representation



The definition of the number *K* of latent classes is an important step of our analysis because it allows us to introduce a methodological dimension in financial product classification. The issues about robustness and reliability of the tests for defining the value of *K* have been widely discussed. We agree with the concerns about the uncritical

use of these indicators but we also believe that they can positively contribute to the actual procedures where the choice of K happens in a completely arbitrary way.

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

$$AIC = -2LL + 2NPar$$

where LL is the log-likelihood function and $NPar$ denotes the number of parameters, while the likelihood ratio test statistic is computed as

$$-2(LL|H_0 - LL|H_1)$$

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 $-2LL$ difference value than the original sample.

The classification of the units into the K latent classes is achieved through the Bayes' theorem. Stocks are classified to the latent class x for which the posterior probability is the highest. In particular, assuming that $g(Z_i \mid X = x)$ are normally distributed with mean $\mu(Z_i \mid X = x)$ and variance equals to one then latent class x is more probable than latent class x' if $\pi_{X=x|C,I}\, g(Z_i \mid X = x) > \pi_{X=x'|C,I}\, g(Z_i \mid X = x')$ which is true if (Bartholomew and Knott, 1999)

$$\sum_{i=1}^{p} Z_i \mu(Z_i \mid X = x) - \frac{1}{2}\sum_{i=1}^{p}(\mu(Z_i \mid X = x))^2 + \log \pi_{X=x|C,I} >$$

$$\sum_{i=1}^{p} Z_i \mu(Z_i \mid X = x') - \frac{1}{2}\sum_{i=1}^{p}(\mu(Z_i \mid X = x'))^2 + \log \pi_{X=x'|C,I} \,.$$

Latent classes thus obtained provide powerful insight on portfolio analysis by evaluating the latent risk – return profile of financial activities.

3 Data and Model Estimation

Our second aim is to analyze the Italian financial markets by means of the latent class specification previously illustrated; in particular, we analyze the monthly return distribution from January 2000 to December 2008 of 209 stocks quoted at the Italian Stock Market. According to our proposal for the statistical analysis of financial variables, we consider the following continuous variables as indicators:

- mean return (M);
- standard deviation (S);
- first percentile (P);
- standard deviation in crisis periods (T);

and the following two categorical variables as covariates:

- economic sector (C);
- market index (I).

The crisis periods we considered are: September-October 2001, July-October 2002, and September-December 2008. The covariate *C* is represented by the 10 economic sectors of the Global Industry Classification Standard (GICS) developed by MSCI and Standard and Poors in 1999 and which is one of the main references in portfolio diversification processes. The 209 stocks analyzed belong to the four main segments of the Italian market represented by the 4 major market indexes which define covariate *I*: S&P-Mib, Midex, All-Stars and Standard.

The estimation of latent class models for different values of *K* allows us to define the number of classes which can better explain the relationships existing among the manifest variables. In our proposal the number *K* of latent classes represents the number of groups which characterizes the new stock's classification.

Table 1 reports the results of the log-likelihood values (*LL*), the number of parameters (*NPar*), the Akaike information criterion (*AIC*), and the bootstrap likelihood ratio test (*Bootstrap -2LL Diff*) from LC model estimation with different number of classes. The latter test provides the comparison between the LC models with *K* and *K* + 1 classes. If the test is significant then adding a further latent class provides a better fit to the data.

The results in Table 1 show, according to the Akaike information criterion (AIC), the best model is the 9-class LC model, thus indicating the presence of nine underlying different groups of stocks. Furthermore, the p-values related to the bootstrap likelihood ratio test are always below 0.05 except for the comparison between the models with 9 and 10 latent classes, underlying that the 9-class LC model provides the best fit to the data.

Table 1: Results from LC models estimation with different number of classes: log-likelihood, number of parameters, Akaike information criterion, bootstrap likelihood ratio test and p-value

Model	***LL***	***NPar***	***AIC***	***Bootstrap -2LL Diff***	***p-value***
1-class	-1895.49	8	3806.98	-	-
2-class	-1712.91	29	3483.83	365.16	0.000
3-class	-1659.09	50	3418.17	107.65	0.000
4-class	-1629.33	71	3400.66	59.51	0.004
5-class	-1601.11	92	3386.22	56.44	0.002
6-class	-1570.73	113	3367.46	60.76	0.000
7-class	-1553.58	134	3375.17	45.89	0.038
8-class	-1528.40	155	3366.80	50.37	0.004
9-class	-1503.00	176	3358.00	50.79	0.000
10-class	-1493.99	197	3381.99	23.36	0.119

Following our approach, the 9-class LC model indicates the presence of 9 different stock's typologies. Traditional procedures and subjective beliefs could suggest a lower number of relevant stock's groups, but, since our proposal includes a test for the choice of *K*, we favour a statistical based method.

The results related to the 9-class LC model estimation are shown in Table 2 which illustrates prior probabilities and the conditional means of each indicator. The latent classes are numbered according to their sizes, that is on the basis of prior probabilities

$\hat{\pi}_{X=x|C,I}$ reported on the first row. Class 1 is the modal group and collects the 18.8% of the stocks, while class 9 is the smallest, with only the 3.8% of the stocks. The detail of prior probabilities indicates the presence of some small groups, for instance classes 7, 8, and 9, and some bigger ones, such as classes 1, 2, and 3 which, if cumulated, cluster the 50% of the stocks.

The nine latent classes are ordered in Table 2 according to the conditional means of indicator *M*. The interpretation of the financial characteristics, i.e. the risk-return profile, of each latent class can be obtained on the basis of the conditional means of the indicators reported in the last four rows of the table.

For example, class 9 contains very few stocks ($\hat{\pi}_{X=9|C,I} = 0.038$) and is characterized by the highest conditional mean for the indicator *M*. However, the evaluation of the three indicators related to the risk allows us to define the class 9 as the group characterized by the highest level of risk: in particular, it shows the highest conditional mean value for indicator *S*.

On the other side of Table 2, class 7 is characterized by a strong negative mean return ($\hat{\mu}(M \mid X = 7) = -3.17$) and a quite high level of risk. Classes 4, 3, 2 and 1 are all characterized by negative values of the conditional means for indicator *M*, but their risk levels are quite different. For example, stocks classified into class 4 are particularly volatile during crisis periods ($\hat{\mu}(T \mid X = 4) = 4.69$) and they are affected by large drops in prices ($\hat{\mu}(P \mid X = 4) = -32.5$). On the other hand, despite its negative mean return, class 2 is characterized by a quite low risk: conditional means for indicators *S*, *P*, and *T* are equal to 7.49, –18.1, and 2.83, respectively. Classes 5 and 6 are both characterized by positive mean returns and very low risk levels, especially for class 6.

Finally, class 8 should require a particular attention: despite its slightly positive mean return ($\hat{\mu}(M \mid X = 8) = 0.01$), this group is strongly affected by big price drops as suggested by the highest value for $\hat{\mu}(P \mid X = x)$.

Table 2: Results related to the 9-class LC model estimation: prior probabilities and conditional means for the indicators *M*, *S*, *P* and *T*.

Profile	**Cl 7**	**Cl 4**	**Cl 3**	**Cl 2**	**Cl 1**	**Cl 8**	**Cl 5**	**Cl 6**	**Cl 9**
$\hat{\pi}_{X=x\|C,I}$	0.065	0.138	0.146	0.166	0.188	0.043	0.112	0.105	0.038
$\hat{\mu}(M \mid X = x)$	-3.17	-0.96	-0.30	-0.24	-0.13	0.01	0.28	0.34	1.71
$\hat{\mu}(S \mid X = x)$	12.70	15.17	11.44	7.49	9.17	14.92	10.54	5.74	19.37
$\hat{\mu}(P \mid X = x)$	-30.7	-32.5	-25.9	-18.1	-21.5	-35.6	-21.9	-13.4	-26.1
$\hat{\mu}(T \mid X = x)$	3.66	4.69	3.49	2.83	3.18	3.76	3.50	2.67	3.82

The characterization of the risk-return profiles of the nine groups of stocks facilitates a correct financial evaluation. On one hand, a profitable investment should avoid classes 7, 4, and, probably, also class 8. On the other hand, an attractive portfolio should include stocks classified to classes 5, 6, and, for a higher level of risk, also those belonging to class 9.

4 Latent Class Model Implications for Portfolio Analysis

The results illustrated in Section 3, related to the LC model estimation, lead, first, to the evaluation of the latent risk-return profile of each financial assets and, second, to a new stock classification based on the latent risk-return profile.

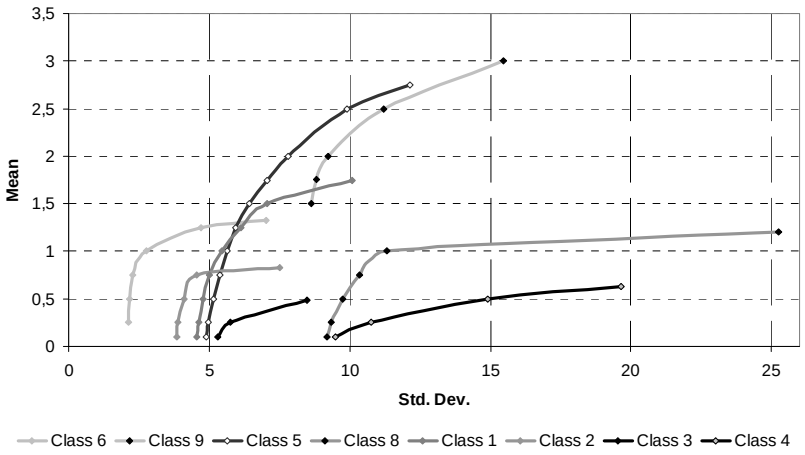
It is possible to further develop the analysis of the latent class model by adding a final step, which allows to create a financial portfolio characterized by an optimal risk-return profile.

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.

In Figure 2 is illustrated, for each latent class described in Table 2, the efficient frontier obtained by using the stocks assigned to each latent class. The comparison of these efficient frontiers shows that the different groups of stocks defined by the LC analysis are heterogeneous one to the other and, for this reason, they are particularly useful for defining effective investment strategies.

On the left side of Figure 2, that is in the lower risk area, it is possible to observe how the best frontiers are achieved by means of class 6, for the lowest mean returns, class 5, for the average mean returns, and class 9, for the highest mean return. Therefore, by jointly using classes 6, 5, and 9, we can propose a set of investment opportunities which are particularly appealing. Furthermore, on the right side of Figure 2, that is in the high risk area, are located the efficient frontiers related to class 4, for lowest mean returns values, and class 8. The efficient frontier related to class 7 is not illustrated in Figure 2 since it leads to negative values of mean return. Finally, it is worth noting how positions of efficient frontiers in Figure 2 are strictly consistent with the latent class characteristics illustrated in Table 2.

Figure 2: Efficient frontiers for the nine latent classes (class 7 is not included since it does not have any positive solutions)



5 Conclusion

The LC analysis allows us to emphasize the relevance of the statistical methodology in financial decision-making processes and, in particular, to introduce a methodological dimension in portfolio analysis.

The stock's classification into homogenous groups under their latent risk-return profile facilitates the choice of a profitable portfolio on the basis of a rigorous statistical procedure. The flexibility of LC models allows to include into the expected return and risk measurement a wide set of information and to overcome the traditional automatic correspondences expected return – mean and risk – standard deviation. Furthermore, this framework can also be used 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 decision.

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

References

1. Akaike, H.: A New Look at the Statistical Model Identification. IEEE Transaction on Automatic Control, **19(6)**, 716-723 (1974)
2. Bartholomew, D.J., Knott, M.: Latent variable models and factor analysis. Kendall's Library of Statistics 7. Oxford University Press, New York (1999)
3. Lazarsfeld, P.F., Henry, N.W.: Latent structure analysis. Houghton Mill, Boston (1968)
4. Lisi, F., Otranto, E.: Clustering mutual funds by return and risk levels. Working Paper CRENoS, **200813**, Centre for North South Economic Research, University of Cagliari and Sassari, Sardinia (2008)
5. Magidson, J., Vermunt, J.K.: Latent class factor and cluster models, bi-plots and related graphics displays. Sociological Methodology, **31**, 223-264 (2001)
6. Moustaki, I., Papageorgiu, I.: Latent class models for mixed variables with applications in archaeometry. Computational Statistics and Data Analysis, **48**, 659-675 (2005)