Macroeconomic nowcasting with big data through the lens of a sparse factor model[1]

Laurent Ferrara (Banque de France)
Anna Simoni (CREST, CNRS, ENSAE, École Polytechnique)

ECB Forecasting Conference
June 19, 2018

[1] The views expressed here are those of the authors and do not necessarily reflect those of the Banque de France.

What do mean by *Big Data* ?

In this paper, we disentangle between *official* and *alternative* datasets:

What do mean by *Big Data* ?

In this paper, we disentangle between *official* and *alternative* datasets:

- *Official data*
  - Released by National Statistical Institutes, Central Banks, International Organizations ...
  - Disaggregated / Granular data (micro and macro)
  - Well structured data

What do mean by *Big Data* ?

In this paper, we disentangle between *official* and *alternative* datasets:

- *Official data*
  - Released by National Statistical Institutes, Central Banks, International Organizations ...
  - Disaggregated / Granular data (micro and macro)
  - Well structured data
- *Alternative data* = Big data
  - Stemming from Google Trends, Web scraping, Scanner data, crowdsourcing with mobile phones ...
  - Data issues: Outliers, Structural breaks, Seasonal patterns ...
  - Reliability of the data and revisions (cf. debate between H. Varian and S. van Norden)

What do mean by *Big Data* ?

From *Official* to *Big Data* = From the 3 V's to the 5 V's

1. Volume
2. Variety (sources, format , ...)
3. Velocity

What do mean by *Big Data* ?

From *Official* to *Big Data* = From the 3 V's to the 5 V's

1. Volume
2. Variety (sources, format , ...)
3. Velocity
4. Variability
5. Veracity

Model characteristics for Big Data

Base equation for nowcasting y_t, for $t = 1, \dots, T$:

$$y_t = \beta_1 x_t^1 + \ldots + \beta_k x_t^k + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2)$$

Model characteristics for Big Data

Base equation for nowcasting y_t, for $t = 1, \ldots, T$:

$$y_t = \beta_1 x_t^1 + \ldots + \beta_k x_t^k + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2)$$

- Big data : large k, $k >> T$

Model characteristics for Big Data

Base equation for nowcasting y_t, for $t = 1, \ldots, T$:

$$y_t = \beta_1 x_t^1 + \ldots + \beta_k x_t^k + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2)$$

- Big data : large k, $k >> T$
- Models have to account for:
  - Computational efficiency
  - Frequency mismatch between y and x^i
  - Ragged-edge data (different reporting lags)

Model characteristics for Big Data

Base equation for nowcasting y_t, for $t = 1, \ldots, T$:

$$y_t = \beta_1 x_t^1 + \ldots + \beta_k x_t^k + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2)$$

- Big data : large k, $k >> T$
- Models have to account for:
  - Computational efficiency
  - Frequency mismatch between y and x^i
  - Ragged-edge data (different reporting lags)
- Standard statistical inference is not a good idea:
  - Too many parameters to estimate
  - High degree of uncertainty in estimates
  - Over-fitting and poor out-of-sample accuracy

Model characteristics for Big Data

- Methods to adress the curse of dimensionality (Giannone, Lenza, Primiceri, 2018)
    1. **Sparse models**: LASSO $=$ Some β_j are constrained to zero
    2. **Dense models**: Dynamic Factor Models $=$ all the variables have a role to play

Model characteristics for Big Data

- Methods to adress the curse of dimensionality (Giannone, Lenza, Primiceri, 2018)
  1. **Sparse models**: LASSO = Some β_j are constrained to zero
  2. **Dense models**: Dynamic Factor Models = all the variables have a role to play
- Are more data always needed ?
  1. **No**: Theoretical evidence by Boivin and Ng (2005) when forecasting with DFM (Dominant vs Dominated Factor)
  2. **No**: Empirical evidence by Barhoumi, Darne, Ferrara (2010) on nowcasting French GDP.
  3. A good idea: Pre-selection of data = Focus on **core datasets**. Examples: Targeted DFM by Bai and Ng (2008), application by Schumacher (2010) on German GDP

Is there a *Big Data hubris?*

Hot debate within institutions (NSIs, CBs, IOs), only few evidence.
2 main issues:

Is there a *Big Data hubris?*

Hot debate within institutions (NSIs, CBs, IOs), only few evidence.
2 main issues:

1. Is there a gain from using *Big Data* in nowcasting/forecasting?
    - No significant gain when controlling by *official* data (hard, soft, financial data), see e.g. Choi and Varian (12) *vs* Li (16)/ Gotz and Knetsch (17, BBK WP)

Is there a *Big Data hubris?*

Hot debate within institutions (NSIs, CBs, IOs), only few evidence.
2 main issues:

1. Is there a gain from using *Big Data* in nowcasting/forecasting?
    - No significant gain when controlling by *official* data (hard, soft, financial data), see e.g. Choi and Varian (12) *vs* Li (16)/ Gotz and Knetsch (17, BBK WP)
2. When is there a gain?
    - Significant gain when there is no information or only fragmented

Is there a *Big Data hubris?*

Hot debate within institutions (NSIs, CBs, IOs), only few evidence.
2 main issues:

1. Is there a gain from using *Big Data* in nowcasting/forecasting?
    - No significant gain when controlling by *official* data (hard, soft, financial data), see e.g. Choi and Varian (12) *vs* Li (16)/ Gotz and Knetsch (17, BBK WP)
2. When is there a gain?
    - Significant gain when there is no information or only fragmented
        - Emerging and Low Income Countries (Carriere-Swallow and Labbe, 13 JoF, Assessing real-time inflation in Venezuela and Argentina by Cavallo and Rigobon at MIT "The Billion Prices Project" ...)
        - Low frequency information: Annual World GDP (Ferrara and Marsilli, 17 World Eco.), Annual National Accounts in LICs ...
        - Lagged information (QNA nowcasts, flash estimates ...)
        - Measuring unobserved variables, e.g. Economic uncertainty (Baker et al., 16 QJE)

What do we do in this paper?

Main questions: Are Google data useful to nowcast EA GDP? And when?

- Estimate common factors of weekly Google data using a sparse factor (SPCA) model
- Integrate those factors into augmented bridge regressions using also official data (hard and soft)
- Nowcast EA GDP on a weekly basis accounting for data releases
- Assess the gain from using Google data in addition to official data
- Compare SPCA with standard approaches (PCA)
- Is data pre-selection based on targeting useful?

Main results

1. We point out the usefulness of Google trends data in nowcasting euro area GDP for the first four weeks of the quarter when there is no information (or rare information) about the state of the economy
2. As soon as official hard data become available (i.e. past GDP and industrial production), that is starting from week 5, then the relative nowcasting power of big data vanishes
3. SPCA does a better job when nowcasting the euro area GDP growth rate when compared with a fully dense approach (PCA), at least at the beginning of the quarter
4. We show that pre-selecting Google data before entering the nowcasting models appears to be a pertinent strategy.

Data

Our goal is to nowcast EA GDP (Y) using a bridge regression involving 3 types of predictors :

$$Y_t = \alpha_0 + \alpha_g' x_{t,g} + \alpha_s' x_{t,s} + \alpha_h' x_{t,h} + \varepsilon_t \tag{1}$$

where:

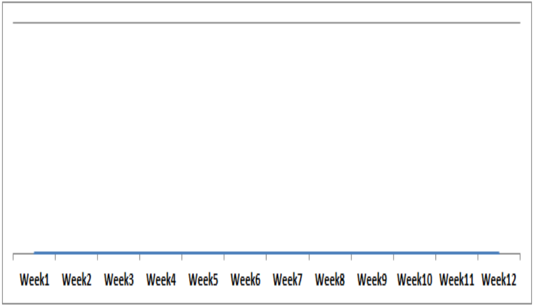
- $x_{t,g}$ is the N_g-vector of variables coming from Google data
- $x_{t,s}$ is the N_s-vector containing *soft* variables (e.g. DG EcFin surveys)
- $x_{t,h}$ is the N_h-vector containing *hard* variables (e.g. IPI)

Data

- **GDP:** EA quarterly GDP growth rate (1995q1 - 2016q3) as available on 24 April 2017 from Eurostat
- **Hard data:** EA monthly IPI growth rate released by Eurostat
- **Soft data:** EA monthly composite index of various sectors : EA Sentiment Index released by DG EcFin
- **Google data:** Google search data are weekly data related to queries performed with Google data, provided by ECB/DGS.
  Data are available for the 6 largest countries (Germany, France, Italy, Spain, Netherlands, Belgium) for 296 sub-categories by countries, that is a total of $N_g = 1776$ variables.

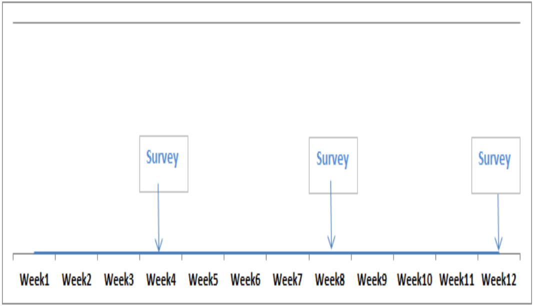
Timeline of data releases within the quarter

Google information is available for the 12 weeks of the quarter



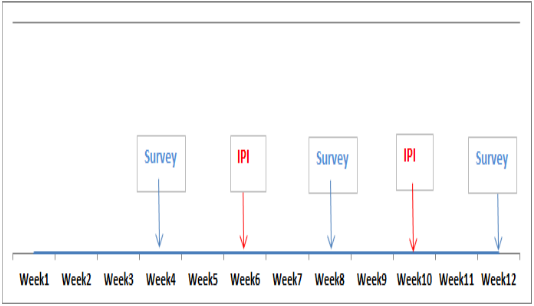
Timeline of data releases within the quarter

Survey data is available at weeks 4, 8 and 12

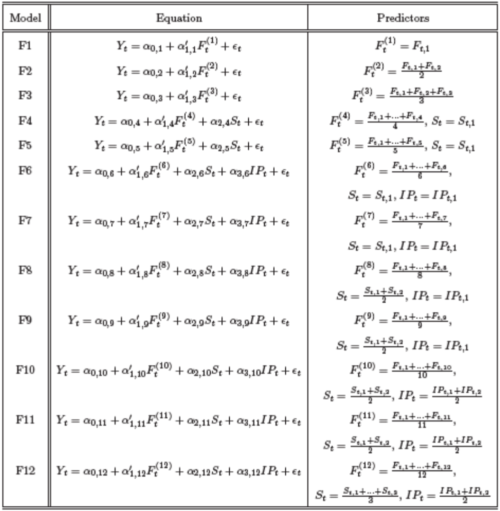


Timeline of data releases within the quarter

IPI information is available at weeks 6 and 10



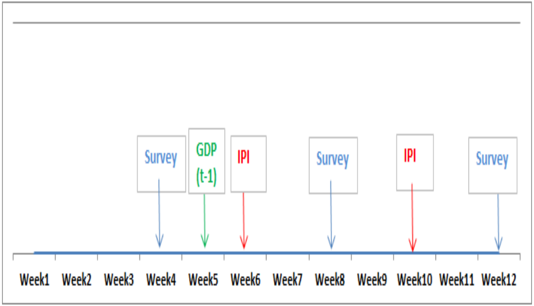
Factor-Augmented Bridge models for the 12 weeks



Model	Equation	Predictors
F1	$Y_t = \alpha_{0,1} + \alpha'_{1,1} F_t^{(1)} + \epsilon_t$	$F_t^{(1)} = F_{t,1}$
F2	$Y_t = \alpha_{0,2} + \alpha'_{1,2} F_t^{(2)} + \epsilon_t$	$F_t^{(2)} = \frac{F_{t,1}+F_{t,2}}{2}$
F3	$Y_t = \alpha_{0,3} + \alpha'_{1,3} F_t^{(3)} + \epsilon_t$	$F_t^{(3)} = \frac{F_{t,1}+F_{t,2}+F_{t,3}}{3}$
F4	$Y_t = \alpha_{0,4} + \alpha'_{1,4} F_t^{(4)} + \alpha_{2,4} S_t + \epsilon_t$	$F_t^{(4)} = \frac{F_{t,1}+\ldots+F_{t,4}}{4}$, $S_t = S_{t,1}$
F5	$Y_t = \alpha_{0,5} + \alpha'_{1,5} F_t^{(5)} + \alpha_{2,5} S_t + \epsilon_t$	$F_t^{(5)} = \frac{F_{t,1}+\ldots+F_{t,5}}{5}$, $S_t = S_{t,1}$
F6	$Y_t = \alpha_{0,6} + \alpha'_{1,6} F_t^{(6)} + \alpha_{2,6} S_t + \alpha_{3,6} IP_t + \epsilon_t$	$F_t^{(6)} = \frac{F_{t,1}+\ldots+F_{t,6}}{6}$, $S_t = S_{t,1}, IP_t = IP_{t,1}$
F7	$Y_t = \alpha_{0,7} + \alpha'_{1,7} F_t^{(7)} + \alpha_{2,7} S_t + \alpha_{3,7} IP_t + \epsilon_t$	$F_t^{(7)} = \frac{F_{t,1}+\ldots+F_{t,7}}{7}$, $S_t = S_{t,1}, IP_t = IP_{t,1}$
F8	$Y_t = \alpha_{0,8} + \alpha'_{1,8} F_t^{(8)} + \alpha_{2,8} S_t + \alpha_{3,8} IP_t + \epsilon_t$	$F_t^{(8)} = \frac{F_{t,1}+\ldots+F_{t,8}}{8}$, $S_t = \frac{S_{t,1}+S_{t,2}}{2}$, $IP_t = IP_{t,1}$
F9	$Y_t = \alpha_{0,9} + \alpha'_{1,9} F_t^{(9)} + \alpha_{2,9} S_t + \alpha_{3,9} IP_t + \epsilon_t$	$F_t^{(9)} = \frac{F_{t,1}+\ldots+F_{t,9}}{9}$, $S_t = \frac{S_{t,1}+S_{t,2}}{2}$, $IP_t = IP_{t,1}$
F10	$Y_t = \alpha_{0,10} + \alpha'_{1,10} F_t^{(10)} + \alpha_{2,10} S_t + \alpha_{3,10} IP_t + \epsilon_t$	$F_t^{(10)} = \frac{F_{t,1}+\ldots+F_{t,10}}{10}$, $S_t = \frac{S_{t,1}+S_{t,2}}{2}$, $IP_t = \frac{IP_{t,1}+IP_{t,2}}{2}$
F11	$Y_t = \alpha_{0,11} + \alpha'_{1,11} F_t^{(11)} + \alpha_{2,11} S_t + \alpha_{3,11} IP_t + \epsilon_t$	$F_t^{(11)} = \frac{F_{t,1}+\ldots+F_{t,11}}{11}$, $S_t = \frac{S_{t,1}+S_{t,2}}{2}$, $IP_t = \frac{IP_{t,1}+IP_{t,2}}{2}$
F12	$Y_t = \alpha_{0,12} + \alpha'_{1,12} F_t^{(12)} + \alpha_{2,12} S_t + \alpha_{3,12} IP_t + \epsilon_t$	$F_t^{(12)} = \frac{F_{t,1}+\ldots+F_{t,12}}{12}$, $S_t = \frac{S_{t,1}+\ldots+S_{t,3}}{3}$, $IP_t = \frac{IP_{t,1}+IP_{t,2}}{2}$

Timeline of data releases within the quarter

GDP for the last quarter is available at week 5



Factor-Augmented Bridge models for the 12 weeks

For weeks $w = 1, \ldots, 3$, the models are of the form:

$$Y_t = \alpha_{0,w} + \alpha'_{1,w} F_t^{(w)} + \varepsilon_{t,w} \tag{2}$$

For week $w = 4$, the model is of the form:

$$Y_t = \alpha_{0,4} + \alpha'_{1,4} F_t^{(4)} + \alpha_{2,4} S_t + \varepsilon_{t,w} \tag{3}$$

For week $w = 5$, the model is of the form:

$$Y_t = \alpha_{0,5} + \alpha'_{1,5} F_t^{(w)} + \alpha_{2,5} S_t + \alpha_{4,w} Y_{t-1} + \varepsilon_{t,w} \tag{4}$$

For weeks $w = 5, \ldots, 12$, the models are of the form:

$$Y_t = \alpha_{0,w} + \alpha'_{1,w} F_t^{(w)} + \alpha_{2,w} S_t + \alpha_{4,w} Y_{t-1} + \alpha_{3,w} IP_t + \varepsilon_{t,w} \tag{5}$$

Sparse PCA approach

The factor model in matrix form:

$$X_g = F\Lambda' + e_g \tag{6}$$

PCA estimates solves the minimization problem:

$$\begin{array}{lccc} \mathsf{min} & & \|X_g - F\Lambda\|_F^2 & \\ \text{s.t.} & \frac{1}{T}F'F = I_k & \text{and} & \Lambda'\Lambda \text{ diagonal} . \end{array} \tag{7}$$

Sparse PCA approach

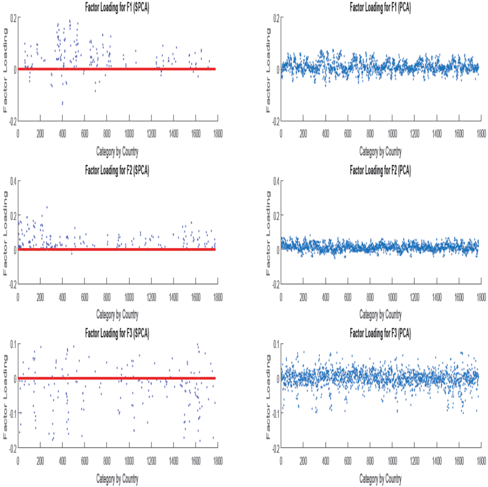
Sparse PCA estimator, proposed by Zou, Hastie, Tibshirani (2006, JCGS), solves:

$$\min_{A,B\in\mathbb{R}^{N_g\times k}} \quad \sum_{t=1}^{T}\|x_t - AB'x_t\|^2 + \mu_1\sum_{j=1}^{k}\|\beta_j\|^2 + \sum_{j=1}^{k}\mu_{2,j}\|\beta_j\|_1 \qquad \text{s.t.} \quad A'A = I_k. \tag{8}$$

where:
$\alpha_j, \beta_j \in \mathbb{R}^{N_g}$,
$A = [\alpha_1, \dots, \alpha_k]$, $B = [\beta_1, \dots, \beta_k]$,
$\|\beta_j\|^2 = \sum_{i=1}^{N_g}\beta_{ji}^2$ and $\|\beta_j\|_1 = \sum_{i=1}^{N_g}|\beta_{ji}|$.

Costs/Benefits: μ_1 and $\mu_{2,1}, \dots, \mu_{2,k}$ have to be estimated, but factors are easier to interpret.

Sparse PCA approach



Preselection of data using SIS approach

- Main idea: Pre-select Google data by targeting GDP (Bai-Ng, 2008 JoE, Schumacher, 2010, EcoLet).

Preselection of data using SIS approach

- Main idea: Pre-select Google data by targeting GDP (Bai-Ng, 2008 JoE, Schumacher, 2010, EcoLet).

- We follow Fan and Lv (2008, JRSS): **Sure Independence Screening** *"all important variables survive after applying a variable screening procedure, with probability tending to 1"*

Preselection of data using SIS approach

- Main idea: Pre-select Google data by targeting GDP (Bai-Ng, 2008 JoE, Schumacher, 2010, EcoLet).

- We follow Fan and Lv (2008, JRSS): **Sure Independence Screening** *"all important variables survive after applying a variable screening procedure, with probability tending to 1"*

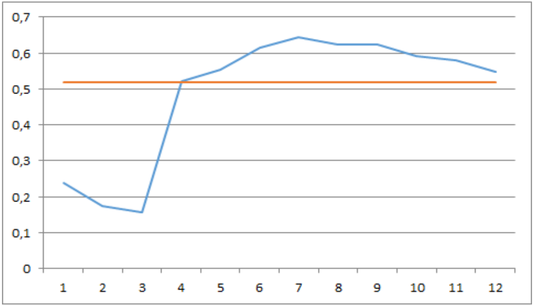
- Basic idea of SIS: only the variables with the highest absolute correlation should be used in modelling (correlation learning, hard thresholding).

Preselection of data using SIS approach

- Only consider $\overline{X}_g$ the $T \times N_g$ matrix of average Google variables as explanatory variables.
- Let $M^* = \{1 \leq i \leq N_g : \beta_g^i \neq 0\}$ be the true sparse model with non-sparsity size $s = |M^*|$.
- Compute $\omega = (\omega_1, \ldots, \omega_{N_g})'$ the vector of absolute marginal correlations of predictors with the response variable Y,
- For any given $\lambda \in]0,1[$, the N_g componentwise magnitudes of the vector ω are sorted in a decreasing order and we define a submodel M_λ such as: $M_\lambda = \{1 \leq i \leq N_g : |\omega_i|$ is among the first $[\lambda N_g]$ largest of all $\}$,
- Under some conditions the sure screening property holds, namely for a given λ:

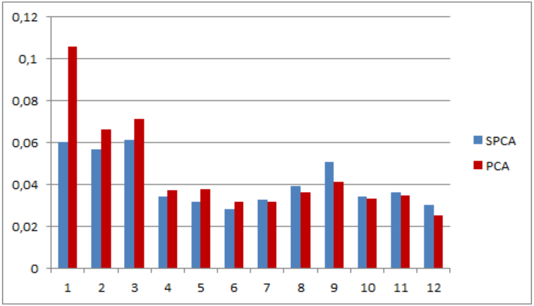
$$P(M^* \subset M_\lambda) \to 1, \qquad N_g \to \infty$$

Example: EA GDP nowcasts for 2016q1

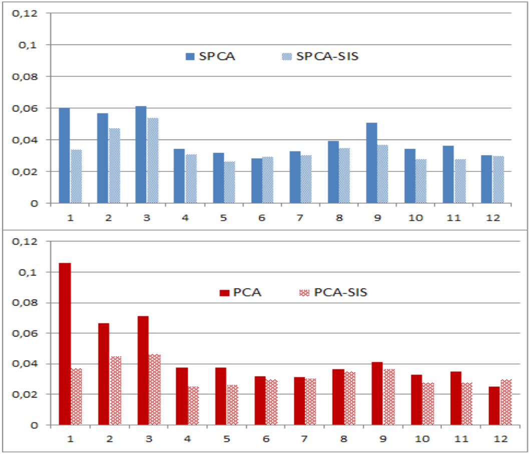


Comparing SPCA and PCA based on MSFEs

Gain from SPCA during the first 3 weeks, but no gain on average over the quarter



Is it worth to preselect Google data?



When is there a gain from using Google data?

We compare MSFEs with those from bridge models with no Google information: There is a gain of using Google data during the first 4 weeks of the quarter.

	Week 4	Week 6	Week 8	Week 10	Week 12	mean
No factors	0.0616	**0.0183**	**0.0259**	**0.0250**	**0.0255**	**0.0313**
SPCA+SIS	**0.0308**	0.0296	0.0350	0.0279	0.0299	0.0332

Conclusion = Main results

1. We point out the usefulness of Google trends data in nowcasting euro area GDP for the first four weeks of the quarter when there is no information (or rare information) about the state of the economy
2. As soon as official hard data become available (i.e. past GDP and industrial production), that is starting from week 5, then the relative nowcasting power of big data vanishes
3. SCPA does a better job when nowcasting the euro area GDP growth rate when compared with a fully dense approach (PCA), at least at the beginning of the quarter
4. We show that pre-selecting Google data before entering the nowcasting models appears to be a pertinent strategy.