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 to 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):

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

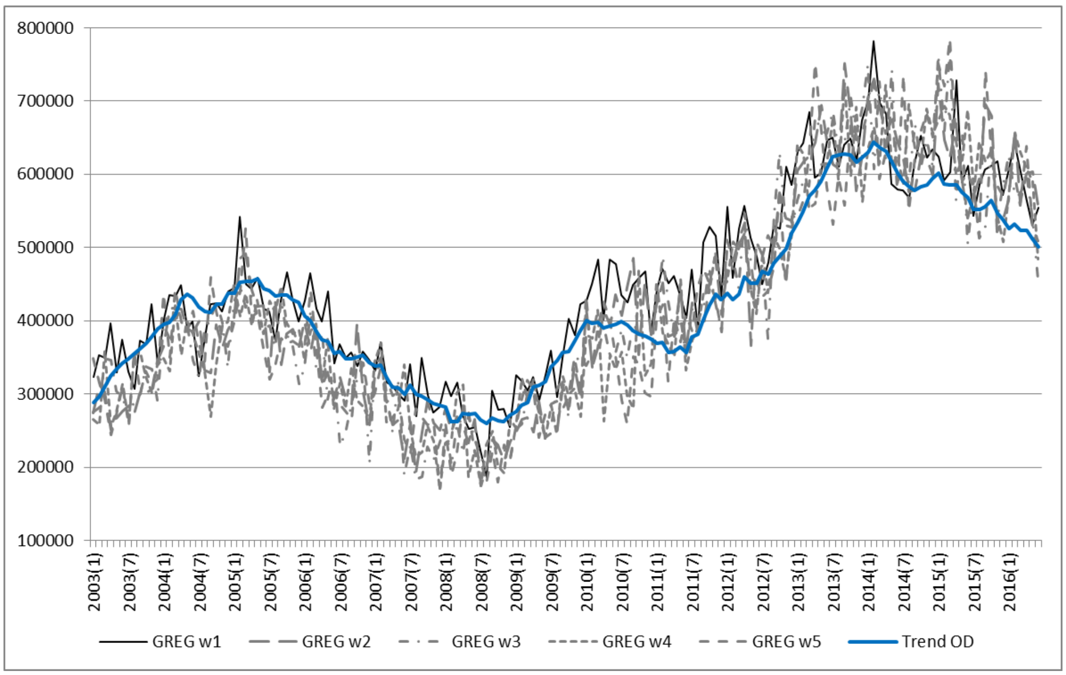
$$\Leftrightarrow$$

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

- 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: $\hat{y}_t^{[j]}$, $j = 1, ..., 5$ in grey
- Filtered trend (level before redesign in 2010) in blue



- Details: van den Brakel and Krieg (2009, 2015)

More timely unemployment figures

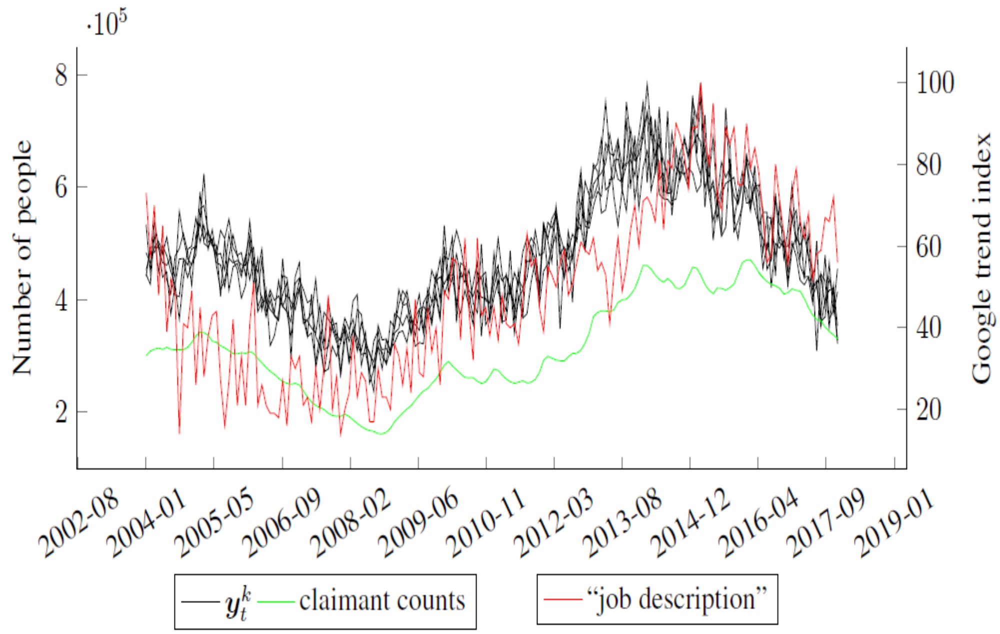
Labour Force Survey

- 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: $\hat{y}_t^{[j]}$, $j = 1, ..., 5$.
- 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

$$\mathbf{x}_t^{[GT]} = \mathbf{\Lambda}\mathbf{f}_t + \boldsymbol{\epsilon}_t \qquad Var(\boldsymbol{\epsilon}_t) = \mathbf{\Psi}$$

$$\mathbf{f}_t = \mathbf{f}_{t-1} + \boldsymbol{\mu}_t$$

  - $\mathbf{x}_t^{[GT]}$: 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)
  - $\mathbf{\Lambda}$: $n \times r$ matrix with factor loadings
  - $\boldsymbol{\epsilon}_t$: n vector with idiosyncratic components / variable specific shocks
  - $\mathbf{\Psi}$: diagonal variance matrix of $\boldsymbol{\epsilon}_t$
  - for identifiability reasons: $E(\boldsymbol{\mu}_t\boldsymbol{\mu}_t') = \mathbf{I}_{[r]}$

- $\mathbf{f}_t$, $\mathbf{\Lambda}$, $\mathbf{\Psi}$ are estimated with Principal Component Analysis applied to the weekly data of GT

Dynamic factor model

- Google trends are aggregated to monthly frequence
- Usual approach: time series model for LFS and CC at a weekly frequency
- Akward for the LFS due to the complexity of the model component for the sampling error
- In this case:

$$\mathbf{x}_t^{q,[GT]} = \frac{1}{q}\sum_{q=0}^{q-1} \mathbf{x}_t^{[GT]}, \qquad t = q, 2q, 3q, etc.$$

Dynamic factor model

Step 2

- State space model for the entire data set

$$\begin{pmatrix} \hat{\mathbf{y}}_t \\ x_t^{[CC]} \\ \mathbf{x}_t^{q,[GT]} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{[5]}(L_t^{[y]} + S_t^{[y]}) \\ L_t^{[CC]} + S_t^{[CC]} \\ \hat{\mathbf{\Lambda}}\mathbf{f}_t \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ I_t \\ \boldsymbol{\epsilon}_t \end{pmatrix}$$

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

$$\mathbf{f}_t = \mathbf{f}_{t-1} + \boldsymbol{\mu}_t$$

$$Cov \begin{pmatrix} \eta_t^{[y]} \\ \eta_t^{[CC]} \\ \boldsymbol{\mu}_t \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & \sigma_{y,CC} & \sigma_{y,f_1} & \dots \\ \sigma_{y,CC} & \sigma_{CC}^2 & 0 & \dots \\ \sigma_{y,f_1} & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & 1 \end{pmatrix}$$

$$\sigma_{y,CC} = \rho_{CC}\sigma_y\sigma_{CC},$$

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

- $\hat{\mathbf{\Lambda}}$, $\hat{\mathbf{\Psi}}$ obtained in step 1 are kept fixed

- $\mathbf{f}_t$ are re-estimated with the Kalman filter

Dynamic factor model

- Strong correlations between trend disturbance terms $\eta_t^{[y]}$, $\eta_t^{[CC]}$ and $\boldsymbol{\mu}_t$ 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:

$$\hat{\mathbf{y}}_t \;=\; \mathbf{1}_{[5]}\left(L_t^{[y]} + S_t^{[y]}\right) + \boldsymbol{\lambda}_t + \mathbf{e}_t$$

2. CC only:

$$\begin{pmatrix} \hat{\mathbf{y}}_t \\ x_t^{[CC]} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{[5]}(L_t^{[y]} + S_t^{[y]}) \\ L_t^{[CC]} + S_t^{[CC]} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ I_t \end{pmatrix}$$

3. GT only

$$\begin{pmatrix} \hat{\mathbf{y}}_t \\ \mathbf{x}_t^{q,[GT]} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{[5]}(L_t^{[y]} + S_t^{[y]}) \\ \hat{\boldsymbol{\Lambda}}\mathbf{f}_t \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ \boldsymbol{\epsilon}_t \end{pmatrix}$$

4. CC+GT:

$$\begin{pmatrix} \hat{\mathbf{y}}_t \\ x_t^{[CC]} \\ \mathbf{x}_t^{q,[GT]} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{[5]}(L_t^{[y]} + S_t^{[y]}) \\ L_t^{[CC]} + S_t^{[CC]} \\ \hat{\boldsymbol{\Lambda}}\mathbf{f}_t \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ I_t \\ \boldsymbol{\epsilon}_t \end{pmatrix}$$

Results

- Resluts 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:

$$\widehat{MSE}(\hat{\mathbf{a}}_{t|t}) = \frac{1}{(T-d)} \sum_{t=d+1}^{T} \mathbf{P}_{t|t}$$

- Nowcast accuracy:

$$\widehat{MSFE}(\hat{\mathbf{a}}_{t|t}) = \frac{1}{h} \sum_{t=T-h+1}^{T} \mathbf{P}_{t|t}$$

Results

- Number of common factors for Google trends: 2
- Correlations trend disturbance terms:

Model	$\hat{\rho}_{1,GT}$	(p-value)	$\hat{\rho}_{2,GT}$	(p-value)	$\hat{\rho}_{CC}$	(p-value)
CC					0.90	(0.0004)
GT	0.43	(0.39)	-0.40	(0.31)		
GT+CC	-0.04	(1.0)	0.05	(1.0)	0.90	(0.0007)

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

Results

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

	model		
	CC	GT	CC+GT
$\widehat{MSE}(L_t^{[y]})$	0.869	0.967	0.869
$\widehat{MSFE}(L_t^{[y]})$	0.715		
$\widehat{MSFE}(L_t^{[y]})$		0.988	0.709
week 1		0.989	0.707
week 2		0.987	0.712
week 3		0.989	0.709
week 4		0.989	0.713
week 5		0.977	0.691

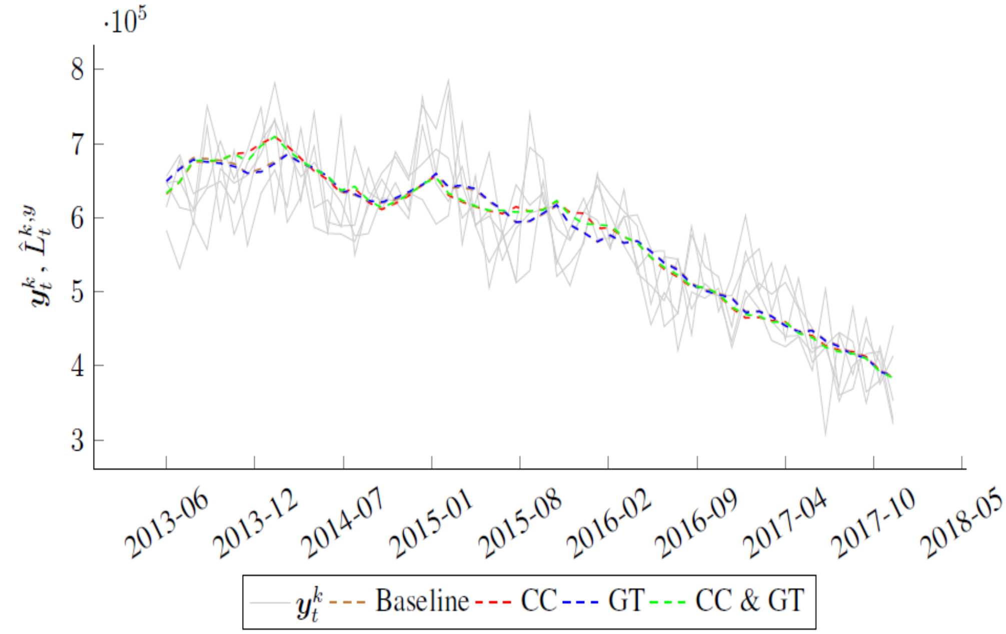
Results

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

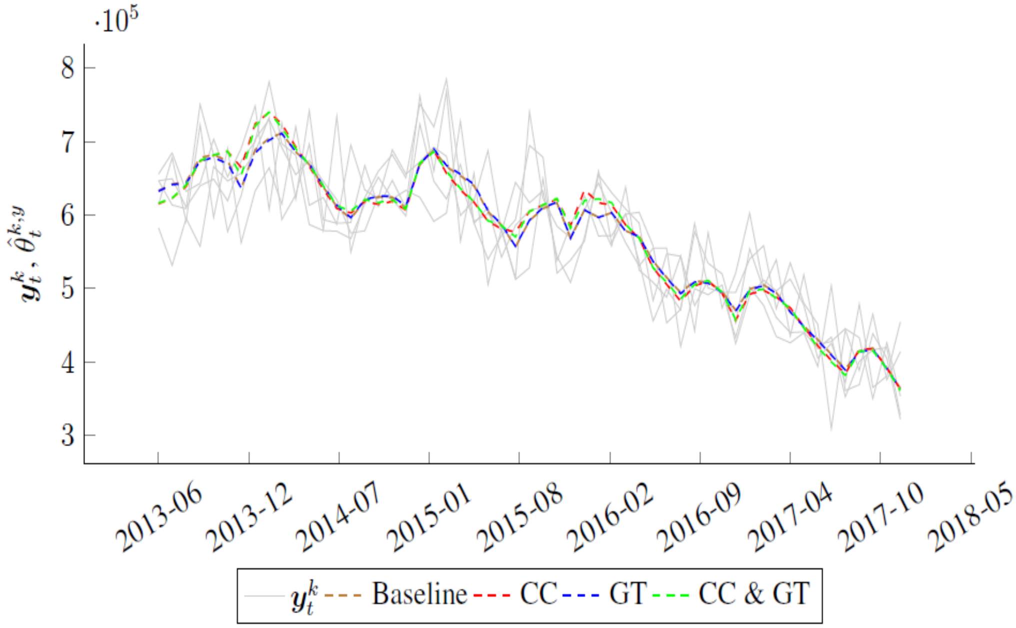
	model		
	CC	GT	CC+GT
$\widehat{MSE}(\theta_t^{[y]})$	0.890	0.977	0.889
$\widehat{MSFE}(\theta_t^{[y]})$	0.729		
$\widehat{MSFE}(\theta_t^{[y]})$		0.953	0.743
week 1		0.953	0.749
week 2		0.953	0.735
week 3		0.955	0.744
week 4		0.956	0.756
week 5		0.943	0.717

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 $t = 3k, k = 1, 2, \dots$ and missing otherwise
- x_t auxiliary series observed for $t = 1, 2, 3, \dots$
- Model for both series defined on a high frequency

$$L_t^z + S_t^{[z]} + I_t^{[z]}, \qquad z \in x, y$$

- $L_t^{[z]}$ 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_t^k (stock variable):

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

- 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: $\mathbf{y}_t = \mathbf{Z}\boldsymbol{\alpha}_t + \mathbf{I}_t$

  with $\mathbf{y}_t = (y_t^k, x_t)'$

- Transition equation: $\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t$

- $\boldsymbol{\alpha}_t = \begin{pmatrix} \boldsymbol{\alpha}_t^y \\ \boldsymbol{\alpha}_t^x \end{pmatrix}$
  - $\boldsymbol{\alpha}_t^y = (L_t^y, R_t^y, L_{t-1}^y, L_{t-2}^y, S_t^y, S_{t-1}^y, S_{t-2}^y)^t$
  - $S_t^y = (\gamma_{1\ t}^y, \tilde{\gamma}_{1\ t}^y, \gamma_{2\ t}^y)$
  - $S_{t-1}^y = (\gamma_{1\ t-1}^y, \gamma_{2\ t-1}^y)$
  - $S_{t-2}^y = (\gamma_{1\ t-2}^y, \gamma_{2\ t-2}^y)$
  - $\boldsymbol{\alpha}_t^x = (L_t^x, R_t^x, \gamma_{1\ t}^x, \tilde{\gamma}_{1\ t}^x, \ldots \gamma_{6\ t}^x)^t$

- $\mathbf{Z} = \begin{pmatrix} \mathbf{z}^y \\ \mathbf{z}^x \end{pmatrix}$
  - $\mathbf{z}^y = (1,0,1,1,\ 1,0,1,0,\ 1,1,1,1,\ \ 0,0,\ 0,0,0,0,0,0,0,0,0,0,0)$
  - $\mathbf{z}^x = (0,0,0,0,\ 0,0,0,0,\ 0,0,0,0,\ \ 1,0,\ 1,0,1,0,1,0,1,0,1,0,1)$

- $\mathbf{T} = BlockDiag(\mathbf{T}^y, \mathbf{T}^x)$
  - $\mathbf{T}^y = \begin{pmatrix} \mathbf{T}_L^y & \mathbf{0}_{[4\times4]} & \mathbf{0}_{[4\times2]} & \mathbf{0}_{[4\times2]} \\ \mathbf{0}_{[4\times4]} & \mathbf{T}_S^y & \mathbf{0}_{[4\times2]} & \mathbf{0}_{[4\times2]} \\ \mathbf{0}_{[2\times4]} & \mathbf{T}_{S-1}^y & \mathbf{I}_{[2]} & \mathbf{0}_{[2\times2]} \\ \mathbf{0}_{[2\times4]} & \mathbf{0}_{[2\times4]} & \mathbf{0}_{[2\times2]} & \mathbf{0}_{[2\times2]} \end{pmatrix}$
  - $\mathbf{T}_L^y = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
  - $\mathbf{T}_S^y = BlockDiag(\mathbf{C}_1, \mathbf{C}_2)$ (See Block 2)
  - $\mathbf{T}_{S-1}^y = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
  - $\mathbf{T}_S^x = BlockDiag(\mathbf{T}_L^x, \mathbf{T}_S^x)$ (See Block 2)
- $\boldsymbol{\eta}_t = \begin{pmatrix} \boldsymbol{\eta}_t^y \\ \boldsymbol{\eta}_t^x \end{pmatrix}$
  - $\boldsymbol{\eta}_t^y = (0, \eta_{R_t}^y, 0, 0, \omega_{1,t}^y, \omega_{1,t}^{*y}, \omega_{2,t}^y, \omega_{2,t}^{*y}, 0, 0, 0, 0)^t$
  - $\boldsymbol{\alpha}_t^x = (0, \eta_{R_t}^x, \omega_{1,t}^x, \omega_{1,t}^{*x}, \ldots \omega_{6,t}^x)^t$

References

Doz, C., Giannone, D., and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics* 164, 188–205.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics* 55, 665–676.

Harvey, A. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge.

Harvey, A. C. and Chung, C. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, A series* 163, 303–339.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics* 9, 163–175.

Schiavoni, C., Palm, F., Smeekes, S., and van den Brakel, J. (2019). *A dynamic factor model approach to incorporate Big Data in state space models for official statistics*. Technical report, Statistics Netherlands.

van den Brakel, J. and Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology* 35, 177–190.

van den Brakel, J. and Krieg, S. (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. *Survey Methodology* 41, 267–296.

van den Brakel, J. and Krieg, S. (2016). Small area estimation with state-space common factor models for rotating panels. *Journal of the Royal Statisitcal Society* 179, 763–791.