When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage

Laurent Ferrara *(SKEMA Business School)*

joint with Anna Simoni *(CREST, CNRS, Ecole Polytechnique, ENSAE)*

Dauphine Seminar - 18 Feb. 2023

Outline:

1. Introduction
2. Methodology
3. Theoretical Properties
4. Empirical Study

Background. I

- Empirical motivation: evaluate the current-quarter nowcast of real economic activity, measured by the growth rate of GDP, by incorporating both official and alternative (high-dimensional) data $\Rightarrow$ crucial for policy-makers to assess economic conditions in real-time.

- Official information is typically released with a reporting lag of several weeks (or months) and is often revised a few months later.

- Usual macroeconomic nowcasting tools integrate standard official macroeconomic information (from National Statistical Institutes, Central Banks, International Organizations):

  (i) hard data (production, sales, employment . . .),

  (ii) opinion surveys (households or companies are asked about their view on current and future economic conditions),

  (iii) financial markets information (generally available on high frequency basis).

Background. II

- **Alternative data** provide a more timely (high-frequency) information. Many examples in the literature: *e.g.* economic News data, scraped data, scanner data, satellite data, data from Internet retailers about daily prices and detailed product attributes on products, Google Trends, Google Search, . . .

- The main research questions related to alternative datasets are:

  (i) When such data do improve nowcasting accuracy?

  (ii) Are they useful even after controlling for official variables, such as opinion surveys or production, generally used by forecasters?

Main contributions:

Our paper:

- proposes a new methodology to deal with high-dimensional alternative data with the purpose of economic nowcasting:

  2-step procedure: targeted preselection + Ridge (Ridge after model selection).

- establishes sure screening property of our procedure;
- establishes in-sample and out-of-sample (OOS) large-sample properties for the proposed method.
- we evaluate optimality of Generalized Cross validation (GCV) to choose the regularisation parameter α for OOS prediction.
- Motivating real data application: Google search data (GSD) for nowcasting GDP growth. We answer questions (i) and (ii) above, *i.e.*

  (i) when do *Google search data* improve nowcasting accuracy?

  (ii) Are *Google search data* useful to nowcast GDP growth even after controlling for official variables?

Outline:

1 Introduction

2 Methodology

3 Theoretical Properties

4 Empirical Study

The nowcasting approach. I

Linear bridge equation to construct Y_t nowcasts by using predictors available at different frequencies:

$$Y_t = \beta_0 + \beta_s' x_{t,s} + \beta_h' x_{t,h} + \beta_g' x_{t,g} + \varepsilon_t, \qquad \mathbf{E}[\varepsilon_t | x_{t,s}, x_{t,h}, x_{t,g}] = 0, \quad (1)$$

where:

- t denotes a given quarter of interest, *e.g.* the 1^{st} quarter of 2005 is dated by $t =$ March2005,
- $x_{t,g}$: N_g-vector of Google search data (GSD) variables (or other alternative data like Economic News, weekly),
- $x_{t,s}$: N_s-vector containing *soft* variables (monthly),
- $x_{t,h}$: N_h-vector containing *hard* variables (monthly).

We are interested in settings where: $N_g \gg T$.

Step 1: Pre-selection. I

- Using all the variables in $x_{t,g}$ is not necessarily a good strategy (low correlation and noise).
- One should target more accurately the choice of variables to the variable Y_t to be nowcast.
- We preselect variables in $x_{t,g}$ by using the **Conditional Sure Independence Screening** (CSIS) method of Bai & Ng (2008) and Barut, Fan & Verhasselt (2016).
- Basic idea of CSIS: measuring the contribution of each variable conditional on a subset of covariates.

Step 1: Pre-selection. II

- The algorithm is the following: (Bai & Ng, 2008)

  1. for each $j = 1, \ldots, N_g$, regress Y_t on a constant, $x_{t,s}$, $x_{t,h}$ and $x_{t,g,j}$, and compute the corresponding t-statistics t_j associated with $x_{t,g,j}$.
  2. select the variables in $x_{t,g}$ that have the absolute value $|t_j|$ largest than a given threshold λ:

$$\widehat{M}_g := \widehat{M}_g(\lambda) := \{1 \leq j \leq N_g : |t_j| > \lambda\}.$$

- We take λ as the $(1-\tau)$-quantile of a $\mathcal{N}(0,1)$ distribution with $\tau \in \{20\%, 10\%, 5\%, 2.5\%, 1\%, 0.5\%\}$.
- Let $N_1 := 1 + N_s + N_h$. For any given λ we have a submodel $\widehat{M}_g$ and we denote

$$\widehat{M} = \widehat{M}(\lambda) := \{1, \ldots, N_1\} \cup \{N_1 + j; j \in \widehat{M}_g(\lambda)\}.$$

Sure Screening property. I

- $N_1 := 1 + N_s + N_h$, $N := N_1 + N_g$,
- $X_t := (1, x'_{t,s}, x'_{t,h}, x'_{t,g})'$, $\kappa_T := K_T B(N_1 + 1) + m_0 K_T^{\rho}/s_0$, with K_T given in Assumption 5 (iii),
- $\beta := (\beta_0, \beta'_s, \beta'_h, \beta'_g)'$.

Assumption 1

(i). $Y_t = \beta'_* X_t + \varepsilon_t$, $t = 1, \ldots, T$, $\mathbf{E}[\varepsilon|X_t] = 0$ *and* $\mathbf{E}[\varepsilon\varepsilon'|X_t] = \sigma^2 I$, *where* $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)'$;

(ii). $\beta_{*j} \neq 0$, $\forall j \leq N_1$ *and* $|\beta_{*g}|_0 \leq s_g^* \leq N_g$;

(iii). $\varepsilon_t|X_t$, $t \geq 1$ *are independent sub-Gaussian random variables.*

The true sparse model is $M^* := \{1, \ldots, N_1\} \cup \{N_1 + j; j \in M_g^*\}$, where $M_g^* := \{1 \leq j \leq N_g : \beta_{*g,j} \neq 0\}$ and $s_g^* = |M_g^*|$.

Sure Screening property. II

Theorem 1 (Sure Screening property)

Let Assumptions 3-5 hold and assume: $\frac{T^{1-2\kappa}}{(\kappa_T K_T)^2} \to \infty$ *and* $T^{-\kappa/2} K_T^{\rho/2} = \mathcal{O}(1)$ *with* $\kappa < 1/2$ *given in Assumption 5 (i). Then, by taking* $\lambda = c_7 T^{1/2-\kappa}$ *for some constant* $c_7 > 0$,

$$P\left(M_g^* \subset \widehat{M}_g(\lambda)\right) \geq 1 - 8s_g^*(N_1+1)\exp\left\{-\frac{\min\{c_2, b_1^2/4, 1/4\}}{\kappa_T^2 K_T^2} T^{1-2\kappa}\right\}$$

$$- 6s_g^* T m_2 e^{-m_0 K_T^\rho} - 2s_g^* T \exp\left\{-\frac{K_T^\rho}{4CK_1^2}\right\} - 2s_g^* \exp\left\{-c_1 T^{1-2\kappa} \min\left\{\frac{c_\epsilon^2}{4K_1^2}, \frac{c_\epsilon}{2K_1}\right\}\right\},$$

where $c_2 := \underline{C}_x^2 c_1^2/(256N_1)$, $m_2 := (N_1 m_1 + \sqrt{s_1}\sqrt{\mathbf{E}[\exp\{4C_m s_0^2 \|\varepsilon\|_{\psi_2}^2\}]})$,

$b_1, C, c_\epsilon, c_1,\ C_m$ *are positive constants,*
$\|\varepsilon\|_{\psi_2} := \max_t \sup_{p\geq 1} p^{-1/2} \left(\mathbf{E}\left[|\varepsilon_t|^p \mid X_t\right]\right)^{1/p}$,
$K_1 := \max_{j\in M_g^*} \max_t \|\widetilde{\varepsilon}_{t,j}^2\|_{\psi_1}$, *with* $\|\cdot\|_{\psi_1}$ *denoting the sub-exponential norm.*

Sure Screening property. III

- If $\log(s_g^* N_1)\frac{(\kappa_T K_T)^2}{T^{1-2\kappa}} < \min\left\{c_4, \frac{b_1}{4}, \frac{1}{4}\right\}$, $\frac{\log(s_g^* T m_2)}{K_T^\rho} < m_0$, $\log(s_g^*) = o\left(T^{1-2\kappa}/\max(K_1^2, K_1)\right)$ then the upper bound in Theorem 1 converges to one.

- If $0 < c < \min\left\{\frac{c_\epsilon^2}{K_1^2}, \frac{c_\epsilon}{K_1}\right\} < C$, for two positive constants c, C, and if $K_T \asymp T^{(1-2\kappa)/A}$ where $A := \max\{4+\rho, 2+3\rho\}$, then

$$P\left(M_g^* \subset \widehat{M}_g(\lambda)\right) \gtrsim 1 - s_g^* m_2 \exp\{-CT^{(1-2\kappa)\rho/A}\}.$$

- In this case, we can deal with an N_g such that $\log(N_g) = o(T^{(1-2\kappa)\rho/A})$.

Step 2: Ridge regression. I

Even after pre-selection, the number of variables in $x_{t,g}$ may remain large compared to the time dimension T.

Hence, we use **Ridge (or Tikhonov) regularization**:

$$\widehat{\beta} := \arg \min_{\beta;\beta_{g,j}=0,j\in\widehat{M}_g^c} \left\{ \frac{1}{T}\sum_{t=1}^{T} \left(Y_t - \beta_0 - \beta_s' x_{t,s} - \beta_h' x_{t,h} - \beta_g' x_{t,g}\right)^2 + \alpha\|\beta\|_2^2 \right\},$$

where $\alpha > 0$ is the shrinkage parameter.

Equivalently, $\widehat{\beta} := (\underline{\widehat{\beta}}_{\widehat{M}}', \mathbf{0}')'$ where

$$\underline{\widehat{\beta}}_{\widehat{M}} = \underline{\widehat{\beta}}_{\widehat{M}}(\alpha) = \left(\frac{1}{T}\sum_{t=1}^{T} X_{t,\widehat{M}} X_{t,\widehat{M}}' + \alpha I_{\widehat{M}} \right)^{-1} \frac{1}{T}\sum_{t=1}^{T} X_{t,\widehat{M}}' Y_t,$$

where $X_{t,\widehat{M}} := (1, x_{t,s}', x_{t,h}', x_{t,g,\widehat{M}_g}')'$ and $x_{t,g,\widehat{M}_g} := \{x_{t,g,j}; j \in \widehat{M}_g\}$.

Step 2: Ridge regression. II

The shrinkage parameter α is chosen by Generalized cross-validation (GCV, see Li (1986, 1987)).

- Select the value of α that minimizes the following quantity:

$$GCV(\alpha) := \frac{\sum_{t=1}^{T}(Y_t - X'_{t,\widehat{M}}\widehat{\beta}_{\widehat{M}}(\alpha))^2}{T\left(1 - tr(X_{t,\widehat{M}}\left(T^{-1}\sum_{t=1}^{T}X_{t,\widehat{M}}X'_{t,\widehat{M}} + \alpha I_{\widehat{M}}\right)^{-1}X'_{t,\widehat{M}}/T)/T\right)^2},$$

  where $tr(\cdot)$ denotes the trace operator.

- Denote:

$$\widehat{\alpha}^{(w)} := \arg\min_{\alpha} GCV^{(w)}(\alpha).$$

Outline:

1. Introduction
2. Methodology
3. Theoretical Properties
4. Empirical Study

Theoretical Properties

For the *Ridge after model selection estimator* and for sparse models:

- we establish an upper bound for the in-sample and out-of-sample prediction error;
- out-of-sample evaluation of $\widehat{\alpha}^{(w)}$.

Out-of-sample evaluation of the selection of α. I

Recall: $\widehat{\alpha}^{(w)} := \arg\min_{\alpha} GCV^{(w)}(\alpha)$.

Aim: evaluate the performance of $\widehat{\alpha}_{\tau}^{(w)}$ for out-of-sample (OOS) prediction which is the objective of nowcasting for central bankers.

- Consider a new copy $(Y_{T+1}, X'_{T+1})'$ of $(Y_t, X'_t)'$ that satisfies:

$$Y_{T+1} = \sum_{j=1}^{s^*} x_{T+1,j}\beta_{*,j} + \varepsilon_{T+1} = \sum_{j=1}^{\widehat{m}+s^*} x_{T+1,j}\beta_{*,j} + \varepsilon_{T+1}$$

  with $\beta_{*,j} = 0$ for every $j \in \{s^*+1, \ldots, \widehat{m}+s^*\}$, $\mathbf{E}[\varepsilon_{T+1}] = 0$ and $Var(\varepsilon_{T+1}) = \sigma^2$.

- Denote: $Y^{(T)} := (Y_1, \ldots, Y_T)'$ and $X^{(T)} := (X_1, \ldots, X_T)'$

Out-of-sample evaluation of the selection of α. II

- Given a selected value $\widehat{\alpha}$, we evaluate its OOS performance by considering the conditional mean squared prediction error given by

$$\rho^2(\alpha; Y^{(T)}, X^{(T)}) := \mathbf{E}[(Y_{T+1} - \widehat{\beta}(\alpha)' X_{T+1})^2 | Y^{(T)}, X^{(T)}, \mathcal{A}],$$

  where $\widehat{\beta}(\alpha)' X_{T+1} = \sum_{j=1}^{\widehat{m}+s^*} x_{T+1,j} \widehat{\beta}_j(\alpha)$, $\widehat{\beta}_j(\alpha)$ is the Ridge after model selection estimator and $\mathcal{A} := \{M_* \subseteq \widehat{M}\}$.

- Denote by $\widehat{m}$ the number of incorrect covariates selected.

- Denote: $\widehat{\Sigma}_{\widehat{M}} := X'_{\widehat{M}} X_{\widehat{M}}/T$, $\Sigma_{\widehat{M}} := \mathbf{E}[X_{t,\widehat{M}} X'_{t,\widehat{M}}]$ and for a matrix A, $\|A\|_{op}$ denotes its operator norm. Moreover, $\mathcal{B} := \left\{ X^{(T)}; \left\| \widehat{\Sigma}_{\widehat{M}} - \Sigma_{\widehat{M}} \right\|_{op} \leq C \sqrt{\frac{\widehat{m}+s^*}{T}} \right\}$ with $0 < C < \infty$ a universal constant.

Out-of-sample evaluation of the selection of α. III

Assumption 2

Assume that:

(i) $\mathbf{E}[\varepsilon_t^4] < C_1$ *for some constant* $0 < C_1 < \infty$;

(ii) $\mathbf{E}[X_{t,\widehat{M}} X'_{t,,\widehat{M}} | X^{(t-1)}, \mathcal{A}] = \Sigma_{\widehat{M}}$ *and* $\Sigma_{\widehat{M}}$ *is bounded almost surely;*

(iii) $P(M^* \nsubseteq \widehat{M}) = o(r_{\mathcal{A},T})$ *where* $r_{\mathcal{A},T}$ *is a non-stochastic sequence independent of* α *that converges to zero as* $T \to \infty$;

(iv) $P(\mathcal{B}^c) = o(r_{\mathcal{B},T})$ *where* $r_{\mathcal{B},T}$ *is a non-stochastic sequence independent of* α *that converges to zero as* $T \to \infty$;

(v) *for any index set* $M \subset \{1, \ldots, N\}$, *there exist a* $w(M) \in \mathbb{R}^{|M|}$ *and* $\gamma > 0$ *such that* $\underline{\beta}_{*,M} = \Sigma_M^{\gamma/2} w(M)$ *and* $\|w(M)\|_2 < \infty$ *(source condition).*

Out-of-sample evaluation of the selection of α. IV

Theorem 2

Let $\alpha > 0$ *and* $\widetilde{\gamma} := \min\{\gamma, 2\}$. *Assume that: (i)* $\alpha \to 0$, *(ii)* $\alpha^2 T \to \infty$, *and (iii)* $\frac{(\widehat{m}+s^*)^2}{T^2} \to 0$. *Then, under Assumptions 3 and 2:*

$$\left|\rho^2(\alpha; Y^{(T)}, X^{(T)}) - GCV(\alpha)\right|$$
$$= \mathcal{O}_p\left(\alpha^{(\gamma\wedge 2)/2} + \frac{1}{\sqrt{\alpha T}} + \frac{(\widehat{m}+s^*)}{T}\left(1 + \frac{1}{\alpha^2 T}\right)\right) + \mathcal{O}_p\left(\max\{r_{\mathcal{A},T}, r_{\mathcal{B},T}\}\right).$$

Moreover, for any constants $0 < \underline{\alpha} < \infty$ *and* $0 < \underline{u} < 1/2$, *and for a sequence* $\overline{\alpha}_T \to 0$ *as* $T \to \infty$ *such that:* $T^{-1/2+\underline{u}}/\overline{\alpha}_T \to 0$, *it holds:*

$$\sup_{\alpha \in [\underline{\alpha} T^{-1/2+\underline{u}}, \overline{\alpha}_T]} \left|\rho^2(\alpha; Y^{(T)}, X^{(T)}) - GCV(\alpha)\right| = \mathcal{O}_p\left(r_T\right) + \mathcal{O}_p\left(\max\{r_{\mathcal{A},T}, r_{\mathcal{B},T}\}\right),$$

where $r_T := \overline{\alpha}_T^{(\gamma\wedge 2)/2} + T^{-(1+\underline{u})/2} + \frac{(\widehat{m}+s^*)}{T}\left(1 + T^{-2\underline{u}}\right)$.

Out-of-sample evaluation of the selection of α. V

Theorem 3

In the setting of Theorem 2, assume that $\alpha \to 0$ *and* $\frac{(\widehat{m}+s^*)}{T} \to 0$. *Consider the minimizers of* $GCV(\alpha)$ *and* $\rho^2(\alpha; Y^{(T)}, X^{(T)})$*:*
$\widehat{\alpha} = \arg\min_{\alpha \in [\underline{\alpha} T^{-1/2+\underline{u}}, \overline{\alpha}_T]} GCV(\alpha)$ *and*
$\alpha^* = \arg\min_{\alpha \in [\underline{\alpha} T^{-1/2+\underline{u}}, \overline{\alpha}_T]} \rho^2(\alpha; Y^{(T)}, X^{(T)})$. *Then,*
(i) $\widehat{\alpha}$ *is as good as* α^* *for out-of-sample prediction in the sense that*

$$|\rho^2(\widehat{\alpha}; Y^{(T)}, X^{(T)}) - \rho^2(\alpha^*; Y^{(T)}, X^{(T)})| = O_p(r_T) + \mathcal{O}_p(\max\{r_{\mathcal{A},T}, r_{\mathcal{B},T}\}),$$

and (ii) the out-of-sample predictive performance can be consistently estimated in the sense that

$$|GCV(\widehat{\alpha}) - \rho^2(\widehat{\alpha}; Y^{(T)}, X^{(T)})| = O_p(r_T) + \mathcal{O}_p(\max\{r_{\mathcal{A},T}, r_{\mathcal{B},T}\}).$$

▶ Simulations

Outline:

1. Introduction
2. Methodology
3. Theoretical Properties
4. Empirical Study

Empirical Study.

- Apply our *Ridge after Model selection* procedure to nowcast GDP growth rate with weekly GSD for three countries/areas:
  - the euro area (EA, hereafter),
  - the U.S.
  - Germany.
- Three various phases of the business cycle:
  - a calm period (2014-16),
  - a period with a sudden downward shift in GDP growth (2017-18),
  - a recession period with large negative growth rates (2008-09).
- *Pseudo* real-time analysis, that is, we account for the release dates of official variables but we do not use vintages of data.
- A robustness check for a *true* real-time nowcasting is carried out on EA data.

Data. I

- **Quarterly GDP growth rate Y_t:** from Eurostat for EA as a whole, from Destatis for Germany and from the BEA for the U.S.
- **Soft data:** a composite index of opinion surveys from various sectors as a proxy for soft variables, denoted by S_t. For Germany and EA: we use the sentiment indexes computed by the European Commission, while we use the ISM index for the U.S. economy.
- **Hard data:** growth rate of the industrial production index, which is the most used measure of hard data by practitioners, denoted by IP_t.
- **Google Search data:**
  - Google search data are weekly data related to queries performed with Google search. $\Rightarrow$ Google Search data $\neq$ Google Trends.
  - The queries are assigned by Google to particular categories using natural language processing methods.

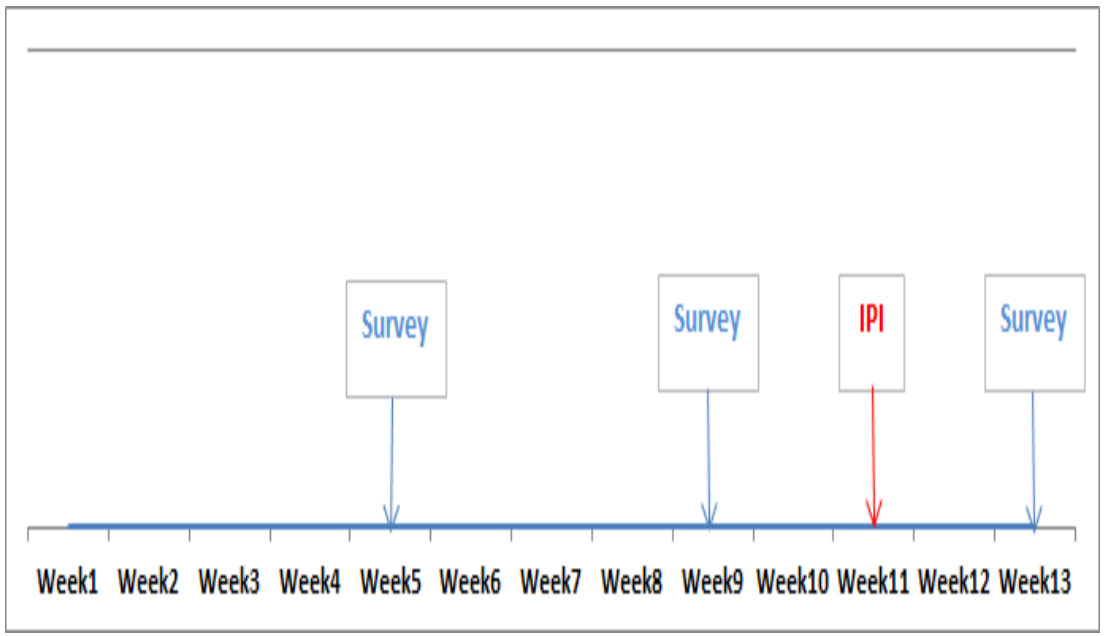
Data. II

- The data are indexes of weekly volume changes of Google queries grouped by category and by country. Data are normalized at 1 at the first week of January 2004.
- Then, the next values indicate the deviation from the first value. No information about the search volume.
- GSD are received and made available by the European Central Bank every Tuesday.
- We consider Google searches for the six main Euro area countries: Belgium, France, Germany, Italy, Netherlands and Spain, and for the U.S. For EA we have at disposal a total of $N_g = 1776$ variables, corresponding to 26 categories and 296 subcategories for each country.
- Original GSD are not seasonally adjusted, thus we take the growth rate over 52 weeks to eliminate the seasonality within the data.

Timeline of data releases within the quarter

In addition to frequency mismatch in the data, another challenge: data on official series and Google search are released with various reporting lags $\Rightarrow$ unbalanced information set at each point in time within the quarter (*ragged-edge database*).

For EA: S_t is available at weeks 5, 9 and 13, and IP_t at week 11.



The relevant information set.

- Because $x_{t,s}$, $x_{t,h}$ and $x_{t,g}$ are sampled over different frequencies and released with various reporting lags the relevant information set for calculating the nowcasts evolves within the quarter (unbalanced data set).
- Let $x_{t,j,i}^{(w)}$, $j \in \{s,h,g\}$, denote the i-th series in vector $x_{t,j}$ released at week $\leq w = 1,\ldots,13$ of quarter t.
- The relevant information set at week w of a quarter t is:

$$\Omega_t^{(w)} := \left\{ x_{t,j,i}^{(w)}, i = 1,\ldots,N_j, j \in \{s,h,g\} \right\}.$$

$\forall t = 1,\ldots,T$, $\forall w = 1,\ldots,13$, for each $\Omega_t^{(w)}$ within a given quarter t, the nowcast is computed as $\widehat{Y}_{t|w} := \mathbf{E}[Y_t|\Omega_t^{(w)}; M_{(w)}]$ based on the model:

$$M_{(w)}: \qquad \mathbf{E}[Y_t|\Omega_t^{(w)}] = \beta_{0,w} + \beta_{s,w}' x_{t,s}^{(w)} + \beta_{h,w}' x_{t,h}^{(w)} + \beta_{g,w}' x_{t,g}^{(w)}, \quad (2)$$

where $\beta_{j,w,i}^{(w)} = 0$ if $x_{t,j}^{(w)} \notin \Omega_t^{(w)}$.

Bridge models for the 13 weeks.

For weeks $w = 1, \ldots, 4$, the models are of the form:

$$Y_t = \beta_{0,w} + \beta'_{g,w} x^{(w)}_{t,g} + \varepsilon_{t,w} \tag{3}$$

From week $w = 5$ to $w = 10$, the models are of the form:

$$Y_t = \beta_{0,w} + \beta'_{g,w} x^{(w)}_{t,g} + \beta_{s,w} S_t + \varepsilon_{t,w} \tag{4}$$

For weeks $w = 11, \ldots, 13$, the models are of the form:

$$Y_t = \beta_{0,w} + \beta'_{g,w} x^{(w)}_{t,g} + \beta_{s,w} S_t + \beta_{h,w} IP_t + \varepsilon_{t,w} \tag{5}$$

Overall evaluation of Google search data.

The main empirical research questions are:

(I). are Google search data informative when there is no official data available for the forecaster?

(II). to what extent Google Search data remain informative when official data become available?

Main empirical results. I

We assess the overall gain from using Google search data when controlling from official data and focus on 3 various periods of time:

- in periods of cyclical stability $2014q1 - 2016q1$ (both pseudo real-time and true real-time);
- in period that exhibits a sharp downturn in GDP series ($2017q1 - 2018q4$)
- during recession periods ($2008q1 - 2009q2$).

For three countries/areas: the euro area (EA), the U.S. and Germany.

Main empirical results. II

- A simple Ridge regression with all variables shows an expected decline of RMSFEs over the 13 weeks of the current quarter

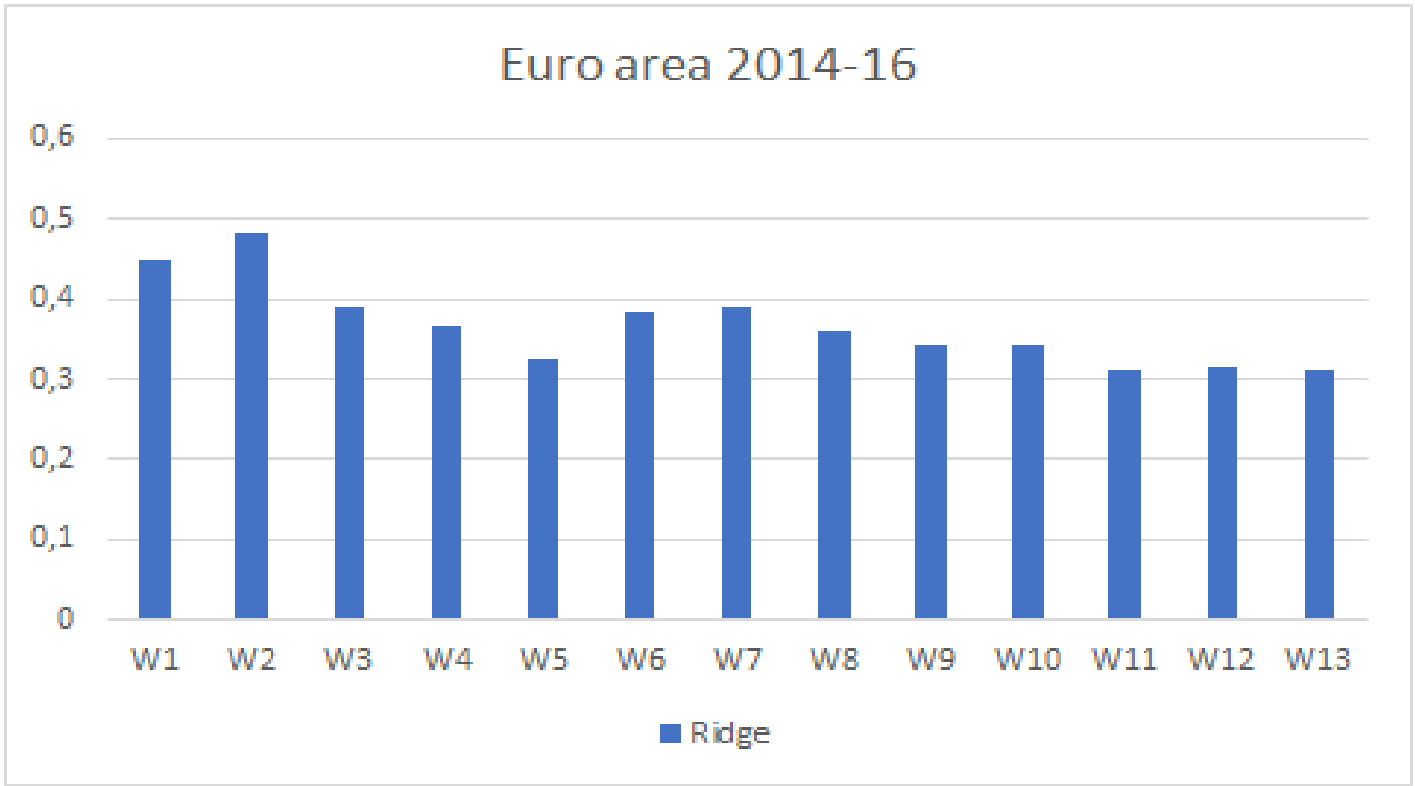


Figure: RMSFEs for simple Ridge regression EA 2014-16

Main empirical results. III

- Pre-selection step leads to a drop in RMSFEs compared to a simple Ridge regression using all variables

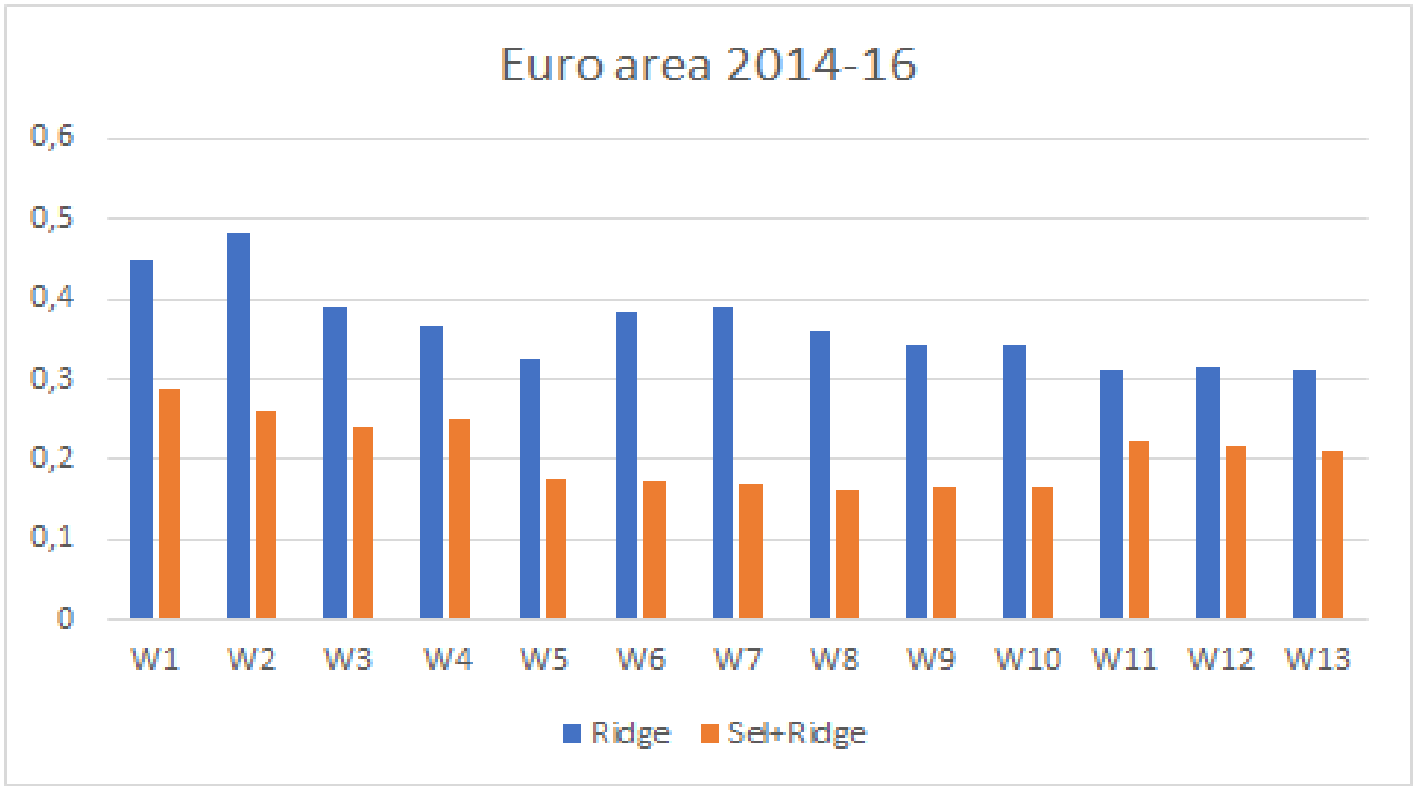


Figure: RMSFEs from simple Ridge regression and from SIS - EA 2014-16

Main empirical results. IV

- Comparison with models (i) without Google data (yellow bars) and (ii) with only Google data (green bars)

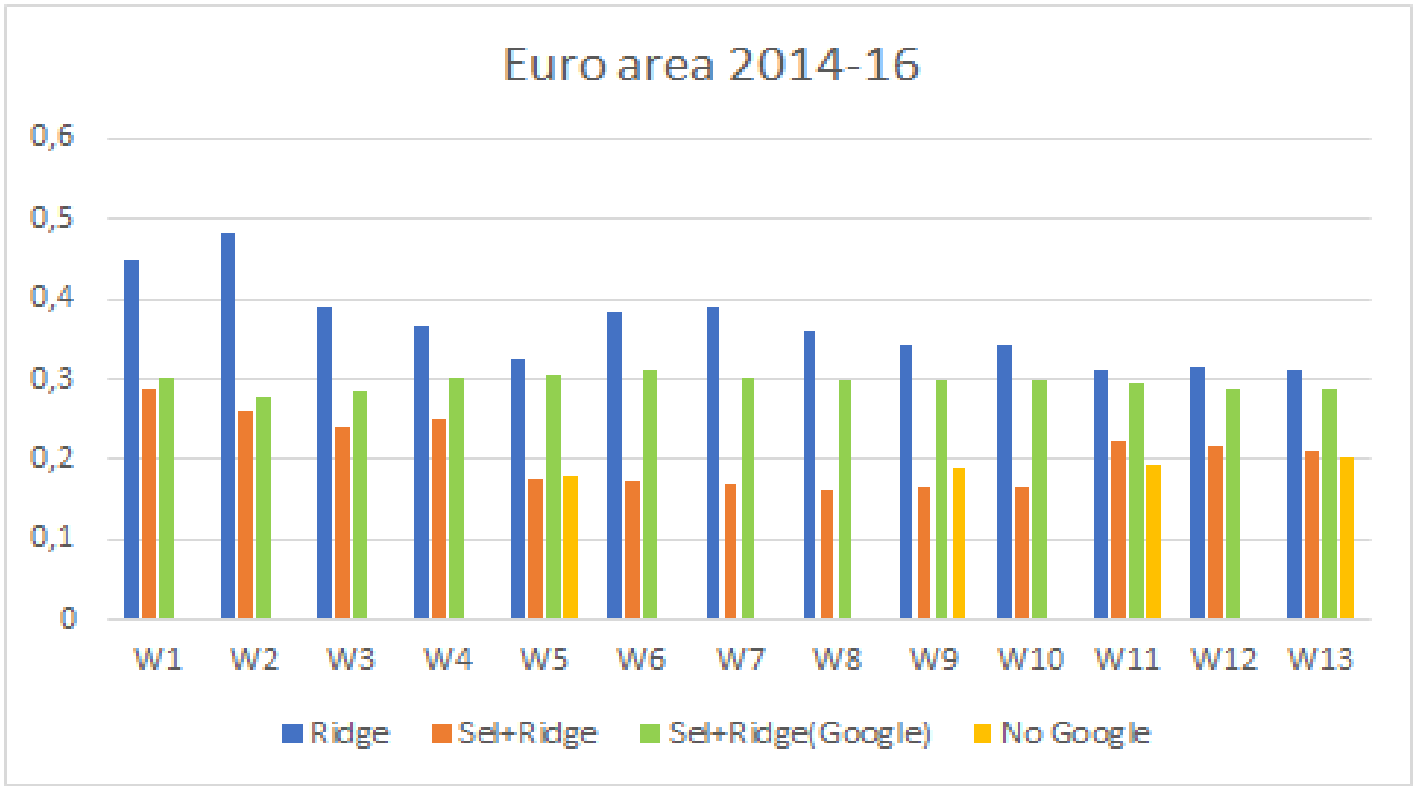


Figure: RMSFEs from all approaches EA 2014-16

Main empirical results. V

- Overall, GSD are useful when trying to nowcast GDP growth
- Combining macroeconomic and GSD variables in the same model appears to be generally fruitful.
  - At the beginning of the quarter, when there is no official information available about the current state of the economy, we show that using only GSD leads to very reasonable Mean Squared Forecasting Errors (MSFEs).
  - As soon as we integrate official macroeconomic information, starting from the fifth week of the quarter, MSFEs decrease reflecting the importance of this type of data in nowcasting.

Main empirical results. VI

- Recession periods present specific patterns: model with only GSD, without any preselection, gives better nowcasting accuracy.

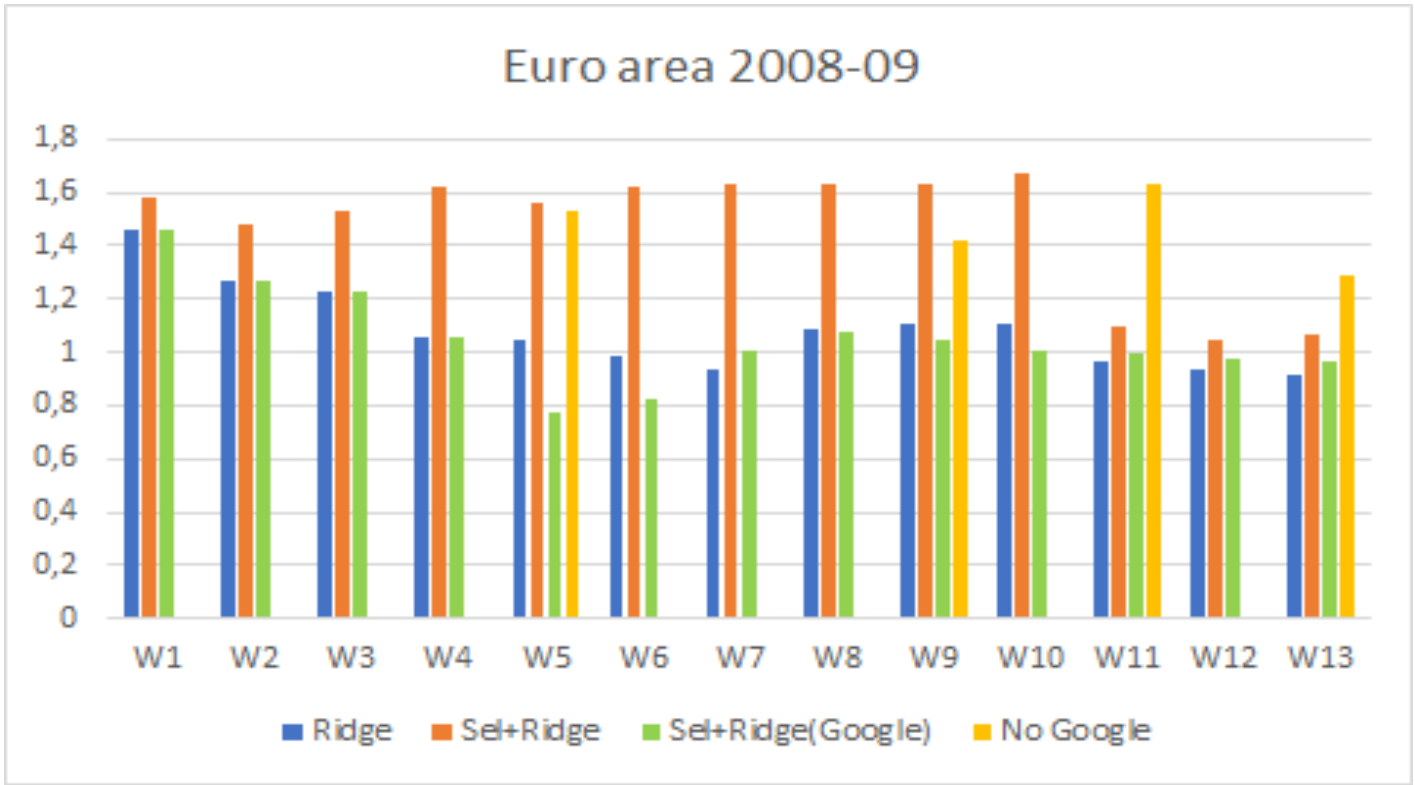


Figure: RMSFEs from all approaches EA during a recession (2008-09)

Conclusions

1. Propose a new methodology to deal with ultra-high dimensional data to evaluate the current-quarter nowcast of real economic activity (*Ridge after Model selection*).

2. Good theoretical properties: Sure Screening property, in-sample, out-of-sample.

3. Empirical analysis: first to use Google search data (which differ from Google Trends data) to nowcast GDP (for EA, U.S. and Germany).

4. We point out the usefulness of GSD for nowcasting when there is no official information available. As soon as official information are available, marginal nowcasting gain from GSD vanishes.

When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage

Laurent Ferrara *(SKEMA Business School)*

joint with Anna Simoni *(CREST, CNRS, Ecole Polytechnique, ENSAE)*

Dauphine Seminar - 18 Feb. 2023

Sure Screening property: Assumptions I.

- $N_1 := 1 + N_s + N_h$, $\quad N := N_1 + N_g$,
- $X_t := (1, x'_{t,s}, x'_{t,h}, x'_{t,g})'$, $\quad X_{t,O} := (1, x'_{t,s}, x'_{t,h})'$,
  $X_{t,O,j} := (X'_{t,O}, x_{t,g,j})'$, $\quad \forall j \in \{1, \ldots, N_g\}$,
- $\beta := (\beta_0, \beta'_s, \beta'_h, \beta'_g)'$.

Assumption 3

(i). $Y_t = \beta'_* X_t + \varepsilon_t$, $t = 1, \ldots, T$, $\mathbf{E}[\varepsilon | X_t] = 0$ *and* $\mathbf{E}[\varepsilon\varepsilon' | X_t] = \sigma^2 I$, *where* $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)'$;

(ii). $\beta_{*j} \neq 0$, $\forall j \leq N_1$ *and* $|\beta_{*g}|_0 \leq s_g^* \leq N_g$;

(iii). $\varepsilon_t | X_t$, $t \geq 1$ *are independent sub-Gaussian random variables.*

The true sparse model is $M^* := \{1, \ldots, N_1\} \cup \{N_1 + j; j \in M_g^*\}$, where $M_g^* := \{1 \leq j \leq N_g : \beta_{*g,j} \neq 0\}$ and $s_g^* = |M_g^*|$.

Sure Screening property: Assumptions II.

- $\forall j \in \{1, \ldots, N_g\}$, define:

$$\widetilde{\beta}_{O,j} := \arg \min_{\beta_O, \beta_{g,j}} \mathbf{E}(Y_t - X'_{t,O}\beta_O - x_{t,g,j}\beta_{g,j})^2,$$

  which is the pseudo-true value of $\beta_{O,j} := (\beta_O^{1'}, \beta_{g,j})' \in \mathbb{R}^{N_1+1}$ in the j-th misspecified model.

- $\mathcal{B} := \{\beta_{O,j}, j = 1, \ldots, N_g; |\beta_{O,j,k}| \leq B, \forall k = 1, \ldots, N_1 + 1\}$ for a large positive constant B is the set over which the Least Squares estimates in 1 are searched.

- Denote: $\sigma^2_{O,j} := \mathbf{E}[(Y_t - X'_{t,O,j}\widetilde{\beta}_{O,j})^2]$.

Sure Screening property: Assumptions III.

Assumption 4

(i). $\{(x'_{t,s}, x'_{t,h}, x'_{t,g})'\}_{t\geq 1}$ *are i.i.d.random vectors in* $\mathbb{R}^{N-1}$;

(ii). $\forall j \in \{1,\ldots,N_g\}$, $\mathbf{E}[\widetilde{\varepsilon}^4_{t,j}|X_{t,O,j}]$ *is bounded, where* $\widetilde{\varepsilon}_{t,j} := (y_t - X'_{t,O,j}\widetilde{\beta}_{O,j})$;

(iii). $\exists$ *two constants* $0 < \underline{C}_x^2 < \overline{C}_x^2 < \infty$ *such that:*

$$\underline{C}_x^2 \leq \min_{1\leq j\leq N_g} \lambda_{\min}\left(\mathbf{E}[X_{t,O,j}X'_{t,O,j}]\right) \leq \max_{1\leq j\leq N_g} \lambda_{\max}\left(\mathbf{E}[X_{t,O,j}X'_{t,O,j}]\right) \leq \overline{C}_x^2;$$

(iv). *let* $\sigma^2_{O,j} := \mathbf{E}[(Y_t - X'_{t,O,j}\widetilde{\beta}_{O,j})^2]$, $\exists$ *two constants* $0 < \underline{\sigma}^2_O < \overline{\sigma}^2_O < \infty$ *such that:*

$$\underline{\sigma}^2_O \leq \min_{1\leq j\leq N_g} \sigma^2_{O,j} \leq \max_{1\leq j\leq N_g} \sigma^2_{O,j} \leq \overline{\sigma}^2_O.$$

Sure Screening property: Assumptions IV.

Assumption 5

(i). *for* $j \in M_g^*$, $\exists$ *two constants* $c_1, c_2 > 0$ *and* $0 < \kappa < 1/2$ *such that:*

$$|cov(Y_t, x_{t,g,j}|X_{t,O})| \geq c_1 T^{-\kappa}, \quad \textit{and} \quad \max_{j \in \{1,\ldots,N_g\}} \mathbf{E}[x_{t,g,j}^2] \leq c_2;$$

(ii). $\mathbf{E}[X_{t,O,j}X_{t,O,j}']/\sigma_{O,j}^2$ *is finite and positive definite, and* $\|\mathbf{E}[X_{t,O,j}X_{t,O,j}']\|_{op}$ *is bounded from above;*

(iii). $\exists \epsilon_T > 0$ *such that* $\forall j \in \{1, \ldots, N_g\}$*:*

$$\sup_{\beta_{O,j} \in \mathcal{B}; \|\beta_{O,j} - \widetilde{\beta}_{O,j}\| \leq \epsilon_T} \frac{1}{2\sigma_{O,j}^2} \left| \mathbf{E}\left[(X_{t,O,j}'\beta_{O,j})^2 - (X_{t,O,j}'\widetilde{\beta}_{O,j})^2 \right] 1\{|x_{t,g,j}| > K_T\} \right| \leq o(N_1/T),$$

where K_T *is an arbitrarily large constant;*

Sure Screening property: Assumptions V.

(iv). $\exists$ positive constants m_0, m_1, s_0, s_1, ρ such that for sufficiently large τ,

$$P\left(|X_{t,j}| > \tau\right) \leq m_1 e^{-m_0 \tau^{\rho}}, \quad \forall j \in \{2, \ldots, N\},$$

and

$$\mathbf{E}\left[e^{2\beta_*' X_t s_0} + e^{-2\beta_*' X_t s_0}\right] \leq s_1;$$

(v). $\forall \beta_{O,j} \in \mathcal{B}$,

$$\mathbf{E}\left[(Y_t - X_{t,O}'\beta_O - x_{t,g,j}\beta_{g,j})^2 - (Y_t - X_{t,O}'\widetilde{\beta}_O - x_{t,g,j}\widetilde{\beta}_{g,j})^2\right] \geq V\|\beta_{O,j} - \widetilde{\beta}_{O,j}\|^2$$

for some positive constant V bounded from below uniformly over $j \in \{1, \ldots, N_g\}$.

In-sample and Out-of-sample Prediction error. I

- Let $X := (X_1, \ldots, X_T)'$ be a $(T \times N)$ matrix.
- Recall: $M^* := \{1 \leq j \leq N : \beta_{*j} \neq 0\}$ with $s^* := |M^*|$, and let M^{*c} denote the complementary set of M^* in $\{1, \ldots, N\}$.
- Denote: $\beta_{M,j} := \beta_j \mathbb{1}\{j \in M\}$.
- For a $\delta \in \mathbb{R}^N$ and given covariates X_t, $t = 1, \ldots, T$, define:
  - $\|\delta\|_{2,T}^2 := \delta' X' X \delta / T$, the prediction norm of δ,
  - $\|\delta\|_0 := \sum_{j=1}^N \mathbb{1}\{\delta_j \neq 0\}$, the ℓ_0-norm of δ,
  - $\|\delta\|_2 := \sqrt{\delta'\delta}$, the **Euclidean norm**.

In-sample and Out-of-sample Prediction error. II

Assumption 6 (Restricted Sparse Eigenvalue condition)

For a given $m < T$, for a $\delta \in \mathbb{R}^N$, with probability $1 - o(1)$,

$$\underline{\varphi}(m)^2 := \min_{\|\delta_{M^{*c}}\|_0 \leq m,\, \delta \neq 0} \frac{\|\delta\|_{2,T}^2}{\|\delta\|_2^2} > 0, \qquad \overline{\varphi}(m) := \max_{\|\delta_{M^{*c}}\|_0 \leq m,\, \delta \neq 0} \frac{\|\delta\|_{2,T}^2}{\|\delta\|_2^2} > 0.$$

In-sample and Out-of-sample Prediction error. III

Theorem 4 (In-sample prediction error)

Suppose that Assumptions 3 (i)-(ii) and 6 are satisfied and that $\varepsilon_t | X_t$ is Gaussian. Then, for every $\epsilon > 0$, there is a constant K_ϵ independent of T such that with probability at least $1-\epsilon$,

$$\|\widehat{\beta}-\beta_*\|_{2,T} \leq \left(K_\epsilon \sqrt{\frac{\widehat{m}\log(N) + (\widehat{m}+s^*)\log(e^2\mu(\widehat{m}))}{T}} + 2\alpha\|\beta_*\|_2 \frac{1}{\underline{\varphi}(\widehat{m})} \right) \mathbb{1}\{M^* \subseteq \widehat{M}\} +$$

$$\left(\frac{K_\epsilon \sigma}{\sqrt{T}} \sqrt{(\widehat{k}+\widehat{m})\left(\log(s^*+\widehat{m}) + \log(e^2\mu(\widehat{k}+\widehat{m}))\right)} + \frac{2\alpha}{\underline{\varphi}(\widehat{m})}\|\beta_*\|_2 + \|\beta_{*,M^*\setminus\widehat{M}}\|_{2,T} \right) \mathbb{1}\{M^* \not\subseteq \widehat{M}$$

where $\mu(\widehat{m}) = \frac{\sqrt{\overline{\varphi}(m)}}{\underline{\varphi}(m)}$, $\widehat{m} := |\widehat{M} \setminus M_*| \, \mathbb{1}\{\widehat{M} \supseteq M_*\}$ *and*
$\widehat{k} := |M_* \setminus \widehat{M}| \, \mathbb{1}\{M^* \not\subseteq \widehat{M}\}$.

The number $\widehat{m}$ is the number of incorrect covariates selected.

In-sample and Out-of-sample Prediction error. IV

Corollary 1 (Coefficient estimation)

Suppose that Assumptions 3 (i)-(ii) and 6 are satisfied and that $\varepsilon_t | X_t$ is Gaussian. Then, for every $\epsilon > 0$, there is a constant K_ϵ independent of T such that with probability at least $1-\epsilon$,

$$\|\widehat{\beta}-\beta_*\|_2 \leq \left(K_\epsilon \sqrt{\frac{\widehat{m}\log(N) + (\widehat{m}+s^*)\log(e^2\mu(\widehat{m}))}{T\underline{\varphi}(\widehat{m})^2}} + 2\alpha\|\beta_*\|_2 \frac{1}{\underline{\varphi}(\widehat{m})^2} \right) \mathbb{1}\{M^* \subseteq \widehat{M}\}$$

$$+ \left(\frac{K_\epsilon \sigma}{\underline{\varphi}(\widehat{m})\sqrt{T}} \sqrt{(\widehat{k}+\widehat{m})\left(\log(s^*+\widehat{m}) + \log(e^2\mu(\widehat{k}+\widehat{m}))\right)} + \frac{2\alpha}{\underline{\varphi}(\widehat{m})^2}\|\beta_*\|_2 + \frac{\|\beta_{*,M^*\setminus\widehat{M}}\|_{2,T}}{\underline{\varphi}(\widehat{m})} \right)$$

$$\times \mathbb{1}\{M^* \not\subseteq \widehat{M}\}.$$

In-sample and Out-of-sample Prediction error. V

Let $(Y_\tau, X_\tau')'$, $\tau > T$, be a new copy of $(Y_t, X_t')'$ that satisfies Assumption 3 *(i)-(ii)* and that is $\underline{||}$ of (Y, X).

Corollary 2 (Out-of-sample prediction error)

Suppose that Assumptions 3 (i)-(ii) and 6 are satisfied and that $\varepsilon_t | X_t$ is Gaussian. Let X_τ be such that $\sum_{j=1}^{\widehat{m}+s^*} X_{\tau,j}^2 < C^2(\widehat{m}+s^*)$ for a constant $0 < C < \infty$. Then, for every $\epsilon > 0$, there is a constant K_ϵ independent of T such that with probability at least $1-\epsilon$,

$$
\begin{aligned}
X_\tau'(\widehat{\beta} - \beta_*) \leq & \left(\sqrt{\widehat{m}+s^*}\right) C \times \\
& \Bigg[\left(K_\epsilon \sqrt{\frac{\widehat{m}\log(N) + (\widehat{m}+s^*)\log(e^2\mu(\widehat{m}))}{T\underline{\varphi}(\widehat{m})^2}} + 2\alpha\|\beta_*\|_2 \frac{1}{\underline{\varphi}(\widehat{m})^2} \right) \mathbb{1}\{M^* \subseteq \widehat{M}\} \\
+ \Bigg(& \frac{K_\epsilon \sigma}{\underline{\varphi}(\widehat{m})\sqrt{T}} \sqrt{(\widehat{k}+\widehat{m})\left(\log(s^*+\widehat{m}) + \log(e^2\mu(\widehat{k}+\widehat{m}))\right)} + \frac{2\alpha}{\underline{\varphi}(\widehat{m})^2}\|\beta_*\|_2 + \frac{\|\beta_{*,M^*\setminus\widehat{M}}\|_{2,T}}{\underline{\varphi}(\widehat{m})} \Bigg) \\
& \times \mathbb{1}\{M^* \not\subseteq \widehat{M}\} \Bigg].
\end{aligned}
$$

In-sample and Out-of-sample Prediction error. VI

Monte-Carlo exercise I.

Aim: to understand how N, T, s^* and the correlation among predictors affect the nowcasting performance in finite sample.

Two exercises:

1. Compare our *Ridge after Model Selection* procedure with the mostly used methods in the macroeconometric nowcasting/forecasting literature (not presented here).
2. Look at the effect of varying N, T, s^* on the in-sample and out-of-sample prediction error.

Monte-Carlo exercise II.

The data are simulated according with the following DGP: $t = 1, \ldots, T$,

$$
\begin{aligned}
y_t &= \gamma' z_t + \beta' x_t + v_t, \qquad z_t = (z_{1,t}, z_{2,t})' \sim \mathcal{N}_2\left(0, \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}\right), \\
\underset{(N\times 1)}{x_t} &= \delta' z_t + u_t, \qquad u_t \sim \mathcal{N}_N(0, \Psi), \qquad v_t \sim \mathcal{N}(0,1)
\end{aligned} \tag{6}
$$

where

- $\gamma = (1, 2)'$,
- Ψ is an $(N \times N)$-full rank covariance matrix. We consider two cases:

  (I). *uncorrelated:* $\Psi = I_N$ and

  (II). *decreasing correlation*: $\Psi = \left(|0.5|^{j-k}\right)_{j,k}$.
- We consider a sparse structure: $\beta_j \sim \mathcal{N}(0,1)$ for $j \leq s^*$ and $\beta_j = 0$ for $j > s^*$.
- $\delta = 0.2\iota$ and $\delta = 0.8\iota$ with ι a $(N \times 2)$-matrix of ones.

Monte-Carlo exercise III.

For N, T, s we consider: $(N = 150, T = 100, s = 105)$,
$(N = 200, T = 150, s = 105)$, $(N = 200, T = 150, s = 110)$ and
$(N = 200, T = 100, s = 110)$.

The split between in-sample and out-of-sample is adjusted to keep the same fraction of observations in the two samples.

Results I.

	$\delta = 0.2\iota$					
	$N = 200, T = 150, s = 105$		$N = 200, T = 150, s = 110$		$N = 200, T = 100, s = 110$	
FDR	MSER	MSFER	MSER	MSFER	MSER	MSFER
	$\Psi = I_N$					
20%	0.9526	0.8104	1.0206	0.8478	0.7831	1.1369
10%	0.9620	0.8412	1.0061	0.8863	0.9226	1.0240
5%	0.9414	0.8364	0.9797	0.8794	0.9667	1.0041
2.5%	0.9343	0.8611	0.9770	0.9030	0.9825	1.0019
1%	0.9130	0.8753	0.9538	0.9220	1.0001	1.0188
0.5%	0.9089	0.8781	0.9485	0.9324	1.0013	1.0146
	$\Psi = \left((0.5)^{\lvert j-k \rvert}\right)_{j,k}$					
20%	0.9643	0.7904	1.0066	0.8154	0.9051	1.1521
10%	0.9657	0.8325	1.0053	0.8724	0.9492	1.0860
5%	0.9236	0.8387	0.9654	0.8812	0.9805	1.0222
2.5%	0.9178	0.8503	0.9642	0.8905	0.9996	1.0216
1%	0.9011	0.8593	0.9477	0.9025	1.0262	1.0287
0.5%	0.8865	0.8607	0.9403	0.9097	1.0301	1.0427

Table: Effect of N, T, s on the in-sample and out-of-sample prediction error. In-sample (MSER) and out-of-sample MSE (MSFER) expressed as ratios with respect to the case $N = 150, T = 100, s = 105$. FDR denotes the percentage of false positive that can be tolerated.

Results II.

	$\delta = 0.8\iota$					
	$N = 200, T = 150, s = 105$		$N = 200, T = 150, s = 110$		$N = 200, T = 100, s = 110$	
FDR	MSER	MSFER	MSER	MSFER	MSER	MSFER
	$\Psi = I_N$					
20%	0.9481	0.8047	1.0185	0.8433	0.7856	1.11799
10%	0.9476	0.8438	0.9955	0.8885	0.9184	1.01455
5%	0.9275	0.8280	0.9718	0.8747	0.9767	0.99942
2.5%	0.9338	0.8477	0.9768	0.8932	1.0001	1.00466
1%	0.9136	0.8666	0.9547	0.9149	0.9993	1.02664
0.5%	0.9097	0.8694	0.9487	0.9248	0.9995	1.01358
	$\Psi = \left((0.5)^{\lvert j-k \rvert}\right)_{j,k}$					
20%	0.9488	0.7826	0.9952	0.8103	0.8957	1.14796
10%	0.9511	0.8283	0.9941	0.8686	0.9463	1.07847
5%	0.9081	0.8314	0.9555	0.8779	0.9928	1.02038
2.5%	0.9018	0.8368	0.9499	0.8777	1.0081	1.01674
1%	0.8987	0.8490	0.9452	0.8942	1.0303	1.03002
0.5%	0.8866	0.8518	0.9373	0.9010	1.0321	1.04643

Table: Effect of N, T, s on the in-sample and out-of-sample prediction error. In-sample (MSER) and out-of-sample MSE (MSFER) expressed as ratios with respect to the case $N = 150, T = 100, s = 105$. FDR denotes the percentage of false positive that can be tolerated.

Phases of the Business Cycle. I

- Models are estimated on a recursive basis starting the last week of March 2005 (because of de-seasonalisation and stationarisation).
- For the period 2014q1-2016q1:
  - Estimation sample: 2005q1 - 2013q3.
  - Nowcasting sample: 2014q1 - 2016q1.
- For the period 2017q1-2018q4:
  - Estimation sample: 2005q1 - 2016q3.
  - Nowcasting sample: 2017q1 - 2018q4.
- For the period 2008q1-2009q2:
  - Estimation sample: 2005q1 - 2007q3.
  - Nowcasting sample: 2008q1 - 2009q2.

Phases of the Business Cycle. II

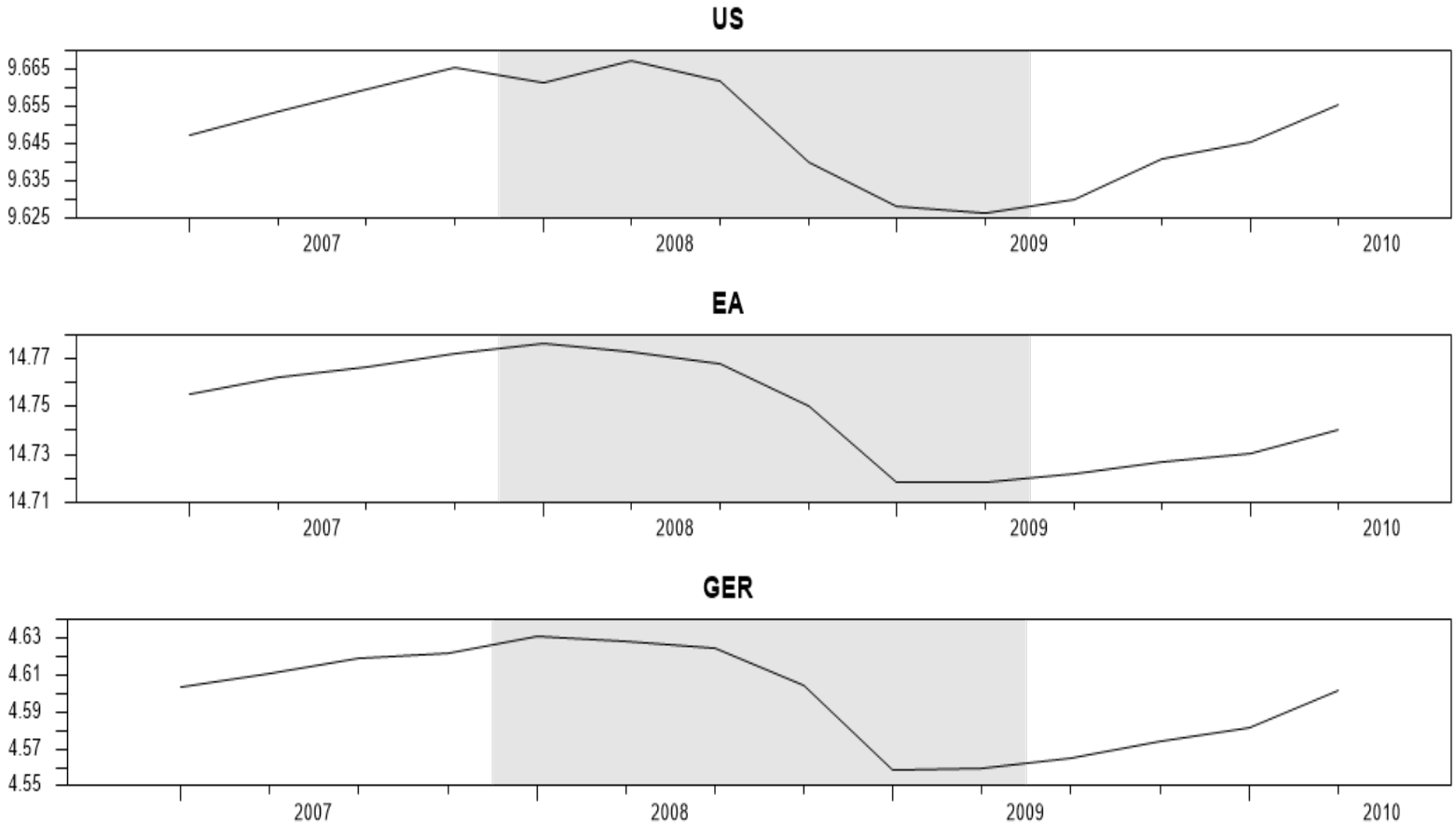


Figure: Log-GDP over sub-period 2008-09

Phases of the Business Cycle. III

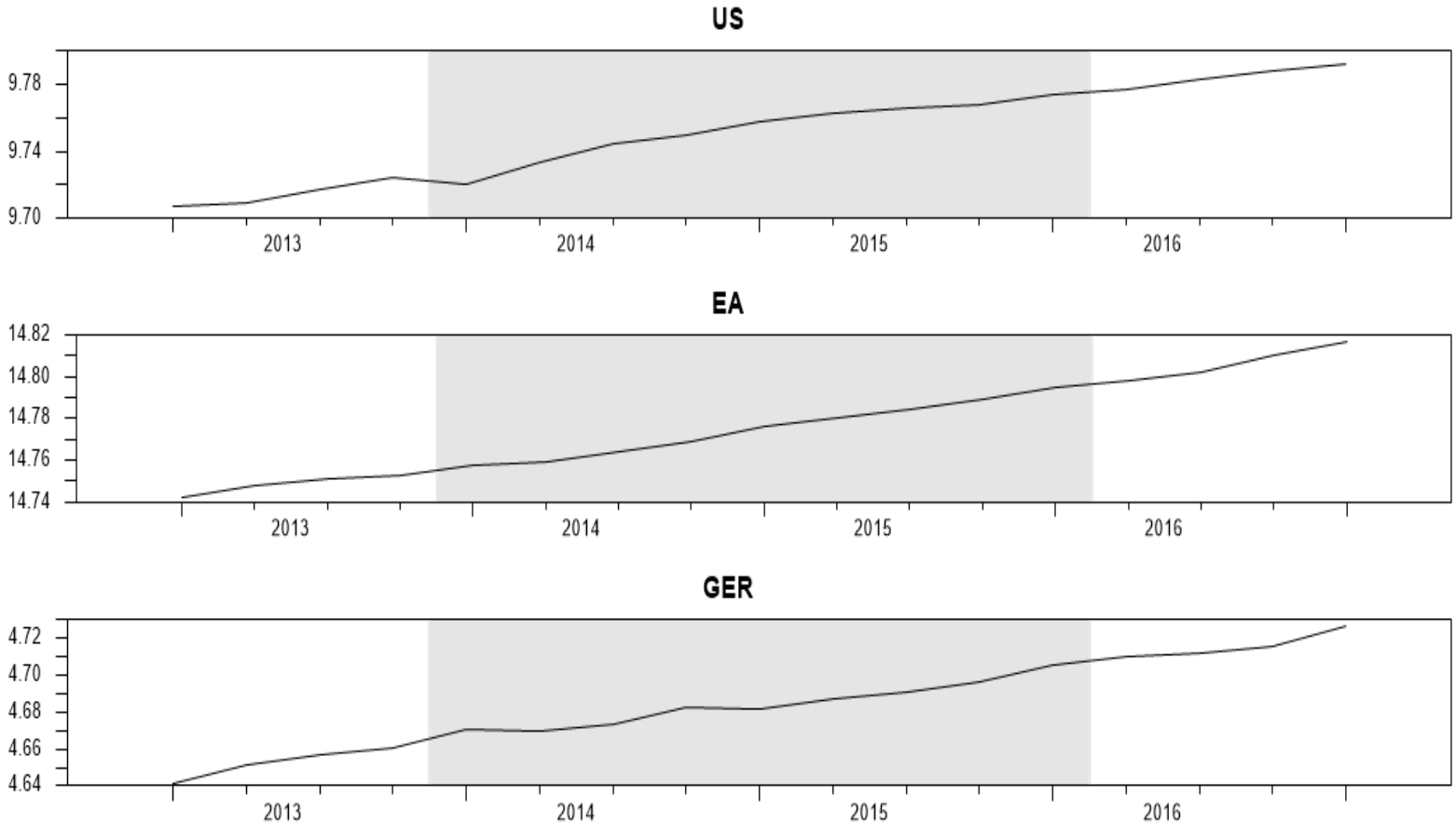


Figure: Log-GDP over sub-period 2014-16

Phases of the Business Cycle. IV

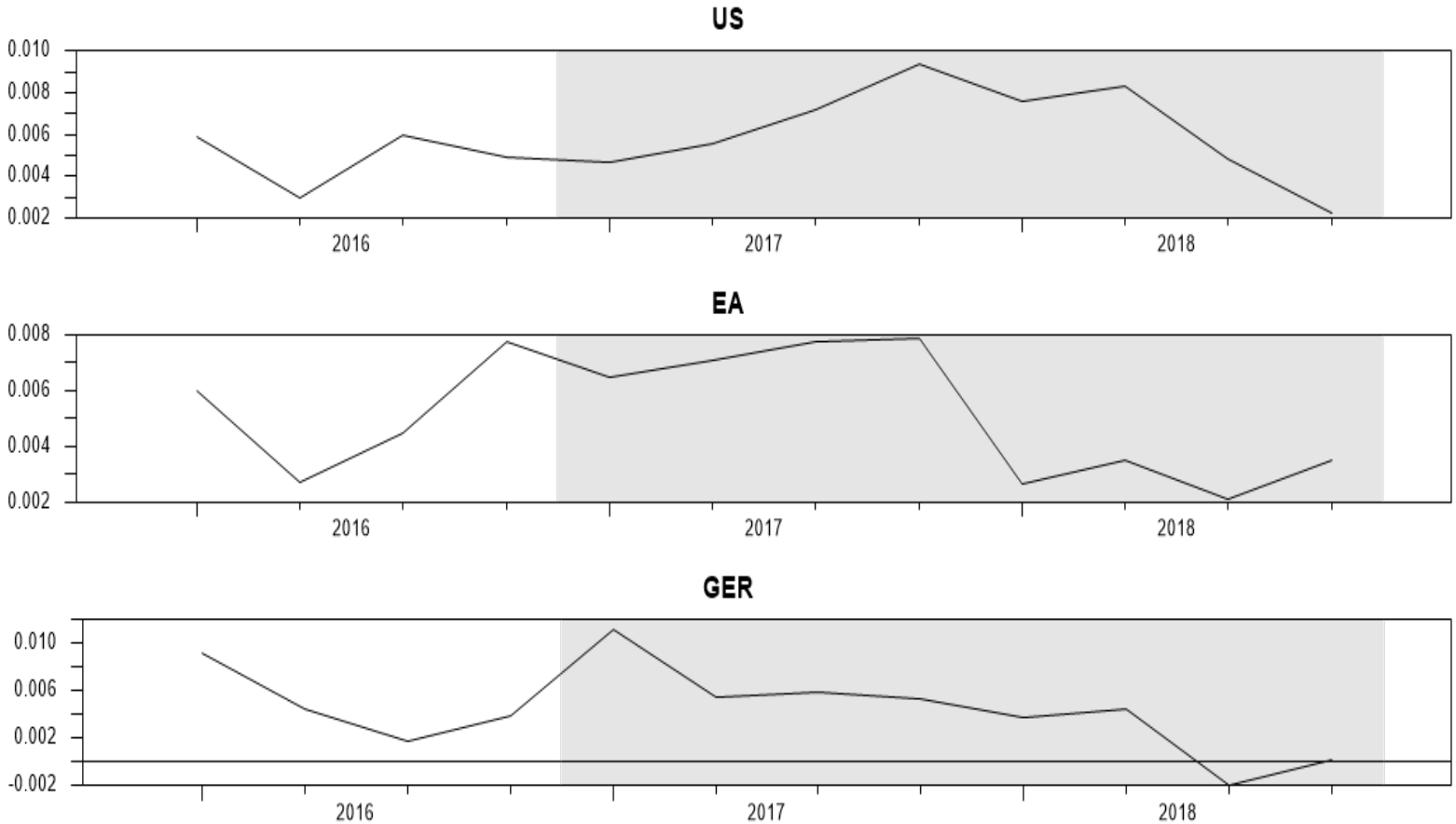
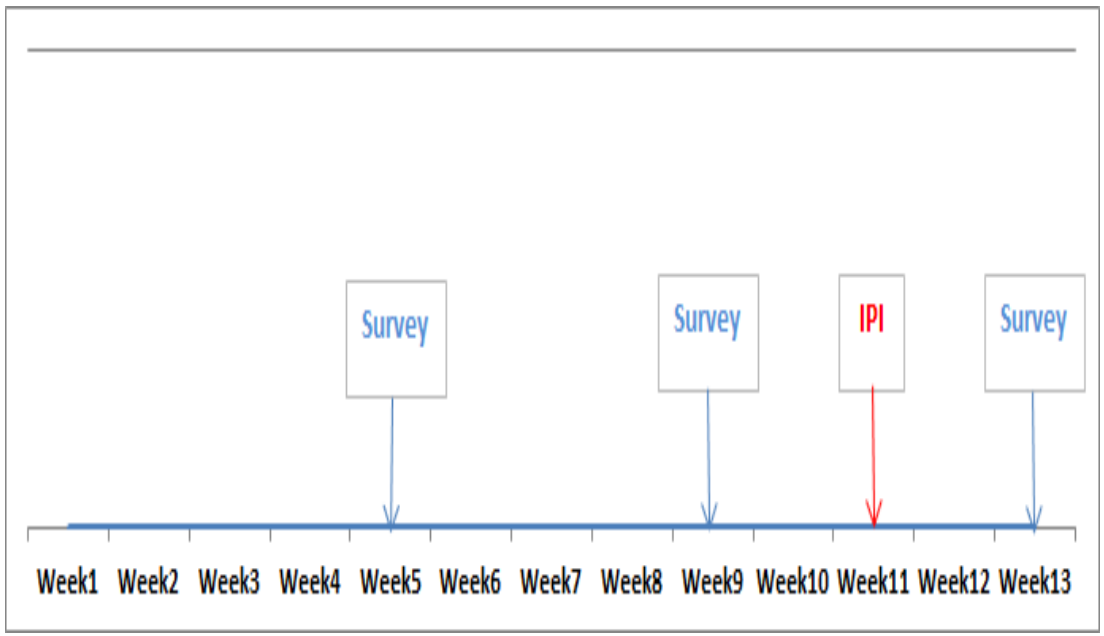


Figure: GDP growth rates over sub-period 2017-18

Timeline of data releases within the quarter

In addition to frequency mismatch in the data, another challenge: data on official series and Google search are released with various reporting lags $\Rightarrow$ unbalanced information set at each point in time within the quarter (*ragged-edge database*).

For EA: S_t is available at weeks 5, 9 and 13, and IP_t at week 11.



The relevant information set.

- Because $x_{t,s}$, $x_{t,h}$ and $x_{t,g}$ are sampled over different frequencies and released with various reporting lags the relevant information set for calculating the nowcasts evolves within the quarter (unbalanced data set).
- Let $x_{t,j,i}^{(w)}, j \in \{s,h,g\}$, denote the i-th series in vector $x_{t,j}$ released at week $\leq w = 1,\ldots,13$ of quarter t.
- The relevant information set at week w of a quarter t is:

$$\Omega_t^{(w)} := \left\{ x_{t,j,i}^{(w)}, i = 1,\ldots,N_j, j \in \{s,h,g\} \right\}.$$

$\forall t = 1,\ldots,T, \forall w = 1,\ldots,13$, for each $\Omega_t^{(w)}$ within a given quarter t, the nowcast is computed as $\widehat{Y}_{t|w} := \mathbf{E}[Y_t|\Omega_t^{(w)}; M_{(w)}]$ based on the model:

$$M_{(w)}: \qquad \mathbf{E}[Y_t|\Omega_t^{(w)}] = \beta_{0,w} + \beta_{s,w}' x_{t,s}^{(w)} + \beta_{h,w}' x_{t,h}^{(w)} + \beta_{g,w}' x_{t,g}^{(w)}, \quad (7)$$

where $\beta_{j,w,i}^{(w)} = 0$ if $x_{t,j}^{(w)} \notin \Omega_t^{(w)}$.

Bridge models for the 13 weeks.

For weeks $w = 1, \ldots, 4$, the models are of the form:

$$Y_t = \beta_{0,w} + \beta'_{g,w} x^{(w)}_{t,g} + \varepsilon_{t,w} \tag{8}$$

From week $w = 5$ to $w = 10$, the models are of the form:

$$Y_t = \beta_{0,w} + \beta'_{g,w} x^{(w)}_{t,g} + \beta_{s,w} S_t + \varepsilon_{t,w} \tag{9}$$

For weeks $w = 11, \ldots, 13$, the models are of the form:

$$Y_t = \beta_{0,w} + \beta'_{g,w} x^{(w)}_{t,g} + \beta_{s,w} S_t + \beta_{h,w} IP_t + \varepsilon_{t,w} \tag{10}$$

Overall evaluation of Google search data. I

The main empirical research questions are:

(I). are Google search data informative when there is no official data available for the forecaster?

(II). to what extent Google Search data remain informative when official data become available?

Overall evaluation of Google search data. II

We estimate the following models:

(a). the nowcasting models $M_{(1)}, \ldots, M_{(13)}$ accounting for the full set of information (Google, hard and soft data) without preselection,

(b). the nowcasting models $M_{(1)}, \ldots, M_{(13)}$ accounting for the full set of information (GSD, hard and soft data) with the preselection (i.e. *Ridge after Model selection* approach),

(c). the nowcasting models $M_{(1)}, \ldots, M_{(13)}$ by using only GSD (i.e. $\beta_{s,w} = \beta_{h,w} = 0$ for every $w = 1, \ldots, 13$) with and without preselection,

(d). the models that only account for hard and soft data (i.e. without GSD, $\beta_{g,w} = 0$ for every $w = 1, \ldots, 13$).

Overall evaluation of Google search data. III

What do we observe?

1. Downward sloping evolution of RMSFEs throughout the quarter stemming from models $M_{(1)}, \ldots, M_{(13)}$ with full information (second row).
2. When using only Google information (third row), we still observe a decline of RMSFEs throughout the quarter but to a much lower extent.
3. When focusing on the beginning of the quarter, models that only integrate Google information provide very reasonable RMSFEs, which are slightly higher than those obtained at the end of the quarter.

Overall evaluation of Google search data. IV

Gain from using our *Ridge after Model Selection* strategy?

- During both *calm* and *sudden shift* periods (*i.e.* 2014-16 and 2017-18), this estimation strategy applied to the whole dataset (hard data, soft data and GSD) generally tends to provide the lowest RMSFEs (second row of Tables).

- By comparing the 1^{st} and 2^{nd} rows of the Tables: outside recession periods, our *Ridge after Model Selection* strategy outperforms a strategy that would skip the data preselection step.

- By comparing the 2^{nd}, 3^{rd} and 4^{th} rows of the Tables: outside recession periods, combining information (i.e. macroeconomic and Google data) generally leads to more accurate nowcasts than those only based on either pure macroeconomic information or pure Google information.

- This result is robust over the different countries/areas and is robust robust to the size of the training sample.

Overall evaluation of Google search data. V

- Recession periods possess a very specific pattern: preselecting data does not necessarily generate lower RMSFEs during those phases of the business cycle:
  - for almost all weeks within the quarter, the Ridge model that only integrates Google data without preselection outperforms other models (3^{rd} rows of the Tables).
  - during a recession, (i) we do not have to preselect data and (ii) GSD provide more accurate information than official macroeconomic data.

Overall evaluation of Google search data: EA (RMSFE).

EA – Nowcasting during 2014q1 – 2016q1

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.4467	0.4816	0.3897	0.3659	0.3239	0.3829	0.3901	0.3609	0.3427	0.3422	0.3103	0.3142	0.3111
Sel + Ridge (Google+ S+IP)	**0.2889**	**0.2607**	**0.2400**	**0.2493**	**0.1747**	**0.1706**	**0.1695**	**0.1608**	**0.1641**	**0.1668**	0.2222	**0.2178**	0.2082
Sel + Ridge (Google)	0.3026	0.2769	0.2841	0.3008	0.3052	0.3107	0.3001	0.2974	0.2984	0.2975	0.2964	0.2867	0.2880
No G					0.1807				0.1897		**0.1928**		**0.2017**

EA – Nowcasting during 2017q1 – 2018q4

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.5592	0.5956	0.5713	0.5642	0.3604	0.3605	0.3186	0.3064	0.5100	0.4836	0.4181	0.4527	0.4788
Sel + Ridge (Google+ S+IP)	**0.3505**	**0.3306**	0.3341	**0.3227**	**0.2330**	**0.2664**	**0.2501**	**0.2415**	**0.2720**	**0.2451**	**0.1316**	**0.1340**	**0.1314**
Sel + Ridge (Google)	0.3760	0.3431	**0.3262**	0.3276	0.3230	0.3167	0.3051	0.2875	0.2894	0.2856	0.2795	0.2763	0.2700
No Google					0.4340				0.4841		0.2871		0.3177

EA – Nowcasting during recession periods

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	**1.4601**	**1.2693**	**1.2268**	**1.0596**	1.0458	0.9831	**0.9340**	1.0843	1.1047	1.1047	**0.9632**	**0.9401**	**0.9101**
Sel + Ridge (Google+ S+IP)	1.5481	1.4771	1.5257	1.6215	1.5581	1.6184	1.6345	1.6313	1.6344	1.6677	1.0953	1.0468	1.0622
Ridge (Google)	**1.4601**	**1.2693**	**1.2268**	**1.0596**	**0.7745**	**0.8267**	1.0072	**1.0732**	**1.0415**	**1.0042**	0.9962	0.9735	0.9657
No Google					1.5269				1.4241		1.6351		1.2888

Overall evaluation of Google search data: U.S. (RMSFE).

U.S. – Nowcasting during 2014q1 – 2016q1

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.5843	0.5177	0.5771	0.5769	0.4588	0.4735	0.4207	0.4213	0.4315	0.4313	0.4479	0.4488	0.4786
Sel + Ridge (Google+ S+IP)	0.4889	**0.4792**	**0.4647**	**0.4670**	0.4101	**0.4277**	**0.3957**	**0.3922**	**0.3948**	**0.3933**	**0.4233**	**0.4273**	0.4509
Sel + Ridge (Google)	**0.4873**	0.4833	0.4829	0.4816	0.4777	0.4740	0.4750	0.4751	0.4745	0.4746	0.4753	0.4749	0.4703
No Google					**0.4062**		0.4061		0.4156		0.4260		**0.4466**

U.S. – Nowcasting during 2017q1 – 2018q4

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.4746	0.4090	0.4625	0.4549	0.2762	0.3021	0.2401	0.2407	0.2963	0.2506	0.2555	0.1791	0.2081
Sel + Ridge (Google+ S+IP)	0.3639	0.3601	**0.3092**	**0.3181**	**0.1735**	**0.1685**	**0.1347**	**0.1330**	**0.1042**	**0.0991**	**0.1187**	**0.1081**	**0.1320**
Sel + Ridge (Google)	**0.3482**	**0.3335**	0.3177	0.3270	0.3168	0.3103	0.3061	0.3055	0.2949	0.2833	0.2770	0.2816	0.2757
No Google					0.2598		0.2255		0.3604		0.3510		0.2979

U.S. – Nowcasting during recession periods

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	**1.0156**	1.0507	**1.0506**	**1.0282**	1.0967	1.1191	1.1185	1.1986	1.2123	1.1884	1.1738	1.1801	1.1217
Sel + Ridge (Google+ S+IP)	1.0320	1.0611	1.0590	1.0521	1.2030	1.2029	1.1929	1.1893	1.1815	1.1751	1.1319	1.1307	1.0396
Ridge (Google)	**1.0156**	1.0507	**1.0506**	**1.0282**	0.9744	**0.9320**	**0.9731**	**0.9991**	**1.0061**	**1.0158**	**1.0204**	**1.0196**	1.2224
No Google					**0.8439**		1.1286		1.0580		1.0659		**0.7828**

Overall evaluation of Google search: Germany (RMSFE).

Germany – Nowcasting during 2014q1 – 2016q1

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.3316	0.3225	0.3296	0.3365	0.3137	0.2905	0.2687	0.2690	0.2695	0.2684	0.2713	0.2754	0.2698
Sel + Ridge (Google+ S+IP)	0.2619	0.2454	**0.2219**	**0.2306**	**0.2378**	0.2406	0.2373	0.2382	0.2429	0.2460	0.2717	0.2794	0.2754
Sel + Ridge (Google)	**0.2265**	**0.2266**	0.2484	0.2436	0.2433	**0.2341**	**0.2310**	**0.2296**	**0.2362**	**0.2372**	**0.2380**	**0.2470**	**0.2453**
No Google					0.3977				0.4208		0.2914		0.4325

Germany – Nowcasting during 2017q1 – 2018q4

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.4889	0.4894	0.4837	0.4777	0.4352	0.4800	0.4649	0.4441	0.5038	0.4927	0.4382	0.4059	0.4181
Sel + Ridge (Google+ S+IP)	0.3814	**0.3794**	**0.3751**	0.3769	0.3831	0.3826	0.3876	0.3898	0.3238	0.3181	0.3141	**0.3088**	**0.2917**
Sel + Ridge (Google)	**0.3812**	0.3800	0.3788	**0.3757**	**0.3711**	**0.3680**	**0.3712**	**0.3584**	**0.3184**	**0.3097**	**0.3141**	0.3187	0.3140
No Google					0.6532				0.6241		0.3632		0.6433

Germany – Nowcasting during recession periods

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	**1.9279**	**1.9100**	**1.9123**	**1.9212**	2.2660	2.3237	2.3365	2.3162	2.2939	2.2790	2.7585	2.7372	2.6581
Sel + Ridge (Google+ S+IP)	1.9289	1.8990	1.9053	1.9018	2.1467	2.1477	2.1469	2.1461	2.1353	2.1382	2.6461	2.6538	2.5886
Ridge (Google)	**1.9279**	**1.9100**	**1.9123**	**1.9212**	**1.9477**	**2.0123**	**2.0200**	**1.9972**	**1.9793**	**1.9603**	**1.9440**	**1.9173**	**1.9054**
No Google					2.1580				2.0660		2.6107		1.9803

Robustness checks

Two robustness checks carried out for the euro area:

1. a true real-time analysis executed by using vintages of GDP and industrial production.

2. controlling for more official macroeconomic series than industrial production and opinion surveys.

A true real-time analysis. I

- Analysis done over the period 2014-16
- we use vintages of data for EA GDP and industrial production (Survey data are generally not revised) and account for the observed timeline of data release as provided by Eurostat.
- When available, we also include lagged GDP growth among the explanatory variables.
- The RMSFEs values compare the same four models (a)-(d) as in the previous pseudo real-time analysis.
- Overall, the results that we have illustrated above for the pseudo real-time exercise still hold in true real-time:

  (i). nowcasting accuracy improves throughout the quarter,

  (ii). Google search data provide valuable information at the beginning of the quarter when there is no official information,

A true real-time analysis. II

(iii). combining macroeconomic information with Google information improves the results,

(iv). our *R*idge after Model Selection strategy outperforms the other approaches in terms of nowcasting accuracy.

EA – True real time – "Google+S+IP", Targeted preselection, Nowcasting Period: $2014q1 - 2016q1$

	$M_{(1)}$	$M_{(2)}$	$M_{(3)}$	$M_{(4)}$	$M_{(5)}$	$M_{(6)}$	$M_{(7)}$	$M_{(8)}$	$M_{(9)}$	$M_{(10)}$	$M_{(11)}$	$M_{(12)}$	$M_{(13)}$
20%	0.4357	0.4785	0.3935	0.3700	0.3628	0.4032	0.3980	0.3709	0.3537	0.3571	0.3367	0.3398	0.3381
10%	0.4357	0.4785	0.3935	0.3700	0.3628	0.4032	0.3980	0.3709	0.3537	0.3571	0.3367	0.3398	0.3381
5%	0.4356	0.4781	0.3933	0.3700	0.3626	0.4033	0.3983	0.3712	0.3539	0.3571	0.3368	0.3399	0.3383
2.50%	0.4398	0.5286	0.4554	0.4073	0.4095	0.4576	0.4544	0.4071	0.3800	0.3668	0.3360	0.3399	0.3398
1%	0.2972	0.2820	0.2972	0.2869	0.2074	0.2162	0.2358	0.2287	0.2226	0.2311	0.2516	**0.2533**	0.2485
0.50%	**0.2741**	**0.2688**	**0.2675**	**0.2684**	**0.1936**	**0.1932**	**0.1932**	**0.1831**	**0.1844**	**0.1859**	**0.2499**	0.2561	**0.2436**

EA – True real time – "Google", unconditional preselection, Nowcasting Period: $2014q1 - 2016q1$

	$M_{(1)}$	$M_{(2)}$	$M_{(3)}$	$M_{(4)}$	$M_{(5)}$	$M_{(6)}$	$M_{(7)}$	$M_{(8)}$	$M_{(9)}$	$M_{(10)}$	$M_{(11)}$	$M_{(12)}$	$M_{(13)}$
20%	0.3707	0.3682	0.3571	0.3397	0.3396	0.3559	0.3586	0.3625	0.3646	0.3691	0.3662	0.3572	0.3524
10%	**0.3309**	**0.2981**	**0.2945**	0.3435	0.3248	0.3371	0.3288	0.3362	0.3397	0.3411	0.3117	0.2901	0.2918
5%	0.3556	0.3342	0.3166	0.3196	0.3231	0.3246	0.3148	0.3135	0.3177	0.3223	0.3191	0.2955	0.2960
2.50%	0.3435	0.3144	0.2880	**0.3010**	**0.3070**	**0.3091**	**0.2966**	**0.2967**	**0.2919**	**0.2920**	**0.2922**	**0.2828**	**0.2817**
1%	0.3463	0.3244	0.2988	0.3119	0.3152	0.3280	0.3197	0.3159	0.3091	0.3094	0.2960	0.2857	0.2853
0.50%	0.3567	0.3280	0.2982	0.3119	0.3161	0.3176	0.3157	0.3148	0.2980	0.2986	0.2927	0.2879	0.2864

EA – True real time – No preselection, Nowcasting Period: $2014q1 - 2016q1$

	$M_{(1)}$	$M_{(2)}$	$M_{(3)}$	$M_{(4)}$	$M_{(5)}$	$M_{(6)}$	$M_{(7)}$	$M_{(8)}$	$M_{(9)}$	$M_{(10)}$	$M_{(11)}$	$M_{(12)}$	$M_{(13)}$
Google+S +IP	0.4357	0.4785	0.3935	0.3700	0.3628	0.4032	0.3980	0.3709	0.3537	0.3571	0.3367	0.3398	0.3381
Google	0.4357	0.4785	0.3935	0.3700	0.4087	0.4472	0.4542	0.4313	0.4198	0.4213	0.4190	0.4087	0.4061
No Google					0.2320				0.2365		0.3283		0.2576

Controlling for additional macroeconomic variables. I

- Aim: checking the robustness of our evaluation about Google search data to a richer macroeconomic information set.
- We include additional macroeconomic series (commonly used in the nowcasting literature) among the covariates $x_{t,s}^{(w)}$ and $x_{t,h}^{(w)}$ in model (7): such as sales, exports or unemployment rate (*Big Official Set*).
- The robustness check is made for two periods: $2014q1 - 2016q4$ and $2017q1 - 2018q4$.
- As performance measure, we look at the ratios between the RMSFEs obtained by using: GSD together with the *Small Official Set* of data (resp. the *Big Official Set* of data) in the numerator (resp. in the denominator). A ratio > 1 indicates that including additional official series improves nowcasting accuracy, and conversely.

Controlling for additional macroeconomic variables. II

Results:

- In $2014q1 - 2016q1$: the inclusion of additional macroeconomic variables when using our Ridge after Model Selection strategy does not generally improve the nowcasting accuracy except for week 4, and week 2 at a lower extent.
- The previous result still holds when there is no preselection of data and when we only account for macroeconomic variables.
- However, when there is a downward shift in the GDP, as in the period $2017q1 - 2018q4$, it seems worth including a larger set of macroeconomic variables, especially in the middle of the quarter (rows four and six of the Table). For the first three weeks and last three weeks, the *Small Official* dataset is preferred.
- If we compare the results obtained without preselection of GSD (row 5^{th}) inclusion of additional macroeconomic variables does not improve the nowcasting accuracy.

Controlling for additional macroeconomic variables. III

- It seems that when economic uncertainty is high, it is useful to keep all the Google search data variables into the models. A possible explanation is that this uncertainty generated by the trade war does not have a strong common impact on all macroeconomic variables and does not adversely affect economic activity across the board.

EA – Robustness check

	$M_{(1)}$	$M_{(2)}$	$M_{(3)}$	$M_{(4)}$	$M_{(5)}$	$M_{(6)}$	$M_{(7)}$	$M_{(8)}$	$M_{(9)}$	$M_{(10)}$	$M_{(11)}$	$M_{(12)}$	$M_{(13)}$
$2014q1 - 2016q1$ Selection	1.0020	1.0538	0.9886	1.2026	0.8444	0.8326	0.7074	0.7021	0.7792	0.7989	1.0060	0.9572	0.8646
$2014q1 - 2016q1$ W/o selection	1	1	1	1.0229	0.8098	0.8497	0.8315	0.8100	0.7913	0.7959	0.7334	0.7709	0.7649
$2014q1 - 2016q1$ W/o Google					0.7223				0.7608		0.5234		0.5538
$2017q1 - 2018q4$ Selection	0.9387	0.9068	0.9281	1.1149	0.8659	1.4850	1.7796	1.5742	1.1361	1.0137	0.5792	0.5265	0.5548
$2017q1 - 2018q4$ W/o selection	1	1	1	1.0605	0.6558	0.6512	0.5936	0.5790	0.8780	0.8744	0.7330	0.7894	0.7839
$2017q1 - 2018q4$ W/o Google					1.3163				1.0340		0.8612		0.9208

Euro Area Variables in the Big Official Set

Name	Definition	Reporting lag (in days)	Transformation
IPI	EA industrial production, ex. construction	42	Δ^6 log
Sales	EA total retail sales volume	35	Δ^6 log
Exports	Extra-EA exports	51	Δ^6 log
INOGE	German industrial new orders index	52	Δ^6 log
cars	EA New passenger car registrations	17	Δ^6 log
turnover	EA Retail trade turnover, deflated	35	Δ^6 log
ur	EA unemployment rate	30	Δ^6
empexp	EA Employment Expectations	-2	None
ESIIND	EA Economic Sentiment Index Industry	-2	None
ESISER	EA Economic Sentiment Index Services	-2	None
IFO	Germany Business Climate Index (IFO)	-8	None
USIP	US Industrial Production	16	Δ^6 log
usconexp	US Consumer Expectations (The Conference Board)	-2	None
usmanexp	US Manufacturing Expectations (Richmond Fed)	-2	None
M3	M3 Monetary Supply	27	Δ^6 log
loans	Loans to other EA residents	27	Δ^6 log
10y	10-year Government Bond Yield	0	Δ^6
3m	3-month EA Interbank Rate	0	Δ^6
neer	EA Broad Nominal Effective Exchange Rate (BIS)	16	Δ^6 log
usdeur	US Dollar vs Euro Bilateral Exchange Rate	0	Δ^6 log
eurostoxx	Euro Stoxx Stock Index	0	Δ^6 log
rawmat	World Market Prices of Raw Materials (HWWA)	16	Δ^6 log

Google categories selected in Step 1. I

EA - Pseudo-real time: Categories Selected $2014q1 - 2016q1$		
Category selected	Subcategory selected	Country
Sensitive Subjects	Death and Tragedy	BE
Sensitive Subjects	Violence and Abuse	DE
Shopping	Mass Merchants and Department Stores	ES
Sports	Combat Sports	ES
Business & Industrial	Printing and Publishing	FR
Business & Industrial	Small Business	FR
News	Journalism & News Industry	FR
Health	Aging & Geriatrics	IT
Jobs & Education	Jobs	IT
Sensitive Subjects	Death and Tragedy	IT
Arts & Entertainment	TV & Video	NL
Beauty & Fitness	Body Art	NL
Finance	Credit & Lending	NL
Jobs & Education	Jobs	NL
Law & Government	Social Services	NL
Sensitive Subjects	Controversial Social Issues	NL
Sensitive Subjects	Violence & Abuse	NL

Google categories selected in Step 1. II

EA - Pseudo-real time: Categories Selected $2017q1 - 2018q4$		
Category selected	Subcategory selected	Country
Arts & Entertainment	TV & Video	BE
Health	Health Foundations & Medical Research	DE
Sensitive Subjects	Violence & Abuse	DE
Shopping	Mass Merchants & Department Stores	ES
Sports	Combat Sports	ES
Business & Industrial	Small Business	FR
News	Journalism & News Industry	FR
Shopping	Auctions	FR
Shopping	Shopping Portals & Search Engines	FR
Health	Aging & Geriatrics	IT
Arts & Entertainment	TV & Video	NL
Beauty & Fitness	Body Art	NL
Sensitive Subjects	Controversial Social Issues	NL
Sensitive Subjects	Violence & Abuse	NL

Google categories selected in Step 1. III

U.S.: Categories Selected $2014q1 - 2016q1$	
Category selected	Subcategory selected
Arts & Entertainment	Movies
Arts & Entertainment	Offbeat

U.S.: Categories Selected $2017q1 - 2018q4$	
Category selected	Subcategory selected
Arts & Entertainment	Movies
Arts & Entertainment	Offbeat
News	Broadcast & Network News
News	
Travel	Specialty Travel

Google categories selected in Step 1. IV

Germany: Categories Selected $2014q1 - 2016q1$	
Category selected	Subcategory selected
Autos & Vehicles	Hybrid & Alternative Vehicles
Autos & Vehicles	Trucks & SUVs
Finance	Financial Planning & Management
Finance	Insurance
Health	Health Education & Medical Training
Hobbies & Leisure	Contests, Awards & Prizes
Law & Government	Military
Sensitive Subjects	
Sensitive Subjects	Recreational Drugs

Germany: Categories Selected $2017q1 - 2018q4$	
Category selected	Subcategory selected
Health	Health Education & Medical Training
Sensitive Subjects	Recreational Drugs

Google Search and Google Trends data. I

- Google Search data are data that European central bank receives weekly (every Tuesday) from Google related to queries done with Google search machines.
- Google Search data is more accurate than Google Trends: much larger samples are used in Google Search data than in Google Trends.
- Google Trends covers more countries and has sub-sub-categories.

Google Trends: let

- $S_{d,r} = V_{d,r}/T_{d,r}$ = search share for a particular keyword in day d, in region r;
- $V_{d,r}$ = number of web searches containing that keyword;
- $T_{d,r}$ = total number of web searches performed through Google in the same day in that area.

Google Search and Google Trends data. II

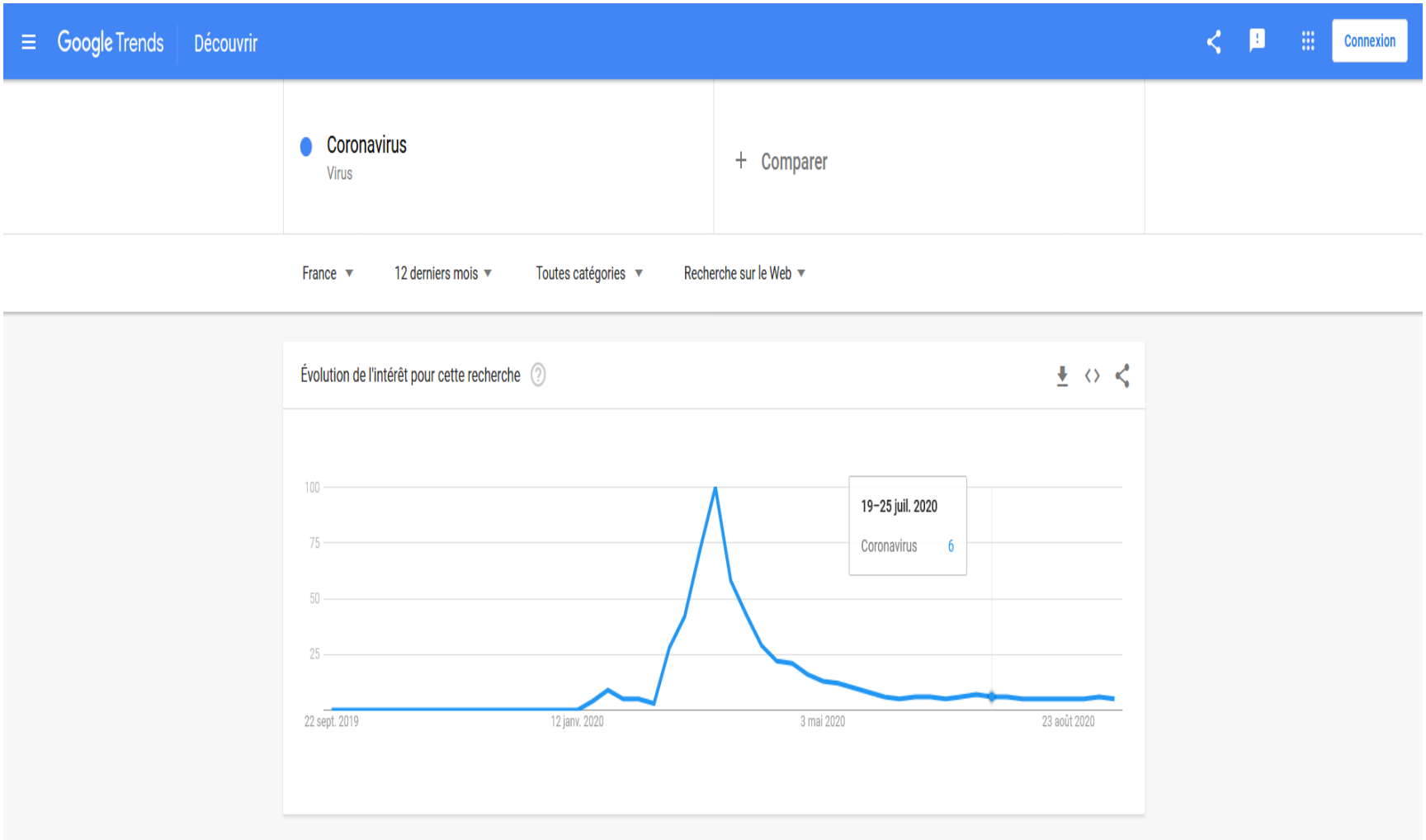
The search share of week w is

$$S_{w,r} = \frac{1}{7} \sum_{d \in w, d=Sun}^{Sat} S_{d,r}.$$

The Google Trends index over weeks in a set $[\underline{w}, \overline{w}]$ is

$$GI_{w,r} = \frac{100}{\max_{i \in [\underline{w}, \overline{w}]} S_{i,r}} S_{w,r}, \quad w \in [\underline{w}, \overline{w}].$$

Google Search and Google Trends data. III



Google Search and Google Trends data. IV

Google assigns queries to particular categories using natural language processing methods.

- Example: the query `[car tire]` would be assigned to category **Vehicle Maintenance** which is a subcategory of **Autos & Vehicles**.
- For Google search there are: 26 categories and 296 subcategories.
- **Google search data** are indexes of weekly volume changes of Google queries grouped by category and by country. Data are normalized at 1 at the first week of January 2004.
- Then, the next values indicate the deviation from the first value.

Google data and related literature

- Choi & Varian (2009, 2012): Google Trends for Consumer Confidence Index, Monthly visitor arrival summary, Initial claims for unemployment benefits, Motor Vehicles and Parts Dealers;
- D'Amuri & Marcucci (2017): Google Trends for unemployment forecast;
- Götz & Knetsch (2019): Google search data for German GDP;
- Li (2019): Google Trends for nowcasting US jobless initial claims and unemployment rate.
- About the importance of pre-selecting for forecasters when a large number of variables is available: *e.g.* Bai and Ng (2008) and Boivin and Ng (2006).