Modelling Mortality: A Bayesian Factor-Augmented VAR (FAVAR) Approach

Yang Lu, Dan Zhu

November 28, 2023
Modelling Mortality: A Bayesian Factor-Augmented VAR (FAVAR) Approach

Yang Lu, Dan Zhu

Motivation

The data

LC

VAR

FAVAR–Combining VAR and Factors

Model Formulation

Bayesian Prior Specification

MCMC

Empirical Analysis

Data

Horsrace Comparison

Empirical Insights

Conclusion

Future Directions

US male population

Figure: Evolution of the log US male mortality rates over time at ages 20, 40, 60, and 80, respectively.
Key observations

- For a given date, the higher the age, the higher the mortality rates.
- There is a general downward trend of the mortality rates over the past eight decades, which is the longevity phenomenon.
- The four time series tend to move together, with the synchronization more pronounced between neighbouring ages. This suggests that they are likely co-integrated.
LC Model—Capturing the downward trend

The original LC model specifies the log mortality rates \( \log m_{x,t} \) are driven by the same factor across different ages:

\[
\log m_{x,t} = a_x + b_x \kappa_t + \varepsilon_{x,t}, \quad \forall x, t,
\]

where \( a_x, b_x \) are age-specific intercept and slope, respectively, and \( \varepsilon_{x,t} \) is an normally distributed i.i.d error term. To project the processes forward, the standard trick is to consider

\[
\kappa_t = \gamma + \kappa_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma^2_\eta).
\]

Several variants of this dynamic factor model, including the two-factor Cairns-Blake-Dowd (CBD) model, have been introduced to the mortality literature.
Takeaway from LC

- Simple and popular
- Standard identification

\[ \sum b_x = 1, \sum \kappa_t = 0; \]

implies \( \kappa_t \) captures the "average" trend.

- Even multiple factors, the observed variables are linear combinations of a few unobservable factors. This may limit their ability to capture complex temporal relationships.
Modelling Mortality: A Bayesian Factor-Augmented VAR (FAVAR) Approach

Yang Lu, Dan Zhu

Motivation

The data

LC

VAR

FAVAR–Combining VAR and Factors

Model Formulation

Bayesian Prior Specification

MCMC

Empirical Analysis

Data

Horsrace Comparison

Empirical Insights

Conclusion

Future Directions

Factor is not enough

Residual—Deviation from the trend

20

40

60

80


-0.6

-0.4

-0.2

0

0.2

0.4

0.6

0.8

20

40

60

80

Figure: Residual of the log US male mortality rates over time at ages 20, 40, 60, and 80, respectively.
The VAR($p$) model, instead of using a low dimensional dynamic factors explaining the high dimensional mortality movements, assumes that

$$\log m_{x,t} = \alpha_{x,0} + \sum_{i=1}^{p} \sum_{j=0}^{d-1} \alpha_{x,x_0+j}^{i} \log(m_{j,t-i}) + \epsilon_{x,t} \quad \forall x, t$$

where $d$ is the total number of ages for which mortality data are available. The model is highly parameterised, that the mortality literature has only considered VAR with one lag case. Even in the one lag case, the number of parameters is $(d+1)d$.

**Problem:** Short mortality data set.
VAR as the workhorse in macro

- In macroeconomics, various economic indicators (such as GDP, inflation, interest rates) are interrelated, and VAR allows for the joint modeling of these variables.
- Banbura et al. (2010) considered Bayesian VAR (130 variables), and demonstrated that VAR forecast better than factor methods (at least in their application).
- Extensions includes VAR-SV, TVP-VAR, TVPSV-VAR and recently Quantile/Distributional VAR.
Benefit of VAR

- VAR models allow for a flexible representation of the dynamic relationships among variables over time. Each variable can be influenced by its own lagged values and the lagged values of other variables in the system.
- Easier interpretation of the IRF
- SVAR formulation—causality.
Combining VAR and a univariate factor, we have

\[
\log(m_{x,t}) = a_x + \sum_{j=0}^{d-1} a_{x,x_0+j} \log(m_{j,t-1}) + b_x \kappa_t + \epsilon_{x,t}, \quad (1)
\]

where \(x_0\) is the lowest age for which mortality rates are observable, \((\epsilon_t)\) is an i.i.d. random vector satisfying:

\[
\epsilon_t = \{\epsilon_{x_0,t}, \ldots, \epsilon_{x_0+d-1,t}\} \sim N(0, \text{diag}(S)), \quad (2)
\]

with \(S = (\sigma_0^2, \ldots, \sigma_{d-1}^2)'\).
The AR(1) factors

The factor \((\kappa_t)\) is unobservable, following the dynamics:

\[
\kappa_t = \gamma_1 + \gamma_2 \kappa_{t-1} + \eta_t,
\]

where \((\eta_t)\) is another i.i.d. sequence and is mutually independent with \((\varepsilon_t)\), following:

\[
\eta_t \sim N(0, \sigma_\eta^2).
\]

Note that in the above model, the factor \(\kappa_t\) is only identified up to an affine transformation. Thus, for identification purpose, we shall let:

\[
b_{x_0} = 1.
\]

We set the initial \(\kappa, \kappa_0\), as a model parameter.
Factor-augmented VAR representation

The above model can be more conveniently represented using a matrix, factor VAR form:

\[ y_t = a + Ay_{t-1} + \kappa_t b + \varepsilon_t, \]

where \( y_t = \{ \log(m_{x_0}, t), \ldots, \log(m_{x_0+d-1}, t) \} \), \( a = \{ a_{x_0}, \ldots, a_{x_0+d-1} \} \), \( b = \{ b_{x_0}, \ldots, b_{x_0+d-1} \} \) are \( d \) dimensional column vectors and the \( d \times d \) matrix \( A \) is given by:

\[
A = \begin{bmatrix}
  a_{x_0,x_0} & \cdots & a_{x_0,x_0+j} & \cdots & a_{x_0,x_0+d-1} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  a_{x_0+i,x_0} & \cdots & a_{x_0+i,x_0+j} & \cdots & a_{x_0+i,x_0+d-1} \\
  \vdots & \vdots & \ddots & \ddots & \vdots \\
  a_{x_0+d-1,x_0} & \cdots & a_{x_0+d-1,x_0+j} & \cdots & a_{x_0+d-1,x_0+d-1}
\end{bmatrix}.
\]
Let us consider the factor-augmented vector \( \tilde{y}_t = [y'_t, \kappa_t]' \), then we can rewrite the system as a VAR:

\[
\tilde{y}_t = \begin{bmatrix} a \\ \gamma_1 \end{bmatrix} + \begin{bmatrix} A \\ 0_{1 \times d} \end{bmatrix} \tilde{y}_{t-1} + \begin{bmatrix} I_d \\ 0 \\ b \end{bmatrix} \tilde{\varepsilon}_t,
\]

where the new error \( \tilde{\varepsilon}_t = (\varepsilon'_t, \eta_t)' \).
Co-integration relationships

- If $\gamma_2$ is between 0 and 1, then process $(\kappa_t)$ is stationary. Then
  - if the eigenvalues of $A$ are all smaller than 1, then process $(y_t)$ is also stationary.
  - if some of the eigenvalues of $A$ are equal to 1, then $Id - A$ is of reduced rank and some of the components of $y_t$ are integrated. Then there might exist co-integration relationships, and the co-integration vectors are the left eigenvectors of matrix $Id - A$.
  - if some of the eigenvalues of $A$ are larger than 1, then some of the components of $y_t$ are geometrically explosive.

- If instead $\gamma_2$ is equal to 1, that is process $(\kappa_t)$ is integrated, then:
  - if the eigenvalues of $A$ are all smaller than 1, then process $(y_t)$ is integrated of order 1 and no co-integration relationship exists.
  - if some of the eigenvalues of $A$ are equal to 1, then process $(y_t)$ is integrated of order 2.
Overparameterized VAR

Since VAR models are parameter-rich, their estimation can be challenging:

- regularization techniques, Lasso/ridge
- bayesian shrinkage priors that consider the baseline VAR model:

\[
y_t = a + Ay_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Sigma), \forall t.
\]

We set

\[
\text{vec}[a'; A'] \sim N(\mu_a, \Sigma_a), \Sigma \sim IW(v, S).
\]

The standard Minnesota type prior is to set \( \mu_a = 0 \) and using \( \Sigma_a \) to shrink the posterior such that less shrinkage on own lag coefficients.
A Shrinkage prior formulation

The standard BVAR in macro shrinks the model to a random walk—lack of cointegration. We consider the case, we the prior mean implies a model

$$y_t = y_{t-1} + \kappa_t + z_t$$

For the prior variance specification, we reduce shrinkage on the main, first and second lower diagonal of $A$

$$\log m_{x,t} = \alpha_{x,0} + \alpha_{x,1} \log m_{x,t-1} + \alpha_{x,2} \log m_{x+1,t-1} + \alpha_{x,3} \log m_{x+2,t-1} + \varepsilon_{x,t}, \quad \forall x$$

subject to the constraints

$$\alpha_{x,1} + \alpha_{x,2} + \alpha_{x,3} = 1, \forall x$$

$$\alpha_{x,k} \geq 0, \forall x, k = 1, 2, 3.$$
Several other possibilities

1. Fix $\gamma_2$ to 1, so that factor $(\kappa_t)$ is constrained to be a random walk. Instead of shrinking matrix $A$ towards $A_0$ with unit eigenvalue, we can shrink it instead towards a matrix whose spectral radius is smaller than 1.

2. Force the spectral radius of $A$ to be equal to 1, while at the same restricting $\gamma_2$ to lie between 0 and 1.
Integrated likelihood

Let us first compute the likelihood function of the observed process \((y_t)\), for given value of the parameter vector \(\theta\). By integrating out the factor path, this likelihood function is equal to:

\[
f(\mathbf{Y}|\theta) = \int \ell(\mathbf{Y}, \kappa|\theta) d\kappa, \tag{6}\]

where the integral is of dimension \(T\), and \(\ell(\mathbf{Y}, \kappa|\theta)\) is the joint likelihood function of the observation \(\mathbf{Y}\) and the latent process \(\kappa\), with \(\mathbf{Y} = [y_2, ..., y_T]'\), \(\kappa = [\kappa_2, ..., \kappa_T]'\). We have:

\[
\log \ell(\mathbf{Y}, \kappa|\theta) = -\frac{T - 1}{2} \sum_{x=x_0}^{d-1} \log(2\pi\sigma_x^2) \tag{7}
\]

\[-\frac{1}{2} \sum_{t=2}^{T} (y_t - a - Ay_{t-1} - b\kappa_t)'\Sigma^{-1} (y_t - a - Ay_{t-1} - b\kappa_t),\]

\[-\frac{1}{2} \sum_{t=2}^{T} (\kappa_t - \gamma_1 - \gamma_2\kappa_{t-1})^2 \sigma_{\eta}^2 - \frac{T - 1}{2} \log(2\pi\sigma_{\eta}^2), \tag{8}\]
Precision sampling of the latent

To sample the latent factor from the conditional distribution \( \kappa | \mathbf{Y}, \theta \), we derive

\[
\kappa | \mathbf{Y}, \theta \sim N(\mu_k, K^{-1})
\]

where

\[
K = (M - \gamma_2 H)'(M - \gamma_2 H) \frac{1}{\sigma^2_\eta} + \begin{bmatrix}
\sigma_k^{-2} & 0_{T-1}' \\
0_{T-1} & b' \Sigma^{-1} b
\end{bmatrix}
\]

\( \mu_k = K^{-1} \begin{bmatrix}
0 \\
(Y - X\alpha)' \Sigma^{-1} b
\end{bmatrix} + \frac{\gamma_1}{\sigma^2_\eta} (M - \gamma_2 H)' \mathbf{1}_{T-1} \)

where

\[
M = [\mathbf{1}_{T-1}, \mathbf{0}_{T-1}]
\]
Data

To illustrate our methodology, let us now estimate the FAVAR model using data from two populations,

▸ the French male
▸ US male general populations.

For each of the two populations, we will also compare the performance of the FAVAR model with that of the LC, as well as Li and Lu (2017) model. Both data are downloaded freely from the Human Mortality Database (HMD).
Data

- For the US male population, mortality rates are observed from the year 1950 till 2017. We use data from 1930 to 2007 to estimate the model, and the observations between 2008 and 2017 for forecast evaluation. Hence, in this example, $T = 79$. We use age range 0-80.

- For the French data from (1950 to 2018), we partition our data into the training set (from 1950 to 2007) and the validation set (from 2008 to 2018), and age range (0-80).
### Table: In-Sample and out of Sample MSE.

<table>
<thead>
<tr>
<th></th>
<th>LC</th>
<th>Li and Lu (2017)</th>
<th>FAVAR (strong)</th>
<th>FAVAR (weak)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In-Sample</strong></td>
<td>0.0042</td>
<td>0.0014</td>
<td>0.0009</td>
<td>0.0008</td>
</tr>
<tr>
<td><strong>One Year</strong></td>
<td>0.0090</td>
<td>0.0014</td>
<td>0.0015</td>
<td>0.0014</td>
</tr>
<tr>
<td><strong>Five Year</strong></td>
<td>0.0132</td>
<td>0.0031</td>
<td>0.0030</td>
<td>0.0031</td>
</tr>
<tr>
<td><strong>Ten Year</strong></td>
<td>0.0161</td>
<td>0.0058</td>
<td>0.0145</td>
<td>0.0192</td>
</tr>
</tbody>
</table>
## French data MSE

<table>
<thead>
<tr>
<th></th>
<th>LC</th>
<th>Li and Lu (2017)</th>
<th>FAVAR(strong)</th>
<th>FAVAR(weak)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Sample</td>
<td>0.0094</td>
<td>0.0044</td>
<td>0.0024</td>
<td>0.0019</td>
</tr>
<tr>
<td>One Year</td>
<td>0.0227</td>
<td>0.0097</td>
<td>0.0080</td>
<td>0.0076</td>
</tr>
<tr>
<td>Five Year</td>
<td>0.0317</td>
<td>0.0086</td>
<td>0.0099</td>
<td>0.0113</td>
</tr>
<tr>
<td>Ten Year</td>
<td>0.0461</td>
<td>0.0159</td>
<td>0.0134</td>
<td>0.0161</td>
</tr>
</tbody>
</table>

**Table**: In-sample and out-of-sample MSE.
## More Comparisons

<table>
<thead>
<tr>
<th>Models</th>
<th>Age: 51-100</th>
<th>Age: 61-100</th>
<th>Age: 51-90</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In-sample</td>
<td>1Y  5Y  10Y</td>
<td>In-sample</td>
</tr>
<tr>
<td>LC</td>
<td>0.036</td>
<td>0.069  0.088</td>
<td>0.111</td>
</tr>
<tr>
<td>VAR</td>
<td>0.010</td>
<td>0.024  0.049</td>
<td>0.139</td>
</tr>
<tr>
<td>Li&amp;Lu</td>
<td>0.028</td>
<td>0.024  0.026</td>
<td>0.040</td>
</tr>
<tr>
<td>FAVAR</td>
<td>0.022</td>
<td>0.024  0.030</td>
<td>0.047</td>
</tr>
<tr>
<td>FDM</td>
<td>0.017</td>
<td>0.023  0.062</td>
<td>0.115</td>
</tr>
</tbody>
</table>

### Fitting the models using data from 1950-2007

<table>
<thead>
<tr>
<th>Models</th>
<th>Age: 51-100</th>
<th>Age: 61-100</th>
<th>Age: 51-90</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In-sample</td>
<td>1Y  5Y  10Y</td>
<td>In-sample</td>
</tr>
<tr>
<td>LC</td>
<td>0.033</td>
<td>0.065  0.085</td>
<td>0.112</td>
</tr>
<tr>
<td>VAR</td>
<td>0.010</td>
<td>0.028  0.041</td>
<td>0.103</td>
</tr>
<tr>
<td>Li&amp;Lu</td>
<td>0.026</td>
<td>0.022  0.026</td>
<td>0.042</td>
</tr>
<tr>
<td>FAVAR</td>
<td>0.020</td>
<td>0.024  0.030</td>
<td>0.049</td>
</tr>
<tr>
<td>FDM</td>
<td>0.016</td>
<td>0.021  0.053</td>
<td>0.106</td>
</tr>
</tbody>
</table>

### Fitting the models using data from 1960-2007

<table>
<thead>
<tr>
<th>Models</th>
<th>Age: 51-100</th>
<th>Age: 61-100</th>
<th>Age: 51-90</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In-sample</td>
<td>1Y  5Y  10Y</td>
<td>In-sample</td>
</tr>
<tr>
<td>LC</td>
<td>0.030</td>
<td>0.063  0.088</td>
<td>0.123</td>
</tr>
<tr>
<td>VAR</td>
<td>0.000</td>
<td>0.028  0.050</td>
<td>0.080</td>
</tr>
<tr>
<td>Li&amp;Lu</td>
<td>0.022</td>
<td>0.024  0.024</td>
<td>0.054</td>
</tr>
<tr>
<td>FAVAR</td>
<td>0.017</td>
<td>0.026  0.035</td>
<td>0.066</td>
</tr>
<tr>
<td>FDM</td>
<td>0.014</td>
<td>0.021  0.052</td>
<td>0.115</td>
</tr>
</tbody>
</table>

Table 3: In-Sample/Out-of-Sample for US Male population. The four numbers presented are RMSE of in-sample fit, forecast of horizon 1,5 and 10 periods.
Why do we have similar results as of Li and Lu(2017)?

**Figure:** US data: Heat map of the joint histogram (based on 10000 iterations of the MCMC) of the $\rho(A)$. Left panel: the model with strong shrinkage; right panel: the model with weak shrinkage.
We deviate from Li and Lu (2017)

**Figure:** US data: Posterior bivariate density of $\gamma_2$ and $\rho(A)$ under the four prior specifications: left panel: the model with strong shrinkage prior; right panel: the model with weak shrinkage prior.
Do we have anything left for the factors?

Figure: US data: Left panel: posterior trajectory of the latent factor over time in the FAVAR model; right panel: the curve of the associated loading factor across different ages. The blue and orange full lines indicate the specifications with strong and weak shrinkage, respectively.
Predictive Density

Figure: US Male Mortality for age
This paper has introduced the (Bayesian) FAVAR model into the mortality forecasting literature, that combines factor and VAR.

A carefully designed prior is given to balance between short-term and long term objective.

An extensive amount of empirical study demonstrates the efficacy of the approach, and more important the information that can extracted from the model.
A more parsimonious formulation of VAR

As you may see that our proposed method did not consistently outperform Li and Lu(2017), this suggest

\[ \log m_{x,t} = \alpha_{x,0} + \sum_{j=1}^{p} \alpha_{x,j} \log m_{x-j+1,t-1} + \beta_x \kappa_t + \epsilon_{x,t}, \forall x, t \]

subject to the constraints

\[ \sum_{j=1}^{p} \alpha_{x,j} = 1, \forall x \]

\[ \alpha_{x,k} \geq 0, \forall x, k = 1, 2, ..j. \]

MCMC: Sampling from truncated distributions.
Causality

There are a vast literature on mortality forecasting

- Factor
- VAR
- FAVAR.

*Causal inference* is crucial for mortality research as it allows us to identify and understand the specific factors or interventions that contribute to changes in mortality rates, enabling the development of targeted and effective public health strategies to improve overall well-being and save lives.
WHAT DRIVES MORTALITY IMPROVEMENTS? AN EMPIRICAL ANALYSIS USING INSTRUMENTAL VARIABLE APPROACH.

\[ \kappa_t = \kappa_{t-1} + x_t \gamma + u_t \]

where \( x_t = \left[ x_{t,1}, x'_{t,2} \right]' \). Introducing instrumental variable, \( z_t \), such that

\[ x_{t,1} = D'_1 z_t + D'_2 x_{t,2} + v_t. \]
Working Process in Causal Mortality


\[ y_t = \mu + \lambda_1 f_{t,1} + \lambda_2 f_{t,2} + z_t \]

where \( f_{t,1} \) is the vector of unobserved factors and \( f_{t,2} \) is a vector of observed macroeconomic variables. Assume that \( f_t = [f'_t, f''_t] \) evolves according to a VAR

\[ f_t = b + \sum_{i=1}^{p} B_i f_{t-j} + \nu_t. \]

The key here is that \( f_{t,2} \) is observed at higher frequency than \( y_t \).

- Mortality data are typically released on an annual basis, yet economic and financial decision-making frequently necessitates more timely updates.
- Investigate the impacts of health expenditure shocks on mortality rates.