Dynamic Factor Analysis with STATA

Alessandro Federici

Department of Economic Sciences

University of Rome *La Sapienza*

alessandro.federici@uniroma1.it

Abstract

The aim of the paper is to develop a procedure able to implement 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.

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 $\boldsymbol{S}$ relative to $\boldsymbol{X(IT,J)}$, where the role of the units is played by the pair "units-times". The matrix $\boldsymbol{S}$, concerning the $\boldsymbol{JxT}$ 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:

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

$^*\boldsymbol{S}_I$ = 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 time.

$^*\boldsymbol{S}_T$ = 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 basic 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: so 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 some indicators able to measure the variability explained by each of the fundamental components described above:

- $*I_t$: share of the regression system variability with respect to overall variability;
- $I(t)$: quality of the representation of the factorial structure at time *t*; it assess how well is modelled each considered year;
- $*I_I$: quality of the factorial representation of the static structure of units;
- I_{IT}: quality of the factorial representation of the differential dynamic of units.

Secondly, an *ad hoc* procedure will be developed in order to implement through STATA the statistical methodology of DFA: starting from the commands of PCA and linear regression model already available in the software package, the program will develop new commands for the estimation of the model and a comprehensive set of post-estimation indicators about the reliability of the results of the assessed statistical framework.

Thirdly, a macroeconomic empirical application of DFA will be presented, 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.

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.