DYNAMIC AND COMPROMISE FACTOR ANALYSIS

Marianna Bolla
Budapest University of Technology and Economics
marib@math.bme.hu
Many parts are joint work with
Gy. Michaletzky, Loránd Eötvös University
and
G. Tusnády, Rényi Institute, Hung. Acad. Sci.

March 5, 2009

Motivation

- Having multivariate time series, e.g., financial or economic data observed at regular time intervals, we want to describe the components of the time series with a smaller number of uncorrelated factors.
- The usual factor model of multivariate analysis cannot be applied immediately as the factor process also varies in time.
- There is a dynamic part, added to the usual factor model, the auto-regressive process of the factors.
- Dynamic factors can be identified with some latent driving forces of the whole process. Factors can be identified only by the expert (e.g., monitary factors) .

Motivation

- Having multivariate time series, e.g., financial or economic data observed at regular time intervals, we want to describe the components of the time series with a smaller number of uncorrelated factors.
- The usual factor model of multivariate analysis cannot be applied immediately as the factor process also varies in time.
- There is a dynamic part, added to the usual factor model, the auto-regressive process of the factors.
- Dynamic factors can be identified with some latent driving forces of the whole process. Factors can be identified only by the expert (e.g., monitary factors) .

Motivation

- Having multivariate time series, e.g., financial or economic data observed at regular time intervals, we want to describe the components of the time series with a smaller number of uncorrelated factors.
- The usual factor model of multivariate analysis cannot be applied immediately as the factor process also varies in time.
- There is a dynamic part, added to the usual factor model, the auto-regressive process of the factors.
- Dynamic factors can be identified with some latent driving forces of the whole process. Factors can be identified only by the expert (e.g., monitary factors) .

Motivation

- Having multivariate time series, e.g., financial or economic data observed at regular time intervals, we want to describe the components of the time series with a smaller number of uncorrelated factors.
- The usual factor model of multivariate analysis cannot be applied immediately as the factor process also varies in time.
- There is a dynamic part, added to the usual factor model, the auto-regressive process of the factors.
- Dynamic factors can be identified with some latent driving forces of the whole process. Factors can be identified only by the expert (e.g., monitary factors) .

Remarks

- The model is applicable to weakly stationary (covariance-stationary) multivariate processes.
- The first descriptions of the model is found in J. F. Geweke, International Economic Review 22 (1977) and in Gy. Bánkövi et. al., Zeitschrift für Angewandte Mathematik und Mechanik 63 (1981).
- Since then, the model has been developed in such a way that dynamic factors can be extracted not only sequentially, but at the same time. For tis purpose we had to solve the problem of finding extrema of inhomogeneous quadratic forms in Bolla et. al., Lin. Alg. Appl. 269 (1998).

The model

The input data are n-dimensional observations $\mathbf{y}(t) = (y_1(t), \ldots, y_n(t))$, where t is the time and the process is observed at discrete moments between two limits $(t = t_1, \ldots, t_2)$. For given positive integer $M < n$ we are looking for uncorrelated factors $F_1(t), \ldots, F_M(t)$ such that they satisfy the following model equations:

1. As in the usual linear model,

$$F_m(t) = \sum_{i=1}^{n} b_{mi} y_i(t), \quad t = t_1, \ldots, t_2;\ m = 1, \ldots, M. \qquad (1)$$

2. The dynamic equation of the factors:

$$\hat{F}_m(t) = c_{m0} + \sum_{k=1}^{L} c_{mk} F_m(t-k), \quad t = t_1 + L, \dots, t_2;\ m = 1, \dots, M, \tag{2}$$

where the time-lag L is a given positive integer and $\hat{F}_m(t)$ is the auto-regressive prediction of the mth factor at date t (the white-noise term is omitted, therefore we use $\hat{F}_m$ instead of F_m).

3. The linear prediction of the variables by the factors as in the usual factor model:

$$\hat{y}_i(t) = d_{0i} + \sum_{m=1}^{M} d_{mi} F_m(t), \quad t = t_1, \ldots, t_2;\ i = 1, \ldots, n. \tag{3}$$

(The error term is also omitted, that is why we use the notation $\hat{y}_i$ instead of y_i.)

The objective function

We want to estimate the parameters of te model:
$\mathbf{B} = (b_{mi})$, $\mathbf{C} = (c_{mk})$, $\mathbf{D} = (d_{mi})$
$(m = 1, \dots, M;\ i = 1, \dots, n;\ k = 1, \dots L)$
in matrix notation (estimates of the parameters c_{m0}, d_{0i} follow from these) such that the objective function

$$w_0 \cdot \sum_{m=1}^{M} \operatorname{var}(F_m - \hat{F}_m)_L + \sum_{i=1}^{n} w_i \cdot \operatorname{var}(y_i - \hat{y}_i) \tag{4}$$

is minimum on the conditions for the orthogonality and variance of the factors:

$$\operatorname{cov}(F_m, F_l) = 0, \quad m \neq l; \quad \operatorname{var}(F_m) = v_m, \quad m = 1, \dots, M \tag{5}$$

where $w_0, w_1, \dots, w_n$ are given non-negative constants (balancing between the dynamic and static part), while the positive numbers v_m's indicate the relative importance of the individual factors.

Notation

In Bánkövi et al., authors use the same weights

$$v_m = t_2 - t_1 + 1, \qquad m = 1, \ldots, M.$$

Denote

$$\bar{y}_i = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} y_i(t)$$

the sample mean (average with respect to the time) of the ith component,

$$\operatorname{cov}(y_i, y_j) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} (y_i(t) - \bar{y}_i) \cdot (y_j(t) - \bar{y}_j)$$

the sample covariance between the ith and jth components, while

$$\operatorname{cov}^*(y_i, y_j) = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} (y_i(t) - \bar{y}_i) \cdot (y_j(t) - \bar{y}_j)$$

the corrected empirical covariance between them.

The trivial parameters

The parameters c_{m0}, d_{0i} can be written in terms of the other parameters:

$$c_{m0} = \frac{1}{t_2 - t_1 - L + 1} \sum_{t=t_1+L}^{t_2} \left(F_m(t) - \sum_{k=1}^{L} c_{mk} F_m(t-k)\right),$$

$$m = 1, \ldots, M$$

and

$$d_{0i} = \bar{y}_i - \sum_{m=1}^{M} d_{mi} \bar{F}_m,$$

$$i = 1, \ldots, n.$$

Further notation

Thus, the parameters to be estimated are collected in the $M \times n$ matrices $\mathbf{B}, \mathbf{D}$, and in the $M \times L$ matrix $\mathbf{C}$.
$\mathbf{b}_m \in \mathbb{R}^n$ be the mth row of matrix $\mathbf{B}$, $m = 1, \dots, M$.

$$Y_{ij} := \operatorname{cov}(y_i, y_j), \qquad i, j = 1, \dots n,$$

and $\mathbf{Y} := (Y_{ij})$ is the $n \times n$ symmetric, positive semidefinite empirical covariance matrix of the sample (sometimes it is corrected).

Delayed time series:

$$z_i^m(t) = y_i(t) - \sum_{k=1}^{L} c_{mk} y_i(t-k), \tag{6}$$

$$t = t_1 + L, \dots, t_2; \quad i = 1, \dots, n; \quad m = 1, \dots, M$$

and

$$Z_{ij}^m := \operatorname{cov}\left(z_i^m, z_j^m\right) =$$

$$= \frac{1}{t_2 - t_1 - L + 1} \sum_{t=t_1+L}^{t_2} \left(z_i^m(t) - \bar{z}_i^m\right) \cdot \left(z_j^m(t) - \bar{z}_j^m\right), \tag{7}$$

$$i, j = 1, \dots n,$$

where $\bar{z}_i^m = \frac{1}{t_2 - t_1 - L + 1} \sum_{t=t_1+L}^{t_2} z_i^m(t)$, $i = 1, \dots, n$; $m = 1, \dots, M$.

The objective function revisited

Let $\mathbf{Z}^m = (Z_{ij}^m)$ be the $n \times n$ symmetric, positive semidefinite covariance matrix of these variables.
The objective function of (4) to be minimized:

$$G(\mathbf{B}, \mathbf{C}, \mathbf{D}) = w_0 \sum_{m=1}^{M} \mathbf{b}_m^T \mathbf{Z}^m \mathbf{b}_m + \sum_{i=1}^{n} w_i Y_{ii} -$$

$$-2 \sum_{i=1}^{n} w_i \sum_{m=1}^{M} d_{mi} \sum_{j=1}^{n} b_{mj} Y_{ij} + \sum_{i=1}^{n} w_i \sum_{m=1}^{M} d_{mi}^2 v_m,$$

where the minimum is taken on the constraints

$$\mathbf{b}_m^T \mathbf{Y} \mathbf{b}_l = \delta_{ml} \cdot v_m, \quad m, l = 1, \ldots, M. \tag{8}$$

Outer cycle of the iteration

Choosing an initial $\mathbf{B}$ satisfying (8), the following two steps are alternated:

1. Starting with $\mathbf{B}$ we calculate the F_m's based on (1), then we fit a linear model to estimate the parameters of the autoregressive model (2). Hence, the current value of $\mathbf{C}$ is obtained.
2. Based on this $\mathbf{C}$, we find matrices $\mathbf{Z}^m$ using (6) and (7) (actually, to obtain $\mathbf{Z}^m$, the mth row of $\mathbf{C}$ is needed only), $m = 1, \dots, M$. Putting it into $G(\mathbf{B}, \mathbf{C}, \mathbf{D})$, we take its minimum with respect to $\mathbf{B}$ and $\mathbf{D}$, while keeping $\mathbf{C}$ fixed.

With this $\mathbf{B}$, we return to the 1st step of the outer cycle and proceed until convergence.

Fixing **C**, the part of the objective function to be minimized in **B** and **D** is

$$F(\mathbf{B},\mathbf{D}) = w_0 \sum_{m=1}^{M} \mathbf{b}_m^T \mathbf{Z}^m \mathbf{b}_m + \sum_{i=1}^{n} w_i \sum_{m=1}^{M} d_{mi}^2 v_m -$$

$$-2 \sum_{i=1}^{n} w_i \sum_{m=1}^{M} d_{mi} \sum_{j=1}^{n} b_{mj} Y_{ij},$$

Taking the derivative with respect to **D**:

$$F(\mathbf{B},\mathbf{D}^{opt}) = w_0 \sum_{m=1}^{M} \mathbf{b}_m^T \mathbf{Z}^m \mathbf{b}_m - \sum_{i=1}^{n} w_i \sum_{m=1}^{M} \frac{1}{v_m} (\sum_{j=1}^{n} b_{mj} Y_{ij})^2.$$

Introducing $V_{jk} = \sum_{i=1}^{n} w_i Y_{ij} Y_{ik}$, $\mathbf{V} = (V_{jk})$, and

$$\mathbf{S}_m = w_0 \mathbf{Z}^m - \frac{1}{v_m} \mathbf{V}, \quad m = 1, \ldots, M$$

we have

$$F(\mathbf{B},\mathbf{D}^{opt}) = \sum_{m=1}^{M} \mathbf{b}_m^T \mathbf{S}_m \mathbf{b}_m. \tag{9}$$

Thus, $F(\mathbf{B}, \mathbf{D}^{opt})$ is to be minimized on the constraints for $\mathbf{b}_m$'s. Transforming the vectors $\mathbf{b}_1, \dots, \mathbf{b}_m$ into an orthonormal set, an algorithm to find extrema of inhomogeneous quadratic forms is to be used.

The transformation

$$\mathbf{x}_m := \frac{1}{\sqrt{v_m}} \mathbf{Y}^{1/2} \mathbf{b}_m, \quad \mathbf{A}_m := v_m \mathbf{Y}^{-1/2} \mathbf{S}_m \mathbf{Y}^{-1/2}, \; m = 1, \dots, M \tag{10}$$

will result in an orthonormal set $\mathbf{x}_1, \dots, \mathbf{x}_M \in \mathbb{R}^n$, further

$$F(\mathbf{B}, \mathbf{D}^{opt}) = \sum_{m=1}^{M} \mathbf{x}_m^T \mathbf{A}_m \mathbf{x}_m,$$

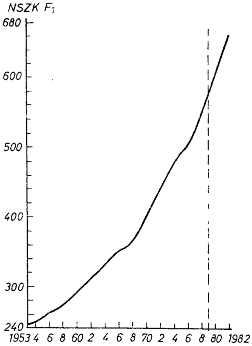
and by back transformation:

$$\mathbf{b}_m^{opt} = \sqrt{v_m} \mathbf{Y}^{-1/2} \mathbf{x}_m^{opt}, \quad m = 1, \dots, M.$$

German Federal Republic, 1953–1982

COP: Consumer Prices
INP: Industrial Production
EMP: Employment
WAG: Wages
EXP: Export
GOC: Goverment Consumption
GFC: Gross Fixed Capital
PRC: Private Consumption
IMP: Imports
GDP: Gross Domestic Product
CPS: Claims on Private Sector
DOC: Domestic Credit
POP: Population, Population

The first dynamic factor

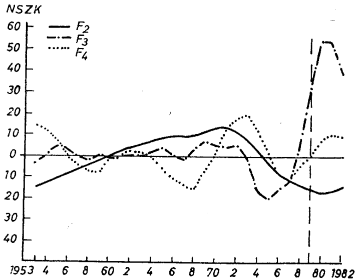


4.3. ábra. Az első dinamikus faktor

Table II.

*) Forrás: The International Financial Yearbook, Washington, 1980. Az eredeti adatokat normáltuk, mindegyik idősort saját 1975. évi értékének 100-adrészével osztottuk.

Further dynamic factors



4.4. ábra. A második, a harmadik és a negyedik dinamikus faktor

Table III.

Factors as linear combinations of variables

4.4. táblázat

Faktorok kifejezése változókkal

Változó	1. faktor	2. faktor	3. faktor	4. faktor
Konstans	–	–206.79	181.07	117.22
FAKT1	–	0.10	–0.86	–5.06
FAKT2	–	–	0.75	3.68
FAKT3	–	–	–	–1.04
COP	–	–1.21	–	–
INP	–	–	–	–
EMP	–	–	0.52	–
WAG	–	–	–1.39	–1.88
EXP	–	–	–	–0.40
GOC	–	–	–	2.18
GFC	–	–	0.54	1.07
PRC	1.51	–	–	5.93
IMP	–	–	–	–
GDP	–	–	–0.40	–
CPS	0.78	–	–	6.01
DOC	–	–	2.95	–
POP	2.60	2.75	–	–

Table IV.

Variables as linear combinations of factors

4.6. táblázat

A változók becslése a faktorokkal

i	Változó	q_i	d_{1i}	d_{2i}	d_{3i}	d_{4i}	R_i
1	COP	−0.61	0.20	−0.25	−0.09	−0.05	0.99987
2	INP	−18.77	0.26	0.68	–	–	0.99661
3	EMP	89.42	0.03	0.73	–	–	0.94562
4	WAG	−71.97	0.35	−0.27	−0.10	0.03	0.99988
5	EXP	−94.41	0.39	−0.66	−0.13	–	0.99770
6	GOC	−90.66	0.37	−0.52	−0.16	0.08	0.99968
7	GFC	−78.68	0.38	0.22	0.29	0.29	0.99861
8	PRC	−74.72	0.35	−0.26	–	−0.02	0.99990
9	IMP	−98.83	0.41	−0.73	–	–	0.99780
10	GDP	−78.59	0.37	−0.21	–	0.06	0.99945
11	CPS	−104.00	0.42	−0.34	0.15	0.11	0.99980
12	DOC	−107.35	0.43	−0.58	0.23	−0.03	0.99989
13	POP	74.87	0.05	0.25	−0.04	−0.02	0.99969

Table V.

Extrema of sums of inhomogeneous quadratic forms

Given the $n \times n$ symmetric matrices $\mathbf{A}_1, \dots, \mathbf{A}_k$ ($k \le n$) we are looking for an orthonormal set of vectors $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ such that

$$\sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{A}_i \mathbf{x}_i \to \text{maximum}.$$

Theoretical solution

By Lagrange's multipliers the $\mathbf{x}_i$'s giving the optimum satisfy the system of linear equations

$$A(\mathbf{X}) = \mathbf{XS} \tag{11}$$

with some $k \times k$ symmetric matrix $\mathbf{S}$, where the $n \times k$ matrices $\mathbf{X}$ and $A(\mathbf{X})$ are as follows:

$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_k), \quad A(\mathbf{X}) = (\mathbf{A}_1\mathbf{x}_1, \dots, \mathbf{A}_k\mathbf{x}_k).$$

Due to the constraints imposed on $\mathbf{x}_1, \dots, \mathbf{x}_k$, the non-linear system of equations

$$\mathbf{X}^T\mathbf{X} = \mathbf{I}_k \tag{12}$$

must also hold.

As $\mathbf{X}$ and the symmetric matrix $\mathbf{S}$ contain alltogether $nk + k(k+1)/2$ free parameters, while the equations (11) and (12) the same number of equations, the solution of the problem is expected. Transform (11) into a homogeneous system of linear equations, to get a non-trivial solution,

$$|\mathbf{A} - \mathbf{I}_n \otimes \mathbf{S}| = 0 \tag{13}$$

must hold, where the $nk \times nk$ matrix $\mathbf{A}$ is a Kronecker-sum $\mathbf{A} = \mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_k$ ($\otimes$ denotes the Kronecker-product). Generalization of the eigenvalue problem: eigenmatrix problem.

Numerical solution

Starting with a matrix $\mathbf{X}^{(0)}$ of orthonormal columns, the mth step of the iteration based on the $(m-1)$th one is as follows $(m = 1, 2, \dots)$:
Teke the polar decomposition

$$A(\mathbf{X}^{(m-1)}) = \mathbf{X}^{(m)} \cdot \mathbf{S}$$

into an $n \times k$ matrix of orthonormal columns and a $k \times k$ symmetric matrix . Let the first factor be $\mathbf{X}^{(m)}$, etc. until convergence.
The polar decomposition is obtained by SVD.
The above iteration is easily adopted to negative semidefinite or indefinite matrices and to finding minima instead of maxima.

COMPROMISE FACTOR ANALYSIS

A method for compromise factor extraction from covariance/correlation matrices corresponding to different strata is introduced.
Compromise factors are independent and on this constraint they explain the largest possible part of the variables' total variance over the strata.
The so-called compromise representation of the strata is introduced. A practical application for parallel factoring of medical data in different strata is also presented.

Application

In biological applications data are frequently derived from different strata, but the observed variables are the same in each of them. We would like to assign scores to the variables – different ones in different strata – in such a way that together with other strata scores they accomplish the best possible compromise between the strata.

In the case of normally distributed data the covariance matrices of the same variables are calculated in each stratum separately. In fact, the data need not be necessarily normally distributed, but it is supposed that the covariance structure somehow reflects the interconnection between the variables. One factor from each stratum is extracted.

The purpose of the compromise factor analysis is similar to that of the discriminant analysis. Here, however, we find a linear combination of the variables for each stratum that obey the orthogonality conditions.

The model

Let $\xi_1, \dots, \xi_k$ be n-dimensional, normally distributed random variables with positive definite covariance matrices $C_1, \dots, C_k$ $(k \le n)$, respectively.
Let us suppose that the mean vectors are zero (otherwise the estimated means are subtracted).

$$\xi_i = f + e_i \qquad (i = 1, \dots, k),$$

where f and e_i $(i = 1, \dots, k)$ are n-dimensional normally distributed random vector variables with zero mean vectors and covariance matrices D and B_i $(i = 1, \dots, k)$, respectively, and D is supposed to be an $n \times n$ diagonal matrix.
e_is are mutually independent of each-other and of f. The random vector variable f can be thought of as a main common factor of ξ_i's while e_i is characteristic to the ith stratum or measurement $(i = 1, \dots, k)$.

Matrix notation

Therefore, $C_i = D + B_i$ and the cross-covariance matrix $E\xi_i\xi_j^T = D$ is the same diagonal matrix with nonnegative diagonal entries for all $i \neq j$.

The observed random vectors $\xi_1, \dots, \xi_k$ may also be repeated measurements for n dependent Gaussian variables in the same population. This kind of linear model can be fitted with the usual techniques, and the maximum likelihood estimate for D is constructed on the basis of a sample taken in k not independent strata or in the case of k times repeated measurements. To test the diagonality of D a likelihood ratio test is used.

The optimum problem

Provided the model fits, we are looking for stochastically independent linear combinations $a_1^T\xi_1, \dots, a_k^T\xi_k$ of the above vector variables such that

$$\sum_{i=1}^{k} \mathrm{Var}\,(a_i^T\xi_i) = \sum_{i=1}^{k} a_i^T C_i a_i \;\rightarrow\; \text{maximum}$$

on the following constraints: the vectors a_is are standardized in such a way that $a_i^T D a_i = 1$ $(i = 1, \dots, k)$.
The constraints together with the independence conditions imply that

$$a_i^T D a_j = \delta_{ij} \qquad (i, j = 1, \dots, k).$$

Numerical algorithm

By means of the transformations $b_i := D^{1/2} a_i$ $(i = 1, \ldots, k)$, the optimization problem is equivalent to

$$\sum_{i=1}^{k} b_i^T (D^{-1/2} C_i D^{-1/2}) b_i \rightarrow \text{ maximum}$$

where the maximization is through all orthonormal systems $b_1, \ldots, b_k \in \mathbb{R}^n$.
Since the $n \times n$ matrices $D^{-1/2} C_i D^{-1/2}$ are symmetric, the algorithm constructed for inhomogeneous quadratic forms is applicable. Let $b_1^*, \ldots, b_k^*$ denote the compromise system of the matrices $D^{-1/2} C_1 D^{-1/2}, \ldots, D^{-1/2} C_k D^{-1/2}$.
Finally, by backward transformations $a_i^* = D^{-1/2} b_i^*$ the linear combinations giving the extremum are obtained.

A medical application

We applied the method for clinical measurements (protein, triglicerin and other organic matter concentration in the urine) of nephrotic patients. We distinguished between three stages of the illness : a no symptoms stage and two nephrotic stages, one of them is an intermediate stage, and in the other the illness has already seriously developed.

First, we tried to perform discriminant analysis for the three above groups, but the difference between them was not really remarkable. We obtained a poor classification, and the canonical variables best discriminating the groups providing the largest ANOVA F-statistics did not show significant difference between the groups.

Instead, our program provides a profile of the variables in each group and remarkable differences in the factor loadings can be observed even in cases when the difference of covariance/correlation matrices is not so evident.

Compromise factor loadings for three nephrotic stages

The total sample consisted of 100 patients.
The results for the three stages:

NO SYMPTOMS INTERMEDIATE NEPHROTIC

AT	-0.104339	-0.151711	-0.068392
PC	-0.151864	$+0.060398$	$+0.062981$
KO2	-0.355027	-0.662945	-0.423931
TG	-0.134190	-0.372486	$+0.781611$
HK	-0.241672	$+0.194526$	$+0.421601$
LK	$+0.496214$	-0.543357	$+0.149016$
PROT	$+0.522984$	$+0.194241$	-0.027665
URIN	-0.493607	$+0.155758$	$+0.001543$
NAK	-0.014336	$+0.005123$	$+0.001286$

Conclusions

In the characterization of the no symptoms stage the variables PROT, LK and URIN play the most important role (former ones positively, while the latter one negatively characterizes the healthy patients).
In the seriously nephrotic stage TG and HK positively, while KO2 negatively characterizes the patients.
In the intermediate stage KO2's effect is also negative (even more than in the case of seriously ill stage), while LK's effect is opposite to that of the no symptoms stage.
Thus, one may conclude that mainly measurements with high loadings in absolute value have to be considered seriously in the diagnosis.