A COMPARATIVE STUDY OF ALTERNATIVE APPROACHES FOR COMMON FACTORS IDENTIFICATION †

Mariano González Sánchez♣

University Cardenal Herrera CEU

Juan M. Nave Pineda

University Cardenal Herrera CEU

Preliminary version: May 2011

ABSTRACT:

For multivariate non-stationary time series modeling is essential to know the number of common factors that define the behavior of the series. The traditional way to approach this problem is to study the cointegration relations among data through tests of the trace or maximum eigenvalue, obtaining the number of stationary long-run relations. Alternatively this problem can be analyzed using dynamic factor models as in Peña and Poncela (2006), estimating in this case the number of all independent common factors, stationary or not, that describe the behavior of data. In this context, we analyze empirically the power of such alternative approaches by applying them to series simulated using known factorial models. The results show that when there are stationary common factors, when the number of observations is reduced and/or when the series have involved more than one cointegration relation, the common factor test is more powerful than the usual cointegration tests. These results together with the greater flexibility of dynamic factor models for identify the load matrix of the DGP make them more suitable for use in multivariate analysis.

KEYWORDS: multivariate analysis, common factors, cointegration, dynamic factor models, stationarity.

JEL: C13, C32, G11.

---

† This work was supported in part by the research project ECO-2009 13616 from the Ministry of Science and Innovation of Government of Spain and the PROMETEO Project 2008/106 from the Government of Generalitat Valenciana (Spain).

♣ Correspondence authors: Department of Economics and Business. Faculty of Law, Business and Politics. University Cardenal Herrera CEU. Luis Vives, 1, 46115 Alfara del Patriarca, Valencia (Spain). Tfno.: +34 961369000. Fax: +34 961369007. Email: mariano.gonzalez@uch.ceu.es and juan.nave@uch.ceu.es

1. INTRODUCTION

Two approaches are usually used in the dynamic multivariate analysis when it involves non-stationary variables: cointegration analysis and dynamic factorial analysis. Cointegration analysis identifies the long-run relationships among variables while dynamic factorial analysis determines the number of independent unobservable common dynamic factors that describe the behavior of the variables set. In the long-run, the solutions of both problems are complementary, because for a set of variables with cointegrating relationships, the number common factors results of the difference between the total number of variables and the number of cointegrating relationships, and vice versa.

However, in the economic literature cointegration techniques are most widely used in the dynamic multivariate context, perhaps due to they are more popular and are present in the standard econometric software. Beyond focusing exclusively in the long-run, others shortcomings may arise when we apply the cointegration methodology, specially when model errors are serially dependent and/or non-Gaussian (Gonzalo, 1994 and Gonzalo and Lee, 1998), therefore alternative approaches are necessary (Li et al., 2009). Additionally, when the number of variables involved increase to achieve an identified system required to restrict it through a priori relations among variables not always available. In this context dynamic factorial analysis becomes more and more useful.

The main objective of this paper is to analyze these two a priori alternative approaches in order to show their ability to get the true number of hidden common factors present in the generating process of a set of variables. For so doing, we specifically run on previously simulated data series one test developed in the dynamic factor analysis framework by Peña and Poncela (Peña and Poncela, 2006) and the traditional cointegration tests, i.e. the trace test and the maximum eigenvalue. We choose the Peña and Poncela [hereafter, PP] test as representative of a set of similar tests developed in the dynamic factor analysis framework from a robust methodology (Forni et al., 2000, 2004 and 2005; Hallin and Liska, 2007). Other tests developed in this framework are not considered either because of their lack of robustness (Park et al., 2009) or because of their high computational cost (Pan and Yao, 2008).

The rest of the paper is structured as follows. In section two we formalize the problem and describe the PP methodology. In section three we describe the simulation experiments and show the results of them. Finally, we summarize the main conclusions.

2. COINTEGRATION vs. FACTORIAL MODEL

In multivariate problems where the observed variables are non-stationary in levels, by taking a stationary transformation of the series may not adequately exploit all the information about the relationships among them. To avoid this, the problem can be analyzed from two complementary perspectives:

1. First, considering the long-run stable and stationary relationships among sets and, using different techniques related to cointegration. Thus, if *X* is the vector of non-stationary variables and ΔX is its stationary transformation, the general model usually arise in the econometric literature as an Vector Error Correction Model or VECM [Engle and Granger (1987)]:

$$\Delta X_t = \Pi X_{t-1} + \sum_{i=1}^{P} \Delta X_{t-i} + \varepsilon_t \tag{1}$$

   So, by well know cointegration tests (maximum eigenvalue and trace test), we can estimate the cointegration rank, or rank of the matrix (Π) in equation (1). Thus, if there are *N* variables and, *m* cointegration relations, then there are *r* = *N*-*m* common factors.

2. The second possibility is to study directly the number of common factors [see Escribano and Peña (1994) and, Gonzalo and Granger (1995)]. In this case, the problem formulation would depend on a number of factors (*F*) and a load matrix (λ):

$$\begin{aligned} X_t &= \lambda F_t + \omega \varepsilon_t \\ F_t &= \rho F_{t-1} + u_t \end{aligned} \tag{2}$$

   Let us note that if ρ is less than 1, in absolute value, the factor is I(0), and will only be I(1), if it is equal to 1.

The relationship between these two approaches can be explained from the decomposition of the Π matrix,

$$\Pi = \alpha \beta^T \tag{3}$$

Where α is the weight matrix of the cointegration relationships for each variables; and β is the coefficient matrix of the cointegration relations.

Using (3) the factorial model (2) could be expressed as:

$$X_t = \beta_\perp \left(\alpha_\perp^T \beta_\perp \right)^{-1} \alpha_\perp^T X_t + \alpha \left(\beta^T \alpha \right) \beta^T X_t \tag{4}$$

And from the expressions (1) and (4) is easy to see the relationship between the two approaches. So we can solve the multivariate problem studying the cointegration relations with the relevant

restrictions, which becomes intractable as the number of variables increase[1], or analyzing the dynamic common factors involved, and this is the point where the test Peña and Poncela (2006) attempts to provide some light on the problem.

Peña and Poncela (2006) express the factorial model (2) as follows:

$$X_t = \mu + \begin{bmatrix} \lambda_1 & \lambda_2 \end{bmatrix} \begin{bmatrix} F_1 & F_2 \end{bmatrix}^T + \varepsilon_t \tag{5}$$

That is, if there are *N* observed variables with the corresponding mean vector (μ), then the common factors explaining may belong to the subset F_1 of factors I(1) or the F_2 subset of common factors I(0). Thus, if the first group is made up r_1 factors and the second r_2, then there must be $N-(r_1+r_2)$ cointegration relationships.

The advantage of this framework is that, regardless of whether the observed series are stationary or not, can identify stationary and non-stationary common factors. It also allows to test whether the cointegration test correctly discriminate between cointegration relations and stationary factors. The test is implemented as follows:

$$\begin{aligned} Y_t &= X_t - \mu \\ \hat{M}(k) &= \left[\sum_{t=k+1}^{T} \left(Y_t Y_t^T \right) \right]^{-1} \bullet \left[\sum_{t=k+1}^{T} \left(Y_t Y_{t-k}^T \right) \right] \bullet \left[\sum_{t=k+1}^{T} \left(Y_{t-k} Y_{t-k}^T \right) \right]^{-1} \bullet \left[\sum_{t=k+1}^{T} \left(Y_{t-k} Y_t^T \right) \right] \end{aligned} \tag{6}$$

According to Peña and Poncela (2006) Theorem-3, the matrix $\hat{M}(k)$ has $N-(r_1+r_2)$ eigenvalues which converge in probability to zero as the sample size (*T*) tends to infinity and the number of lags increases $k = 0, 2, \ldots, K$ such that K/T tends to zero.

In this way, the test is to sort the eigenvalues (*h*) of the matrix $\hat{M}(k)$, and obtain their sum as follows:

$$S_{r=r_1+r_2} = (T-k) \sum_{j=1}^{N-r} \log\left(1-h_j\right) \tag{7}$$

This sum behaves asymptotically as a Chi-square distribution with $(N-r)^2$ degrees of freedom.

Once estimated the number of common factors, Peña and Poncela (2006) use an EM algorithm to determine the load matrix of the expression (5), using as initial values (E, Expected) the eigenvectors (*V*) of the following matrix:

$$C(1) = \frac{1}{T^{2d+d'}} \sum_{t=2}^{T} \left(Y_{t-1} Y_t^T \right) \tag{8}$$

[1] When *N* increases the determination of matrices α and β is not possible without add restrictions to transform the problem in a defined system (reduced-rank regression). But sometimes it is not possible to know a priori relations among these variables.

As shown in (8) would operate with one lag, *d* is the order of integration of the observed series and *d'* would be worth 0 if *d*>0, or 1 if it is equal to zero. Thus, once ordered from highest to lowest eigenvalues, the initial value of the loading matrix would be the eigenvectors associated with the first *r* non-zero eigenvalues. Similarly, the initial values of the factors would be obtained from these eigenvectors and the observed variables as follows:

$$\begin{aligned} \hat{\lambda}_0 &= V_r \\ \hat{F}_0 &= V_r^T Y \end{aligned} \tag{9}$$

Once recovered the initial values of the factors ($\hat{F}_0$), they propose to use a contrast of stationarity (eg, Augmented Dickey-Fuller) to determine which factors are I(1) and what I(0).

From the initial values (Expected) the final results (Maximization) are estimated by applying the Kalman filter on the following Space of State expression:

$$\begin{aligned} &Eq.\,Mesure \\ &\begin{bmatrix} x_{1,t} \\ \vdots \\ x_{N,t} \end{bmatrix} = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,K} \\ \vdots & \ddots & \vdots \\ \beta_{N,1} & \cdots & \beta_{N,K} \end{bmatrix} \bullet \begin{bmatrix} f_{1,t} \\ \vdots \\ f_{K,t} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \vdots \\ \varepsilon_{N,t} \end{bmatrix} \qquad \begin{bmatrix} \varepsilon_{1,t} \\ \vdots \\ \varepsilon_{N,t} \end{bmatrix} \sim \Phi(0,\Sigma) \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_N^2 \end{bmatrix} \\ &Eq.\,Transition \\ &\begin{bmatrix} f_{1,t+1} \\ \vdots \\ f_{K,t+1} \end{bmatrix} = \begin{bmatrix} \rho_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho_K \end{bmatrix} \bullet \begin{bmatrix} f_{1,t} \\ \vdots \\ f_{K,t} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ \vdots \\ u_{K,t} \end{bmatrix} \qquad \begin{bmatrix} u_{1,t} \\ \vdots \\ u_{K,t} \end{bmatrix} \sim \Phi(0,\mathrm{I}) \end{aligned} \tag{10}$$

Where the parameters ρ is one when the test indicated that the factors are I(1).

3. EMPIRICAL STUDY

To measure the PP-test power related to traditional cointegration approach, we calculate for four different scenarios 100 variables time series (N=100) with 5000 data (T=5000) from factors previously simulated and known load matrix, then both approaches will be applied on and compare the results with the actual values of the data generating processes.

a) Scenario I: one common factor I(1) for each two variables, i.e., there are 50 common factors and therefore 50 cointegration relations.
b) Scenario II: one common factor I(1) for each four variables, i.e., there are 25 common factors and cointegration relationships are 75.
c) Scenario III: one factor I(1) and other I(0) for each four variables, i.e., there are 50 common factors.
d) Scenario IV: two factors I(1) for each four variables, i.e., there are 50 common factors.

The contrasts are made by varying the number of individuals (which will vary the number of total factors) and the sample size (to check PP-Theorem 3). The aim is to determine the number of lags that test-PP obtained the correct number of factors, this also applies for two different confidence levels of the Chi-square distribution, 95% and 99%.

The simulated processes (from loading matrix) are as follows, taking only 2 or 4 of total 100 variables for each scenario, i.e., the variables grouped by common factors and scenario:

a) Scenario I:

$$\begin{bmatrix} z_{1,t} \\ z_{2,t} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} f_t + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \qquad f_t = \sum_{j=1}^{t} u_j \qquad [\varepsilon_1, \varepsilon_2, u] \sim i.i.d.\Phi(0,1) \tag{11}$$

b) Scenario II:

$$\begin{bmatrix} z_{1,t} \\ z_{2,t} \\ z_{3,t} \\ z_{4,t} \end{bmatrix} = \begin{bmatrix} 1 \\ 0.25 \\ 0.5 \\ 0.75 \end{bmatrix} f_t + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \\ \varepsilon_{4,t} \end{bmatrix} \qquad f_t = \sum_{j=1}^{t} u_j \qquad [\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, u] \sim i.i.d.\Phi(0,1) \tag{12}$$

c) Scenario III:

$$\begin{bmatrix} z_{1,t} \\ z_{2,t} \\ z_{3,t} \\ z_{4,t} \end{bmatrix} = \begin{bmatrix} 1 & 0.25 \\ 0.75 & 0.5 \\ 0.5 & 0.75 \\ 0.25 & 1 \end{bmatrix} \begin{bmatrix} f_{1,t} & f_{2,t} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \\ \varepsilon_{4,t} \end{bmatrix} \qquad f_{1t} = f_{1t-1} + u_{1,t} \qquad f_{2t} = 0.8 f_{2t-1} + u_{2,t} \qquad [\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, u_1, u_2] \sim i.i.d.\Phi(0,1) \tag{13}$$

d) Scenario IV:

$$\begin{bmatrix} z_{1,t} \\ z_{2,t} \\ z_{3,t} \\ z_{4,t} \end{bmatrix} = \begin{bmatrix} 1 & 0.25 \\ 0.75 & 0.5 \\ 0.5 & 0.75 \\ 0.25 & 1 \end{bmatrix} \begin{bmatrix} f_{1,t} & f_{2,t} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \\ \varepsilon_{4,t} \end{bmatrix} \qquad f_{1t} = f_{1t-1} + u_{1,t} \qquad f_{2t} = f_{2t-1} + u_{2,t} \qquad [\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, u_1, u_2] \sim i.i.d.\Phi(0,1) \tag{14}$$

First, we show a statistical summary of the simulated values for each of the four scenarios:

[Insert here Table-1, 2, 3 and 4]

As shown in statistical tables, variables and factors are neither normal nor stationary, except for the factor I(0) of the Scenario III. It also highlights that the variables have a minimum in the ADF test for stationarity nearest being accepted, that is, despite generating factors are I(1), in some cases, the variables constructed from these may seem I(0), at least 95% confidence level.

Then we show the results of cointegration and PP-test, the errors are shown in Table-5, i.e., the difference between the number of real factors and the number of factors estimated (for cointegration analysis is estimated by the difference between the number of variables and the cointegration rank):

[Insert here Table-5]

From the above results, using the cointegration test, include the following:

- The trace test is more consistent than maximum eigenvalue.
- The result depends mainly on the degrees of freedom, i.e., by increasing the number of lags cointegrating rank converges to the correct value, but there is a limitation in function the number of observations and individuals. Specifically, when the numbers of observations decreases, the best results are achieve with lower lags, since this way the degrees of freedom are higher.
- When there is one common factor and sufficient degrees of freedom, the cointegration tests determine the correct number of factors without mistakes. However, when the series are two-cointegrated, the performance declines. Thus, we suggest that if variables have different number of common factors, these tests do not allow to discriminate correctly, and if one of them is I(0), not is detected.

Therefore, the strength of these tests is when there is one stationary common factor.This, together with the impossibility of estimating cointegration relationships accurately, or load matrix associated with each common factor, is more than enough to justify the study of tests on common factors, as the PP-test.

The highlights of these results using the PP-test are:

- When *T* is much bigger than *N*, the test converges quickly to the correct number of factors, including a single lag. Therefore, the test is valid for long time series.
- As *T* is close to *N*, the errors can be remedied by increasing the number of lags, or changing the level of confidence (increase), or a mixture of both.
- If *N* is greater than *T*, only by increasing the number of lags we reach the correct value, provided that the number of individuals *N* is not much higher than *T* (temporal size). This is due again to the degrees of freedom.
- Note that, when there is a common factor I(0), which did not detect cointegration test, the PP test does it, in fact, is more consistent and for the same set of variables, the number of cointegration common factors, i.e., the series are two-cointegrated, three-cointegrated, …

In summary, the test is a perfect complement to the traditional analysis of cointegration, when there are more than two variables and is intended to determine the number of common factors and/or these variables participate in more than one cointegration relationship.

The modus operandi then would see the convergence of the result to vary the level of trust and the number of lags. For example, on the Scenario IV, we performed three tests with different *N* and *T*:

[Insert here Table-6]

A final issue that remains to be analyzed is whether the methodology proposed by PP, allows to recovering the true factors, testing if they are stationary, and determining the corresponding loading matrices. This is important not only for being the culmination of reducing the problem, but also because it would identify better the groups of cointegrated variables and its common factors, i.e., there could be, for example, two groups of four cointegrated variables cointegration each one, that trace test detect it, but we do not indicate either the exact number of common factors, or whether they are stationary.

To test this, we taken, of the simulated samples, T = 1000, a number of variables equal to M. These M variables are 2 of Scenario I (a common factor I(1)), and 4 of Scenario IV[2] (two common factors I(1)), therefore, M=6 variables[3], with two cointegration relations, the first with a non-stationary factor and the second with two non-stationary factors.

As expected, given previous results, the cointegration tests identified correctly the number of common factors, but do not tell us that there are two groups of variables, because the cointegration rank is three, so we could even think that there are three sets of two cointegrated variables each one. The results were as follows:

[Insert here Table 7]

Similarly, the PP test, regardless of the number of lags and confidence level, identified three factors for 1-lag, with a test-value of 11.62 [and p-value=0.236]. In addition, the ADF test of the 3 factors recovered, from the product of the matrix transpose of the eigenvectors associated with the three largest eigenvalues and the observed variables, resulted in three factors I(1):

[Insert here Table 8]

Thus, both methods agree on the number of factors, but it is interesting to know if the methodology proposed by PP to recover factors and grouping variables. The parameters (sigmas and betas of loading matrix) obtained were:

[Insert here Table 9]

Graphically, start, estimated-EM and true factors are as follow:

[2] We have not taken variables of Experiment-III because of the trace test did not identify the second factor stationary.
[3] Take a few variables is due to better convergence, and especially that our interest is solely to establish and recover the factors in a set of variables with different number of common factors.

[Insert here Figure 1a, 1b, 1c]

As can be seen, the estimated factor after the maximization is more similar to the original one. Finally, we present for the start factors and estimated factors, as suggested by Peña and Poncela (2006) [PP], the mean absolute error [MAE] and root mean square error [RMSE]:

[Insert here Table 10]

4. CONCLUSIONS

When a univariate analysis turns out to be non-stationary variable the practice solves the problem by some kind of transformation of the original series. In contrast, when the problem is multivariate, operate the same way may involve loss of information relevant to the data modeling. In these cases, the cointegration analysis and the determination of the cointegration relations allow modeling the relations of long-term equilibrium among variables involved in the problem.

However, when variables are present in more than one cointegration relationship, estimating the parameters of the cointegration relations is conditioned by the restrictions imposed on the problem which is often complex without prior information; in other case the system is not identified. Furthermore, estimate the correct number of independent factors when one of them is stationary is also difficult since the tests usually used are not able to distinguish between relations of cointegration and stationary factors. In this context, we have studied, on simulated series from known DGP, the power of cointegration tests and common factors test proposed by Peña and Poncela (2006).

The results for both methodologies are similar except when there are stationary common factors where the cointegration tests do not identify the correct number of factors. But, when data size decrease and/or the cointegration relationships involve more than two variables the Peña and Poncela methodology results overcomes the cointregration ones. Finally, an added advantage of Poncela and Peña methodology compared to cointegration analysis is that allows a better approximation to the DGP.

REFERENCES

- Alexander, C.; Giblin, I. and Weddington, W. (2001). Cointegration and Asset Allocation: A New Active Hedge Fund Strategy. INQUIRE Europe Conference. Brighton.
- Davidson, J. (1998,a). Structural relations, cointegration and identification: some simple results and their application. Journal of Econometrics, 87: 87-113.
- Davidson, J. (1998,b). A Wald test of restrictions on the cointegrating space based on Johansen's estimator. Economics Letters, 59: 183-187.
- Engle, R. F. and Granger, C.W.J. (1987). Cointegration and error corection: Representation, estimation and testing. Econometrica, 55: 251-276.
- Escribano, A. and Peña, D. (1994). Cointegration and common factors. Journal of Time Series Analysis, 15: 577-586.
- Forni, M.; Hallin, M.; Lippi, M. and Reichlin, L. (2000). The generalized dynamic factor model: Identification and estimation. The Review of Economics and Statistics 82, 540-554.
- Forni, M.; Hallin, M.; Lippi, M. and Reichlin, L. (2004). The generalized dynamic-factor model: consistency and rate. Journal of Econometrics, 119, 231-255.
- Forni, M.; Hallin, M.; Lippi, M. and Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. Journal of the American Statistical Association, 100, 830-840.
- Gonzalo, J. (1994). Five alternative methods of estimating long-run equilibrium relationships Journal of Econometrics 60, 203-233.
- Gonzalo, J. and Granger, C.W.J. (1995). Estimation of common long-memory components in cointegrated systems. Journal of Business and Economic Statistics, 13: 27-35.
- Gonzalo, J. and Lee, T. H. (1998). Pitfalls in testing for long-run relationships. Journal of Econometrics 86, 129-154.
- Hallin, M. and Liska, R. (2007). Determining the number of factors in the general dynamic factor model. Journal of American Statistical Association, 102, 603-617.
- Li, Q.; Pan, J. and Yao, Q. (2009). On determination of cointegration ranks. Statistics and its interface, 2 (1). pp. 45-56.
- Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. Biometrika, 95, 365-379.
- Park, B.; Mammen, E.; Hardle, W. and Borak, S. (2009). Modelling dynamic semiparametric factor models. Journal of the American Statistical Association, 104, 284-298.

- Peña, D. and Poncela, P. (2006). Nonstationary dynamic factor analysis. Journal of Statistical Planning and Inference, 136(4) :1237-1257.

TABLES

Table 1. Statiscal Summary of Scenario I

	FACTORS			VARIABLES		
STATISTICAL	MINIMUM	MEAN	MÁXIMUM	MINIMUM	MEAN	MÁXIMUM
Mean	-81.3281	-3.8161	83.7614	-81.3337	-2.6927	83.7586
Std.Devn.	12.8389	23.8037	48.5269	6.4926	18.3969	48.5567
Skewness	-1.1633	-0.0060	1.0944	-1.1588	-0.0034	1.0913
Excess Kurtosis	-1.3959	-0.7468	1.0699	-1.3930	-0.7351	1.0588
Minimum	-143.7454	-48.3838	1.7816	-144.0226	-38.8694	2.6938
Maximum	2.6683	41.9840	157.1158	3.1923	30.2335	156.9546
Normality test Chi^2(2):	139.040	560.370	4824.900	100.69	438.16	2997.01
ADF test	-2.743	-1.695	0.116	-3.191*	-1,759	-0.0019

Table 2. Statistical Summary of Scenario II

	FACTORS			VARIABLES		
STATISTICAL	MINIMUM	MEAN	MÁXIMUM	MINIMUM	MEAN	MÁXIMUM
Mean	-111.8510	9.0179	67.4385	-111.8279	3.3383	67.4331
Std.Devn.	11.2978	18.6616	50.9810	2.9889	12.5300	50.9969
Skewness	-1.5257	0.2169	1.2449	-1.5112	0.2141	1.2435
Excess Kurtosis	-1.4509	-0.5292	2.8060	-1.4497	-0.5234	2.7727
Minimum	-183.6206	-40.4077	-0.0282	-184.8709	-20.5148	0.2120
Maximum	3.5753	52.8598	159.7909	2.2901	28.8126	160.1546
Normality test Chi^2(2):	42.7750	295.1900	4706.6000	34.3890	279.6050	4679.5000
ADF test	-2.5330	-2.0030	-0.3368	-3.093*	-1.927	-0.2922

Table 3. Statistical Summary of Scenario III

	FACTORS I(1)			FACTORS I(0)			VARIABLES		
STATISTICAL	MINIMUM	MEAN	MÁXIMUM	MINIMUM	MEAN	MÁXIMUM	MINIMUM	MEAN	MÁXIMUM
Mean	-81.3281	-3.8161	83.7614	-0.0319	-0.0017	0.0586	-81.3327	-2.6930	83.7523
Std.Devn.	12.8389	23.8037	48.5269	0.9766	0.9994	1.0231	6.5593	18.4069	48.5641
Skewness	-1.1633	-0.0060	1.0944	-0.0585	0.0068	0.0742	-1.1595	-0.0033	1.0904
Excess Kurtosis	-1.3959	-0.7468	1.0699	-0.2623	-0.0293	0.1189	-1.3933	-0.7232	1.0621
Minimum	-143.7454	-48.3838	1.7816	-4.5500	-3.5407	-3.1001	-144.1497	-40.0419	1.9830
Maximum	2.6683	41.9840	157.1158	3.0937	3.5573	4.4922	3.2835	31.4876	157.1545
Normality test JB Chi^2(2):	52.1860	379.6850	4824.9000	0.0165**	1.1163**	16.307*	39.2140	366.5400	4786.3000
ADF test	-2.6740	-1.7830	0.1112	-71.76**	-29.53**	-24.06**	-3.392*	-1.8010	0.0383

Table 4. Statistical Summary of Scenario IV

	FACTORS I(1)			FACTORS I(1)			VARIABLES		
STATISTICAL	MINIMUM	MEAN	MÁXIMUM	MINIMUM	MEAN	MÁXIMUM	MINIMUM	MEAN	MÁXIMUM
Mean	-111.8510	9.0179	67.4385	-68.9979	-8.1259	98.8278	-123.9412	1.6492	114.5438
Std.Devn.	11.2978	18.6616	50.9810	11.8151	22.9448	50.6171	11.7244	23.4704	53.6941
Skewness	-1.5257	0.2169	1.2449	-0.9858	-0.1853	0.8649	-1.2764	-0.0584	1.2535
Excess Kurtosis	-1.4509	-0.5292	2.8060	-1.1794	-0.5126	0.8638	-1.6293	-0.7070	1.4736
Minimum	-183.6206	-40.4077	-0.0282	-164.2842	-54.0195	-0.7104	-194.7119	-46.8430	0.4409
Maximum	3.5753	52.8598	159.7909	-0.7699	27.1990	162.7531	-1.0514	48.0840	189.1853
Normality test JB Chi^2(2):	42.7750	295.1900	4706.6000	33.115	375.76	2948.4	18.7410	320.7350	4943.8000
ADF test	-2.4330	-2.0030	-0.3368	-2.556	-1.757	0.3456	-3.093*	-1.7120	0.2362

Table 5. Errors Number of Factors by Cointegration and PP-test

ERRORS COINTEGRATION TEST						ERRORS PP-TEST					
Scenario I						**Scenario I**					
T/N	*8*	*20*	*40*	*80*	*100*	*T/N*	*8*	*20*	*40*	*80*	*100*
25	-3.00	ngl	ngl	ngl	ngl	*25*	2.00	-7.00	1.00	23.00	41.00
50	0.00	-1.00	ngl	ngl	ngl	*50*	1.00	0.00	-12.00	0.00	3.00
100	0.00	1.00	-6.00	ngl	ngl	*100*	0.00	0.00	2.00	-25.00	-29.00
250	0.00	0.00	-2.00	-6.00	2.00	*250*	0.00	0.00	1.00	3.00	1.00
500	0.00	0.00	0.00	-1.00	0.00	*500*	0.00	0.00	0.00	1.00	1.00
1000	0.00	0.00	0.00	0.00	0.00	*1000*	0.00	0.00	0.00	0.00	1.00
2500	0.00	0.00	0.00	0.00	0.00	*2500*	0.00	0.00	0.00	0.00	0.00
5000	0.00	0.00	0.00	0.00	0.00	*5000*	0.00	0.00	0.00	0.00	0.00
Scenario III						**Scenario III**					
T/N	*8*	*20*	*40*	*80*	*100*	*T/N*	8.00	20.00	40.00	80.00	100.00
25	0.00	ngl	ngl	ngl	ngl	*25*	3.00	-7.00	4.00	23.00	41.00
50	1.00	-2.00	ngl	ngl	ngl	*50*	2.00	3.00	1.00	0.00	3.00
100	1.00	-2.00	-3.00	ngl	ngl	*100*	1.00	2.00	4.00	0.00	2.00
250	2.00	0.00	-1.00	0.00	-8.00	*250*	0.00	1.00	2.00	6.00	3.00
500	2.00	5.00	5.00	0.00	-3.00	*500*	0.00	0.00	1.00	3.00	6.00
1000	2.00	5.00	10.00	15.00	14.00	*1000*	0.00	0.00	0.00	0.00	1.00
2500	2.00	5.00	10.00	20.00	25.00	*2500*	0.00	0.00	0.00	0.00	0.00
5000	2.00	5.00	10.00	20.00	25.00	*5000*	0.00	0.00	0.00	0.00	0.00
Scenario II						**Scenario II**					
T/N	*8*	*20*	*40*	*80*	*100*	*T/N*	*8*	*20*	*40*	*80*	*100*
25	-1.00	ngl	ngl	ngl	ngl	*25*	1.00	-1.00	0.00	3.00	16.00
50	0.00	1.00	ngl	ngl	ngl	*50*	0.00	-1.00	0.00	0.00	-1.00
100	0.00	0.00	-15.00	25.00	ngl	*100*	0.00	-1.00	0.00	0.00	-1.00
250	0.00	0.00	0.00	0.00	-8.00	*250*	0.00	0.00	-1.00	-5.00	-13.00
500	0.00	0.00	0.00	-1.00	3.00	*500*	0.00	0.00	0.00	0.00	0.00
1000	0.00	0.00	0.00	0.00	0.00	*1000*	0.00	0.00	0.00	0.00	0.00
2500	0.00	0.00	0.00	0.00	0.00	*2500*	0.00	0.00	0.00	0.00	0.00
5000	0.00	0.00	0.00	0.00	0.00	*5000*	0.00	0.00	0.00	0.00	0.00
Scenario IV						**Scenario IV**					
T/N	*8*	*20*	*40*	*80*	*100*	*T/N*	*8*	*20*	*40*	*80*	*100*
25	1.00	ngl	ngl	ngl	ngl	*25*	3.00	-6.00	1.00	23.00	41.00
50	1.00	-3.00	ngl	ngl	ngl	*50*	1.00	3.00	-11.00	0.00	3.00
100	0.00	-2.00	-3.00	ngl	ngl	*100*	0.00	0.00	2.00	0.00	-1.00
250	0.00	0.00	2.00	-3.00	-8.00	*250*	0.00	0.00	1.00	4.00	2.00
500	0.00	0.00	0.00	0.00	-4.00	*500*	0.00	0.00	0.00	0.00	4.00
1000	0.00	0.00	0.00	1.00	2.00	*1000*	0.00	0.00	0.00	0.00	0.00
2500	0.00	0.00	0.00	0.00	2.00	*2500*	0.00	0.00	0.00	0.00	0.00
5000	0.00	0.00	1.00	0.00	0.00	*5000*	0.00	0.00	0.00	0.00	0.00

Table 6. Additional results for PP test

T=500 N=100 Fact=50						T=25 N=40 Fact=20						T=100 N=100 Fact=50							
Level/Lag	*1*	*2*	*3*	*5*	*10*	*Level/Lag*	*1*	*2*	*3*	*4*	*5*	*Level/Lag*	*1*	*2*	*3*	*5*	*10*	*15*	*18*
99.99%	44	39	37	37	35	*99.99%*	24	23	22	***20***	19	*99.99%*	98	51	51	51	51	51	44
99%	45	41	39	38	36	*99%*	24	23	22	21	19	*99%*	98	51	51	51	51	51	44
95%	46	41	39	39	37	*95%*	24	23	22	21	19	*95%*	98	51	51	51	51	51	44
90%	46	42	40	39	37	*90%*	24	23	22	21	19	*90%*	98	52	52	52	52	52	44
85%	46	42	40	39	38	*85%*	24	23	22	21	19	*85%*	98	52	52	52	52	52	44
80%	46	42	40	40	38	*80%*	24	23	22	21	19	*80%*	98	52	52	52	52	52	44
75%	47	42	40	40	38	*75%*	24	23	22	21	19	*75%*	98	52	52	52	52	52	44
70%	47	43	41	40	38	*70%*	24	23	22	21	***20***	*70%*	98	52	52	52	52	52	44
60%	47	43	41	40	38	*60%*	24	23	22	21	***20***	*60%*	98	52	52	52	52	52	44
50%	47	43	41	41	39	*50%*	24	23	22	21	***20***	*50%*	98	52	52	52	52	52	44
25%	48	44	42	41	39	*25%*	24	23	22	21	***20***	*25%*	98	52	52	52	52	52	44
10%	49	45	42	42	40	*10%*	24	23	22	21	***20***	*10%*	98	52	52	52	52	52	44
1%	***50***	46	44	43	42	*1%*	24	23	22	21	***20***	*1%*	98	52	52	52	52	52	44

Note: In bold, italics and are shaded where the test resulted in the correct number of factors. ***Fact*** is number of factors.

Table 7. Cointegración tests in the portfolio *M*

Variables	Lag	Fact	Rank	Trace test	Prob	Max test	prob	Trace test	prob	Max test	prob
6	1	3	3	34.33	[0.060]	25.97	[0.012]*	34.12	[0.064]	25.81	[0.013]*
6	2	3	3	27.51	[0.267]	20.43	[0.089]	27.18	[0.284]	20.18	[0.097]
6	3	3	3	29.42	[0.185]	22.53	[0.044]*	28.89	[0.206]	22.12	[0.051]

Table 8. Test ADF on Common Factors

Statistics	Fact1	Fact2	Fact3
t-ADF	-1.5060	-1.9130	-2.7630
D-lag	2	2	2
AIC	0.0000	0.0000	0.0343

Table9. Parameters estimated and statistics significance

	Sigma		Loading Matrix						START loadings			TRUE loadings		
Variables	parameters	t-value	β(•,1)	t-value	β(•,2)	t-value	β(•,3)	t-value	Fact1	Fact2	Fact3	Fact1	Fact2	Fact3
Var1	1.5222	9.6732	-0.035	-0.3216	0.0526	0.3112	0.8883	6.3469	-0.3278	0.0755	1	0	0	1
Var2	0.8391	2.9653	-0.006	-0.047	0.0346	0.2655	0.4756	3.2234	-0.1651	0.0383	0.5061	0	0	0.5
Var3	1.4767	13.707	0.3481	3.4737	0.8864	5.7134	0.0742	0.597	0.0682	1	-0.0523	0.25	1	0
Var4	1.4590	12.988	0.5456	2.8362	0.6383	4.4934	0.0622	0.4961	0.3749	0.5552	0.0634	0.5	0.75	0
Var5	1.3841	11.49	0.739	5.8397	0.3847	2.3846	0.0255	0.1558	0.6837	0.1117	0.1653	0.75	0.5	0
Var6	1.3459	10.584	0.9418	5.0997	0.1303	1.9478	0.0038	0.0323	1	-0.3388	0.2675	1	0.25	0

Note: In order to facilitate the comparison, we present the initial weights (resulting from the eigenvectors of the autocovariance matrix with one lag), the estimates after the EM algorithm, and remember the true weights that generated the series.

Table 10. Mean Absolute Error and Root Mean Square Error of estimations

ESTIMATIONS	Factor 1		Factor 2		Factor 3	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
START	16.1086	19.3675	11.1100	12.9020	6.8305	7.9816
ESTIMATED-PP	14.6399	16.2678	4.3636	5.4261	1.1430	1.4386

FIGURES

Figure 1a. Start, Estimated (EM) and True Factor-1

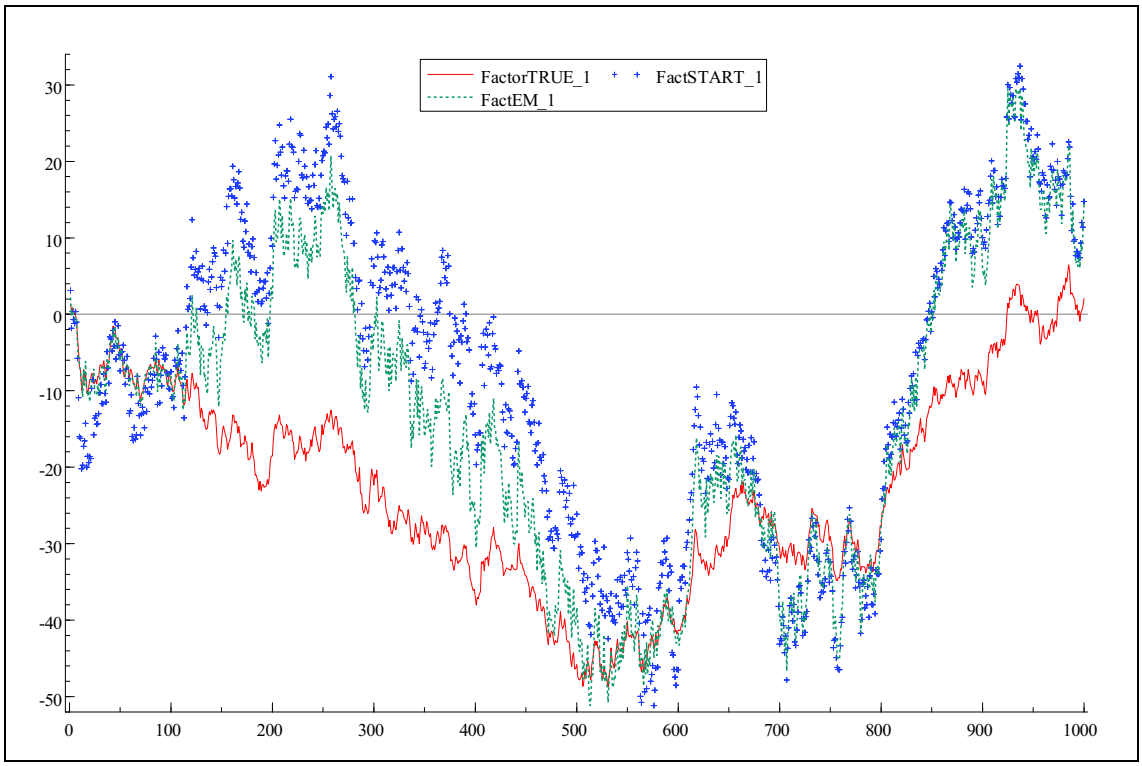

Figure 1b. Start, Estimated (EM) and True Factor-2

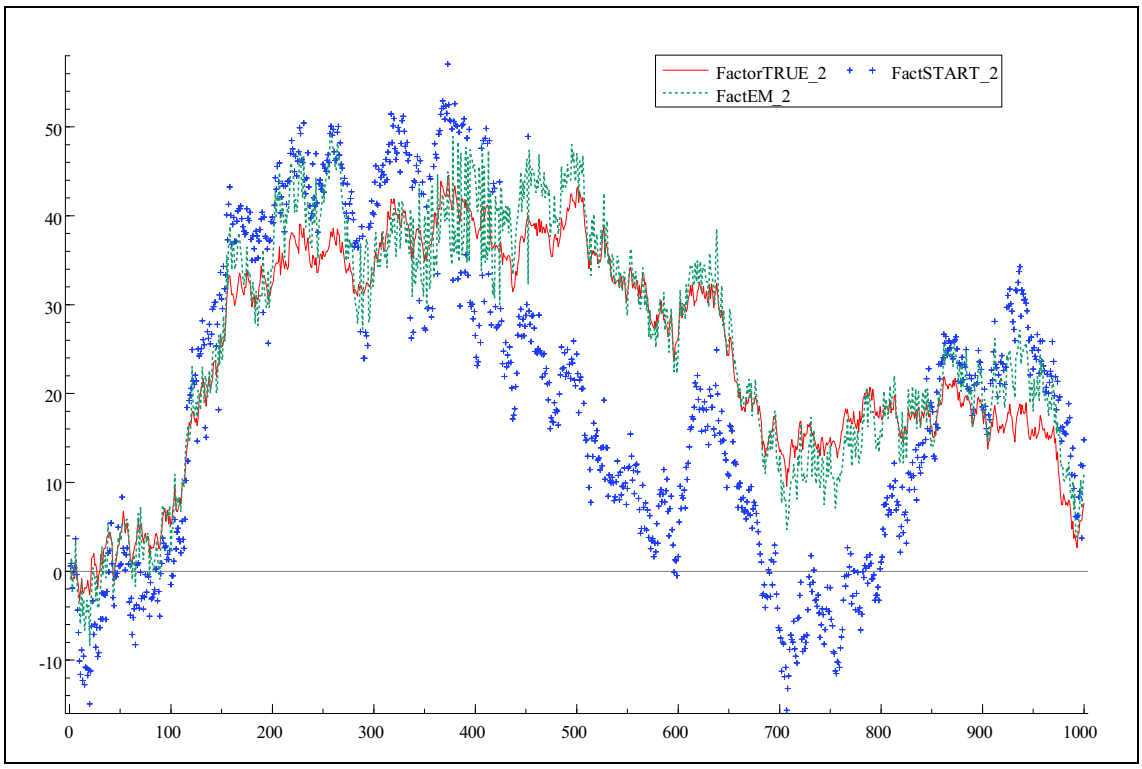

Figure 1c. Start, Estimated (EM) and True Factor-3