Dynamic Factor Analysis with STATA

Alessandro Federici*

Department of Economic Sciences

University of Rome *La Sapienza*

alessandro.federici@uniroma1.it

Andrea Mazzitelli

Department of National Accounts and Social Processes Analysis

University of Rome *La Sapienza*

mazzitelliandrea@libero.it

Abstract

The paper develops a procedure able to implement the Dynamic Factor Analysis in STATA: this methodology manages to combine, from a descriptive point of view (not probabilistic), the cross-section analysis through Principal Component Analysis and the time series dimension of data through linear regression model. An *ad hoc* step by step procedure is presented in order to highlight the wide range of field of application where this (not well known yet in the Anglo-Saxon literature) statistical framework may be successfully applied.

1. Introduction

The aim of the paper is to develop a procedure able to implement the Dynamic Factor Analysis (DFA henceforth) in STATA. DFA is a statistical multiway analysis technique[1], where quantitative "units x variables x times" arrays are considered:

$$\boldsymbol{X(I,J,T)} = \{x_{ijt}\},\ i=1\ldots I,\ j=1\ldots J,\ t=1\ldots T,$$

where i is the unit, j the variable and t the time.

Broadly speaking, this kind of methodology manages to combine, from a descriptive point of view (not probabilistic), the cross-section analysis through Principal Component Analysis (PCA henceforth) and the time series dimension of data through linear regression model.

The paper is organized as follows: firstly, a theoretical overview of the methodology will be provided in order to show the statistical and econometric framework that the procedure of STATA will implement.

---

* Referring author.

[1] That is a methodology where three or more indexes are simultaneously analysed.

Secondly, an *ad hoc* step by step procedure will be presented in order to implement through STATA the statistical methodology of DFA, with the goal of providing an overall innovation index for the OECD countries.

Finally, some concluding remarks will be drawn in order to highlight the wide range of field of application where this (not well known yet in the Anglo-Saxon literature) statistical framework may be successfully applied.

2. A theoretical overview

The DFA framework has been introduced and developed by Coppi and Zannella (1978), and then re-examined by Coppi et al. (1986) and Corazziari (1997): in this paper the original approach will be followed.

The goal of the methodology is to decompose the variance and covariance matrix ***S*** relative to ***X(IT,J)***, where the role of the units is played by the pair "units-times". The matrix ***S***, concerning the ***JxT*** observations over the *I* units, may be decomposed into the sum of three distinct variance and covariance matrices:

$$\boldsymbol{S} = {}^{*}\boldsymbol{S}_I + {}^{*}\boldsymbol{S}_T + \boldsymbol{S}_{IT}, \qquad (1)$$

where:

${}^{*}\boldsymbol{S}_I$ = matrix of the static structure of the units = matrix of variance and covariance of the average of the units with respect to time. It reflects the variability of the relational structure of the units, independently from the time dimension.

${}^{*}\boldsymbol{S}_T$ = matrix of the average dynamic of the system = variance and covariance matrix of the average of the times. It mirrors the variability, due to the time dimension, of the average of the units, independently from the dynamic of the single units.

$\boldsymbol{S}_{IT}$ = matrix of the differential dynamic of the single units = variance and covariance matrix of the interactions between units and times. It reflects the variability due to the difference between the dynamic of the overall average of the units, that is the average dynamic, and the dynamic of the single units.

On the basis of the fundamental decomposition of total variability (1), the generic element x_{ijt} may be considered as the sum of four distinct components:

$$x_{ijt} = \bar{x}_{\bullet j \bullet} + \left(\bar{x}_{ij\bullet} - \bar{x}_{\bullet j \bullet}\right) + \left(\bar{x}_{\bullet jt} - \bar{x}_{\bullet j \bullet}\right) + \left(x_{ijt} - \bar{x}_{ij\bullet} - \bar{x}_{\bullet jt} + \bar{x}_{\bullet j \bullet}\right), \qquad (2)$$

where:

$\bar{x}_{\bullet j \bullet}$ = overall average of the single variable;

$\left(\bar{x}_{ij\bullet} - \bar{x}_{\bullet j\bullet}\right)$= effect due to the static structure of the units;

$\left(\bar{x}_{\bullet jt} - \bar{x}_{\bullet j\bullet}\right)$= effect due to average dynamic;

$\left(x_{ijt} - \bar{x}_{ij\bullet} - \bar{x}_{\bullet jt} + \bar{x}_{\bullet j\bullet}\right)$= effect due to the differential dynamic, that is the interaction between units and times.

The relation (2) represents a two-factor model for the variance analysis: the model that will be implemented in the empirical section of the work, the so-called Model 1 of the DFA, considers the different components of (2) and the relative elements of total variability (1) in terms of PCA and a linear regression model.

Model 1 of DFA is based on the following decomposition of total variability into two components:

$$\boldsymbol{S} = (^{*}\boldsymbol{S}_I + \boldsymbol{S}_{IT}) + {}^{*}\boldsymbol{S}_T = \boldsymbol{S}_T + {}^{*}\boldsymbol{S}_T, \qquad (3)$$

where $\boldsymbol{S}_T$ is the average dispersion matrix within times (*within* variability), modelled through PCA, while $^{*}\boldsymbol{S}_T$ represents the variability between times (*between* variability), modelled through a linear regression model:

$$\bar{x}_{\bullet jt} = a_j + b_j t + e_{jt}\,, j = 1\ldots J;\ t = 1\ldots T, \qquad (4)$$

where the residuals satisfy the following condition:

$$\operatorname{cov}\left(e_{jt}, e_{j't'}\right) = \begin{cases} w_j & j = j';\ t = t' \\ 0 & \text{otherwise} \end{cases}. \qquad (5)$$

Condition (5) has to be taken into consideration because the relationship between the j variables has to be explained in this model only by the factorial part, that is by PCA relative to $\boldsymbol{S}_T$ matrix: the average dynamic of the system is distinct from the average dynamic of the single variables.

Relation (3) is a contraction of the fundamental decomposition (1), due to the aggregation of two sources of variability: the static structure and the differential dynamic. In order to asses the explaining capability of the model, we may take into account two indicators able to measure the variability explained by each of the fundamental components described above:

- $I(t)$: quality of the representation of the factorial structure at time t; it assesses how well is modelled each considered year;
- I_T: it assesses the overall quality of the representation of the model with respect the variability of matrix $\boldsymbol{S}_T$.

3. A procedure for the computation of an innovation index

In this section of the work a step by step procedure for the computation of an overall innovation index for 13 OECD countries[2] in the period 1992-2000 will be presented.

In particular, three indicators[3] will be considered:

- Potential innovation (var1): investment in knowledge (% of GDP).
- Quantitative innovation (var2): number of patents (% of total population).
- Qualitative innovation (var3): high-technology exports (% of total manufacturing exports).

Data must be filled into the data editor as in the case of panel data: firstly all the observation of time t_0, then the ones of time t_1, etc.; of course the order of the units must be the same for all times: in this case we have for each variable an array of 13*9 = 117 observations.

The first step of the procedure is the standardization of the considered variables:

```
foreach  x of varlist  var1 var2 var3 {
    egen z`x' = std(`x')
    mkmat z`x'
}
```

Then the standardized variables are stored in a matrix A:

```
matrix A = zvar1, zvar2, zvar3
```

Second step is the computation of $\boldsymbol{S}_T$, the average dispersion matrix within times (*within* variability):

$$\boldsymbol{S}_T = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{S}(t)\,, \qquad (6)$$

where $\boldsymbol{S}(t)$ is the variance and covariance matrix at time t. The dispersion within times jointly mirrors two effects: the effect due to the static structure of data and the one due to the differential dynamic of the single units. Matrix $\boldsymbol{S}_T$ is computed through the following commands:

```
matrix ST=J(3,3,0)

forvalues i=1(13)117 {
    matrix C=A[`i'..(`i'+13-1),1...]
    svmat C
    matrix accum cov = C1-C3, deviations noconstant
    matrix cov=cov/(r(N)-1)
    matrix ST=ST+cov
    drop C1-C3
}
```

[2] Australia, Canada, Finland, France, Germany, Italy, Japan, Netherlands, Spain, Sweden, Switzerland, United Kingdom and United States.
[3] Taken from OECD Factbook 2005: www.oecd.org.

```
mat list ST

symmetric ST[3,3]
             C1           C2           C3
C1   1.0043003
C2   .57612239   1.0039619
C3   .58495502   .63996449   1.0378201
```

Once derived $\boldsymbol{S}(t)$, its eigenvalues and the relative eigenvectors have to be calculated:

```
matrix symeigen eigenvectors eigenvalues = ST

mat list eigenvalues

eigenvalues[1,3]
             e1           e2           e3
r1    2.216949   .44864888   .38048445
```

In this case there is only one eigenvalues greater than one. The explained variability by each of them is given by:

```
mat D= diag(eigenvalues)

mat explained_variability = eigenvalues/trace(D)

mat list explained_variability

explained_variability[1,3]
             e1           e2           e3
r1   .72780337   .14728718   .12490945
```

So the first eigenvector associated to the first eigenvalues embodies more than the 70% of total variability of matrix $\boldsymbol{S}_T$. The cumulative proportion of explained variability is given by:

```
mat cumulative_variability = explained_variability

forvalues i=2/3 {
    mat cumulative_variability[1,`i'] = cumulative_variability[1,`i'] + cumulative_variability[1,`i'-1]
}

mat list cumulative_variability

cumulative_variability[1,3]
             e1           e2           e3
r1   .72780337   .87509055            1
```

The eigenvectors of the matrix $\boldsymbol{S}_T$ are the following:

```
mat list eigenvectors

eigenvectors[3,3]
             e1           e2           e3
C1    .56063818    .82671191   -.04724664
C2    .57874153   -.35039072    .73639975
C3    .59223566   -.44019741   -.67489493
```

The generic unit *i* may be represented in the common factorial space through its scores on the derived axes:

$$c_{ih} = \left(\bar{z}_i - \bar{z}. \right)' \cdot \boldsymbol{a}_h , \qquad (7)$$

where $\bar{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_{it}$, $i = 1 \ldots I$, $\bar{z}. = \frac{1}{I}\sum_{i=1}^{I} \bar{z}_i$ and $z_{it}' = \left(z_{i1t}, ..., z_{iJt} \right)$, $i = 1 \ldots I$, $t = 1 \ldots T$.

The vector of overall averages $\bar{z}.$ is given by:

```
mat overall_average= J(1,1,0)

foreach i of varlist  zvar1 zvar2 zvar3{
     mat Q=diag( `i')
     mat overall_average = overall_average,trace(Q)/rowsof(Q)
}

mat overall_average = overall_average[1...,2...]

mat list overall_average

overall_average[1,3]
          c2            c3            c4
r1   1.688e-09   7.323e-10  -5.453e-09
```

Of course the overall average of the variables is equal to zero because of the initial standardization.

The vector of the average of each unit in the sample $\bar{z}_i$ is given by:

```
mat zi=J(1,3,0)

forvalues i = 1/13 {
     mat z=J(1,3,0)
     forvalues k = `i'(13)117 {
          mat z =z+A[`k',1...]
     }
     mat z=z/7
     mat zi= zi \z
}

mat zi = zi[2...,1...]

mat list zi

zi[13,3]
              c1            c2            c3
r105  -.36437429  -1.3176396  -1.2938373
r93    .70310751  -1.1905136  -1.1370869
r94    .49662875  -.59518197   .70766644
r95   -.15924507  -.05556321  -.43400007
r96   -.26855738   -.6270179   .52494324
r97   -2.0418458  -1.3863045  -1.2279188
r98    .02294206   1.2751781   1.139758
r99   -.14709928   .2436944   .03185565
r100  -2.1147206  -1.4738533  -1.5524952
r101   1.3711271   .18957337   1.2259558
```

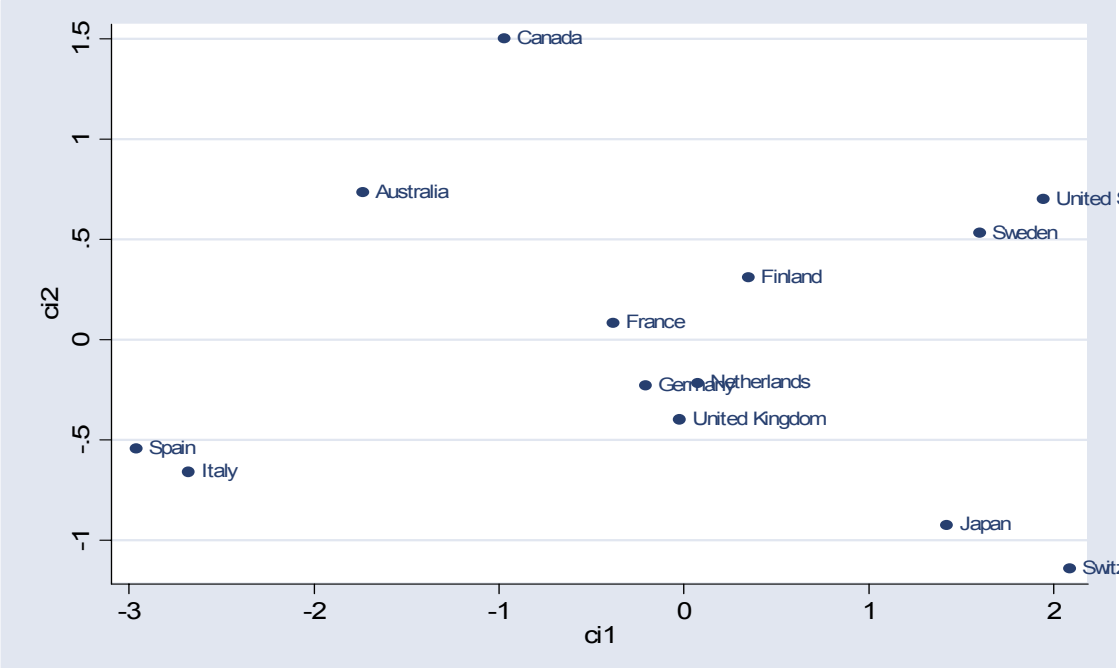
```
r102   .25371259    1.143059    2.170599
r103  -.40216129   1.0698365  -.69744799
r104   1.6140432   1.7192891   .07921359
```

The final computing of the scores is given by:

```
mat ci=J(1,3,0)

forvalues i=1/13 {
     mat cih=(zi[`i',1...]-overall_average)*eigenvectors
     mat ci = ci\cih
}

mat ci = ci[2...,1...]

mat list ci

ci[13,3]
          c1           c2           c3
r1  -1.7331115   .72999996  -.07988977
r1  -.96823416    1.498955   -.1424992
r1   .35307782    .3076022  -.93935639
r1  -.37846592   .07886474   .25951151
r1  -.20255471  -.23339699  -.80332892
r1   -2.674266  -.66174332  -.09568771
r1   1.4258661  -.92956265   .16873999
r1   .07743265  -.22101976   .16490722
r1  -2.9580162  -.54843588   .06233938
r1   1.6044749   .52743983  -.75257069
r1   2.0892828  -1.1462621  -.63516497
r1  -.01936173  -.40031752   1.2775322
r1   1.9468314   .69705619   1.1363651
```

The derived scores allow us to represent the static structure of the units:

```
svmat ci

twoway (scatter ci2 ci1, mlabel(country))
```



The diferential dynamic of the units is expressed by the "trajectories" traced by each unit on the common factorial space: they are given by the scores at each time *t*:

$$c_{iht} = \left(z_{it} - \bar{z}_{.t} \right)' \cdot \boldsymbol{a}_h \text{, } h = 1 \ldots k, \, t = 1 \ldots T, \qquad (8)$$

where $\bar{z}_{.t} = \frac{1}{I} \sum_{i=1}^{I} z_{it}$. The matrix $\bar{z}_{.t}$ of the average of the variables for each time *t* is given by:

```
mat zt = J(1,3,0)

forvalues t=1(13)117{
     mat z2=J(1,3,0)
     local t2=`t'+12
     forvalues k=`t'/`t2' {
          mat z2= z2+A[`k',1...]
     }
     mat z2= z2/13
     mat zt=zt\z2
}

mat zt=zt[2...,1...]

mat list zt

zt[9,3]
          c1           c2           c3
r13  -.23253518  -.31561093  -.27603422
r26  -.18675477   -.2478984  -.26258295
r39  -.21945506  -.20847082  -.15207311
r52  -.18021475  -.14932952  -.07305269
r65  -.09519413  -.06447453   .06276746
r78  -.04287366   .03409439   .15785342
r91   .11408752   .16351963   .19599714
r104   .29720904   .31351571   .16814389
r117   .54573101   .47465448    .178981
```

The computation of the trajectories of each unit is given by:

```
forvalues i = 1/13 {
     mat unit`i' = J(1,3,0)
     forvalues k = `i'(13)117 {
          mat cht = A[`k',1...]
          mat unit`i' = unit`i' \ cht
      }
     mat unit`i' = unit`i'[2...,1...]
     mat unit`i'= (unit`i'-zt)*eigenvectors
     mat list unit`i'
}

unit1[9,3]
          e1           e2           e3
r13   -1.034368   .67919196  -.11281426
r26  -1.1414165   .49211058  -.08860214
r39  -1.0445375   .57488974   .01781159
r52  -1.1526866   .60176445    .0352827
r65  -1.2825544   .69241238   .08038011
r78  -1.4126207   .60867353  -.02135901
```

```
r91  -1.5644423    .6199642  -.10268546
r104  -1.6797483   .47400546  -.14902313
r117  -1.8194063   .36698745  -.21821881

...

unit13[9,3]
          e1            e2            e3
r13  1.9338931  .58829222  1.1275012
r26  1.7309441  .56131935  1.0157729
r39   1.656433  .55027825  1.0457995
r52  1.7034963  .67055284  .94033538
r65  1.8528796  .66913385  .88860325
r78  1.8726567  .80361414  .98476942
r91  1.7937259  .67278811   1.081159
r104  1.7704496  .56947539  1.0735729
r117  1.8054068  .85680889  .97912953
```

Because the power of explanation of the total variability of the first component is satisfactory, the scores of each unit for each time *t* on the first axis provide the wanted overall innovation index. For a graphical representation of the trajectories a trend variable is needed:

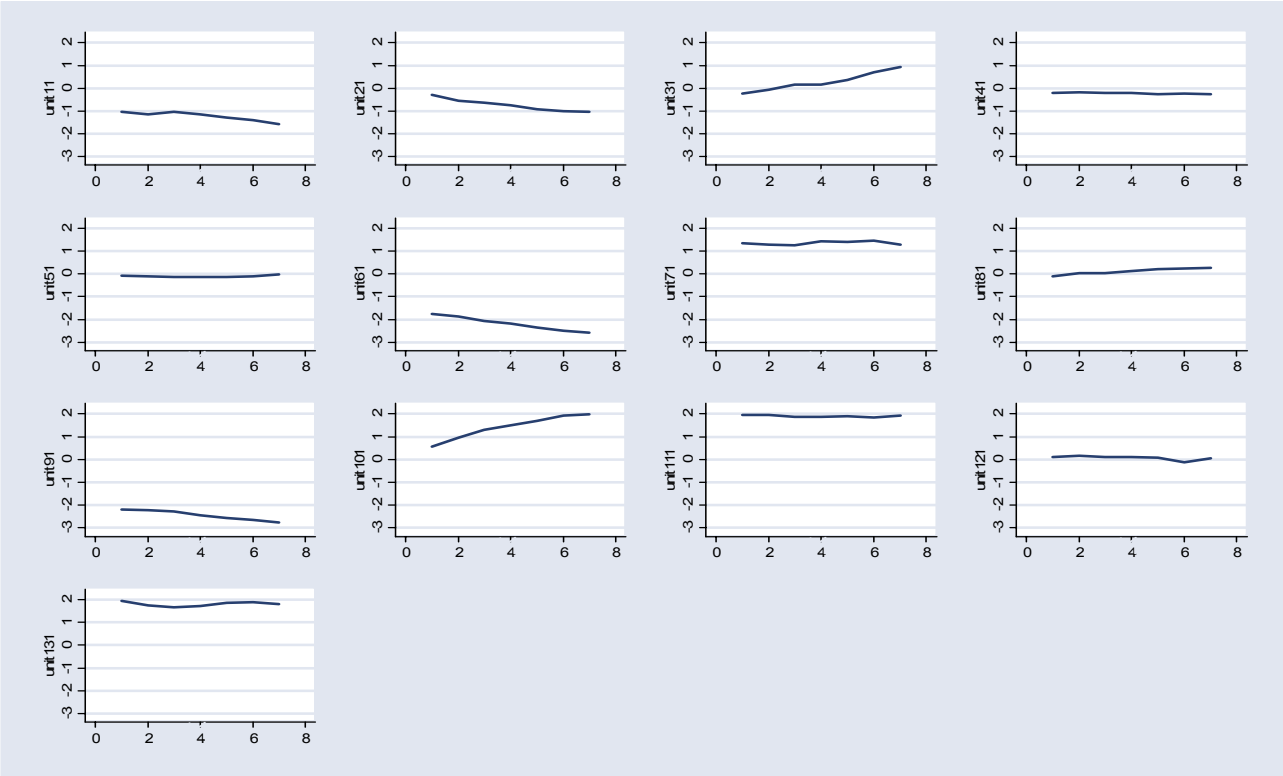
```
mat t=J(9,1,1)

forvalues i=2/9 {
    mat t[`i',1]= t[`i',1]+ t[`i'-1,1]
}

svmat t
```

Now is possible to compile a graph for the trajectory of each country:

```
forvalues i = 1/13 {
    twoway (line unit`i'1 t), xtitle(, size(zero) color(ltbluishgray)) legend(off) nodraw
    graph save unit`i', replace
}
```

The combined graph is the following:

```
gr combine unit1.gph unit2.gph unit3.gph unit4.gph unit5.gph unit6.gph unit7.gph unit8.gph
unit9.gph unit10.gph unit11.gph unit12.gph unit13.gph, xcom ycom
```



The best performances are for units 7 (Japan), 10 (Sweden), 11 (Switzerland) and 13 (United States), which show the highest values on the first components for all the considered times.

As regards the average dynamic of the system we consider the matrix $*\boldsymbol{S}_T$, regressing over time the average value of the single variables for each time t[4]:

```
svmat zt

forvalues i=1/3 {
     regress zt`i' t
}
```

```
Source       |       SS           df       MS                  Number of obs   =         7
-------------+----------------------------------                F(  1,     5)   =     22.52
Model        |  .075285295         1  .075285295               Prob > F        =    0.0051
Residual     |  .016711732         5  .003342346               R-squared       =    0.8183
-------------+----------------------------------                Adj R-squared   =    0.7820
Total        |  .091997027         6  .015332838               Root MSE        =    .05781

------------------------------------------------------------------------------
zt1          |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
t1           |   .0518533   .0109256     4.75   0.005      .023768    .0799385
_cons        |   -.327833   .0488609    -6.71   0.001     -.4534341   -.202232
------------------------------------------------------------------------------
```

The high value of the (adjusted) R-squared show that the average dynamic of the system is well captured by the proposed model.

[4] In order to save space, only the result of the regression relative to the first variable is reported. The results of the other two regressions are available upon request and their quality (proxied by the value of the (adjusted) R-

As regards the quality of the representation of the factorial structure with respect the observed overall variability for each time *t*, we can compute the following index:

$$I(t) = \frac{\sum_{h=1}^{k} \boldsymbol{a}'_h \, \boldsymbol{S}(t) \, \boldsymbol{a}_h}{\mathrm{tr}\left[\boldsymbol{S}(t)\right]}. \qquad (9)$$

For a global assessment of the esplicative power of the Model 1 of Dynamic Factor Analysis with respect the variability of the matrix $\boldsymbol{S}_T$, we can compute the following index:

$$I_T = \frac{\sum_{h=1}^{k} \boldsymbol{a}'_h \, \boldsymbol{S}_T \, \boldsymbol{a}_h}{\mathrm{tr}\left[\boldsymbol{S}_T\right]}. \qquad (10)$$

As regards the first series of indexes, we can recall the initial procedure adopted for the computation of the matrix $\boldsymbol{S}_T$. First of all the computation of the single eigenvectors is needed:

```
forvalues i=1/3 {
    mat e`i' = J(3,1,0)
}

forvalues i=1/3 {
    forvalues j=1/3 {
        mat e`i'[`j',1]=eigenvectors[`j',`i']
    }
}
```

Then we can compute the series of indexes $I(t)$:

```
local t=1

matrix ST=J(3,3,0)

forvalues i=1(13)117 {
    matrix num = J(1,1,0)
    matrix C=A[`i'..(`i'+13-1),1...]
    svmat C
    matrix accum cov = C1-C3, deviations noconstant
    matrix cov=cov/(r(N)-1)
    matrix ST=ST+cov
    drop C1-C3
    forvalues k = 1/3 {
        mat num = num + (e`k'' * cov * e`k')
    }
    mat I_`t' = num/trace(cov)
    mat list I_`t'
    local t =`t'+1
}
```

Then for the global index I_T:

```
mat ST=ST/7
```

```
matrix numIT = J(1,1,0)

forvalues k=1/3 {
    mat numIT = numIT + (e`k'' * ST  * e`k')
}

mat IT = (numIT)/trace(ST)

mat list IT
```

All the computed indexes we derive show the highest value of 1: it is likely that it is due to the low number of the considered units, variables and years in this instructive application. Anyway we can conclude that the fit of the Model 1 of Dynamic Factor Analysis to the considered data is highly satisfactory and provide useful insights about the innovative performance of the sampled countries over time when the three considered variables are jointly analysed.

4. Conclusions

Regardless the specific economic field chosen for the application in this paper, the Dynamic Factor Analysis may be applied in the case of a higher number of units, variables and years. The presented methodology is a powerful statistical framework for the analysis of three-dimensions arrays, that is the assessment of the behaviour of a sample of units over time when a given number of correlated variables are jointly considered.

For this reason, even if the Dynamic Factor Analysis is a not well known technique yet (at least in the Anglo-Saxon literature), its application in a wide range of sectors is growing, as for example in the economic, biomedical and environmental fields.

References

Coppi, R. (1986). *Analysis of three-way data matrices based on pairwise relation measures*. In *Compstat 1986 – Proceedings in computational statistics*, Physica-Verlag, Wien.

Coppi, R. (1988). *Simultaneous analysis of a set of multiway contingency tables*. In *Data Analysis and Informatics, V* (E. Diday ed.), North Holland, Amsterdam.

Coppi, R.; Di Ciaccio, A. (1994). *Multiway data analysis: software and application*. Special issue of *Computational statistics and data analysis*, vol. 18, North Holland, Amsterdam.

Coppi, R; Bolasco, S. (1989). *Multiway data analysis*. North Holland, Amsterdam.

Coppi, R; Zannella, F. (1978). *L'analisi fattoriale di una serie temporale multipla relative allo stesso insieme di unità statistiche*. In Atti della XXIX Riunione Scientifica della SIS, Bologna.

Corazziari, I. (1997). *Dynamic factor analysis*. In *Atti del convegno dell'IFCS, sezione italiana* (Pescara, 3-4 luglio 1997).

Gifi, F. (1990). *Nonlinear multivariate analysis*. J. Wiley, New York.

Law, H.G.; Snyder, C.W. Jr; Hattie, J.A.; McDonald, R.P. (1984). *Researh methods for multimode data analysis*. Preger, New York.