Workshop on Survey Methodology:

Big data in official statistics

Block 5: Dynamic factor models for nowcasting

20 May 2019,
Brazilian Network Information Center (NIC.br),
São Paulo, Brazil

Jan van den Brakel
Statistics Netherlands and Maastricht University
Introduction

Introduction:

• Block 4: Bivariate STM

• Combine time series observed with a repeated survey with an auxiliary series.
  – Improve survey estimates
  – Estimation in real time or nowcasting

• But what if there are $n$ auxiliary series?

• Results in a high dimensionality problem (deteriorated prediction power of a model)

• Dynamic Factor Models (Doz et al., 2011)

• Illustrating example: nowcasting unemployed labour force with Google trends
Labour Force Survey

- Monthly, quarterly and annual figures labour force
- Rotating panel design
- Monthly samples observed 5 times at quarterly intervals
- Problems:
  - Sample size too small for monthly figures with GREG estimator
  - Rotation Group Bias
  - Discontinuities due to a major survey redesign
- Solution: 5 dimensional structural time series model (Pfeffermann, 1991)
• Each month: 5 independent samples

• Gives 5 direct estimates $\hat{y}_t^{[j]}$, $j = 1, \ldots, 5$ for population parameter (e.g. unemployed labour force).

• Monthly figures: 5-dimensional state space model (Pfeffermann, 1991):

$$
\begin{pmatrix}
\hat{y}_t^{[1]} \\
\hat{y}_t^{[2]} \\
\vdots \\
\hat{y}_t^{[5]}
\end{pmatrix} =
\begin{pmatrix}
1 \\
1 \\
\vdots \\
1
\end{pmatrix}
\begin{pmatrix}
L_t^{[y]} + S_t^{[y]} + I_t^{[y]} \\
\lambda_t^{[1]} \\
\lambda_t^{[2]} \\
\vdots \\
\lambda_t^{[5]}
\end{pmatrix} +
\begin{pmatrix}
\beta^{[1]} \delta_t^{[1]} \\
\beta^{[2]} \delta_t^{[2]} \\
\vdots \\
\beta^{[5]} \delta_t^{[5]}
\end{pmatrix} +
\begin{pmatrix}
e_t^{[1]} \\
e_t^{[2]} \\
\vdots \\
e_t^{[5]}
\end{pmatrix}
$$

$\Leftrightarrow$

$$\hat{y}_t = 1_{[5]} \left( L_t^{[y]} + S_t^{[y]} + I_t^{[y]} \right) + \Delta \beta + \lambda_t + e_t$$

• Used by Statistics Netherlands since 2010 to produce official monthly figures about the labour force.
Figure illustrates official monthly unemployed labour force figures:

- General regression estimates monthly unemployed labour force at the national level: \( \hat{y}_t^{[j]} \), \( j = 1, \ldots, 5 \) in grey

- Filtered trend (level before redesign in 2010) in blue

More timely unemployment figures

Labour Force Survey

• Figures month $t$ published in $t + 1$

• How to improve:
  – accuracy
  – timeliness

• Potential auxiliary information for unemployment
  – Claimant counts (register): for month $t$ available in $t + 1$
  – Google trends: weekly or daily frequency.

• Google trends potentially useful to estimate unemployment in real time
**Auxiliary series unemployment**

Figure illustrates:

- **Black**: general regression estimates monthly unemployed labour force per wave at the national level: $\hat{y}_t^{[j]}, j = 1, \ldots, 5$.

- **Green**: Claimant counts

- **Red**: Google trend for the search term “job description”

- In this application about 80 Google trends
Auxiliary series unemployment

Issues

- High dimensionality problem:
  - Cannot include 80 series with separate trends, seasonals etc
  - Large models with many parameters result in reduced prediction power

- Mixed frequency series: observations become available at different moments in time resulting in time series with ”jagged” ends (observations are partially missing at the end of the series)

- Solution: dynamic factor model with a two-step estimator proposed by:
  - Giannone et al. (2008)
  - Doz et al. (2011)
Dynamic factor model

Step 1

- Estimate the common factors in the Google trends

\[
\begin{align*}
\mathbf{x}_t^{[GT]} &= \Lambda \mathbf{f}_t + \epsilon_t \\
\text{Var}(\epsilon_t) &= \Psi
\end{align*}
\]

\[
\mathbf{f}_t = \mathbf{f}_{t-1} + \mathbf{\mu}_t
\]

- \( \mathbf{x}_t^{[GT]} \): \( n \) vector with auxiliary series / Google trends assumed to be I(1) (weekly frequency)
- \( \mathbf{f}_t \): \( r \) vector with common factors \( r << n \) assumed to be I(1)
- \( \Lambda \): \( n \times r \) matrix with factor loadings
- \( \epsilon_t \): \( n \) vector with idiosyncratic components / variable specific shocks
- \( \Psi \): diagonal variance matrix of \( \epsilon_t \)
- For identifiability reasons: \( E(\mathbf{\mu}_t \mathbf{\mu}_t') = \mathbf{I}_{[r]} \)

- \( \mathbf{f}_t, \ \Lambda, \ \Psi \) are estimated with Principal Component Analysis applied to the weekly data of GT
Dynamic factor model

- Google trends are aggregated to monthly frequency

- Usual approach: time series model for LFS and CC at a weekly frequency

- Awkward for the LFS due to the complexity of the model component for the sampling error

- In this case:

\[ x_{t}^{q,[GT]} = \frac{1}{q} \sum_{q=0}^{q-1} x_{t}^{[GT]}, \quad t = q, 2q, 3q, etc. \]
Dynamic factor models for nowcasting

Dynamic factor model

Step 2

- State space model for the entire data set

\[
\begin{pmatrix}
\hat{y}_t \\
x_t^{[CC]} \\
x_t^{[GT]} \\
\end{pmatrix}
=\begin{pmatrix}
1 \cdot [5 \cdot (L_t^{[y]} + S_t^{[y]})] \\
L_t^{[CC]} + S_t^{[CC]} \\
\hat{\Lambda}_f_t \\
\end{pmatrix}
+ \begin{pmatrix}
\lambda_t \\
0 \\
0 \\
\end{pmatrix}
+ \begin{pmatrix}
e_t \\
\end{pmatrix}
= \begin{pmatrix}
1 \\
0 \\
0 \\
\end{pmatrix}
+ \begin{pmatrix}
e_t \\
\end{pmatrix}
\]

\[L_t^{[z]} = L_{t-1}^{[z]} + R_{t-1}^{[z]} \quad R_t^{[z]} = R_{t-1}^{[z]} + \eta_t^{[z]} \quad z = (y, CC)\]

\[f_t = f_{t-1} + \mu_t\]

\[
\text{Cov} \begin{pmatrix}
\eta_t^{[y]} \\
\eta_t^{[CC]} \\
\mu_t \\
\end{pmatrix} = \begin{pmatrix}
\sigma_y^2 & \sigma_{y,CC} & \sigma_{y,f_1} & \cdots \\
\sigma_{y,CC} & \sigma_{CC}^2 & 0 & \cdots \\
\sigma_{y,f_1} & 0 & 1 & \cdots \\
\vdots & \vdots & \vdots & 1 \\
\end{pmatrix}
\]

\[\sigma_{y,CC} = \rho_{CC} \sigma_y \sigma_{CC},\]

\[\sigma_{y,f_1} = \rho_{1,GT} \sigma_y\]

- \(\hat{\Lambda}, \hat{\Psi}\) obtained in step 1 are kept fixed

- \(f_t\) are re-estimated with the Kalman filter

10
Dynamic factor model

- Strong correlations between trend disturbance terms \( \eta_t[y], \eta_t^{CC} \) and \( \mu_t \) improves accuracy trend LFS \( L_t[y] \)

- Examples where claimant count series are used to improve accuracy of monthly unemployment figures based on Labour Force Survey data:
  - Harvey and Chung (2000) UK LFS
  - van den Brakel and Krieg (2016) Dutch LFS

- Google trends are added to estimate \( L_t[y] \) in real time
Results

Models:

1. Baseline model: model used in production using the LFS component only:

\[ \hat{y}_t = 1_{[5]} \left( L_t^{[y]} + S_t^{[y]} \right) + \lambda_t + e_t \]

2. CC only:

\[
\begin{pmatrix}
\hat{y}_t \\
\mathbf{x}_{t}^{[CC]} \\
\end{pmatrix} = \begin{pmatrix}
1_{[5]} (L_t^{[y]} + S_t^{[y]}) \\
L_t^{[CC]} + S_t^{[CC]} \\
\end{pmatrix} + \begin{pmatrix}
\lambda_t \\
0 \\
\end{pmatrix} + \begin{pmatrix}
e_t \\
I_t \\
\end{pmatrix}
\]

3. GT only

\[
\begin{pmatrix}
\hat{y}_t \\
\mathbf{x}_{t}^{q,[GT]} \\
\end{pmatrix} = \begin{pmatrix}
1_{[5]} (L_t^{[y]} + S_t^{[y]}) \\
\hat{\Lambda}f_t \\
\end{pmatrix} + \begin{pmatrix}
\lambda_t \\
0 \\
\end{pmatrix} + \begin{pmatrix}
e_t \\
\epsilon_t \\
\end{pmatrix}
\]

4. CC+GT:

\[
\begin{pmatrix}
\hat{y}_t \\
\mathbf{x}_{t}^{[CC]} \\
\mathbf{x}_{t}^{q,[GT]} \\
\end{pmatrix} = \begin{pmatrix}
1_{[5]} (L_t^{[y]} + S_t^{[y]}) \\
L_t^{[CC]} + S_t^{[CC]} \\
\hat{\Lambda}f_t \\
\end{pmatrix} + \begin{pmatrix}
\lambda_t \\
0 \\
0 \\
\end{pmatrix} + \begin{pmatrix}
e_t \\
I_t \\
\epsilon_t \\
\end{pmatrix}
\]
Results

• Results based on the period January 2004 until December 2017 (168 months)

• Out-of-sample nowcasts based on the last 56 months:
  
  – nowcast for $t$: LFS and CC missing, only GT available
  
  – Hyperparameter estimates based available information in $t$

• Estimation accuracy:

  \[
  \hat{MSE}(\hat{a}_{t|t}) = \frac{1}{(T - d)} \sum_{t=d+1}^{T} P_{t|t}
  \]

• Nowcast accuracy:

  \[
  \hat{MSFE}(\hat{a}_{t|t}) = \frac{1}{h} \sum_{t=T-h+1}^{T} P_{t|t}
  \]
Results

- Number of common factors for Google trends: 2

- Correlations trend disturbance terms:

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\rho}_{1,GT}$ (p-value)</th>
<th>$\hat{\rho}_{2,GT}$ (p-value)</th>
<th>$\hat{\rho}_{CC}$ (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>0.43 (0.39)</td>
<td>-0.40 (0.31)</td>
<td>0.90 (0.0004)</td>
</tr>
<tr>
<td>GT</td>
<td>0.04 (1.0)</td>
<td>0.05 (1.0)</td>
<td>0.90 (0.0007)</td>
</tr>
<tr>
<td>GT+CC</td>
<td>0.90 (0.0007)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

p-value: LR test $H_0 : \rho_x = 0$
Results

Results trend $L_t^{[y]}$ relative to baseline model

<table>
<thead>
<tr>
<th>model</th>
<th>CC</th>
<th>GT</th>
<th>CC+GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{MSE}(L_t^{[y]})$</td>
<td>0.869</td>
<td>0.967</td>
<td>0.869</td>
</tr>
<tr>
<td>$\hat{MSFE}(L_t^{[y]})$</td>
<td>0.715</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{MSFE}(L_t^{[y]})$</td>
<td></td>
<td>0.988</td>
<td>0.709</td>
</tr>
<tr>
<td>week 1</td>
<td>0.989</td>
<td>0.707</td>
<td></td>
</tr>
<tr>
<td>week 2</td>
<td>0.987</td>
<td>0.712</td>
<td></td>
</tr>
<tr>
<td>week 3</td>
<td>0.989</td>
<td>0.709</td>
<td></td>
</tr>
<tr>
<td>week 4</td>
<td>0.989</td>
<td>0.713</td>
<td></td>
</tr>
<tr>
<td>week 5</td>
<td>0.977</td>
<td>0.691</td>
<td></td>
</tr>
</tbody>
</table>
Results

Results signal $\theta_t^{[y]} = L_t^{[y]} + S_t^{[y]}$ relative to baseline model

<table>
<thead>
<tr>
<th>model</th>
<th>CC</th>
<th>GT</th>
<th>CC+GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{MSE}(\theta_t^{[y]})$</td>
<td>0.890</td>
<td>0.977</td>
<td>0.889</td>
</tr>
<tr>
<td>$\hat{MSFE}(\theta_t^{[y]})$</td>
<td>0.729</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{MSFE}(\theta_t^{[y]})$</td>
<td></td>
<td>0.953</td>
<td>0.743</td>
</tr>
<tr>
<td>week 1</td>
<td>0.953</td>
<td>0.749</td>
<td></td>
</tr>
<tr>
<td>week 2</td>
<td>0.953</td>
<td>0.735</td>
<td></td>
</tr>
<tr>
<td>week 3</td>
<td>0.955</td>
<td>0.744</td>
<td></td>
</tr>
<tr>
<td>week 4</td>
<td>0.956</td>
<td>0.756</td>
<td></td>
</tr>
<tr>
<td>week 5</td>
<td>0.943</td>
<td>0.717</td>
<td></td>
</tr>
</tbody>
</table>
Results

Nowcast trend $L_t^{[y]}$

Nowcast signal $\theta_t^{[y]} = L_t^{[y]} + S_t^{[y]}$
Results

Model diagnostics:

- Test on standardized innovations of LFS

Software:

- R
Conclusions

- Dynamic factor model to include large sets of auxiliary series in parsimonious model (avoids high dimensionality problems)

- Strongest contribution in this application comes from claimant counts

- Effect of the selected Google trends is minor

- Details: Schiavoni et al. (2019)
Extension

Model for mixed frequencies

- Time series repeated survey quarterly basis
- Auxiliary series on a monthly frequency
- Temporal disaggregation
- Define time series model for the survey at the highest frequency
- Stock variables: quarterly observation is the mean over three months
- Flow variables: quarterly observation is the total over three months
Extension

Bivariate model:

- $y_t^k$ sample survey observed if $t = 3k, k = 1, 2, \ldots$ and missing otherwise
- $x_t$ auxiliary series observed for $t = 1, 2, 3, \ldots$
- Model for both series defined on a high frequency

\[ L_z^z + S_t^z + I_t^z, \quad z \in x, y \]

- $L_t^z$ for example a smooth trend

- Model the correlation between the slope disturbance terms $\eta_t^y$ and $\eta_t^x$ (see Block 3)

- Measurement equation $x_t$:

\[ x_t = L_t^x + S_t^x + I_t^x, \]

- Measurement equation $y_t^k$ (flow variable):

\[ y_t^k = \sum_{j=0}^{2} (L_{t-j}^y + S_{t-j}^y + I_{t-j}^y), \]
• Measurement equation $y^k_t$ (stock variable):

$$y^k_t = \frac{1}{3} \sum_{j=0}^{2} (L_{t-j}^y + S_{t-j}^y + I_{t-j}^y),$$

• Seasonal component quarterly series: only the first two frequencies can be estimated (Harvey, 1989)

$$S_{t}^y = \sum_{j=1}^{2} \gamma_{jt}^y$$

• Can be applied in a similar way to a dynamic factor model

• Efficient approach for nowcasting: Kalman filter produces predictions for the missing values
Extension

State space representation:

- Measurement equation:  \( \mathbf{y}_t = \mathbf{Z}\alpha_t + \mathbf{I}_t \)
  
  with \( \mathbf{y}_t = (y^k_t, x_t)' \)

- Transition equation:  \( \alpha_t = \mathbf{T}\alpha_{t-1} + \mathbf{\eta}_t \)

\[
\begin{align*}
\alpha_t &= \begin{pmatrix} \alpha^y_t \\ \alpha^x_t \end{pmatrix} \\
- \alpha^y_t &= (L^y_t, R^y_t, L^y_{t-1}, L^y_{t-2}, S^y_t, S^y_{t-1}, S^y_{t-2})^t \\
- S^y_t &= (\gamma^y_1 t, \tilde{\gamma}^y_1 t, \gamma^y_2 t) \\
- S^y_{t-1} &= (\gamma^y_1 t-1, \gamma^y_2 t-1) \\
- S^y_{t-2} &= (\gamma^y_1 t-2, \gamma^y_2 t-2) \\
- \alpha^x_t &= (L^x_t, R^x_t, \gamma^x_t, \tilde{\gamma}^x_1 t, \ldots \gamma^x_6 t)^t \\
\end{align*}
\]

- \( \mathbf{Z} = \begin{pmatrix} \mathbf{z}^y \\ \mathbf{z}^x \end{pmatrix} \)
  
  - \( \mathbf{z}^y = (1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) \)
  
  - \( \mathbf{z}^x = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1) \)
Block 5: Dynamic factor models for nowcasting

- \( T = \text{BlockDiag}(T^y, T^x) \)

\[
T^y = \begin{pmatrix}
T^y_L & 0_{[4 \times 4]} & 0_{[4 \times 2]} & 0_{[4 \times 2]} \\
0_{[4 \times 4]} & T^y_S & 0_{[4 \times 2]} & 0_{[4 \times 2]} \\
0_{[2 \times 4]} & T^y_{S-1} & I_{[2]} & 0_{[2 \times 2]} \\
0_{[2 \times 4]} & 0_{[2 \times 4]} & 0_{[2 \times 2]} & 0_{[2 \times 2]}
\end{pmatrix}
\]

\[
T^y_L = \begin{pmatrix}
1 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0
\end{pmatrix}
\]

- \( T^y_S = \text{BlockDiag}(C_1, C_2) \) (See Block 2)

- \( T^y_{S-1} = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0
\end{pmatrix} \)

- \( T^x_S = \text{BlockDiag}(T^x_L, T^x_S) \) (See Block 2)

- \( \eta_t = \begin{pmatrix}
\eta^y_t \\
\eta^x_t
\end{pmatrix} \)

- \( \eta^y_t = (0, \eta^y_{R_t}, 0, 0, \omega^y_{1,t}, \omega^*_y_{1,t}, \omega^y_{2,t}, \omega^*_y_{2,t}, 0, 0, 0, 0)^t \)

- \( \alpha^x_t = (0, \eta^x_{R_t}, \omega^x_{1,t}, \omega^*_x_{1,t}, \ldots \omega^x_{6,t})^t \)
References


