5

Michele Costa

The determination of the number of factors in a factor model.

Serie Ricerche n.8

Dipartimento di Scienze Statistiche "Paolo Fortunati"
Università degli studi di Bologna
1992

* Support for this research was provided by MURST grant 40 % 1990-91: "Modelli a variabili latenti: problemi di costruzione e trattamento di dati di panel e test di specificazione negli schemi di formazione dei prezzi delle attività finanziarie".
I'm grateful to Sandra Marciatori for computing assistance.

Finito di stampare nel mese di Gennaio 1993
presso le Officine Grafiche Tecnoprint
Via del Legatore 3, Bologna

1 - Introduction

Factor analysis is closely related to unobservability problems, and especially to the problem of variables that "do not correspond directly to anything that is likely to be measured" (Griliches, 1977). Indeed the factor analysis model specifies a set of linear relations in which p observable variables are determined by k unobservable factors and p error terms.

The determination of the "true" number of factors is the first problem to be solved in the selection of the "true" factor model

$$X = f\Lambda + U$$

where
$X_{N\times p}$ is the matrix of the observable data;
$U_{N\times p}$ is the matrix of the errors;
$\Lambda_{k\times p}$ is the matrix of the factor loadings;
$f_{N\times k}$ is the matrix, with $k < p$, of the factors;
N is the number of observations of the series used.

The identification of a stable factor structure is traditionally done by means of the likelihood ratio and, more recently, through other methods, as information criteria and cross-validation.

The purpose of this paper is to study the contribute of likelihood ratio, of some information criteria and of cross-validation to the determination of the "true" number of factors by a simulation.

One of the most relevant results obtained is that all the methods considered show a tendency to underestimate the "true" k. Akaike's information criterion and cross-validation seem to identify the factor structure more accurately than Hannan-Quinn and Schwarz's criteria.

2 - The likelihood ratio

The possibility to test the number of factors is one of the principle reasons for the success of the procedure of maximum likelihood (Lawley and Maxwell 1971; Kim and Mueller, 1983).

With uncorrelated factors (Anderson, 1984) the data covariance matrix Σ can be expressed as

$$\Sigma = \Lambda'\Lambda + \Psi$$

where Ψ is the diagonal variance matrix $p \times p$ of the errors terms U: $E(U'U) = \Psi$.

If the data are normally distributed, the log likelihood of the model is

$$\log L = -\tfrac{1}{2}Np\log(2\pi) - \tfrac{1}{2}N\log|\Lambda'\Lambda + \Psi| - \tfrac{1}{2}Np$$

The null hypothesis used to test the number of factors is

$$H_0\text{:} \quad \Sigma_p = \Lambda_k'\Lambda_k + \Psi_k$$

and corresponds to

$$H_0\text{:} \quad k \textit{ factors are sufficient (or minus)}$$

against

$$H_1\text{:} \quad p \textit{ factors are needed.}$$

where Σ_p indicates the covariance matrix of the p observable variables X, Λ_k and Ψ_k are, respectively, the factor loadings matrix and the errors terms variance matrix of a k factor model.

The usual statistic used to test the number of factors is the log likelihood ratio

$$LR = N^*(\log|\Lambda_k'\Lambda_k + \Psi_k| - \log|\Sigma_p|)$$

where the number of observations is corrected by Bartlett's formula $N^* = N - (2p + 4k + 11)/6$.

The log likelihood ratio is distributed as χ^2 with $((p-k)^2 - p - k)/2$ degrees of freedom.

Conway and Reinganum (1988) indicate cross-validation as an alternative solution for the determination of the number of factors. Cross-validation can be considered as a two stage procedure. In the first stage maximum likelihood estimates of the parameters are calculated in a sample of p variables X. In the second stage the estimates obtained are not compared with the respective sample variance matrix Σ, but with a Σ^* of another sample of p variables X, in order to isolate the stable factor structure from the random components.

The log likelihood ratio

$$LR = N^*(\log|\Lambda'\Lambda + \Psi| - \log|\Sigma|)$$

is therefore modified into

$$CV = N^*(\log|\Lambda'\Lambda + \Psi| - \log|\Sigma^*|) + N^*(tr((\Lambda'\Lambda + \Psi)^{-1}\Sigma^*) - p)$$

From the way it is derived, cross-validation statistic can be interpreted as an out of sample-likelihood ratio.

3 - Information criteria

Akaike's information criterion is probably the most relevant and famous as for the comparison and selection between different models and is constructed on log likelihood

$$AIC = -2\log\max L + 2h$$

where h is the number of the model's free parameters.

The factor loadings matrix Λ contains kp parameters to be estimated, while the diagonal variance covariance matrix Ψ has p non zero elements: therefore the sample variance covariance matrix Σ contains $p(k+1)$ parameters to be estimated. Moreover, in order to guarantee the identification of the model the following condition is required

$$\Gamma = \Lambda'\Psi^{-1}\Lambda \qquad \Gamma \;\; \textit{diagonal}$$

which gives $k(k-1)/2$ additional constraints. The number of free parameters in an orthogonal factor model is thus

$$p(k+1) - \tfrac{1}{2}k(k-1)$$

and the form of AIC is

$$AIC(k) = \{Np\log(2\pi) + N\log|\Lambda\Lambda' + \Psi| + Np\} + \{2(p(k+1) - k(k-1)/2)\}$$

The first term can be interpreted as a goodness-of-fit measure, while the second gives a growing penalty to increasing numbers of parameters, according to the parsimony principle.

In the choice of the model a minimisation rule is used to select the model with the minimum Akaike information criterion value.

Following the modification of FPE (Final Prediction Error) proposed by Bhansali and Downham (1977), in 1980 Smith and Spiegelhalter suggested to modify the AIC by transforming the second term into a generic αh:

$$AIC = -2\log\max L + \alpha h$$

Still in the context of likelihood based procedures, in 1978 Schwarz proposed the alternative information criterion

$$SCH = -\log\max L + \tfrac{1}{2}h\log N$$

that, unlike AIC, considers the length N of the time series and is therefore less favourable to factors inclusion.

In 1979 Hannan and Quinn suggested another information criterion, based, as the precedent, on the minimisation of $-\log\max L + hC$

$$HQ = -2\log\max L + 2h\,c\,\log\log N \qquad c > 1$$

4 - A simulation

The purpose of this paper is to illustrate some results obtained on simulated data, for which the factor structure is perfectly known. The different methods, illustrated in the previous paragraphs, are applied to the simulated data and the indications of the number of factors are compared with the true value k, which is a priori known.

The following model is used to obtain the new simulated variables X^*

$$X^* = f^*\Lambda^* + U^*$$

where

f^* is the $N \times k$ matrix of the new factors, obtained by random extraction from a standardized normal distribution

U^* is the $N \times p$ matrix of error terms, randomly extracted from a standardized normal distribution too

Λ^* is the $k \times p$ matrix of factor loadings, obtained from a factor analysis of a sample of p assets returns randomly extracted from a set of 100 assets returns daily quoted at Milan stock exchange from 1986 to 1989.

The various methods illustrated above are thus applied to samples of p=20, p=30, p=40 simulated variables X^* to analyze the influence of variations in the number of original variables.

For each value of p different lengths N of the time series analyzed were considered, in order to study how variations of N can influence the number of factors detected. Specifically the cases N=100, N=200, N=1000, N=5000 were considered.

Finally, in the simulations three different factor structures were analyzed in order to evaluate the chosen criteria for different values of k, specifically the cases k=1, k=5, k=10.

In order to generate the k independent factors f^* a matrix of dimension 5000 x 10, corresponding to the maximum value of N and k, was randomly extracted from a standardized normal distribution. For other values of N and k appropriate submatrices were extracted from this matrix: for example, for the case of k=1 and N=200 the relative submatrix $f^*_{200\times 1}$ contains the first 200 rows of the first column of the $f^*_{5000\times 10}$.

To obtain the factor loadings Λ^* three samples of 20 assets returns, three samples of 30 and three of 40 were randomly and independently extracted from a set of 100. For each sample a factor analysis was performed, three times with k=1 to obtain $\Lambda^*_{20\times 1}, \Lambda^*_{30\times 1}, \Lambda^*_{40\times 1}$, three times with k=5 to obtain $\Lambda^*_{20\times 5}, \Lambda^*_{30\times 5}, \Lambda^*_{40\times 5}$ and the last three times with k=10 to obtain $\Lambda^*_{20\times 10}, \Lambda^*_{30\times 10}, \Lambda^*_{40\times 10}$.

The factor loadings Λ^* and the factors f^* are assumed as fixed. Having thus obtained the term $f^*\Lambda^*$, the p simulated variables X^* are obtained by p random extractions of the error terms vector U^*.

The factor structure is so a priori known as k are the columns of Λ., and the variability of the X^* is entirely attributable to the different determinations of the vector U^*: it's also possible to compare the indications given by the different criteria with the true and known k.

Summarizing, for k=1 three matrix Λ^{*} were randomly and independently calculated, one of dimension 1 x 20 from a sample of 20 assets for the case p=20, one of dimension 1 x 30 from a sample of 30 assets for the case p=30 and the last of dimension 1 x 40 from a sample of 40 assets for the case p=40.

The ensuing three models are the following:

$$\begin{aligned} p=20 \qquad & X^{*}_{N\times 20} = f^{*}_{N\times 1}\,\Lambda^{*}_{1\times 20} + U^{*}_{N\times 20} \\ p=30 \qquad & X^{*}_{N\times 30} = f^{*}_{N\times 1}\,\Lambda^{*}_{1\times 30} + U^{*}_{N\times 30} \\ p=40 \qquad & X^{*}_{N\times 40} = f^{*}_{N\times 1}\,\Lambda^{*}_{1\times 40} + U^{*}_{N\times 40} \end{aligned}$$

For each model 100 extractions of U^{*} are considered, thus obtaining 100 samples of simulated variables X^{*}, for each value of N.

Therefore for k=1, p=20 and N=200, 100 samples of 20 variables X^{*} are considered and so for each combinations of k, p and N.
The same as for k=1 is repeated for k=5 and for k=10.

The root mean squared error (RMSE) was calculated to compare the different methods

$$S = \left(\frac{1}{100}\sum_{i=1}^{100}(k_i^{*}-k)^2\right)^{\frac{1}{2}}$$

k^{*} is the number of factors indicated by the generic method, k is the true number of factors underlying the simulated variables X^{*} and i indicates the generic i-th sample. Obviously S is calculated for each method and in general method A is better than method B if $S_A < S_B$, as S measures the distance between the true k and the empirical k^{*} and so the smaller S the better approximation of k one obtains through k^{*}.

The bias

$$D = \frac{\sum_{i=1}^{100} k_i^{*}}{100} - k$$

indicates the direction of the RMSE and is negative when the method underestimates the true number of factors and positive when k is overestimated. The bias is calculated in order to complete the informations about the distribution of the k^{*} around k, indeed the RMSE indicates only the distance between k^{*} and k; information on the sign of this distance are given by the bias.

The results obtained by transforming the AIC in:

$$AIC3 = -2\log \max L + 3h$$

and

$$AIC4 = -2\log \max L + 4h$$

are not particularly brilliant, because AIC3 and AIC4 generally converge to the true value k more slowly than AIC.

The following tables show S_{AIC},S_{AIC3}, S_{AIC4}, D_{AIC}, D_{AIC3} and D_{AIC4} for the different cases considered.

In this and the next tables the values below 0,05 are set to 0.

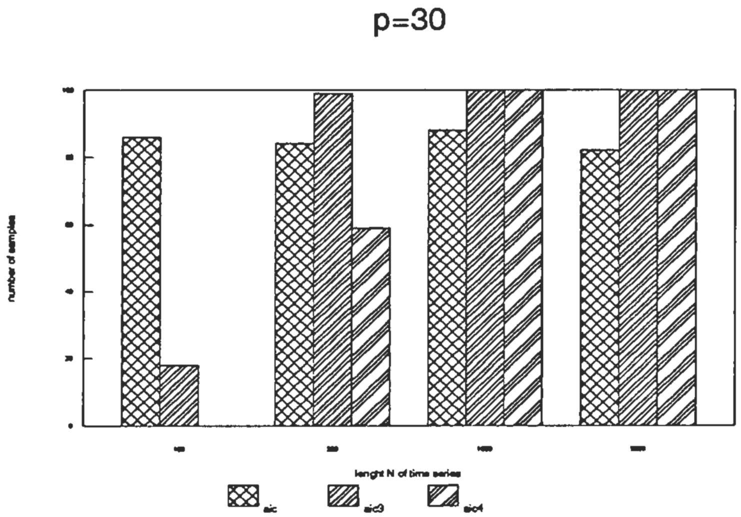
k=1		AIC			AIC3			AIC4		
N		*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*
100	S	0,3	0,3	0,3	0	0	0	0	0	0
	D	0,1	0,1	0,1	0	0	0	0	0	0
200	S	0,4	0,3	0,2	0	0	0	0	0	0
	D	0,1	0,1	0	0	0	0	0	0	0
1000	S	0,5	0,5	0,5	0	0	0	0	0	0
	D	0,2	0,2	0,2	0	0	0	0	0	0
5000	S	0,4	0,5	0,5	0	0	0	0	0	0
	D	0,1	0,2	0,2	0	0	0	0	0	0

Tab. 1 - Values of the RMSE and of the bias for AIC, AIC3 and AIC4 when k=1.

When k=1 AIC3 and AIC4 are slightly better than AIC, even if AIC doesn't strongly depart from the true k. Besides the number p of simulated variables doesn't seem to influence the results.

k=5		AIC			AIC3			AIC4		
N		*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*
100	S	1,0	0,4	1,0	2,6	1,6	2,9	3,3	2,8	4,0
	D	-0,7	0	-0,5	-2,5	-1,3	-2,8	-3,3	-2,8	-4,0
200	S	0,5	0,6	0,4	1,1	0,1	0,2	1,9	0,8	0,5
	D	-0,1	0,2	0,1	-0,9	0	0	-1,8	-0,5	-0,3
1000	S	0,4	0,3	0,3	0	0	0	0	0	0
	D	0,2	0,1	0,1	0	0	0	0	0	0
5000	S	0,6	0,4	0,4	0	0	0	0	0	0
	D	0,2	0,1	0,1	0	0	0	0	0	0

Tab. 2 - Values of the RMSE and of the bias for AIC, AIC3 and AIC4 when *k*=5.



Pic. 1 - Number of samples in which AIC, AIC3 and AIC4 indicate $k^* = 5$ when *k*=5.

When *k*=5 AIC3 and AIC4 are initially much worse than AIC but, by increasing *N*, AIC shows a tendency to overestimate the number of factors and, on the contrary, AIC3 and AIC4 converge to the true value *k*=5. Furthermore, getting from 20 to 40 variables, the true value of *k* is more easily detected.

k=10		AIC			AIC3			AIC4		
N		*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*
100	S	2,8	2,6	1,8	5,8	6,3	5,5	7,8	8,1	7,2
	D	-2,5	-2,3	-1,4	-5,7	-6,2	-5,4	-7,8	-8,1	-7,2
200	S	1,1	0,7	0,4	2,3	2,1	1,7	3,9	4,0	3,9
	D	-0.8	-0,2	0	-2,2	-1,9	-1,5	-3,7	-3,9	-3,7
1000	S	0,3	0,3	0,4	0,1	0	0	0,1	0	0
	D	0,1	0,1	0,2	0	0	0	0	0	0
5000	S	0,3	0,5	0,4	0	0,2	0	0	0	0
	D	0,1	0,2	0,2	0	0	0	0	0	0

Tab. 3 - Values of the RMSE and of the bias for AIC, AIC3 and AIC4 when *k*=10.

When *k*=10 the situation of *k*=5 is repeated and AIC seems to be generally better than AIC3 and AIC4 which strongly underestimate the number of factors.

As for variations of α in AIC, variations of *c* in Hannan-Quinn criterion bring to different methods

$$HQ1 = -2\log\max L + 2h \log\log N$$
$$HQ2 = -2\log\max L + 2h\ 2\log\log N$$
$$HQ3 = -2\log\max L + 2h\ 3\log\log N$$
$$HQ4 = -2\log\max L + 2h\ 4\log\log N$$

and the relative results are illustrated in the following tables.

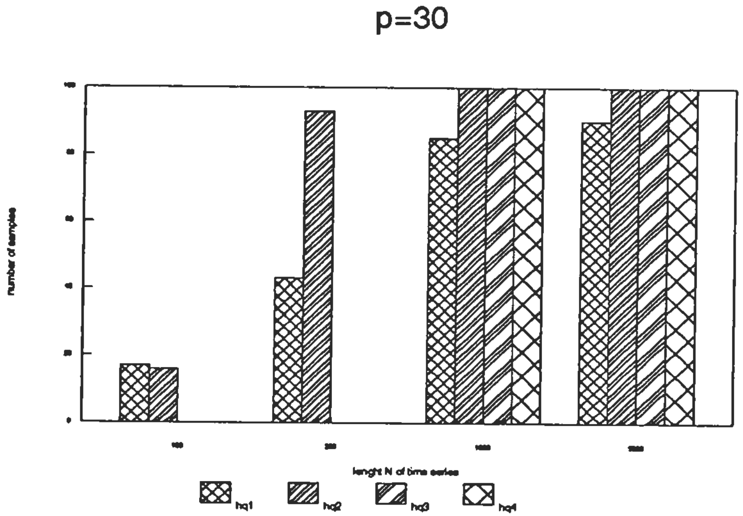
k=1		HQ1			HQ2			HQ3			HQ4		
N		*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*
100	S	1,8	2,2	2,7	0	0	0	0	0	0	0	0	0
	D	1,3	1,6	2,1	0	0	0	0	0	0	0	0	0
200	S	1,2	0,9	1,1	0	0	0	0	0	0	0	0	0
	D	0,8	0,6	0,7	0	0	0	0	0	0	0	0	0
1000	S	0,5	0,6	0,6	0	0	0	0	0	0	0	0	0
	D	0,2	0,3	0,3	0	0	0	0	0	0	0	0	0
5000	S	0,3	0,3	0,2	0	0	0	0	0	0	0	0	0
	D	0,1	0,1	0,1	0	0	0	0	0	0	0	0	0

Tab. 4 - Values of the RMSE and of the bias for HQ1, HQ2, HQ3 and HQ4 when *k*=1.

When k=1 only HQ1 shows difficulties in detecting the only factor and HQ2, HQ3 and HQ4, even with only 100 observations, are optimal indicators.

k=5		HQ1			HQ2			HQ3			HQ4		
N		*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*
100	S D	1,3 0,7	2,8 1,6	5,7 5,7	2,6 -2,5	1,7 -1,4	2,3 -2,2	3,6 -3,6	3,0 -3,0	4,0 -4,0	3,9 -3,9	3,3 -3,3	4,0 -4,0
200	S D	0,7 0,3	1,4 0,9	1,3 1,0	1,3 -1,2	0,3 -0,1	0,3 -0,1	2,6 -2,5	2,3 -2,1	1,1 -1,0	3,3 -3,3	3,0 -3,0	1,9 -1,9
1000	S D	0,5 0,2	0,4 0,2	0,5 0,2	0 0	0 0	0 0	0 0	0 0	0 0	0,2 0	0 0	0 0
5000	S D	0,4 0,1	0,2 0,1	0,2 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0

Tab. 5 - Values of the RMSE and of the bias for HQ1, HQ2, HQ3 and HQ4 when k=5.



Pic. 2 - Number of samples in which HQ1, HQ2, HQ3 and HQ4 indicate $k^* = 5$ when k=5.

When k=5 the indications of HQ2, HQ3 and HQ4 are more differentiate and HQ2 seems to converge to the true value k more quickly than the other types of Hannan-Quinn criterion. When a larger number of factor is present in the model, HQ1's goodness improves sensibly.

k=10		HQ1			HQ2			HQ3			HQ4		
N		*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*
100	S D	1,6 -1,1	0,9 0,2	0,9 0,7	6,1 -5,9	6,4 -6,4	5,6 -5,6	8,4 -8,4	8,7 -8,7	7,9 -7,9	9,0 -9,0	9,0 -9,0	8,8 -8,7
200	S D	0,9 -0,4	0,7 0,4	0,7 0,5	2,7 -2,6	2,5 -2,4	2,4 -2,3	6,1 -6,0	5,9 -5,9	5,7 -5,7	7,9 -7,9	8,6 -8,5	7,3 -7,2
1000	S D	0,3 0,1	0,4 0,1	0,4 0,2	0 0	0 0	0 0	0,3 -0,1	0 0	0 0	0,9 -0,7	0,3 -0,1	0 0
5000	S D	0,2 0,1	0,4 0,2	0,2 0,1	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0

Tab. 6 - Values of the RMSE and of the bias for HQ1, HQ2, HQ3 and HQ4 when k=10.

When k=10 the situation of k=5 is confirmed for HQ2, HQ3 and HQ4; yet HQ1 seems to be better than HQ2.

The results related to Akaike's, Hannan-Quinn's (c=2), Schwarz's, information criteria, cross-validation and log likelihood ratio are reported in the following tables.

k=1		AIC			HQ2			SCH			CROSS			LR		
N		*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*
100	S D	0,3 0,1	0,3 0,1	0,3 0,1	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0,4 0,1	0,3 0,1	0,8 0,3
200	S D	0,4 0,1	0,3 0,1	0,2 0	0 0	0 0	0 0	0 0	0 0	0 0	0,1 0	0 0	0 0	0,4 0,1	0,3 0,1	0,4 0,1
1000	S D	0,5 0,2	0,5 0,2	0,5 0,2	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0,4 0,1	0,3 0,1	0,3 0,1
5000	S D	0,4 0,1	0,5 0,2	0,5 0,2	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0,3 0,1	0,2 0	0,1 0

Tab. 7 - Values of the RMSE and of the bias for AIC, HQ2, SCH, CROSS and LR when k=1.

It's interesting to note how with k=1 Schwarz's information criterion, cross-validation and log likelihood **ratio**, **as** AIC and HQ2, can detect the presence of the only factor with **an optimal** approximation.

k=5		AIC			HQ2			SCH			CROSS			LR		
N		*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*
100	S	1,0	0,4	1,0	2,6	1,7	2,3	3,6	3,0	4,0	1,8	0,6	1,2	1,6	1,1	2,3
	D	-0,7	0	-0,5	-2,5	-1,4	-2,2	-3,6	-3,0	-4,0	-1,4	-0,2	-0,8	-1,4	-0.8	0,7
200	S	0,5	0,6	0,4	1,3	0,3	0,3	2,8	2,6	1,3	0,7	0	0,1	1,0	0,7	0,5
	D	-0,1	0,2	0,1	-1,2	-0,1	-0,1	-2,8	-2,6	-1,2	-0,4	0	0	-0,6	0	0
1000	S	0,4	0,3	0,3	0	0	0	0,1	0	0	0,1	0	0	0,4	0,3	0,3
	D	0,2	0,1	0,1	0	0	0	0	0	0	0	0	0	0,1	0,1	0,1
5000	S	0,6	0,4	0,4	0	0	0	0	0,1	0	0	0,1	0	1,2	0,2	0,3
	D	0,2	0,1	0,1	0	0	0	0	0	0	0	0	0	0,4	0	0,1

Tab. 8 - Values of the RMSE and of the bias for AIC, HQ2, SCH, CROSS and LR when *k*=5.

When *k*=5 AIC and cross-validation seem to be the best methods and they converge to the true value more quickly than the other ones.

Schwarz's criterion shows a strong tendency to underestimate the true number of factors.

k=10		AIC			HQ2			SCH			CROSS			LR		
N		*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*	*p=20*	*p=30*	*p=40*
100	S	2,8	2,6	1,8	6,1	6,4	5,6	8,5	8,7	7,9	3,6	4,1	3,5	3,2	3,1	2,6
	D	-2,5	-2,3	-1,4	-5,9	-6,4	-5,6	-8,4	-8,7	-7,9	-3,0	-3,7	-3,1	-3,0	-2,9	-2,4
200	S	1,1	0,7	0,4	2,7	2,5	2,4	6,7	6,5	5,9	1,1	1,1	0,6	1,4	1,2	1,1
	D	-0,8	-0,2	0	-2,6	-2,4	-2,3	-6,6	-6,4	-5,9	-0,7	-0,6	-0,2	-1,1	-0,9	-0,9
1000	S	0,3	0,3	0,4	0	0	0	0,7	0	0	0,3	0,1	0	0,5	0,2	0,2
	D	0,1	0,1	0,2	0	0	0	-0,4	0	0	0,1	0	0	0,2	0	0,1
5000	S	0,3	0,5	0,4	0	0	0	0	0	0	0,2	0,4	0	1,0	0,2	0,2
	D	0,1	0,2	0,2	0	0	0	0	0	0	0,1	0,1	0	1,0	0	0

Tab. 9 - Values of the RMSE and of the bias for AIC, HQ2, SCH, CROSS and LR when *k*=10.

When *k*=10 the situation for *k*=5 is repeated again and AIC and cross-validation are the best methods.

5 - The number of factors in the financial market

The choice of the number of factors represents a crucial point in the theory of financial markets and expecially in two of the most important assets returns models.

On one side the Capital Asset Pricing Model (CAPM) of Sharpe (1964) and Lintner (1965) assumes that only one factor can explain the assets returns; on the other the Arbitrage Pricing Theory (APT) of Ross (1976) states that *k* factors underlie the market.

Following the CAPM the return of the *i*-th asset is characterized by

$$E(r_i) = r_0 + (E(r_m) - r_0)\beta_i$$

where

r_0 is the risk free rate;

r_m is the return of the market portfolio;

$\beta_i = cov(r_i, r_m)/var(r_i)$.

The resulting market model is

$$r_{it} = \alpha_i + \beta_i r_{mt} + \varepsilon_{it}$$

where

$\alpha_i = (1 - \beta_i)r_0$;

ε_{it} is an error term.

The APT assumes that the generating model of the *i*-th asset is

$$E(r_i) = r_0 + \sum_{j=1}^{k} \lambda_{ij} y_j$$

where y_j is the premium for risk associated with the factor j and the coefficients λ_{ij} are estimated from the model

$$r_{it} = E(r_i) + \sum_{j=1}^{k} \lambda_{ij} f_{jt} + u_{it}$$

where

f_{jt} is the value at time *t* of the latent factor *j*;

u_{it} is an error term.

In order to discriminate between CAPM and APT it is necessary to determinate the number of factors; and this is the aim of this paper.

6 - Conclusions

In this paper a simulation study is performed to compare different methods for choosing the number k of factors in a factor model. The definitions of considered methods are given in the next table, in which the last column contains the average RMSE

$$\overline{S} = \frac{1}{36} \sum_{k,N,p}^{36} \left(\frac{1}{100} \sum_{i=1}^{100} (k_i^{*} - k)^2 \right)^{\frac{1}{2}}$$

with k = 1, 5, 10; N = 100, 200, 1000, 5000; p = 20, 30, 40.

Method	$\overline{S}$
$CV = N^{*}(\log\lvert\Lambda'\Lambda+\Psi\rvert - \log\lvert\Sigma^{*}\rvert) + N^{*}(tr((\Lambda'\Lambda+\Psi)^{-1}\Sigma^{*}) - p)$	0,55
$AIC = -2\log \max L + 2h$	0,63
$LR = N^{*}(\log\lvert\Lambda'\Lambda+\Psi\rvert - \log\lvert\Sigma\rvert)$	0,81
$AIC3 = -2\log \max L + 3h$	0,90
$HQ2 = -2\log \max L + 2h\ 2\log\log N$	0,95
$HQ1 = -2\log \max L + 2h\ \log\log N$	0,98
$AIC4 = -2\log \max L + 4h$	1,34
$HQ3 = -2\log \max L + 2h\ 3\log\log N$	1,64
$SCH = -\log \max L + 0,5h\ \log N$	1,73
$HQ4 = -2\log \max L + 2h\ 4\log\log N$	1,98

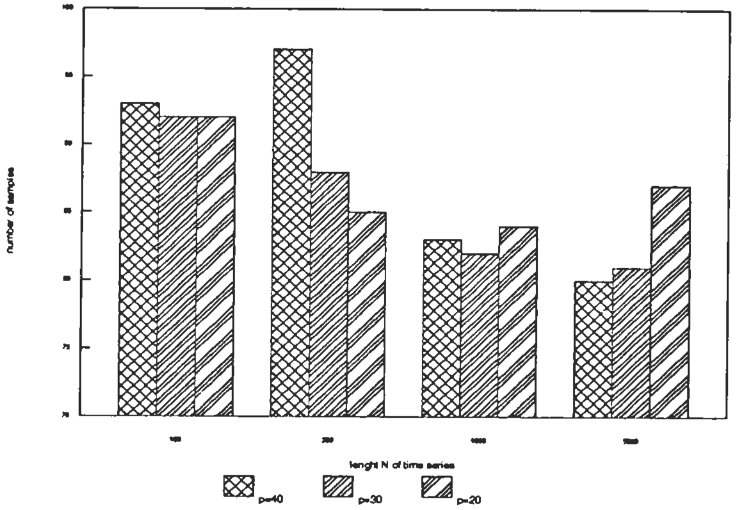
Tab. 10 - Methods for the determination of k and values of the medium RMSE.

Cross-validation and AIC have minimum $\overline{S}$ value and also seem to be, in complex, the more accurate methods. On the contrary, modifications of AIC don't improve the results ($\overline{S}_{AIC} < \overline{S}_{AIC3}$, $\overline{S}_{AIC} < \overline{S}_{AIC4}$) as modifications of HQ don't seem to obtain better indications ($\overline{S}_{HQ2} < \overline{S}_{HQ1}$, $\overline{S}_{HQ2} < \overline{S}_{HQ3}$, $\overline{S}_{HQ2} < \overline{S}_{HQ4}$). Values of 3 or 4 for α in AIC and for c in HQ bring to a strong underestimation of the true value of k. Schwarz's information

criterion too, underestimates sensibly the number of factors, particularly when the length N of the time series is very large. The usual test for the number of factors, the log likelihood ratio, is, after cross-validation and AIC, the best method.

A further consideration is that the goodness of the different methods is a function of the number k of "true" factors underlying the simulated variables.

Indeed in the case of k=1 all methods analyzed indicate the right value k=1 with the exceptions of AIC, HQ1 and LR. However for AIC and LR the distances from the exact value are quite small. In the following picture the results related to AIC are illustrated: as N increases AIC gets worse.



Pic. 3 - Number of samples in which AIC indicates $k^{*} = 1$ when k=1.

With only 100 observations there are already generally good indications and the dimension p seems to be not particularly relevant. This result shows how, when the true model contains only one factor, information criteria and cross validation can detect it with a good precision.

In the case k=5 the situation is more complex and the length N of the time series is particularly relevant: asymptotically, indeed, all methods converge to the true value k=5. However, it is important to emphasize that AIC and cross-validation converge more quickly.

The situation for k=10 is similar to the one for k=5: AIC and cross-validation show the best performance.

From the sign of the bias, reported in the next table, one can observe how the minus prevails thus meaning a stronger tendency to underestimate rather than to overestimate the true number of factors.

	-	0	+
AIC	22	8	70
AIC3	28	72	-
AIC4	33	67	-
HQ1	6	-	94
HQ2	33	67	-
HQ3	36	64	-
HQ4	39	61	-
SCH	36	64	-
CV	28	64	8
LR	25	22	53

Tab. 11 - Percentual cases in which bias is negative, null or positive.

Concluding one can affirm that when only one factor constitutes the factor model a small number of observations is sufficient to detect it. When, on the contrary, more factors underlie the observed variables, cross-validation and AIC seem to be the more appropriate indicators.

7 - References

H. Akaike (1979), A bayesian analysis of the minimum AIC procedure, *Annals of the Institue of Statistical Mathematics A*, 30, 9-14.

H. Akaike (1987), Factor analysis and AIC, *Psychometrika*, 52, 3, 317-332.

T.W. Anderson (1984), An Introduction to Multivariate Statistical Analysis, New York, Wiley.

M.S. Bartlett (1950), Tests of Significance in Factor Analysis, *British Journal of Mathematical and Statistical Psychology*, 3, 77-85.

P. Bekker (1989), Identification in restricted factor models and the evaluation of rank conditions, *Journal of Econometrics*, 41, 5-16.

R.J. Bhansali, D.Y. Downham (1977), Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion, *Biometrika*, 64, 3, 547-551.

H. Bozdogan, D.E. Ramirez (1987), An expert model selection approach to determine the best pattern structure in factor analysis models, in Multivariate statistical modeling and data analysis, D. Reidel Publishing Company, 35-60.

D.E. Conway, M.R. Reinganum (1988), Stable Factors in Security Returns: Identification Using Cross-Validation, *Journal of Business & Economic Statistics*, 6, 1-15.

Z. Griliches (1977), Errors in variables and other unobservables, in AignerD., Goldberger A. (1977), Latent variables in socio-economic models, North-Holland.

E.J. Hannan, B.G. Quinn (1979), The determination of the order of an autoregression, *Journal of the Royal Statistical Society B*, 41, 2, 190-195.

J. Kim, C.W. Mueller (1983), Factor Analysis, London, Sage.

D.N. Lawley, A.E. Maxwell (1971), Factor Analysis as a Statistical Method, London, Butterworths.

J. Lintner (1965), The Valuation of Risky Assets and the Selection of Risk Investments in Stock Prtofolios and Capital Budgets, *Review of Economics and Statistics*, 47, pp. 13-37.

S. Ross (1976), The Arbitrage Theory of Capital Asset Pricing, Journal of Economic Theory, 13, 341-360.