DYNAMIC FACTOR ANALYSIS

Marianna Bolla

Budapest University of Technology and Economics

marib@math.bme.hu

Partly joint work with

Gy. Michaletzky, Loránd Eötvös University

and

G. Tusnády, Rényi Institute, Hung. Acad. Sci.

October 22, 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., monetary 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., monetary factors).
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., monetary factors).
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., monetary factors).
The model is applicable to weakly stationary (covariance-stationary) multivariate processes.


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 this 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 $y(t) = (y_1(t), \ldots, y_n(t))$, 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,

$$F_m(t) = \sum_{i=1}^{n} b_{mi} y_i(t), \quad t = t_1, \ldots, t_2; \quad m = 1, \ldots, M. \quad (1)$$
2. The dynamic equation of the factors:

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

where the time-lag \( L \) is a given positive integer and \( \hat{F}_m(t) \) is the auto-regressive prediction of the \( m \)th 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:

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

(The error term is also omitted, that is why we use the notation $\hat{y}_i$ instead of $y_i$. )
We want to estimate the parameters of the model: 
\( \mathbf{B} = (b_{mi}), \mathbf{C} = (c_{mk}), \mathbf{D} = (d_{mi}) \) 
\((m = 1, \ldots, M; i = 1, \ldots, n; k = 1, \ldots, L)\)
in matrix notation (estimates of the parameters \( c_{m0}, d_{0i} \) follow from these) such that the objective function

\[
 w_0 \cdot \sum_{m=1}^{M} \text{var} \left( F_m - \hat{F}_m \right) + \sum_{i=1}^{n} w_i \cdot \text{var} \left( y_i - \hat{y}_i \right)
\]  

(4)
is minimum on the conditions for the orthogonality and variance of the factors:

\[
 \text{cov} (F_m, F_l) = 0, \quad m \neq l; \quad \text{var} (F_m) = v_m, \quad m = 1, \ldots, M
\]  

(5)
where \( w_0, w_1, \ldots, w_n \) 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.
In Bánkövi et al., authors use the same weights
\[ v_m = t_2 - t_1 + 1, \quad m = 1, \ldots, M. \]

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 \( i \)th component,

\[ \text{cov} (y_i, y_j) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} (y_i(t) - \bar{y}_i) \cdot (y_j(t) - \bar{y}_j) \]
the sample covariance between the \( i \)th and \( j \)th components, while

\[ \text{cov}^* (y_i, y_j) = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} (y_i(t) - \bar{y}_i) \cdot (y_j(t) - \bar{y}_j) \]
the corrected empirical covariance between them.
The parameters $c_{m0}$, $d_{0i}$ can be written in terms of the other parameters:

$$c_{m0} = \frac{1}{t_2 - t_1 - L + 1} \sum_{t=t_1+L}^{t_2} (F_m(t) - \sum_{k=1}^{L} c_{mk} F_m(t - k)), \quad m = 1, \ldots, M$$

and

$$d_{0i} = \bar{y}_i - \sum_{m=1}^{M} d_{mi} \bar{F}_m, \quad i = 1, \ldots, n.$$
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 $\mathbf{C}$.

$b_m \in \mathbb{R}^n$ be the $m$th row of matrix $\mathbf{B}$, $m = 1, \ldots, M$.

$$Y_{ij} := \text{cov}(y_i, y_j), \quad i, j = 1, \ldots, n,$$

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

\[ z_{im}^m(t) = y_i(t) - \sum_{k=1}^{L} c_{mk} y_i(t - k), \quad (6) \]

\[ t = t_1 + L, \ldots, t_2; \quad i = 1, \ldots, n; \quad m = 1, \ldots, M \]

and

\[ Z_{ij}^m := \text{cov} (z_{im}^m, z_{jm}^m) = \]

\[ = \frac{1}{t_2 - t_1 - L + 1} \sum_{t=t_1+L}^{t_2} (z_{im}^m(t) - \bar{z}_{im}^m) \cdot (z_{jm}^m(t) - \bar{z}_{jm}^m), \quad (7) \]

where \( \bar{z}_{im}^m = \frac{1}{t_2 - t_1 - L + 1} \sum_{t=t_1+L}^{t_2} z_{im}^m(t), \quad i = 1, \ldots, n; \quad m = 1, \ldots, M. \)
Let $Z^m = (Z_{ij}^m)$ be the $n \times n$ symmetric, positive semidefinite covariance matrix of these variables. The objective function of (4) to be minimized:

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

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

where the minimum is taken on the constraints

$$b_m^T Y b_l = \delta_{ml} \cdot v_m, \quad m, l = 1, \ldots, M.$$  
(8)
Choosing an initial $B$ satisfying (8), the following two steps are alternated:

1. Starting with $B$ 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 $C$ is obtained.

2. Based on this $C$, we find matrices $Z^m$ using (6) and (7) (actually, to obtain $Z^m$, the $m$th row of $C$ is needed only), $m = 1, \ldots, M$. Putting it into $G(B, C, D)$, we take its minimum with respect to $B$ and $D$, while keeping $C$ fixed.

With this $B$, 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

\[
F(B, D) = w_0 \sum_{m=1}^{M} b_m^T Z^m b_m + \sum_{i=1}^{n} w_i \sum_{m=1}^{M} d_{mi}^2 v_m - \\
-2 \sum_{i=1}^{n} w_i \sum_{m=1}^{M} d_{mi} \sum_{j=1}^{n} b_{mj} Y_{ij},
\]

Taking the derivative with respect to $D$:

\[
F(B, D^{opt}) = w_0 \sum_{m=1}^{M} b_m^T Z^m b_m - \sum_{i=1}^{n} w_i \sum_{m=1}^{M} \frac{1}{v_m} \left( \sum_{j=1}^{n} b_{mj} Y_{ij} \right)^2.
\]

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

\[
S_m = w_0 Z^m - \frac{1}{v_m} V, \quad m = 1, \ldots, M
\]

we have

\[
F(B, D^{opt}) = \sum_{m=1}^{M} b_m^T S_m b_m. \tag{9}
\]
Thus, $F(B, D^{opt})$ is to be minimized on the constraints for $b_m$'s. Transforming the vectors $b_1, \ldots, b_m$ into an orthonormal set, an algorithm to find extrema of inhomogeneous quadratic forms is to be used.

The transformation

$$x_m := \frac{1}{\sqrt{v_m}} Y^{1/2} b_m, \quad A_m := v_m Y^{-1/2} S_m Y^{-1/2}, \quad m = 1, \ldots, M$$

(10)

will result in an orthonormal set $x_1, \ldots, x_M \in \mathbb{R}^n$, further

$$F(B, D^{opt}) = \sum_{m=1}^{M} x_m^T A_m x_m,$$

and by back transformation:

$$b^{opt}_m = \sqrt{v_m} Y^{-1/2} x^{opt}_m, \quad m = 1, \ldots, M.$$
Hungarian Republic, 1993–2007

VARIABLES OF THE MODEL

Gross Domestic Product (1000 million HUF) – GDP
Number of Students in Higher Education – EDU
Number of Hospital Beds – HEALTH
Industrial Production (1000 million HUF) – IND
Agricultural Area (1000 ha) – AGR
Energy Production (petajoule) – ENERGY
Energy Import (petajoule) – IMP
Energy Export (petajoule) – EXP
National Economic Investments (1000 million HUF) – INV
Number of Scientific Publications – PUBL
Figure 1. The Factor Process
## Table: Estimation of the Factors

<table>
<thead>
<tr>
<th>Time</th>
<th>factor 1.</th>
<th>factor 2.</th>
<th>factor 3.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>264.977</td>
<td>51.972</td>
<td>-45.760</td>
</tr>
<tr>
<td>1994</td>
<td>259.219</td>
<td>52.811</td>
<td>-44.986</td>
</tr>
<tr>
<td>1995</td>
<td>259.846</td>
<td>53.737</td>
<td>-45.308</td>
</tr>
<tr>
<td>1996</td>
<td>266.300</td>
<td>53.996</td>
<td>-46.514</td>
</tr>
<tr>
<td>1997</td>
<td>261.073</td>
<td>55.183</td>
<td>-45.988</td>
</tr>
<tr>
<td>1998</td>
<td>262.468</td>
<td>55.033</td>
<td>-46.456</td>
</tr>
<tr>
<td>1999</td>
<td>261.569</td>
<td>55.879</td>
<td>-46.562</td>
</tr>
<tr>
<td>2000</td>
<td>262.729</td>
<td>55.729</td>
<td>-47.168</td>
</tr>
<tr>
<td>2001</td>
<td>261.258</td>
<td>55.788</td>
<td>-47.138</td>
</tr>
<tr>
<td>2002</td>
<td>261.933</td>
<td>55.781</td>
<td>-47.337</td>
</tr>
<tr>
<td>2003</td>
<td>261.361</td>
<td>54.962</td>
<td>-47.418</td>
</tr>
<tr>
<td>2004</td>
<td>261.529</td>
<td>54.896</td>
<td>-47.736</td>
</tr>
<tr>
<td>2005</td>
<td>261.107</td>
<td>53.557</td>
<td>-47.833</td>
</tr>
<tr>
<td>2006</td>
<td>261.118</td>
<td>52.758</td>
<td>-48.254</td>
</tr>
<tr>
<td>2007</td>
<td>260.925</td>
<td>51.465</td>
<td>-48.401</td>
</tr>
</tbody>
</table>
Table: Factor Loadings (matrix $B$)

<table>
<thead>
<tr>
<th></th>
<th>factor 1.</th>
<th>factor 2.</th>
<th>factor 3.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP</td>
<td>38.324</td>
<td>-2.541</td>
<td>-6.116</td>
</tr>
<tr>
<td>EDU</td>
<td>-1.775</td>
<td>5.725</td>
<td>0.015</td>
</tr>
<tr>
<td>HEALTH</td>
<td>10.166</td>
<td>0.837</td>
<td>-1.650</td>
</tr>
<tr>
<td>IND</td>
<td>-0.261</td>
<td>0.255</td>
<td>-0.107</td>
</tr>
<tr>
<td>AGR</td>
<td>6.146</td>
<td>2.919</td>
<td>-1.124</td>
</tr>
<tr>
<td>ENERGY</td>
<td>24.082</td>
<td>4.592</td>
<td>-4.054</td>
</tr>
<tr>
<td>IMP</td>
<td>1.560</td>
<td>-1.209</td>
<td>-0.213</td>
</tr>
<tr>
<td>EXP</td>
<td>-3.907</td>
<td>-0.233</td>
<td>0.615</td>
</tr>
<tr>
<td>INV</td>
<td>2.864</td>
<td>0.038</td>
<td>-0.510</td>
</tr>
<tr>
<td>PUBL</td>
<td>-0.608</td>
<td>0.197</td>
<td>0.089</td>
</tr>
</tbody>
</table>
### Variables Estimated by The Factors (matrix D)

<table>
<thead>
<tr>
<th></th>
<th>factor 1.</th>
<th>factor 2.</th>
<th>factor 3.</th>
<th>Constant term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP</td>
<td>-0.108</td>
<td>-0.025</td>
<td>-0.677</td>
<td>-0.670</td>
</tr>
<tr>
<td>EDU</td>
<td>-0.142</td>
<td>0.145</td>
<td>-0.877</td>
<td>-8.637</td>
</tr>
<tr>
<td>HEALTH</td>
<td>0.115</td>
<td>-0.132</td>
<td>0.656</td>
<td>16.250</td>
</tr>
<tr>
<td>IND</td>
<td>-0.898</td>
<td>-0.187</td>
<td>-5.784</td>
<td>-14.690</td>
</tr>
<tr>
<td>AGR</td>
<td>0.021</td>
<td>0.005</td>
<td>0.137</td>
<td>6.809</td>
</tr>
<tr>
<td>ENERGY</td>
<td>0.085</td>
<td>-0.038</td>
<td>0.543</td>
<td>10.055</td>
</tr>
<tr>
<td>IMP</td>
<td>-0.098</td>
<td>-0.152</td>
<td>-0.868</td>
<td>0.311</td>
</tr>
<tr>
<td>EXP</td>
<td>-0.516</td>
<td>-0.931</td>
<td>-1.840</td>
<td>109.915</td>
</tr>
<tr>
<td>INV</td>
<td>-0.209</td>
<td>0.026</td>
<td>-1.341</td>
<td>-6.779</td>
</tr>
<tr>
<td>PUBL</td>
<td>-0.061</td>
<td>0.121</td>
<td>-0.484</td>
<td>-9.867</td>
</tr>
</tbody>
</table>
**Table:** Dynamic Equations of The Factors (matrix C)

<table>
<thead>
<tr>
<th>Timelag</th>
<th>factor 1.</th>
<th>factor 2.</th>
<th>factor 3.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.000</td>
<td>0.001</td>
<td>-0.000</td>
</tr>
<tr>
<td>1</td>
<td>0.069</td>
<td>0.283</td>
<td>0.117</td>
</tr>
<tr>
<td>2</td>
<td>0.473</td>
<td>1.644</td>
<td>0.495</td>
</tr>
<tr>
<td>3</td>
<td>0.205</td>
<td>0.229</td>
<td>0.141</td>
</tr>
<tr>
<td>4</td>
<td>0.251</td>
<td>-1.168</td>
<td>0.258</td>
</tr>
</tbody>
</table>
Given the \( n \times n \) symmetric matrices \( A_1, \ldots, A_k \) \((k \leq n)\) we are looking for an orthonormal set of vectors \( x_1, \ldots, x_k \in \mathbb{R}^n \) such that

\[
\sum_{i=1}^{k} x_i^T A_i x_i \rightarrow \text{maximum}.
\]
Theoretical solution

By Lagrange’s multipliers the \(x_i\)’s giving the optimum satisfy the system of linear equations

\[
A(X) = XS
\]  

(11)

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

\[
X = (x_1, \ldots, x_k), \quad A(X) = (A_1x_1, \ldots, A_kx_k).
\]

Due to the constraints imposed on \(x_1, \ldots, x_k\), the non-linear system of equations

\[
X^TX = I_k
\]

(12)

must also hold.
As $X$ and the symmetric matrix $S$ contain altogether $nk + k(k + 1)/2$ free parameters, while the equations (11) and (12) contain 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,

$$|A - I_n \otimes S| = 0$$

must hold, where the $nk \times nk$ matrix $A$ is a Kronecker-sum $A = A_1 \oplus \cdots \oplus A_k$ ($\otimes$ denotes the Kronecker-product).

Generalization of the eigenvalue problem: **eigenmatrix problem**.
Numerical solution

Starting with a matrix $X^{(0)}$ of orthonormal columns, the $m$th step of the iteration is as follows ($m = 1, 2, \ldots$):

Take the polar decomposition

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

into an $n \times k$ suborthogonal matrix (a matrix of orthonormal columns) and a $k \times k$ symmetric matrix ($k \leq n$). Let the first factor be $X^{(m)}$, etc. until convergence. In fact, the trace of $S^{(m)}$ converges to the optimum of the objective function.

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.
References


