Statistica Applicata Vol. 9, n. 2, 1997

MULTIVARIATE ANALYSIS OF FINANCIAL DATA (*)

Giuseppe Cavaliere, Michele Costa

Dipartimento di Scienze Statistiche, Università degli Studi di Bologna.

Abstract

Financial data are strongly characterised by large movements which affect simultaneously most assets. Although the fundamental sources of these variations are unobservable they can be estimated by using factor analysis. In this paper we show that the approximate factor model represents a more efficient solution for the multivariate analysis of financial data. Moreover the approximate factor model explains the relation between unobservable factors, interpreted as sources of non diversifiable risk, and the financial assets. An empirical example is given by using daily returns from the Milan Stock Exchange in the nineties.

Keywords: Financial data, factor analysis, approximate factor model.

1. INTRODUCTION

The debate on the exploratory analysis of financial data, although it has ancient roots, is still greatly topical and involves a large number of researchers. Within statistical methods based on latent variables, factor analysis represents the main reference and is used for the empirical test of the most important financial market models. In detail the comparison between single-factor models, such as the CAPM by Sharpe (1964) and Lintner (1965), and multifactor models, such as the APT by Ross (1976), has aroused great interest [Brown (1989), Bray (1994)]. In the case of financial data, however, one of the strongest hypotheses of the factor model, that is the independence of error terms, is totally inadequate in measuring empirical reality. In this study, following a debate started by Chamberlain and Rothschild in 1983, financial data are analysed by resorting to an approximate factor model in

(*) *Paper presented at the "Giornate di analisi dei dati multidimensionali", Napoli, 30–31 ottobre 1995. Financial support provided by 60% and 40% MURST grants is gratefully acknowledged. For official purpose only we specify that par. 2.1, 3.2 and 4 are by G. Cavaliere and par. 1, 2.2, 3.1 and 3.3 are by M. Costa.*

which the hypothesis of independence of error terms is not included. In this case, classical inference procedures are no longer valid and it is necessary to specify appropriate tests.

2. FACTOR MODELS FOR FINANCIAL DATA

The objective of determining the fundamental sources of (non diversifiable) risk is central to the theory of financial markets. These few but relevant risky elements are generally not immediately measurable because they are unobservable. Factor analysis represents the natural statistical solution for this problem.

2.1. THE FACTOR MODEL

In the standard representation of the factor model

$$\mathbf{x}_t = \mu +_k \mathbf{B}\mathbf{f}_t + \varepsilon_t$$

where $\mathbf{x}_t$ is the p x 1 vector of the observable variables at time t, μ is the p x 1 mean vector, ${}_k\mathbf{B}$ is the p x k factor loading matrix, $\mathbf{f}_t$ is the k x 1 vector of k orthogonal common latent factors at time t and ε_t is the p x 1 vector of the specific factors, independent upon the common factors, we usually refer to a diagonal variance-covariance matrix of ε_t, i.e. $E(\varepsilon_t\varepsilon_t') =_k \Psi$. Thus, the variance–covariance matrix Σ of the observable variables $\mathbf{x}_t$ can be expressed as

$$\Sigma =_k \mathbf{B}_k\mathbf{B}' +_k \Psi.$$

Within the maximum likelihood estimation of the factor model, parameters are obtained, under the hypothesis of joint normality of $\mathbf{f}$ and ε, by maximising the function

$$l(k) = -\frac{1}{2}pT\log(2\pi) - \frac{1}{2}T\log\left({}_k\hat{\mathbf{B}}_k\hat{\mathbf{B}}' +_k \hat{\Psi}\right) - \frac{1}{2}\mathrm{tr}\left(\mathbf{S}\left({}_k\hat{\mathbf{B}}_k\hat{\mathbf{B}}' +_k \hat{\Psi}\right)^{-1}\right)$$

where ${}_k\hat{\mathbf{B}}$ and ${}_k\hat{\Psi}$ indicate, respectively, the estimates of ${}_k\mathbf{B}$ and ${}_k\mathbf{\Psi}$ under the hypothesis of a k factor model, $\mathbf{S}$ is the sample variance–covariance matrix and T is the number of observations. The main methods for determining of the number of factors are obtained by means of such a procedure.

In particular, the likelihood ratio test is based on the statistics

$$LR(k) = T^*\left(\log\left|{}_k\hat{\mathbf{B}}_k'\hat{\mathbf{B}}' +_k \hat{\Psi}\right| - \log|\mathbf{S}|\right),$$

where $T^* = T-(2p+4k+11)/6$ is the adjusted number of observations. Under the null hypothesis, $LR(k)$ is approximately distributed as a χ^2 with $[(p-k)^2-p-k]/2$ degrees of freedom.

Among information criteria, Akaike (1987) suggests choosing the value of k which minimises the quantity

$$AIC(k) = -2 \max l(k) + 2h$$

where $h = p(k + 1) - k(k - 1)/2$ is the number of parameters in k factor models; alternatively, Schwarz (1978) proposes a more parsimonious criterion where the choice of the factor dimension takes place by minimising

$$SIC(k) = -\max l(k) + 0.5\, h \log T$$

In financial markets a common factor is a factor common to all financial activities, while in factor analysis we mean common factor any element of **f**. Therefore, there are two sets of common factors, one relating to financial markets and the other relating to factor analysis. However, a relation exists between these two sets: financial market common factors are a subset of common factors of factor analysis.

The difficulties in the selection of the factor dimension met by the classical methods can be ascribed to this difference.

A solution can be found if we isolate, among the common factors of the factor analysis, only those which influence all observable variables. From a methodological point of view, this corresponds to a generalisation of the factor model which allows a more general structure for the variance–covariance matrix of the error terms.

2.2. THE APPROXIMATE FACTOR MODEL

An important generalisation of the factor model (Chamberlain and Rothschild, 1983) suggests weakening the hypothesis of diagonality of the variance–covariance matrix of error terms by resorting to a matrix not necessarily diagonal. Therefore it is assumed that

$$E(\varepsilon_t \varepsilon_t') = \mathbf{G}$$

is not necessarily diagonal but such that $\lim_{p \to \infty} {}_G\lambda_{max} < +\infty$, where ${}_G\lambda_{max}$ is the largest eigenvalue of **G**. Moreover, supposing that the matrix **BB'** has exactly k unbounded eigenvalues for $p \to \infty$ the definition of approximate k–factor model is obtained.

In the above situation, the classical methods for determining the number of factors cannot be applied. In order to get information about the factor structure, a number of authors [Brown (1989), Luedecke (1984), Trzcinka (1986)] suggest analysing the behaviour of the eigenvalues of matrix **S** when the number of assets included in the model diverges. Connor and Korajczyk (1993) show that, when p increases, all the sample eigenvalues grow linearly with p, although the first k grow at a faster rate. Thus, by investigating the eigenvalue sequences it is possible to get some information about k by simply examining the trend of the eigenvalues as p tends to infinity.

A further investigation of the problem of determining the number of pervasive factors, within the framework of the approximate factor model, is suggested by Connor and Korajczyk (1993) who propose a test based on the principal components of S, whose asymptotic properties are shown by Chamberlain and Rothschild (1983) and by Connor and Korajczyk (1986, 1988).

The test starts from the intuition that, if k is the true number of factors, the inclusion of further factors does not significantly reduce the average cross-sectional residual variance.

By using Connor and Korajczyk's procedure it is possible to distinguish factors which are common to all assets from factors which influence only particular subsets. In comparison with standard maximum likelihood factor analysis, which cannot discriminate between these two categories of latent factors, there is an indubitable improvement.

3. ANALYSIS OF FINANCIAL DATA

The analysis of financial data performed in this paper is carried out on the daily returns of 245 securities, quoted at Milan Stock Exchange between January 1990 and December 1994. Returns are computed as price percentage variation, ignoring dividends. During the analysis, returns are expressed as deviations from their time average. Following the structure of the previous section the data analysis is also divided into two parts: at the beginning data are analysed by means of the classical methodology of maximum likelihood factor analysis and afterwards the approximate factor model is introduced. Finally, the relation between latent factors and stock returns is investigated.

3.1. DATA ANALYSED BY FACTOR ANALYSIS

In order to study the influence of the number of observed variables on the results of factor analysis, four nested sets, which consist of 20, 50, 100 and 245 assets respectively, are considered. The likelihood ratio test (LR), Akaike's information criterion (AIC) and Schwarz's information criterion (SIC) are computed on the basis of maximum likelihood estimates. The values taken by the Schwarz's information criterion are multiplied by 2 in order to allow a better comparison with Akaike's information criterion. With $p = 20$ and $p = 245$, the results are reported in Tab. I.

In the sample formed by the first 20 assets, the likelihood ratio test indicates the presence of 5 factors, Akaike's information criterion suggests 6 and Schwarz's

3. By increasing the number of assets, a considerable increase in the number of factors suggested by the different methods is observed. With $p \geq 50$, the likelihood ratio test and Akaike's information criterion indicate at least 10 factors. Schwarz's criterion suggests 5 factors when $p = 50$, 7 when $p = 100$ and 8 when $p = 245$.

Tab. I: The number of factors in the factor model.

	AIC		SIC		LR	p(LR)		
	$p = 20$	$p = 245$	$p = 20$	$p = 245$	$p = 20$		$p = 245$	
$k = 1$	57380	729703	57585	732216	1337	0.00	64097	0.00
$k = 2$	56663	726867	56965	730631	586	0.00	60962	0.00
$k = 3$	56338	723829	56733	728839	229	0.00	57645	0.00
$k = 4$	56312	721749	56794	728000	169	0.00	55225	0.00
$k = 5$	56285	719715	56849	727201	111	0.20	52852	0.00
$k = 6$	56278	718229	56919	726946	74	0.78	50994	0.00
$k = 7$	56289	716662	57002	726605	58	0.85	49065	0.00
$k = 8$	56301	715277	57081	726440	44	0.90	47308	0.00
$k = 9$	56315	714157	57156	726536	34	0.90	45802	0.00
$k = 10$	56325	713284	57222	726873	23	0.94	44528	0.00

The analysis of financial data carried out through classical factor analysis methods gives raise to a number of factors which varies according to the method used. Thus, even in the case of financial data, the feature of Akaike's criterion to select more factors than the likelihood ratio test and the propensity of Schwarz's criterion to more parsimonious evaluations are confirmed. Moreover, the number of latent factors is positively related to the number of assets analysed.

From these well–known remarks it is possible to work out a general indication about the stability of the latent factor structure underlying the markets. Financial models, in fact, assert the existence of a stable factor structure, characterised by k latent factors and always valid for all financial assets on the market. On the other hand factor analysis does not seem to verify this hypothesis and suggests a variety of k. It follows that either the stability hypothesis is false or factor analysis does not perform an adequate test.

3.2. DATA ANALYSED BY THE APPROXIMATE FACTOR MODEL

A possible solution can be found by referring to the approximate factor model which, by weakening the hypothesis of diagonality of the variance-covariance matrix of error terms, enables to distinguish between latent factors, which are common to all financial activities, and latent factors only related to particular subsets.

The stability hypothesis appears to be rejected only when the number k of latent factors, which are common to all financial assets, changes by varying the selection method or other conditions, as the sample size p or the number T of observations. If k does not suffer any modification the stability hypothesis is maintained, as suggested by the financial markets theory.

The analysis of the eigenvalues of the correlation matrix of observed variables gives some information about the stability of the factor structure and the number of factors which influence the whole assets market. Figure 1*a* shows the behaviour of the first 6 sample eigenvalues, when the number of assets varies from 20 to 245, while in figure 1*b* eigenvalues from the second to the sixth are reported.

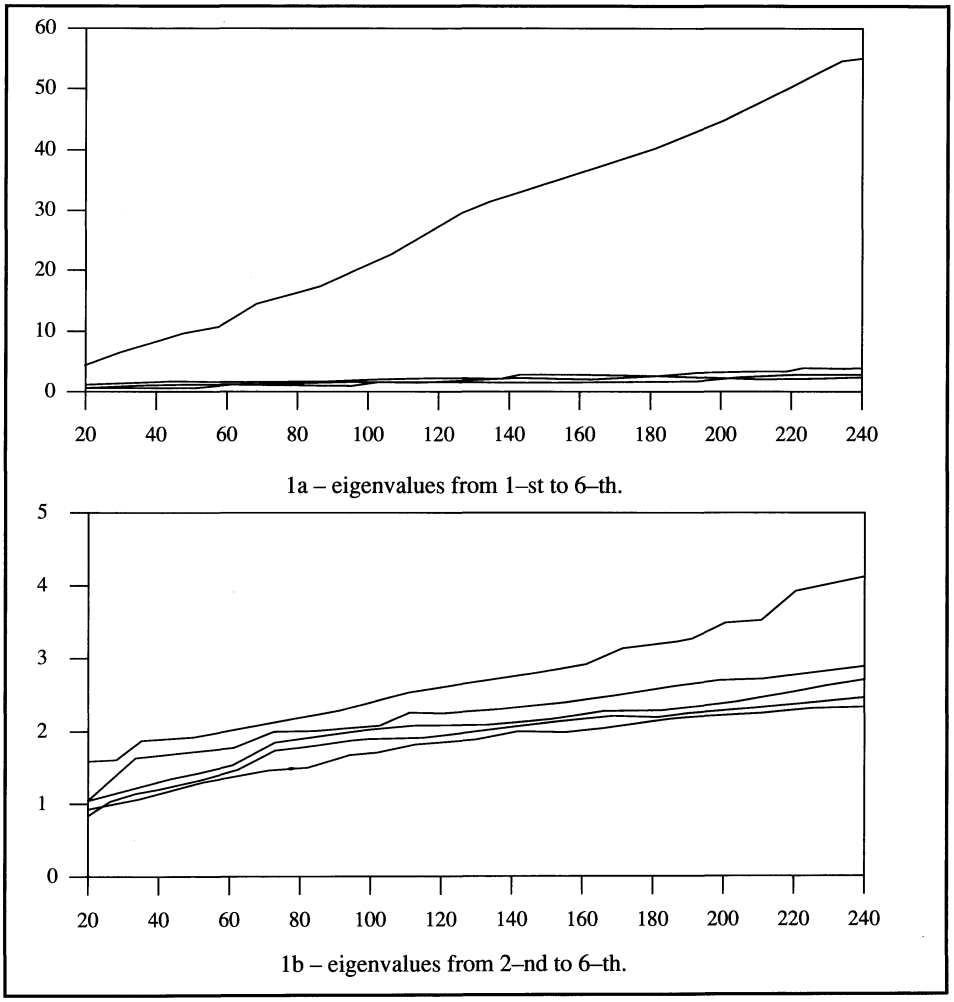

Fig. 1: Eigenvalues of the correlation matrix of the observed variables (p = 20,...,245).

From Fig. 1*a* it arises that all the sample eigenvalues grow as p increases; however, the first dominates the others, both in magnitude and in growth rate. Fig. 1*b* shows the different dynamics between the second eigenvalue and the following ones.

Moreover, eigenvalues analysis is completed by dividing the data in Fig. 1 by the number p of assets. Fig. 2 shows the results.

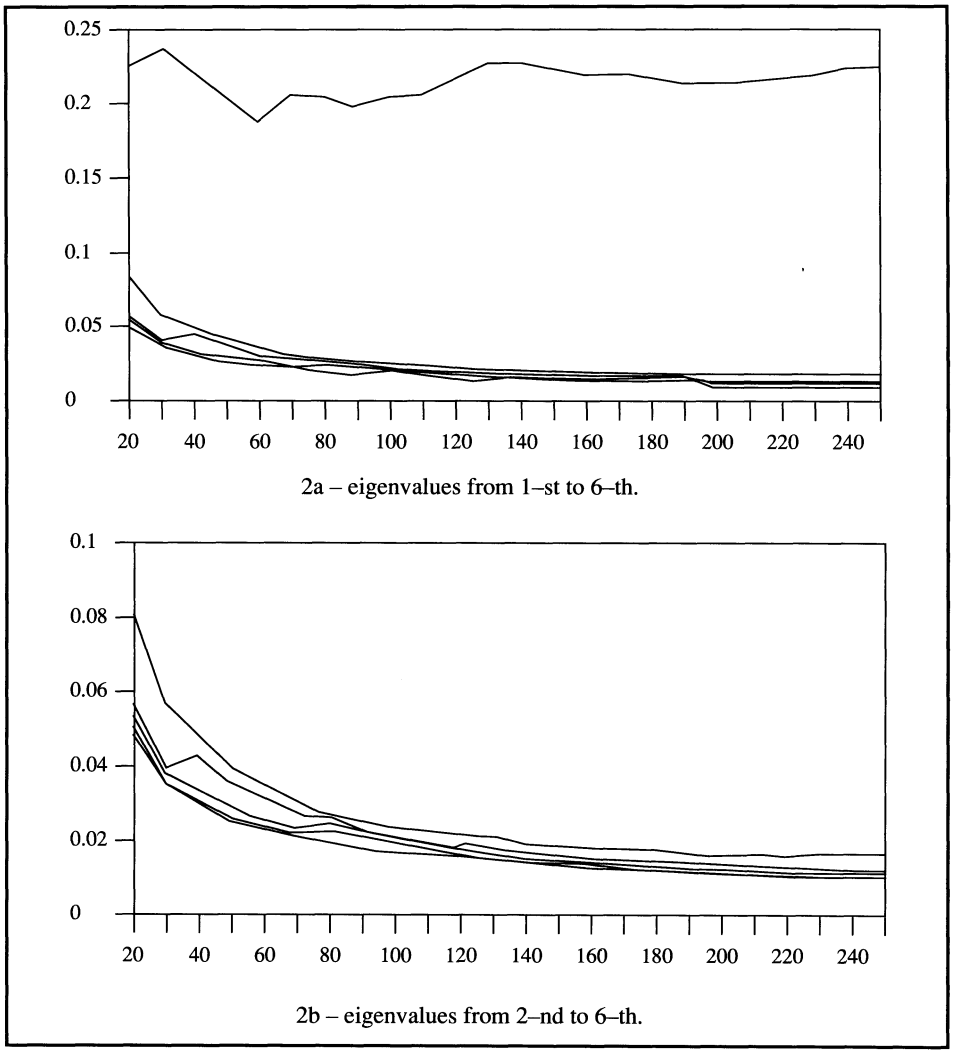

Fig. 2: Eigenvalues divided by the number of assets (p = 20,...,245).

The first eigenvalue (Fig. 2*a*) shows a larger growth rate than all the rest and the second (Fig. 2*b*) keeps on differing from the group of successive eigenvalues.

Thus, eigenvalues analysis seems to indicate one or at most two latent factors. This conclusion does not agree with the results obtained through factor analysis, which leads to a clearly grater estimation of k. Finally, the mere investigation of these results does not allow an effective interpretation of the dynamics of the second larger eigenvalue and the application of more powerful methods is required.

A good solution is provided by Connor and Korajczyk's test, which allows the valuation of the residual variance after the extraction of k pervasive factors. Fig. 3 shows the average of the residual variance (measured as a percentage of the total variance) for different values of p. The upper line corresponds to a single-factor model, while the lower line refers to the five–factor model.

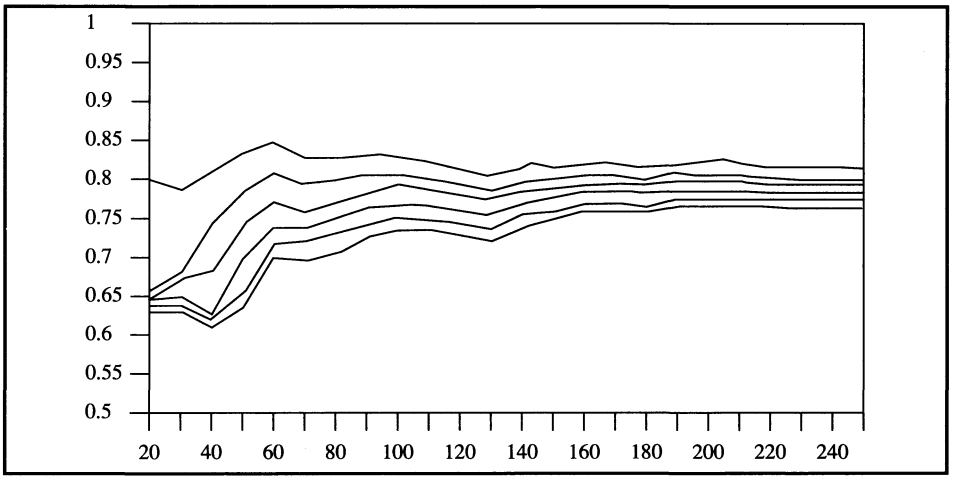

Fig. 3: Average residual variance.

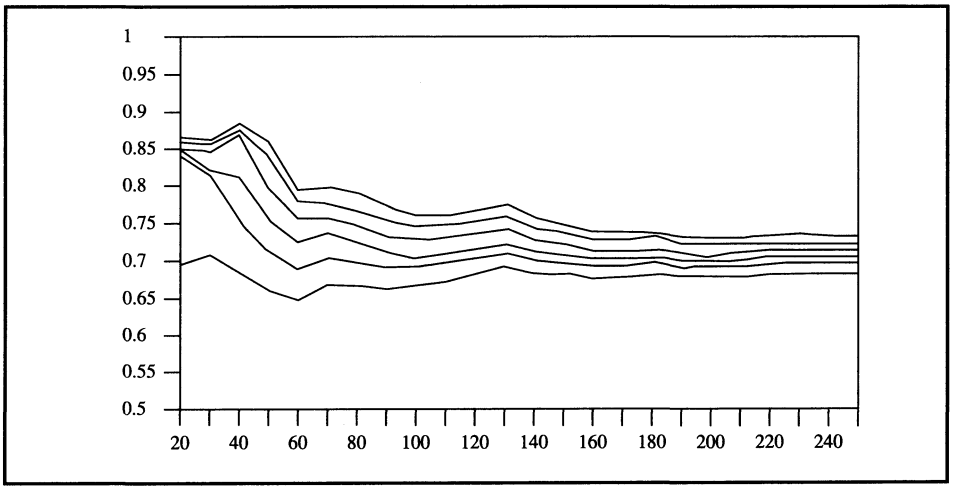

Fig. 4: Average explained variance.

Fig. 4 shows the average of the explained variance (measured as a percentage of the total variance) for different values of p. The upper line corresponds to a five-factor model, while the lower line refers to the single–factor model.

It is interesting to notice that, as p becomes larger, the differences between the models decrease. When p increases, the explanatory capacity of the first factor does not change, while the second factor is characterised by a decreasing explanatory power. Tab. II and Tab. III illustrate in detail the variations of the average residual variance for different values of p and k.

Tab. II: Average residual variance varying p and k.

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
$p=20$	4.93	3.99	3.32	3.23	3.19	3.21	3.18	3.14	3.12	3.11
$p=50$	5.72	4.75	4.46	4.20	3.95	3.68	3.60	3.53	3.48	3.41
$p=100$	5.94	4.93	4.79	4.73	4.65	4.52	4.41	4.31	4.23	4.13
$p=245$	5.75	4.68	4.60	4.56	4.50	4.46	4.42	4.37	4.33	4.29

Tab. III: Decrease of average residual variance from k–1 to k factors.

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
$p=20$	–	19.07	16.79	2.71	1.24	– 0.63	0.93	1.26	0.64	0.32
$p=50$	–	16.96	6.11	5.83	5.95	6.84	2.17	1.94	1.42	2.01
$p=100$	–	17.00	2.84	1.25	1.69	2.80	2.43	2.27	1.86	2.36
$p=245$	–	18.61	1.71	0.87	1.32	0.89	0.90	1.13	0.92	0.92

Going from zero to one factor, a considerable reduction of the residual variance (about 19%) is obtained, while a smaller reduction is obtained if more factors are added.The importance given to the residual variance is justified because Connor and Korajczyk's test is based on this measure. In fact, the null hypothesis tested by this procedure is that there is no significant decrease in the residual variance when $k + 1$ factors instead of k factors are considered. Tab. IV shows for different values of p and k the values of Connor and Korajczyk's test statistics, which is asymptotically standard normally distributed under H_0.

Tab. IV: Connor and Korajczyk's test for the null hypothesis of k factors against the alternative of k+1 factors.

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
$p=20$	1.92	1.65	1.22	1.06	0.74	1.07	0.81	0.41	0.41	0.71
$p=50$	1.97	0.85	0.98	0.89	1.47	1.14	1.14	1.08	1.59	1.41
$p=100$	2.58	0.16	0.14	0.37	0.58	0.51	0.59	0.64	0.53	0.31
$p=245$	5.17	1.29	1.30	1.54	1.63	1.82	1.91	1.97	1.97	1.91

Using a 5 percent significance level only the first factor seems to be significant and therefore the factor structure latent the Italian Stock Market seems to be characterised by a single factor. If other factors exist, their effects are related only to small subsets of the whole economy and they are absorbed by the residual variance–covariance matrix **G**.

3.3. LATENT FACTORS AND STOCK RETURNS

To complete the analysis of financial data the robustness of the factor structure is examined by observing, for particular subsets of assets, the dynamics of non zero elements of the residual variance–covariance matrix by varying the number of latent factors. Moreover, the first 20, 50 and 100 assets are analysed.

The first factor seems to be a proxy of the market portfolio. It is a well-diversified linear combination of all assets and its effect is proved on most assets. The introduction of the second factor causes a reduction in the number of non zero elements of **G**, corresponding to the residual covariances between different assets. However, this reduction follows a very regular and precise scheme: the introduction of the factors behind the first does not cause a general and undifferentiated decrease of non zero elements of **G**, but only a cancellation of the covariances related to particularly small (about 2–3) subsets of assets. This characterisation of the link between latent factors and financial assets is a result of great relevance. Only in the context of an approximate factor model it is possible to obtain such a purpose. Tab. V indicates the groups of assets related to the different factors.

Tab. V: Relation between latent factors and stock returns.

factor	*stock returns*		
	$p = 20$	$p = 50$	$p = 100$
I	market portfolio proxy	market portfolio proxy	market portfolio proxy
II	6, 7, 8	33, 34, 35	92, 93, 94, 12, 23, 25
	national airways company	banking	Fiat, insurance, banking
III	9, 10	6, 7, 8	67, 68, 72
	insurance	national airways company	banking
IV	4, 5	9, 10	33, 34, 35
	building	insurance	banking
V	11, 12	45, 46	61, 62
	insurance, mechanics	chemical	financial
VI	**G** is diagonal	49, 50	6, 7, 8
		building	national airways company
VII		25, 30, 31	9, 10
		banking	insurance
VIII		41, 42	89, 90
		paper	financial
IX		12, 38, 39	45, 46, 60
		insurance, fiancial	chemical, cement
X		–	–

4. CONCLUSIONS

On the basis of eigenvalues analysis and Connor and Korajczyk's test, the existence of no more than two latent factors common to all assets is pointed out. Besides, a number of different factors, each of which only ascribable to specific groups of assets, can be identified. Some classical methods of selection of the factor model, such as the likelihood ratio test and Akaike's and Schwarz's information criteria, point out a number of factors significantly higher and greatly varying according to the number of assets included in the model. The following table shows the number of factors indicated by some methods according to different values of p.

Tab. VI: Number of factors determined by some methods.

	$p = 20$	$p = 50$	$p = 100$	$p = 245$
Likelihood Ratio test	5	≥ 10	≥ 10	≥ 10
Akaike	6	≥ 10	≥ 10	≥ 10
Schwarz	3	5	7	8
Eigenvalues analysis	–	–	–	1 – 2
Connor–Korajczyk	1	1	1	1

A first result allows to state that financial data impose a particular block–diagonal structure on the variance–covariance matrix of error terms of the factor model. The modified specification of the factor model, where $E(\varepsilon_t \varepsilon'_t)$ admits non–zero elements out of its diagonal, then represents a more efficient solution compared to the classical factor model. Moreover, the classical methods of selection of the factor model, adding specific factors relative only to distinctive subsets to latent factors common to all assets, overestimate the dimension of the factor structure of stock returns. Finally, it is possible to establish a relation between these specific factors and the elements of the variance-covariance matrix of error terms.

REFERENCES

AKAIKE H. (1987), Factor analysis and AIC, *Psychometrika, 52,* 317–332.

BRAY M. (1994), The APT is not robust: factor structures and factor pricing, *London School of Economics Discussion Paper,* n. 179.

BROWN S.J. (1989), The number of factors in security returns, *Journal of Finance,* 44, 1247–1262.

CHAMBERLAIN G. and ROTHSCHILD M. (1983), Arbitrage and mean variance analysis on large assets markets, *Econometrica,* 51, 1281–1304.

CONNOR G. and KORAJCZYK R.A. (1986), Performance measurement with the arbitrage pricing theory: a new framework for analysis, *Journal of Financial Economics,* 15, 373–394.

CONNOR G. and KORAJCZYK R.A. (1988), Risk and return in an equilibrium APT: Application of a new test methodology, *Journal of Financial Economics*, 21, 255–289.

CONNOR G. and KORAJCZYK R.A. (1993), A test for the number of factors in an approximate factor model, *Journal of Finance*, 48, 1263–1291.

LITNER J. (1965), Security prices, risk and maximal gains from diversification, *Journal of Finance*, 20, 587–616.

LUEDECKE B.P. (1984), An empirical investigation into arbitrage and large asset markets, Ph.D. dissertation, University of Wisconsin.

PANETTA F. (1992), La struttura fattoriale del mercato azionario italiano, *Ricerche applicate e modelli per la politica economica*, Banca d'Italia, I, 391–420.

ROSS S.A. (1976), The arbitrage theory of capital asset pricing, *Journal of Economic Theory*, 13, 341–360.

SHARPE W.F. (1964), Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance*, 19, 425–442.

SCHWARZ G. (1978), Estimating the dimension of a model, *Annals of Statistics*, 6, 461–464.

TRZCINKA C. (1986), On the number of factors in the arbitrage pricing theory, *Journal of Finance*, 41, 347–368.

L'ANALISI MULTIVARIATA DI DATI FINANZIARI

Riassunto

I dati finanziari sono fortemente caratterizzati da ampie fluttuazioni che coinvolgono contemporanemente molteplici attività. Sebbene le componenti fondamentali di tali variazioni non siano direttamente osservabili, tuttavia possono essere stimate facendo ricorso ai metodi dell'analisi fattoriale. In questo lavoro si dimostra come il modello fattoriale approssimato rappresenti una soluzione ancora più efficiente per l'analisi multivariata dei dati finanziari. Inoltre, il modello fattoriale approssimato è in grado di spiegare la relazione tra i fattori non osservabili, interpretabili come sorgenti di rischio non diversificabile, e le attività finanziarie. L'analisi di dati finanziari viene, infine, effettuata con riferimento ai rendimenti azionari giornalieri della Borsa valori di Milano negli anni novanta.