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

Laurent Ferrara, Anna Simoni

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

Document de Travail
Working Paper
2020-11

Laurent Ferrara
Anna Simoni

When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage[1]

Laurent Ferrara*
SKEMA Business School
EconomiX, U. Paris Nanterre

Anna Simoni[2]
CREST
CNRS

December 13, 2019

Abstract

We analyse whether, and when, a large set of Google search data can be useful to increase GDP nowcasting accuracy once we control for information contained in official variables. We put forward a new approach that combines variable pre-selection and Ridge regularization and we provide theoretical results on the asymptotic behaviour of the estimator. Empirical results on the euro area show that Google data convey useful information for pseudo-real-time nowcasting of GDP growth during the four first weeks of the quarter, when macroeconomic information is lacking. However, as soon as official data become available, their relative nowcasting power vanishes. In addition, a true real-time analysis confirms that Google data constitute a reliable alternative when official data are lacking.

Keywords: Nowcasting, Big data, Google search data, Sure Independence Screening, Ridge Regularization.

[1]We would like to thank Roberto Golinelli, Michele Lenza, Francesca Monti, Giorgio Primiceri, Simon Sheng, Hal Varian and the participants of the 10th ECB Conference on *Macro forecasting with large datasets*, *Data Day@HEC*, and *Bocconi-Banque de France alternative datasets* conferences for useful comments. We would like to thank Per Nymand-Andersen (ECB) for sharing the Google dataset as well as Dario Buono and Rosa Ruggeri-Cannata (Eurostat) for sending the real-time euro area GDP data. We are grateful to Vivien Chbicheb for outstanding research assistance. Anna Simoni gratefully acknowledges financial support from ANR-11-LABEX-0047 and *Fondation Banque de France* for the hospitality. A first version of this paper was circulated under the title: *Macroeconomic nowcasting with big data through the lens of a targeted factor model*. This research project started while the first author was working for the Banque de France.

*Skema Business School - University Cote d'Azur, and EconomiX, University Paris Nanterre, e-mail: laurent.ferrara@skema.edu

[2]CREST, CNRS, École Polytechnique, ENSAE - 5, avenue Henry Le Chatelier, 91120 Palaiseau, France, e-mail: simoni.anna@gmail.com

1 Introduction

Big sets of alternative data are now widely used by practitioners for short-term macroeconomic forecasting and nowcasting purposes. Seminal papers on the use of alternative datasets against this backdrop tend to show evidence of a sizeable gain when using such data (see for example the use of Google data by Choi and Varian [2009] or Choi and Varian [2012]). In this paper, we ask the question whether such data are still useful when controlling for official variables, such as opinion surveys or production, generally used by forecasters. And if so, when exactly are those alternative data actually adding a gain in nowcasting accuracy, both in quasi and true real-time frameworks. In this respect, we focus on Google search data and assess their ability to provide useful information to nowcast the euro area quarterly GDP growth rate. Using a new approach mixing variable selection and ridge regression, for which we provide asymptotic results as regards the estimator, we empirically show that Google data are indeed useful, but only when official data are not available to practitioners, that is during the first four weeks at the beginning of the quarter. After this initial period, the marginal gain of those data tends to disappear as soon as official variables become available. Those results hold in real-time, meaning that Google data can be used by practitioners when official data are lacking. In addition to this, we explore the usefulness of Google search data during recession periods. We show that Google search data overperform official statistics to nowcast euro area GDP during the *Great Recession* period from 2008q1 to 2009q2.

Nowcasting GDP growth is extremely useful for policy-makers to assess macroeconomic conditions in real-time. The concept of macroeconomic nowcasting has been popularized by Giannone et al. [2008] and differs from standard forecasting approaches in the sense it aims at evaluating current macroeconomic conditions on a high-frequency basis. The idea is to provide policy-makers with a real-time evaluation of the state of the economy ahead of the release of official Quarterly National Accounts, that always come out with a delay. For example, the New York Fed and the Atlanta Fed have recently developed new tools in order to evaluate US GDP quarterly growth on a high-frequency basis[1]. The tool developed by the Atlanta Fed, referred to as *GDPNow*, is updated 6 to 7 times per month, while the NY Fed's tool is updated every Friday. With reference to countries other than the US, many papers have put forward econometric modelling to nowcast GDP growth in advanced countries (see among others Frale et al. [2010] or Kuzin et al. [2011] for the euro area, Aastveit and Trovik [2012] for Norway or Bragoli [2017] for Japan), as well as in emerging countries (see for example Modugno et al.

[1]See the websites https://www.newyorkfed.org/research/policy/nowcast and https://www.frbatlanta.org/cqer/research/gdpnow.aspx

[2016] for Turkey or Bragoli et al. [2015] for Brazil). Some researchers have also proposed approaches to nowcast economic output at a global level in order to assess on a regular high-frequency basis world economic conditions (see Golinelli and Parigi [2014] or Ferrara and Marsilli [2018]).

In the existing literature, GDP nowcasting tools integrate standard official macroeconomic information stemming, for instance, from National Statistical Institutes, Central Banks, International Organizations. Typically, three various sources of official data are considered: (i) hard data, like production, sales, employment, (ii) opinion surveys (households or companies are asked about their view on current and future economic conditions), and (iii) financial markets information (generally available on high frequency basis). However, more recently, a lot of emphasis has been put on the possible gain that forecasters can get from using alternative sources of high-frequency information, sometimes referred to as *Big Data* (see for example Varian [2014], Giannone et al. [2017] or Buono et al. [2018]). Various sources of *Big Data* have been used in the recent literature such as for example web scraped data, scanner data or satellite data. One of the main source of alternative data is Google search and seminal papers on the use of such data for forecasting are the ones by Choi and Varian [2009] and Choi and Varian [2012] (see also Scott and Varian [2015] who combine Kalman filters, spike-and-slab regression and model averaging to improve short-term forecasts).

Overall, empirical papers show evidence of some forecasting power for Google data, at least for some specific macroeconomic variables such as consumption (Choi and Varian [2012]), unemployment rate (D'Amuri and Marcucci [2017]), building permits (Coble and Pincheira [2017]) or car sales (Nymand-Andersen and Pantelidis [2018]). However, when correctly compared with other sources of information, the jury is still out on the gain that economists can get from using Google data for forecasting and nowcasting. For example, Vosen and Schmidt [2011] show that Google Trends data lead to an accuracy gain when compared with business surveys to forecast the annual growth rate of US household consumption. But some other papers tend to show that the gain in forecasting using Google data is very weak when other sources of information are accounted for in the analysis. For example, Goetz and Knetsch [2019] estimate German GDP using simultaneously both official and Google data on a monthly basis and show that adding Google data only leads to limited accuracy gains. However, they provide some evidence that those data can be a potential alternative to survey variables. We also refer to Li [2016] on this issue. Overall, the literature tends to point out that Google data can be extremely useful when economists do not have access to information or when information is fragmented, as for example when dealing with emerging economies (see Carriere-Swallow and Labbe [2013]) or low-income developing countries (Narita and Yin

[2018]).

In this paper, we estimate both pseudo real-time and true real-time nowcasts for the euro area quarterly GDP growth between 2014q1 and 2016q1 by plugging Google data into the analysis, in addition to official variables on industrial production and opinion surveys, commonly used to assess for GDP growth. The Google data that we get are indexes of weekly volume changes of Google searches in the six main euro area countries organized by keywords about different topics which are gathered in 26 broad categories such as auto and vehicles, finance, food and drinks, real estate, etc. Those broad categories are then split into a total of 269 sub-categories per country, leading to a total of 1776 variables for all the six countries[2]. Our objective is to assess whether Google search data convey some gain in nowcasting accuracy and when. in this respect, we put forward a new approach relying on a bridge equation explaining GDP growth by a few official variables, as proposed by Angelini et al. [2011], but which also integrates variables selected from a large set of Google data. More precisely, we pre-select Google variables by targeting GDP growth in the vein of Bai and Ng [2008] but with a different approach. Pre-selection is implemented by using the Sure Independence Screening method put forward by Fan and Lv [2008] enabling to preselect the Google variables the most related to GDP growth before entering the bridge equation. After pre-selection, we use Ridge regularization to estimate the bridge equation as the number of pre-selected variables may still be large. We provide new theoretical results showing the asymptotic properties of the estimator for this model that combines variable pre-selection and Ridge regularization.

Five main stylized facts come out from our empirical analysis. First, we point out the usefulness of Google search data for nowcasting euro area GDP for the first four weeks of the quarter when there is no available official information about the state of the economy. Indeed, we show that at the beginning of the quarter, Google data provide an accurate picture of the GDP growth rate. Against this background, this means that such data are a good alternative in the absence of official information and can be used by policy-makers. Second, we get that as soon as official data become available, that is, for the euro area, starting from the fifth week of the quarter, the gain from using Google data for GDP nowcasting rapidly vanishes. This result contributes to the debate on the use of big data for short-term macroeconomic assessment when controlling for standard usual macroeconomic information. Third, we show that pre-selecting Google

[2]See for example Bontempi et al. [2018] for a detailed description of this dataset, in a different framework

data before entering the nowcasting models appears to be a pertinent strategy in terms of nowcasting accuracy. Indeed, this approach enables to retain only Google variables that have some link with the targeted variable. This result confirms previous analyses that have been done when dealing with large datasets through dynamic factor models (see e.g. Bai and Ng [2008] or Schumacher [2010]). Fourth, we carry out a true real-time analysis by nowcasting euro area GDP growth rate using the official Eurostat timeline and vintages of data. We show that the three previous results still hold in real-time, in spite of an expected increase in the size of errors, suggesting that Google search data can be effectively used in practice to help the decision-making process. Finally, we evaluate to what extent Google search data are useful during recession periods. We empirically show that, for the *Great Recession* period from 2008q1 to 2009q2, nowcasts based on Google data, with and without official data, overperform nowcasts based on official data only.

The rest of the paper is organized as follows. In Section 2 we describe the model we consider for nowcasting, the Sure Independence Screening (SIS) approach to pre-select the data, as well as the Ridge regularization. In Section 3, we provide new theoretical results about the convergence of the estimator against this background. Section 4 describes the structure of the Google search data used for nowcasting. The empirical results are presented in Section 5 and Section 6 concludes.

2 Methodology

2.1 The nowcasting approach

In order to get GDP nowcasts, we focus on linear bridge equations that link quarterly GDP growth rates and monthly economic variables. The classical bridging approach is based on linear regressions of quarterly GDP growth on a small set of key monthly indicators as for example in Diron [2008]. In our exercise, in addition to those monthly variables, we also consider Google data, available at a higher frequency, and we aim at assessing their nowcasting power. More precisely, Google data are available on a weekly basis, providing thus additional information when official information is not yet available. Even if Google data are not on average extremely correlated with the GDP growth rate, we are going to show that they still provide accurate GDP nowcasts if conveniently treated.

Therefore, we assume that we have three types of data at disposal: *soft* data, such as

opinion surveys, *hard* data, such as industrial production or sales, and data stemming from Google search machines. Let t denote a given quarter of interest identified by its last month, for example the first quarter of 2005 is dated by $t = \text{March2005}$. A general model to nowcast the growth rate of any macroeconomic series of interest Y_t for a specific quarter t is the following, for $t = 1, \ldots, T$:

$$Y_t = \beta_0 + \beta_s' x_{t,s} + \beta_h' x_{t,h} + \beta_g' x_{t,g} + \varepsilon_t, \qquad \mathbf{E}[\varepsilon_t | x_{t,s}, x_{t,h}, x_{t,g}] = 0, \tag{2.1}$$

where $x_{t,s}$ is the N_s-vector containing *soft* variables, $x_{t,h}$ is the N_h-vector containing *hard* variables, $x_{t,g}$ is the N_g-vector of variables coming from Google search and ε_t is an unobservable shock. In our empirical analysis Y_t is the quarterly GDP growth rate of the euro area. Because variables $x_{t,s}$, $x_{t,h}$ and $x_{t,g}$ are sampled over different frequencies (monthly and weekly, respectively), the relevant dataset for calculating the nowcast evolves within the quarter. We asssume in the remaining of this paper that a given quarter is made up of thirteen weeks. Thus, by denoting with $x_{t,j}^{(w)}$, $j \in \{s, h, g\}$, the j-th series released at week $w = 1, \ldots, 13$ of quarter t, we denote the relevant information set at week w of a quarter t by

$$\Omega_t^{(w)} := \{x_{t,j}^{(w)}, j \in \{s, h, g\} \text{ such that } x_{t,j} \text{ is released at } w\}.$$

For simplicity, we keep in $\Omega_t^{(w)}$ only the observations relative to the current quarter t and do not consider past observations. While $x_{t,g}$ is in $\Omega_t^{(w)}$ for every $w = 1, \ldots, 13$, the other variables are in the relevant information set only for the weeks corresponding to their release and so the dataset is unbalanced.

To explicitly account for the different frequencies of the variables, we replace model (2.1) by a model for each week w such that:

$$\begin{aligned} \widehat{Y}_{t|w} &= \mathbf{E}[Y_t | \Omega_t^{(w)}], \qquad t = 1, \ldots, T \quad \text{and} \quad w = 1, \ldots, 13 \\ \text{and} \quad \mathbf{E}[Y_t | \Omega_t^{(w)}] &= \beta_{0,w} + \beta_{s,w}' x_{t,s}^{(w)} + \beta_{h,w}' x_{t,h}^{(w)} + \beta_{g,w}' x_{t,g}^{(w)} \end{aligned} \tag{2.2}$$

where $x_{t,j}^{(w)} = 0$ if $x_{t,j}^{(w)} \notin \Omega_t^{(w)}$. For instance, as the first observation of industrial production relative to the current quarter t is only released in week 9, then we set $x_{t,h}^{(w)} = 0$ for every $w = 1, \ldots, 8$. The bridge equation (2.2) exploits weekly information to obtain more accurate nowcasts of quarterly GDP growth.

For variables for which we have more than one release per quarter we consider the sample average of all the observations in the quarter. So, for Google variables, $x_{t,g}^{(w)} :=$

$\sum_{v=1}^{w} x_{t,g,(v)}/w$ where $x_{t,g,(v)}$ denotes the Google variable released at week v of quarter t. We define in a similar way $x_{t,s}^{(w)}$ and $x_{t,h}^{(w)}$ if they are released more than once in quarter t. We refer to Table 1 in the Annex for a detailed description of the models for each week.

2.2 Pre-selection of Google data

The recent literature on nowcasting and forecasting with large datasets comes to the conclusion that using the largest available dataset is not necessarily the optimal approach when aiming at nowcasting a specific macroeconomics variable such as GDP, at least in terms of nowcasting accuracy. Indeed, the problem arises because we have too many variables and using all the variables would only add noise in the estimation process. For example, against the background of bridge equations augmented with dynamic factors, Barhoumi et al. [2010] empirically show that factors estimated on a small database lead to competitive results for nowcasting French GDP compared with the most disaggregated data. From a theoretical point of view, Boivin and Ng [2006] suggest that larger databases lead to poor forecast when idiosyncratic errors are cross-correlated or when the forecasting power comes from a factor that is dominant in a small database but is dominated in a larger dataset. An empirical way to circumvent this issue is to target more accurately the variable to be nowcast. For example, Bai and Ng [2008] show that forming targeted predictors enables to improve the accuracy of inflation forecasts while Schumacher [2010] shows that targeting German GDP within a dynamic factor model is a performing strategy.

In this respect, all the categories and subcategories in the Google search data are not necessarily correlated with the GDP growth that we want to nowcast. Therefore, using all the variables in the Google search dataset is not necessarily a good strategy because one would pay the price of dealing with ultra-high dimensionality without increasing the nowcasting accuracy as measured by the Mean Squared Forecasting Error (MSFE). For this reason we pre-select Google data before performing the nowcast, that is, we consider a procedure enabling to pre-select the subset of variables in the Google search dataset that are the most relevant for GDP growth nowcasting. These ones are the variables that are the most "related" with the variable Y_t and that capture much of the variability in GDP growth. In a second step, we will use a Ridge regularization to estimate models (2.2) by using the selected subset of Google data. As explained in Section 2.3 below, a regularization technique is required because the number of selected variables can still be large while not ultra-high.

While in our empirical analysis we have tried several pre-selection procedures, it turns out that the innovative approach put forward by Fan and Lv [2008] appears to

provide interesting and intuitive results. This approach is referred to as *Sure Independence Screening*, or SIS hereafter. Sure screening refers to the property that *all important variables survive after applying a variable screening procedure with probability tending to 1* (see Fan and Lv [2008], p. 853). The basic idea of this approach is based on correlation learning and relies on the fact that only the variables with the highest absolute correlation with the GDP should be used in modelling.

Let us start from the standard linear regression equation (2.1) with only the standardized N_g Google variables as explanatory variables, that is $\beta_0 = \beta_s = \beta_h = 0$ in equation (2.1). Let Y denote the T-vector of quarterly GDP growth: $Y := (Y_1, \ldots, Y_T)'$. We compute $\omega := (\omega_1, \ldots, \omega_{N_g})'$, the vector of marginal correlations of predictors with the response variable Y_t, such as

$$\omega = \overline{X}_g' Y, \tag{2.3}$$

where $\overline{X}_g := (x_{1,g}^{(13)'}, \ldots, x_{T,g}^{(13)'})'$ is the $T \times N_g$ matrix of average Google data – where for each quarter we average over the thirteen weeks of this specific quarter, *i.e.* $x_{t,g}^{(13)} := \sum_{w=1}^{13} x_{t,g,(w)}/13$ – that have been centered and standardized columnwise. The average over each quarter is taken to make the weekly Google data comparable to the quarterly GDP growth data in terms of frequency. For any given $\lambda \in]0,1[$, the N_g components of the vector ω are sorted in a decreasing order and we define a submodel $\widehat{M}_g$ such as:

$$\widehat{M}_g = M_g(\lambda) := \{1 \leq j \leq N_g : |\omega_j| \text{ is among the first } [\lambda T] \text{ largest of all}\},$$

where $[\lambda T]$ denotes the integer part of λT. Since only the order of componentwise magnitudes of ω is used, this procedure is invariant under scaling and thus it is identical to selecting predictors using their correlations with the response. This approach is an easy way to filter out Google variables with the weaker correlations with GDP growth rate so that we are left with $d = [\lambda T] < T$ Google variables. The empirical choice of the hyperparameter λ is discussed in subsection 3.3. An important feature of the SIS procedure is that it uses each covariate $x_{t,g,j}$ independently as a predictor to decide how useful it is for predicting Y_t.

This method is desirable because it has the sure screening property, that is, with probability tending to one, all the important variables in the true model are retained after applying this method. Let $N := 1 + N_s + N_h + N_g$, $\beta := (\beta_0, \beta_s', \beta_h', \beta_g')'$ and let $M^* := \{1 \leq j \leq N : \beta_j \neq 0\}$ be the true sparse model with non-sparsity size $s^* = |M^*|$. Moreover, let $M_g^* := \{1 \leq j \leq N_g : \beta_{g,j} \neq 0\}$ be the subset of the true sparse model containing only the indices of the active Google variables with size $s_g^* = |M_g^*|$ and remark that $\{1 + N_s + N_h + j;\, j \in M_g^*\} \subset M^*$. The $N_g - s_g^*$ variables whose index is not included in M_g^* can also be correlated with Y via linkage to the predictors

contained in the true sparse model M_g^*. Finally, denote $\widehat{M} = \widehat{M}(\lambda) := \{1 \leq j \leq (1 + N_s + N_h)\} \cup \{1 + N_s + N_h + j;\, j \in \widehat{M}_g\}$. Fan and Lv [2008, Theorem 1] show that under normality of ε_t and other conditions (see Fan and Lv [2008, Conditions 1-4]) the sure screening property holds, namely for $\lambda = cT^{-\theta}$ where $c > 0$ is a constant and $\theta < 1 - 2\kappa - \tau$ with κ and τ defined in Fan and Lv [2008, Conditions 3-4]:

$$P(M^* \subset \widehat{M}) = 1 - O(\exp\{-CT^{1-2\kappa}/\log(T)\}).$$

In particular, SIS can reduce the dimension from N_g to $[\lambda T] = O(T^{1-\theta}) < T$ for some $\theta > 0$ and the reduced model $\widehat{M}$ still contains all the variables in the true model M^* with a probability converging to one as $T \to \infty$.

As an alternative to the SIS procedure one could use the Lasso to pre-select Google variables. Let $\widehat{\beta}_{lasso}^{(w)}$ denote the lasso estimator obtained by solving the following minimization problem for each week w:

$$\widehat{\beta}_{lasso}^{(w)} := \arg\min_{\beta_g} \left\{ \frac{1}{T} \sum_{t=1}^{T} \left(Y_t - \beta_0 - \beta_g' x_{t,g}^{(w)} \right)^2 + \lambda \|\beta_g\|_1 \right\}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm and λ is a penalty level. The model selected by the Lasso for each week is given by $\widehat{M}_g := support(\widehat{\beta}_{lasso}^{(w)}) = \{j : \widehat{\beta}_{lasso,j}^{(w)} \neq 0\}$.

2.3 Ridge regression

Google search data have an extremely large dimension, with the number of variables much larger than the number of observations (i.e. $N_g \gg T$, sometimes referred to as *fat datasets*). Therefore, when using Google search data for nowcasting one has to deal with such high dimensionality. Even after implementing one of the pre-selection described in subsection 2.2, the number of Google variables may remain large compared to the time dimension T. Therefore, one needs to use a machine learning technique suitable to treat fat datasets.

One of the most popular ways to deal with a large number of covariates and possible problems of multicollinearity is the Ridge regression (also known as Tikhonov regularisation). Let $\beta := (\beta_0, \beta_s', \beta_h', \beta_g')'$ and denote $X_{t,\widehat{M}} := (1, x_{t,s}', x_{t,h}', x_{t,g,\widehat{M}_g}')'$, where $x_{t,g,\widehat{M}_g} = \{x_{t,g,j}; j \in \widehat{M}_g\}$ is the vector containing only the selected Google variables and where for simplicity we omit the superscript '(w)'. Ridge regression estimates β in equation (2.2) by minimizing a penalized residuals sum of squares where the penalty is given by the Euclidean squared norm $\|\cdot\|_2$. Our procedure consists in first pre-selecting

data by using either the SIS method or the Lasso and then, in a second step, we apply the Ridge regularisation to the selected model $\widehat{M}$. By using model (2.2) for each week $w \in \{1, \ldots, 13, \}$ we define the Ridge estimator after model selection, which we call *Ridge after model selection* estimator, as:

$$\begin{aligned}\widehat{\beta}^{(w)} &= \widehat{\beta}^{(w)}(\alpha) \\ &:= \arg \min_{\beta; \beta_{g,j}=0, j \in \widehat{M}_g^c} \left\{ \frac{1}{T} \sum_{t=1}^{T} \left(Y_t - \beta_0 - \beta_s' x_{t,s}^{(w)} - \beta_h' x_{t,h}^{(w)} - \beta_g' x_{t,g}^{(w)} \right)^2 + \alpha \|\beta\|_2^2 \right\}, \quad (2.4)\end{aligned}$$

where $\alpha > 0$ is a regularization parameter that tunes the amount of shrinkage. Without loss of generality, we can assume that the selected elements of $x_{t,g}^{(w)}$ corresponding to the indices in $\widehat{M}_g$ are the first elements of the vector. Then, we can write $\widehat{\beta}^{(w)}$ as $\widehat{\beta}^{(w)} = (\widehat{\beta}_{1:|\widehat{M}|}^{(w)\prime}, \mathbf{0}')'$ where

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

$\mathbf{0}$ is the $(N - |\widehat{M}|)$-dimensional column vector of zeros, and I is the $|\widehat{M}|$-dimensional identity matrix. This is the estimator we are going to use in our empirical analysis. The empirical choice of the hyperparameter α is a crucial issue because it has an important impact on the nowcasting accuracy. We discuss this choice in Section 4.3. In the next section we present the theoretical properties of the *Ridge after model selection* estimator.

3 Theoretical Properties

In this section we present theoretical properties of the Ridge after model selection estimator. This estimator has not been considered in the literature before. SIS pre-selection has been coupled with the SCAD method of Fan and Li [2001] and the Dantzig selector in Candes and Tao [2007] by Fan and Lv [2008] who establish consistency of the corresponding estimator. The Lasso pre-selection has been coupled with the Least square and the Ridge estimator by Liu and Yu [2013] who also establish consistency. In particular, the latter consider the case where $P(M^* = \widehat{M}) \to 1$ as $T \to \infty$. Asymptotic properties for the out-of-sample prediction error associated with the Ridge estimator without model selection have been analysed in Giannone et al. [2008] and Carrasco and Rossi [2016] while asymptotic properties for the in-sample prediction error are well known in the inverse problems literature, see *e.g.* Carrasco et al. [2007] and Florens and

Simoni [2016] for a Bayesian interpretation of the Ridge estimator. Here, we establish an upper bound for both the in-sample and out-of-sample prediction error associated with the Ridge after model selection estimator. This upper bound gives the rate of covergence as $N, T \to \infty$.

Let $\beta := (\beta_0, \beta_s', \beta_h', \beta_g')'$ and $X_t := (1, x_{t,s}', x_{t,h}', x_{t,g}')'$ for $t = 1, \ldots, T$, be N-dimensional column vectors where we have eliminated the week w index for simplicity, and let $X := (X_1, \ldots, X_T)'$ be a $(T \times N)$ matrix. Recall the definition $M^* := \{1 \leq j \leq N : \beta_j \neq 0\}$ with $s^* := |M^*|$, and let M^{*c} denote the complementary set of M^* in $\{1, \ldots, N\}$. Remark that, if we denote by N_1 the dimension of $(\beta_0, \beta_s', \beta_h')'$, then $M^* = \{1, \ldots, N_1\} \cup \{N_1 + j; j \in \{1, 2, \ldots, N_g\}$ and $\beta_{g,j} \neq 0\}$. For a vector $\beta \in \mathbb{R}^N$ and an index set $M \subset \{1, \ldots, N\}$, denote $\beta_{M,j} := \beta_j \mathbb{1}\{j \in M\}$ and, for a $(T \times N)$ matrix X denote by X_M the $(T \times |M|)$ matrix made of the columns of X corresponding to the indices in M and by $X_{t,M}$ the transpose of the t-th row of X_M. Thus, β_M has zero outside the set M. We denote by P_X (resp. P_{X_τ}) the conditional probability given the covariates X (resp. X and X_τ). For a vector $\delta \in \mathbb{R}^N$ and given covariates X_t, $t = 1, \ldots, T$, define the squared prediction norm of δ as $\|\delta\|_{2,T}^2 := \delta' X' X \delta / T$, the ℓ_0-norm of δ as $\|\delta\|_0 := \sum_{j=1}^N \mathbb{1}\{\delta_j \neq 0\}$ and the Euclidean norm is denoted by $\|\delta\|_2 := \sqrt{\delta'\delta}$.

We now state the assumptions that we use to derive the theoretical results. For simplicity, we leave implicit the dependence of each model on the week w.

ASSUMPTION A.1. *(i) Assume that $Y_t = \beta_*' X_t + \varepsilon_t$, $t = 1, \ldots, T$, with β_* the true value of β, $\beta := (\beta_0, \beta_s', \beta_h', \beta_g')'$ and $X_t := (1, x_{t,s}', x_{t,h}', x_{t,g}')'$ both N-dimensional vectors, and let $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ be independent for $t = 1, \ldots, T$. (ii) $\beta_{*g} = (\beta_{*g,1}, \ldots, \beta_{*g,s_g^*}, \beta_{*g,s_g^*+1}, \ldots, \beta_{*g,N_g})'$ with $\beta_{*g,j} \neq 0$ for $j = 1, \ldots, (s^* - N_1)$ and $\beta_{*j} = 0$ for $j = s^* + 1, \ldots, N_g$.*

Assumption A.1 *(i)* states that the true model is linear with Gaussian errors, while Assumption A.1 *(ii)* states that the subvector of the true β_* corresponding to the Google variables is s_g^*-sparse. Next, we introduce an assumption which is known in the literature as a restricted sparse eigenvalue condition on the empirical Gram matrix $(X_{\widehat{M}}' X_{\widehat{M}})/T$, see *e.g.* Belloni and Chernozhukov [2013].

ASSUMPTION A.2. *For a given $m < T$,*

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

We also define the condition number associated with the empirical Gram matrix $(X_{\widehat{M}}' X_{\widehat{M}})/T$:

$$\mu(\widehat{m}) = \frac{\sqrt{\phi(\widehat{m})}}{\kappa(\widehat{m})}, \qquad \text{where } \widehat{m} := |\widehat{M} \setminus M_*| \mathbb{1}\{\widehat{M} \supseteq M_*\}.$$

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

We start by establishing an upper bound on the in-sample prediction error. Its proof is provided in Appendix B.

THEOREM 3.1 (In-sample prediction error). *Suppose that Assumptions A.1 and A.2 are satisfied and let $\widehat{M}$ be the model selected in the first step. Let $\widehat{\beta}^{(w)}$ be the Ridge estimator defined in (2.4). Then, for every $\epsilon > 0$, there is a constant K_ϵ independent of T such that with P_X-probability at least $1-\epsilon$,*

$$\|\widehat{\beta}-\beta_*\|_{2,T} \leq \left(K_\epsilon \sqrt{\frac{\widehat{m}\log(N) + (\widehat{m}+s^*)\log(e^2\mu(\widehat{m}))}{T}} + 2\alpha\|\beta_*\|_2 \frac{1}{\kappa(\widehat{m})} \right) \mathbb{1}\{\widehat{M} \supseteq M^*\}$$
$$+ \left(\frac{K_\epsilon \sigma}{\sqrt{T}} \sqrt{\widehat{k}\log(s^*) + \widehat{k}\log(e^2\mu(0))} + \frac{2\alpha}{\kappa(0)}\|\beta_*\|_2 + \|\beta_{*,M^*\setminus\widehat{M}}\|_{2,T} \right) \mathbb{1}\{\widehat{M} \subset M^*\}.$$

The theorem is stated in terms of conditional probability given covariates X and selected model $\widehat{M}$. We could eliminate the conditioning on X by adding an assumption about boundedness of the second moment of each component of X. For both Lasso and SIS pre-selection methods the probability of the event $\{\widehat{M} \supset M^*\}$ converges to 1 (and so the probability of $\{\widehat{M} \supseteq M^*\}$). We remark that if $\widehat{M} \subset M^*$ we get a bias term given by $\|\beta_{0,M^*\setminus\widehat{M}}\|_{2,T}$. This is intuitive since the second-step Ridge estimator is always biased for the components in $M^* \setminus \widehat{M}$.

The next corollary establishes an upper bound for the Euclidean norm of $(\widehat{\beta} - \beta_*)$.

Corollary 3.1 (Coefficient estimation)**.** *Suppose that Assumptions A.1 and A.2 are satisfied and let $\widehat{M}$ be the model selected in the first step. Let $\widehat{\beta}^{(w)}$ be the Ridge estimator defined in (2.4). Then, for every $\epsilon > 0$, there is a constant K_ϵ independent of T such that with P_X-probability at least $1-\epsilon$,*

$$\|\widehat{\beta} - \beta_*\|_2 \leq \left(K_\epsilon \sqrt{\frac{\widehat{m}\log(N) + (\widehat{m}+s^*)\log(e^2\mu(\widehat{m}))}{T\kappa(\widehat{m})^2}} + 2\alpha\|\beta_*\|_2 \frac{1}{\kappa(\widehat{m})^2} \right) \mathbb{1}\{\widehat{M} \supseteq M^*\}$$
$$+ \left(\frac{K_\epsilon \sigma}{\kappa(\widehat{m})\sqrt{T}} \sqrt{\widehat{k}\log(s^*) + \widehat{k}\log(e^2\mu(0))} + \frac{2\alpha}{\kappa(\widehat{m})\kappa(0)}\|\beta_*\|_2 + \frac{\|\beta_{*,M^*\setminus\widehat{M}}\|_{2,T}}{\kappa(\widehat{m})} \right) \mathbb{1}\{\widehat{M} \subset M^*\}.$$

Compared to the upper bound for the in-sample prediction error, every term in the upper bound in Corollary 3.1 has an additional factor of $1/\kappa(\widehat{m})$. As seen in Assumption A.2, $\kappa(\widehat{m})$ has to be interpreted as the smallest restricted eigenvalue of the empirical Gram matrix and so it can be small when N is large. Therefore, the upper bound in Corollary 3.1 can be larger than the upper bound in Theorem 3.1.

In the next theorem we establish an upper bound for the out-of-sample prediction

error.

Corollary 3.2 (Out-of-sample prediction error)**.** *Suppose that Assumptions A.1 and A.2 are satisfied and let $\widehat{M}$ be the model selected in the first step. Let $\widehat{\beta}^{(w)}$ be the Ridge estimator defined in (2.4). 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 P_{X_τ}-probability at least $(1-\epsilon)$,*

$$X_\tau'(\widehat{\beta} - \beta_*) \leq (\sqrt{\widehat{m}+s^*})C \times \Big[\left(K_\epsilon \sqrt{\frac{\widehat{m}\log(N) + (\widehat{m}+s^*)\log(e^2\mu(\widehat{m}))}{T\kappa(\widehat{m})^2}} + 2\alpha\|\beta_*\|_2 \frac{1}{\kappa(\widehat{m})^2} \right) \mathbb{1}\{\widehat{M} \supseteq M^*\} + \left(\frac{K_\epsilon \sigma\sqrt{s^*}}{\kappa(\widehat{m})\sqrt{T}} \sqrt{\widehat{k}\log(s^*) + \widehat{k}\log(e^2\mu(0))} + \frac{2\alpha\sqrt{s^*}}{\kappa(\widehat{m})\kappa(0)}\|\beta_*\|_2 + \frac{\sqrt{s^*}}{\kappa(\widehat{m})}\|\beta_{*,M^*\setminus\widehat{M}}\|_{2,T} \right) \mathbb{1}\{\widehat{M} \subset M^*\} \Big].$$

The upper bound for the out-of-sample prediction error is larger than the upper bound for the in-sample prediction error. This is because X_τ has dimension N which is large. However, thanks to the pre-selection, this dimension is reduced from N to $(\widehat{m}+s^*)$ which gives the factor outside the square bracket in the upper bound in Corollary 3.2. Hence, we do not need to assume that $\|X_\tau\|_2 = O_p(1)$ as *e.g.* in Carrasco and Rossi [2016].

4 Design of the empirical analysis

This section first describes the data used in the empirical analysis. Then, it describes how to deal with the various reporting lags. Finally, we propose a way to select both the hyperparameters λ and α involved in the estimation procedure.

4.1 Data

Our objective in this paper is to assess the role of Google data for nowcasting the euro area GDP, especially to assess (i) if these big data are relevant when there is no official data available for the forecaster and (ii) to what extent these data provide useful information when official data become available. In this respect, the variable Y_t in model (2.1)-(2.2) that we target is the quarterly growth rate of the real euro area GDP, stemming from Eurostat. The official data that we consider are of two kinds: industrial production for the euro area as a whole provided by Eurostat, which is a global measure

of hard data and is denoted by IP_t, and a composite index of opinion surveys from various sectors computed by the European Commission (the so called *euro area Sentiment Index*) denoted by S_t.

Our big dataset covers Google searches for the six main euro area countries: Belgium, France, Germany, Italy, Netherlands and Spain. We have at disposal a total of $N_g = 1776$ variables, corresponding to 26 categories and 296 subcategories for each country. Google search data are data related to queries performed with Google search. The data are indexes of weekly volume changes of Googles queries grouped by category and by country. Data are normalized at 1 at the first week of January 2004 which is the first week of availability of these data. Then, the following values indicate the deviation from the first value. However, there is no information about the search volume. Google data are weekly data that are received and made available by the European Central Bank every Tuesday. Original data are not seasonally adjusted, thus we take the growth rate over 52 weeks to eliminate the seasonality within the data.

We use data from 20 March 2005 (twelfth week of the first quarter) until 29 March 2016 (thirteen week of the first quarter). We split the sample in two parts and use data starting from the first week of January 2014 for the out-of-sample analysis.

4.2 Dealing with various reporting lags

An important feature of all these data is that they are released with various reporting lags, leading thus to non-balanced information dataset at each point in time within the quarter. In the literature, this issue is refereed to as *ragged-edge database* (see Angelini et al. [2011]). For instance, Google search data are weekly data available every Tuesday, while the soft and hard data are monthly data available at the end of every month and at the middle of the third month of the quarter, respectively. Treating weekly data is particularly challenging as the number of entire weeks present in every quarter is not always the same, and a careful analysis has to be done when incorporating these data. In addition, there is a frequency mismatch in the data as the explained variable is quarterly and the explanatory variables are either weekly (Google data) or monthly (hard and soft variables). In order to account for the various frequencies and the timing at which the predictive variables become available, we adopt the strategy to consider a different model for every week of the quarter as described in Section 2.1. Thus we end up with thirteen models given in equation (2.2), each model including the variables available at this date.

As regards the dates of availability, we mimic the exact release dates as published by Eurostat. This means that the first survey of the quarter, referring to the first month, typically arrives in week 5. Then, the second survey of the quarter, related to the second month, is available in week 9. Industrial production for the first month of the quarter is only available about 45 days after the end of the reference month, that is generally in week 11. Finally, the last survey, related to the third month of the quarter, is available in week 13. A scheme of the release timeline is presented in Figure 1.

To construct the variable $x_{t,g}^{(w)}$ in equation (2.2) containing Google search data for the w-th week, we take the sample average of the Google variables from week 1 to week w of the quarter t. Let us denote by $x_{t,g,(w)}$ the Google variable available at week w of period t not averaged. Hence, $x_{t,g}^{(w)} = \sum_{v=1}^{w} x_{t,g,(v)}/w$ and pre-selection is applied to this average variable. Take for instance $w = 3$ (*i.e.* Model 3 which is used at week 3), then $x_{t,g}^{(3)}$ is equal to $(x_{t,g,(1)} + x_{t,g,(2)} + x_{t,g,(3)})/3$.[3] The other variables in equation (2.2) denote, respectively: Y_t the euro area GDP growth rate, $x_{t,s}^{(w)}$ the monthly data from surveys, available at the end of each month, and $x_{t,h}^{(w)}$ denotes the growth rate of the index of industrial production, available about 45 days after the end of the reference month. Because of the frequency mismatch within the whole dataset, the thirteen models include a different number of predictors, as we have explained above. As regards the survey, $x_{t,s}^{(w)}$, and the industrial production, $x_{t,h}^{(w)}$, we impose the following specific structure which mimics the data release explained above, and that will be used throughout our exercise. The variable $x_{t,s}^{(w)}$ is not present in models 1 to 4 because the survey is not available in the

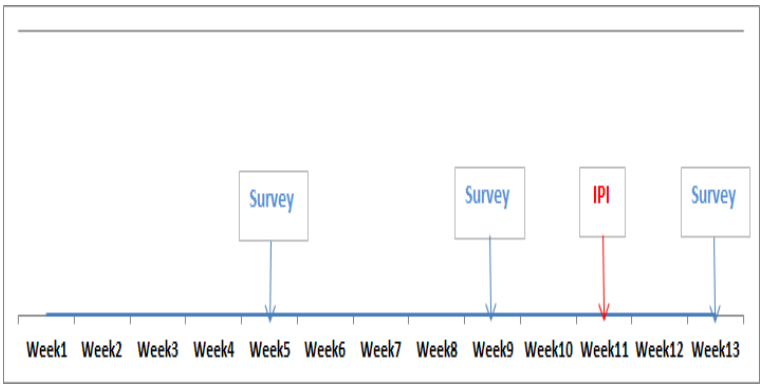

Figure 1: Timeline of data release in the pseudo real-time exercise within the quarter.

[3]In our empirical analysis, we also test models that do not use the average over weeks of Google search data as explanatory variables, but instead, Google search data for each new weeks is considered as the variable for the quarter. Results clearly point that models integrating averaged Google search data give smaller Mean Squared Forecasting Errors than models that do not use the averaged Google search data.

first four weeks of the quarter, so that $\beta_{t,s}^{(1)} = \beta_{t,s}^{(2)} = \beta_{t,s}^{(3)} = \beta_{t,s}^{(4)} = 0$. Then, for models 5 to 8, $x_{t,s}^{(w)}$ is the value of the survey for the first month of the quarter: $x_{t,s}^{(w)} = S_{t,1}$ where $S_{t,i}$ denotes the variable S_t referring to the i-th month of quarter t. In models 9 to 12, $x_{t,s}^{(w)}$ will be equal to the average of the survey data available at the end of the first and second month of the quarter: $x_{t,s}^{(w)} = (S_{t,1} + S_{t,2})/2$. Last, in model 13, $x_{t,s}^{(w)}$ is the average of the survey data over the quarter: $x_{t,s}^{(w)} = (S_{t,1} + S_{t,2} + S_{t,3})/3$. Similarly, the variable $x_{t,h}^{(w)}$ is not present in models 1 to 10 (so that $\beta_{t,h}^{(1)} = \ldots = \beta_{t,h}^{(10)} = 0$), and in models 11 to 13, $x_{t,h}^{(w)}$ will be the value of the growth rate of the index of industrial production IP_t for the first month of the quarter.

The idea of having thirteen models is that a researcher will use one of these ones to nowcast the current-quarter values of Y_t depending on the current week of the quarter. For instance, to nowcast the current-quarter value of Y_t at the end of week 2, the Model $w = 2$ will be used. In Table 1 in the Annex, we give the thirteen models based on equation (2.2) and we denote them by $M1, \ldots, M13$.

One of the main issue in the literature on big data is to know whether and when such alternative data are able to bring an additional gain with respect to standard types of variables, like hard and soft data. To contribute to the existing literature on this issue, we have also estimated nowcasting models without including the vector of variables selected from the Google search data. That is, these models only include as predictors the survey and the growth rate of the index of industrial production (i.e. $\beta_{t,g}^{(w)} = 0$ in equation (2.2)). We have in total four such models, one for each release of data of these two variables within the quarter, denoted $NoGoogle_1, \ldots, NoGoogle_4$ in Table 2 in the Annex, that will be used for comparison purposes.

An additional issue with the reporting lags concerns the release of GDP figures. In fact, the first GDP assessment is generally released about 45 days after the end of the reference quarter, but sometimes the delay may be longer. For instance, GDP figures for the first quarter of 2014 were only released on the 4^{th} of June 2014. For this reason if one wants to nowcast in real-time GDP growth for 2014q2 it is not possible to use the estimated model with the data available up to 2014q1 because one does not observe the GDP for 2014q1. Instead, one has to use the estimated parameters computed with the data available up to 2013q4. Because of this, we impose a gap of two quarters between the sample used for fitting the model (training sample) and the sample used for the out-of-sample analysis (test data). For coherence, we use this structure in both the pseudo-real-time and the true real-time analysis.

Another issue concerns the inclusion of lagged GDP among the explanatory variables. Because of the delay in the release of the GDP we cannot include the lagged GDP as explanatory variable in every nowcasting model. In addition to this, the GDP is not released at a fixed date, meaning that the release is different at every period (every quarter and every year). For these reasons we have not included the lagged GDP among the explanatory variables in the thirteen models (2.2) for the pseudo-real-time analysis. On the other hand, for the true real-time analysis we have exploited the additional information arising from lagged GDP and have included it among the explanatory variables when it is available. We provide in Table 3 in the Annex an overview of the dates at which specific GDP figures are released as well as the indication of the time from which we can include the lagged GDP among the explanatory variables and the arrival times of new vintages. We have used this calendar to construct our real-time analysis.

In fact we carry out two types of real-time analysis: (I) a true real-time analysis which includes the lagged GDP growth when it is available, and (II) a true real-time analysis which does not include the lagged GDP growth. The latter is meant for comparison with the pseudo-real-time analysis which does not include lagged GDP values.

4.3 Selection of the tuning parameters λ and α

To construct our Ridge after model selection estimator for the thirteen models (2.2) one has to set two tuning parameters: λ and α. The empirical choice of the latter is crucial because it has an important impact on the nowcasting accuracy. We select them by using a data-driven method based on a grid-search procedure on the training sample corresponding to the specific nowcasting period we are considering. Selection of (λ, α) is made for each of the thirteen models and for each nowcasting period. Hence, in total we have $13 * 9 = 117$ values for the pair (λ, α).

Let τ denote the last quarter of the training sample. In our analysis, we consider for λ a grid of 99 equispaced values in $(0, 1]$, denoted by Λ. Then, the selection is made sequentially: for each value of λ in the grid we select for α in model w and for the nowcasting period $\tau + 1$ the value $\widehat{\alpha}_\tau^{(w)}(\lambda)$ that solves $\widehat{\alpha}_\tau^{(w)}(\lambda) := \arg\min_{\alpha \in \mathcal{A}} \widehat{Q}_\tau^{(w)}(\alpha; \lambda)$, where $\mathcal{A}$ is a grid of equispaced values in $[0, a]$, for some $a > 0$, and $\widehat{Q}_\tau^{(w)}(\alpha; \lambda)$ is a criterion to be defined below.

Once a value $\widehat{\alpha}_\tau^{(w)}(\lambda)$ is selected for each value of λ in the grid, we select the value of λ that minimizes the MSFE of the nowcast of the GDP growth of the last quarter τ of the training sample obtained by using the selected $\widehat{\alpha}_\tau^{(w)}(\lambda)$. That is, in model w we select the value $\widehat{\lambda}_\tau^{(w)} := \arg\min_{\lambda \in \Lambda} (Y_\tau - X'_{\tau, \widehat{M}} \widehat{\beta}^{(w)}(\widehat{\alpha}_\tau^{(w)}(\lambda)))^2$ where $\widehat{M} = \widehat{M}(\lambda)$.

For the choice of α we propose to use two criteria. The first one is based on the

Generalized cross-validation (GCV) technique (see Li [1986, 1987]) whose idea is to choose a value for α for which the MSFE is as small as possible. This technique has recently been used by Carrasco and Rossi [2016] in a forecasting setting. The idea is to select the value of α that minimizes the following quantity:

$$\widehat{Q}_{\tau}^{(w)}(\alpha;\lambda) = \frac{\tau^{-1}\sum_{t=1}^{\tau}(Y_t - X'_{t,\widehat{M}}\widehat{\beta}^{(w)})^2}{\left(1-\tau^{-1}tr(\widehat{R}_{\tau}(\alpha))\right)^2}$$

where τ denotes the last quarter of the training sample, $tr(\cdot)$ denotes the trace operator and $\widehat{R}_{\tau}(\alpha)$ is given by

$$\widehat{R}_{\tau}(\alpha) = X_{\widehat{M}}\left(\tau^{-1}\sum_{t=1}^{\tau}X_{t,\widehat{M}}X'_{t,\widehat{M}} + \alpha I\right)^{-1}\tau^{-1}X'_{\widehat{M}}.$$

The second criterion is based on the idea of looking at the norm of the residuals with respect to the normal equations, this is known as *error free method* in the inverse problem literature, see Engl et al. [2000]. Hence, the criterion to be minimised is :

$$\widehat{Q}_{\tau}^{(w)}(\alpha;\lambda) = \frac{1}{\alpha^2}\left\|\frac{\sum_{t=1}^{\tau}X_{t,\widehat{M}}Y_t}{\tau} - \frac{\sum_{t=1}^{\tau}X_{t,\widehat{M}}X'_{t,\widehat{M}}\widehat{\beta}^{(w)}_{\widehat{M},2}}{\tau}\right\|_2^2$$

where $\widehat{\beta}^{(w)}_{\widehat{M},2}$ is the 2-times iterated Tikhonov (Ridge) estimator.

5 Empirical Results

In this section we present the results of our empirical exercises aiming at nowcasting the euro area GDP growth using various types of data sources. In the following the notation $M1,\ldots,M13$ and $NoGoogle_1,\ldots,NoGoogle_4$ refer to the notation and models defined in Tables 1 and 2 in the Annex, respectively.

5.1 Overall evaluation of Google search data

This section is split into three parts. First, we look at the accuracy gains stemming from using Google data when controlling for standard official macroeconomic data, by comparing nowcasts obtained with and without such data, in a pseudo real-time exercise. Then we look at the effects of pre-selecting Google data before estimating Ridge regressions. Third, we perform a true real-time analysis.

5.1.1 Is there a gain from using Google data, and when ?

In this subsection we compare the evolution over the quarter of weekly Root MSFEs (RMSFE) stemming from the nowcasting models, with and without Google search data. We do this exercise in pseudo-real time, that is, by using historical data but by accounting for their ragged-edge nature. To evaluate the impact of Google search data on current-quarter nowcasts of the GDP growth, we make two types of comparisons. First, we estimate the thirteen nowcasting models by using only Google data, that is, $x_{t,s}^{(w)} = x_{t,h}^{(w)} = 0$ for every $w = 1, \ldots, 13$ in Equation (2.2). Second, to assess the marginal gain of integrating Google data, we compare the four models that only account for hard and soft data (i.e. without Google data) with the corresponding models given by (2.2) accounting for the full set of information (Google, Survey and Industrial Production). More precisely, we directly compare the following pairs of models: $(NoGoogle_1, M5)$, $(NoGoogle_2, M9)$, $(NoGoogle_3, M11)$, and $(NoGoogle_4, M13)$. The results of these comparisons are reported in Figure 2 below and in Table 4 in the Annex. The estimation has been conducted by using Ridge regularization coupled with the SIS pre-selection approach as described in Sections 2.2 and 2.3. Figure 3 is similar to Figure 2 but with pre-selection conducted by using the Lasso. The corresponding RMSFE values are reported in Table 5 in the Annex. The tuning parameter of the Lasso is automatically chosen by cross-validation in the range $(0, 5)$. This range is clearly arbitrary and the results slightly change when we change the range for λ. We have also computed the *Ridge after Lasso selection* estimator by constraining the λ to be chosen such that the number of selected Google search categories does not exceed the number selected with the SIS procedure. These results are shown in Figure 10 and in Table 6 in the Annex.

The first striking feature that we observe in Figure 2 is the downward sloping evolution of RMSFEs stemming from the models with full information (Google, Industrial Production and Survey) over the quarter. This is in line with what could be expected from nowcasting exercises when integrating more and more information throughout the quarter (see Angelini et al. [2011]). When using Google information only (light gray bars), we still observe a decline but to a much lower extent and the RMSFEs stay above 0.25 even at the end of the quarter. However, when focusing on the beginning of the quarter, models that only integrate Google information provide reasonable RMSFEs that do not exceed 0.30 (see Figure 2). This result shows that Google search data possess a informational content that can be valuable for nowcasting GDP growth for the first four weeks

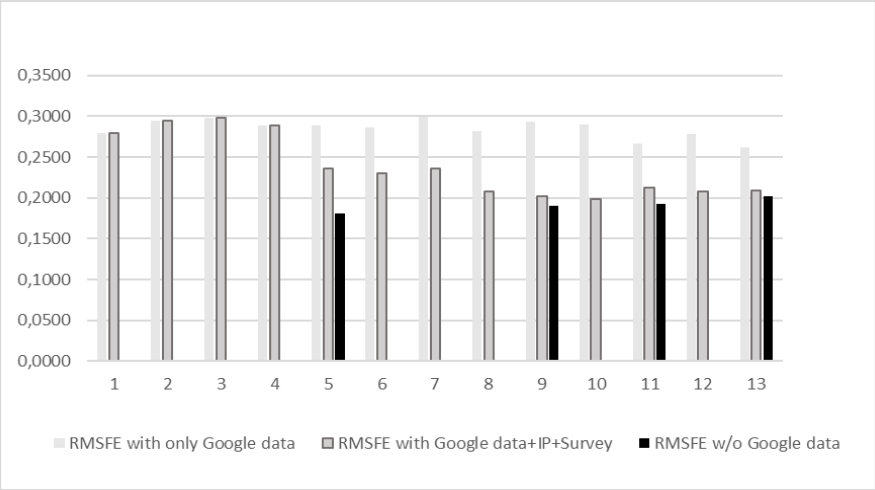

Figure 2: The importance of Google data. Pseudo-real-time analysis with pre-selection of Google data by SIS method. RMSFEs from: (i) models M1 - M13 with only variables extracted from Google data (in light gray), (ii) models M1 - M13 with all the variables (S_t, IP_t and Google data) (in gray), (iii) models with only official variables $NoGoogle_1$ - $NoGoogle_4$ (in black).

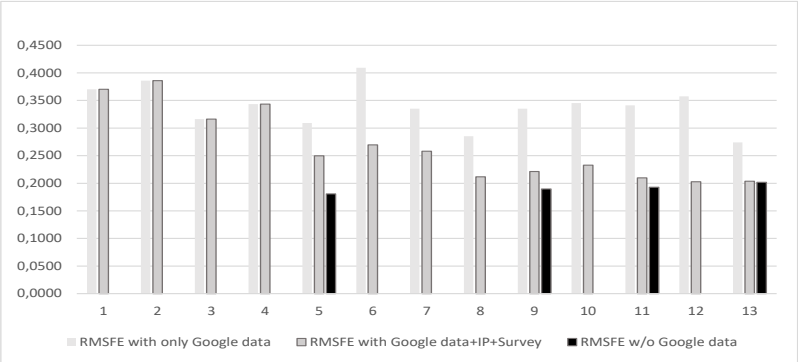

Figure 3: The importance of Google data. Pseudo-real-time analysis with pre-selection of Google data by Lasso. RMSFEs from: (i) models M1 - M13 with only variables extracted from Google data (in light gray), (ii) models M1 - M13 with all the variables (S_t, IP_t and Google data) (in gray), (iii) models with only official variables $NoGoogle_1$ - $NoGoogle_4$ (in black).

of the quarter, when there is no other available official information about the current state of the economy. When information about the first survey of the quarter arrives, that is in week 5, the model that only incorporates Google data clearly suddenly underperfoms. Looking at Table 4, we see that the RMSFE goes from 0.2887 in week 4 when only Google information is used to 0.2361 in week 5 when the full information model is used. In addition, we note that a simple model, only accounting for official hard and

soft information, leads to a much lower RMSFE in week 5 (equal to 0.1807, see Table 4 in the Annex). Comparing black bars and gray bars in Figure 2 clearly shows evidence that there is no additional gain of adding Google data to the model starting from week 5; a simple model with only hard and soft information cannot indeed be outperformed.

As a robustness check, we compare the SIS pre-selection procedure with the Lasso procedure, that we consider as an alternative pre-selection approach, standard in this literature (see Section 2.2). Results for RMSFEs obtained using this approach are presented in Figure 3. We get an overall similar pattern as the one obtained with the SIS approach in the sense that RMFSEs declined over the quarter when more information is integrated. We also observe a clear shift in week 5 when we integrate the first survey of the quarter, RMSFE is going down to 0.2498 from 0.3435 in week 4 (see results in Table 5 in the Annex). However, when compared with the SIS approach, we note that for the first four weeks of the quarter, RMSFEs are much higher than those obtained with the SIS approach, for any model.

5.1.2 Is it worth to pre-select Google data?

As mentioned in Section 2.2, the literature suggests that it could be useful to first pre-select a sub-sample of Google data before estimating the thirteen models given in (2.2). In this respect, various approaches have been put forward in order to target *ex ante* the variable of interest (see *e.g.* Bai and Ng [2008] or Schumacher [2010], against the background of bridge equations augmented with dynamic factors). In this section we present the performance of our Ridge after model selection estimator for nowcasting GDP growth described in Section 2.2, compared with a standard Ridge regularization approach without any pre-selection.

The idea of the SIS pre-selection method is to identify *ex ante*, among the initial large dataset, the Google variables that have the highest absolute correlation with the targeted variable, namely the GDP growth rate. First, let us have a look at the relationship between the number of selected variables through the SIS procedure and the absolute correlation between each Google variable and the GDP growth rate at the same quarter. We recall that for the Google variables we take the average over each quarter, see Section 2.2. This relationship is described in Figure 4. We clearly observe an inverse non-linear relationship, with a kind of plateau starting from an absolute correlation of about 0.25. Indeed, most of Google variables present an absolute correlation with current GDP growth rate lower than 0.30. Thus it seems useful to only focus on a core dataset with the highest correlations.

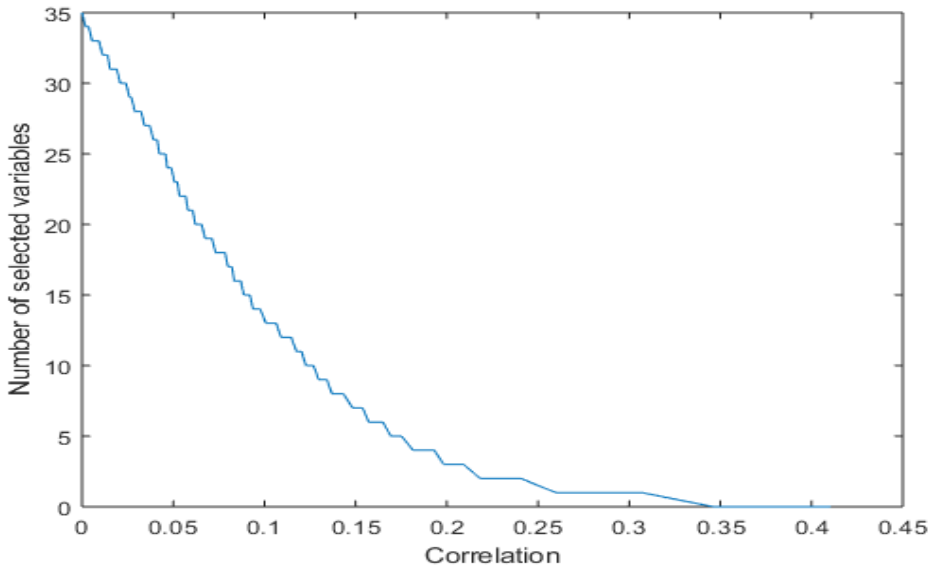

Figure 4: Plot of the number of selected Goggle variables by the SIS method versus the correlation with current GDP (computed for the first training sample).

We then analyse the nowcasting performances of bridge regressions that use the SIS pre-selection approach, as well as the one of bridge regressions that use the Lasso pre-selection approach, associated with Ridge regularization. Figure 5 presents the evolution over the 13 weeks of the quarter of RMSFEs stemming from bridge models estimated using Google search data and Ridge regression coupled or not with the SIS and Lasso pre-selection approaches. We clearly see that using a pre-selection approach (light and dark gray bars, similar to the gray bars in Figure 2) allows for an overall improvement in nowcasting accuracy. A striking result is that the RMSFE is lower for all the weeks when a pre-selection approach is used. Moreover, when pre-selection is implemented, RMSFEs evolve over the quarter in a more smoother way. For example, without any pre-selection, we observe that in week 6 the RMSFE jumps to 0.3829, from 0.3239 in week 5. Table 7 in the Annex reports the exact values of the RMSFEs with and without pre-selection. When comparing the two pre-selection approaches, namely SIS and Lasso, we observe that for the first weeks of the quarter SIS approach leads to lower RMSFEs, especially as regards the first two weeks. This result is noteworthy as, as pointed out in the previous sub-section, the first weeks are those of interest for the use of Google data. The overall gain underlines the need for pre-selecting data using a targeted approach.

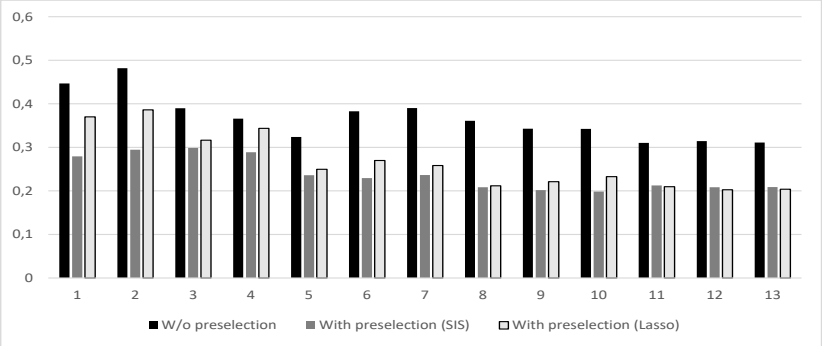

Figure 5: Pseudo-real-time: is it worth to preselect? Evolution over the 13 weeks of the quarter of the RMSFEs stemming from bridge models using Google data, S_t and IP_t estimated from Ridge regularization with and without SIS or Lasso pre-selection approaches.

5.1.3 A true real-time analysis

In this subsection, we carry a true real-time analysis by using vintages of data for GDP and industrial production[4] and by accounting for the observed timeline of data release as provided by Eurostat. As regards the dates of the GDP releases, there is a large heterogeneity from one period to the other. When available, we also include the lagged GDP growth among the explanatory variables of the nowcasting models. Figure 3 in the Annex gives the exact weeks in the out-of-sample period 2014q1-2016q1 where the lagged GDP growth is included in the real-time analysis.

In Figure 6 we show that pre-selecting Google data is still worth in real-time. Indeed, RMSFEs obtained from models integrating pre-selected Google data are systematically lower (light bars), for all weeks, than those obtained without any pre-selection (dark bars). The corresponding RMSFE values are reported in Table 8.

In Figure 7, we show the impact of Google search data on GDP growth nowcasting accuracy in the context of a true real-time nowcasting analysis. The corresponding RMSFE values are reported in Table 9 in the Annex. Similarly to the pseudo real-time exercise, we get that during the first 4 weeks of the quarters, when only Google information is available, RMSFEs are quite reasonable. This fact is reassuring about the reliability of the real-time use of Google search data when nowcasting GDP. However,

[4]Survey data are generally not revised.

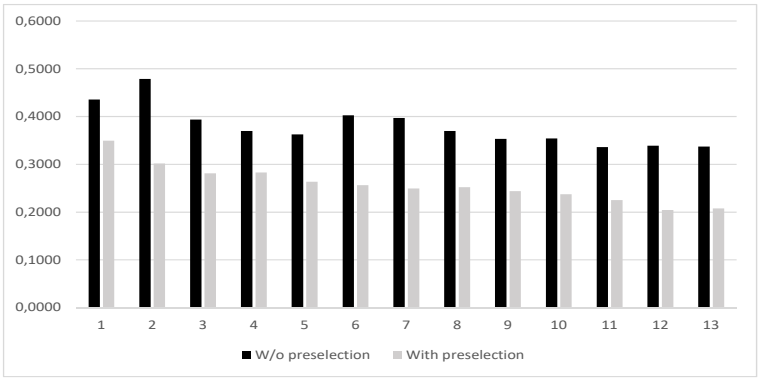

Figure 6: True real-time analysis: is it worth to preselect? Evolution over the 13 weeks of the quarter of the RMSFEs stemming from bridge models using Google data, S_t, IP_t, and lagged GDP growth estimated from Ridge regularization with and without SIS pre-selection approaches. The models include the lagged GDP growth when it is available.

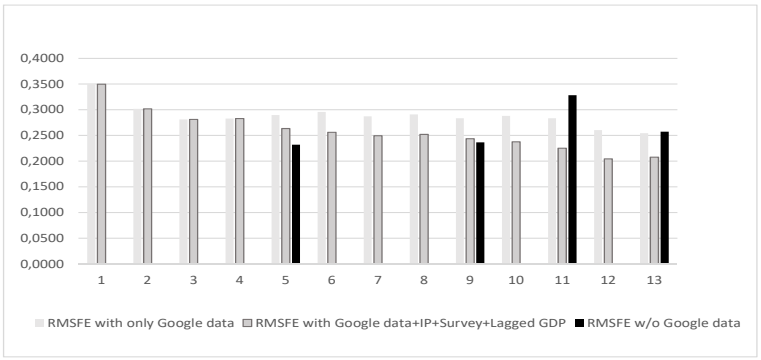

Figure 7: The importance of Google data. True Real-time analysis with pre-selection of Google data. RMSFE from: (i) models M1 - M13 with only variables extracted from Google data (in light gray), (ii) models M1 - M13 with all the variables (S_t, IP_t, $laggedGDP$ and Google data) (in gray), (iii) models $NoGoogle_1$ - $NoGoogle_4$ (in black).

starting from week 5, as soon as the first survey of quarter is released, the marginal gain of using Google data instantaneously vanishes. [5]

[5]There is an exception in week 11, where it is surprising to note that the integration of surveys, past GDP value and industrial production tend to suddenly increase the RMSFEs, in opposition to what can be expected from previous empirical results. This stylized has to be further explored.

Finally, in order to compare the results of the real-time analysis with the ones from the pseudo-real-time analysis, we compute GDP growth nowcasts without including the lagged GDP growth among the explanatory variables. The results are given in Figure 8. The corresponding RMSFE values are reported in Table 10. We see that both analyses lead to a similar shape in the evolution of RMSFEs within the quarter, although, as expected, the uncertainty around weekly nowcasts is a bit higher in real time.

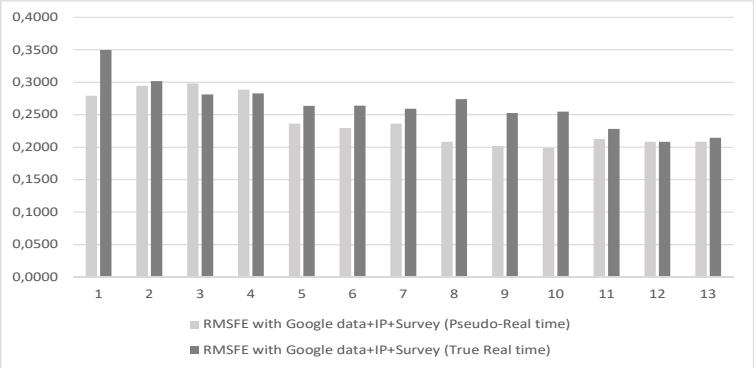

Figure 8: Pseudo-Real-time versus True Real-time analysis (with pre-selection). Comparison of RMSFEs within the quarter from pseudo-real-time (in light gray) and true real-time (in gray) analysis. The true real-time analysis does not include lagged GDP growth among the explanatory variables.

5.2 Are Google search data useful during recession periods?

As an additional result, we evaluate in this subsection to what extent Google search data are useful during recession periods in their ability to reduce the MSFEs of euro area GDP growth nowcasting. Especially, we focus on the *Great Recession* period from 2008q1 to 2009q2. All the euro area economies have been largely negatively affected by the adverse financial shock during this specific period of time. The research question for us is to check whether Google data do present a specific pattern during this major event, in spite of a relatively low number of quarters under consideration (6 quarters).

In this respect, we compute the RMSFEs for the euro area GDP growth, over the 13 weeks of each quarter, stemming from three various approaches: (i) Ridge regression with all available information (i.e.: Google data as well as lagged Survey, Survey, and IPI), (ii) Ridge regression with all information retained after the SIS pre-selection and (iii) Ridge regression with only Google data. Pre-selection of variables using the SIS is

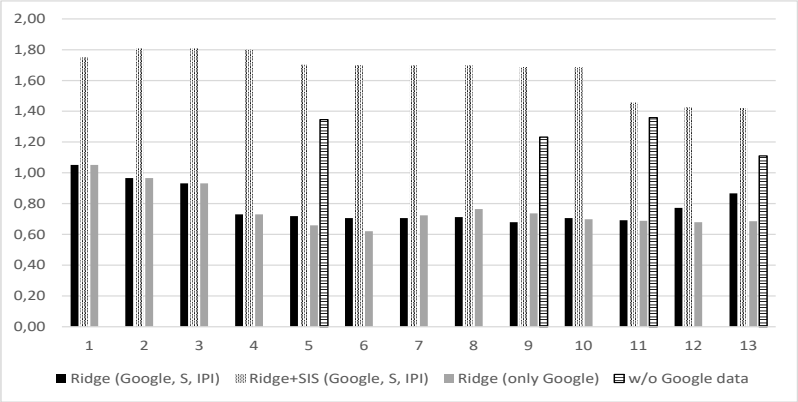

Figure 9: Nowcasting during recession periods. RMSFE from: (i) models M1 - M13 with Google data, Survey, and IPI without pre-selection (in black), (ii) models M1 - M13 with pre-selected Google data, Survey, and IPI (in dotted gray), (iii) models M1 - M13 with Google data only without pre-selection (in gray), (iv) models with only official variables $NoGoogle_1$ - $NoGoogle_4$ (in black lines).

specifically carried out for this recession.

Figure 9 and Table 11 in the Annex report the RMSFE for the three approaches that we consider in this exercise. As a first result, when comparing nowcasts obtained by using only Google data (in grey) and by using all available data (in dark), we observe relatively similar results on average over the 13 quarters. In particular, we see that the nowcast based on Google search data outperforms nowcast without Google data. A striking result is that Google-based nowcasting during the Great Recession performs better when data are not pre-selected using the SIS approach, in opposition to what has been observed in the previous analysis over the whole sample. Indeed, RMSFEs are about twice higher when we pre-select variables (light grey bars). This result suggests that during a recession phase, a broader information set is needed to adequately assess the state of the economy, while a core dataset is likely to be sufficient during expansions.

Leaving aside the relatively low number of quarters in this robustness check, we get the interesting result that pre-selecting data among a large database before entering nowcasting models does not appear to be an efficient strategy during recession periods, in opposition to expansions. We point out here an asymmetric result between expansion and recessions phases that would be worth to further study.

6 Conclusions

Large sets of alternative variables have gained in popularity among macroeconomists when trying to assess the current state of the economy on a high-frequency basis. However, the jury is still out as regards the marginal gain of those data when controlling for available official variables such as production, sale or opinion surveys.

In this paper, we ask the question whether, and when, a large set of Google variables can be useful to nowcast euro area GDP growth when controlling for official information conveyed by opinion surveys and industrial production. Because Google search data are high dimensional, in the sense that the number of variable is large compared to the time series dimension, there is a price to pay for using them: first, we need to reduce their dimensionality from ultra-high to high by using a screening procedure and, second, we need to use a regularized estimator to deal with the pre-selected variables. In this respect, we put forward a new approach combining variable pre-selection and Ridge regularization enabling to account for a large database. Especially, we implement the Sure Independent Screening approach put forward by Fan and Lv [2008] enabling to retain only the Google variables that are the most correlated with the targeted variable, that is GDP growth rate. We provide theoretical results on the asymptotic behaviour of the estimator against this background of this new combined approach.

Five salient facts emerge from our empirical analysis. First, against the background of a pseudo real-time analysis, we point out the usefulness of Google search data in nowcasting euro area GDP growth rate for the first four weeks of the quarter when there is no information about the state of the current quarter. We show that at the beginning of the quarter, Google data indeed provide an accurate picture of the GDP growth rate. Second, as soon as official data become available, that is starting from week 5 with the release of the first opinion survey of the quarter, then the relative nowcasting power of Google data rapidly vanishes. Third, we show that pre-selecting Google data before entering the nowcasting models appears to be a pertinent strategy in terms of nowcasting accuracy. This result confirms previous results obtained with bridge equations augmented with dynamic factor (see e.g. Bai and Ng [2008]). Fourth, we show that when using Google search data in the context of a true real-time analysis, the three previous salient facts remain valid. This result argues in favor of the use of Google search data at the beginning of the quarter, when there is no official information available about the current quarter, for real-time policy-making. Finally, we evaluate to what extent Google search data are useful during recession periods. We show that for the *Great Recession* period from 2008q1 to 2009q2 nowcasts that account for Google data information clearly outperform those only based on official data.

References

K. Aastveit and T. Trovik. Nowcasting norwegian GDP: the role of asset prices in a small open economy. *Empirical Economics*, 42(1):95–119, 2012.

E. Angelini, G. Camba-Mendez, D. Giannone, L. Reichlin, and G. Ruenstler. Short-term forecasts of euro area gdp growth. *Economic Journal*, 14:C25–C44, 2011.

J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304 – 317, 2008. Honoring the research contributions of Charles R. Nelson.

K. Barhoumi, O. Darne, and L. Ferrara. Are disaggregate data useful for forecasting french gdp with dynamic factor models ? *Journal of Forecasting*, 29(1-2):132–144, 2010.

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 05 2013.

J. Boivin and S. Ng. Are more data always better for factor analysis? *Journal of Econometrics*, 132:169–194, 2006.

M.-E. Bontempi, M. Frigeri, R. Golinelli, and M. Squadrini. Uncertainty, perception and internet. Technical report, 2018.

D. Bragoli. Nowcasting the japanese economy. *International Journal of Forecasting*, 33 (2):390–402, April 2017.

D. Bragoli, L. Metelli, and M. Modugno. The importance of updating: Evidence from a Brazilian nowcasting model. *OECD Journal: Journal of Business Cycle Measurement and Analysis*, 2015(1):5–22, 2015.

D. Buono, G. Kapetanios, M. Marcellino, G. L. Mazzi, and F. Papailias. Big data econometrics: Nowcasting and early estimates. Technical Report 82, Working Paper Series, Universita Bocconi, 2018.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 12 2007.

M. Carrasco and B. Rossi. In-sample inference and forecasting in misspecified factor models. *Journal of Business & Economic Statistics*, 34(3):313–338, 2016.

M. Carrasco, J.-P. Florens, and E. Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. volume 6, Part B of *Handbook of Econometrics*, pages 5633 – 5751. Elsevier, 2007.

Y. Carriere-Swallow and F. Labbe. Nowcasting with Google trends in an emerging market. *Journal of Forecasting*, 32(4):289–298, 2013.

H. Choi and H. Varian. Predicting initial claims for unemployment insurance using Google trends. *Google Technical Report*, 2009.

H. Choi and H. Varian. Predicting the present with google trends. *Google Technical Report*, 2012.

D. Coble and P. Pincheira. Nowcasting building permits with Google Trends. MPRA Paper 76514, University Library of Munich, Germany, 2017.

F. D'Amuri and J. Marcucci. The predictive power of Google searches in forecasting unemployment. *International Journal of Forecasting*, 33:801–816, 2017.

M. Diron. Short-term forecasts of euro area real gdp growth: An assesment of real-time performance based on vintage data. *Journal of Forecasting*, 27:371–390, 2008.

H. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Kluwer Academic, Dordrecht, 2000.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society B*, 70:849–911, 2008.

L. Ferrara and C. Marsilli. Nowcasting global economic growth: A factor-augmented mixed-frequency approach. *The World Economy*, 2018.

J.-P. Florens and A. Simoni. Regularizing priors for linear inverse problems. *Econometric Theory*, 32(1):71–121, 2016.

C. Frale, M. Marcellino, G. L. Mazzi, and T. Proietti. Euromind: a monthly indicator of euro area economic conditions. *Journal of the Royal Statistical Society, Series A*, 174(2):439–470, 2010.

D. Giannone, L. Reichlin, and D. Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676, May 2008.

D. Giannone, M. Lenza, and G. Primiceri. Economic predictions with big data: The illusion of sparsity. *mimeo*, 2017.

T. Goetz and T. Knetsch. Google data in bridge equation models for german gdp. *International Journal of Forecasting*, 35(1):45–66, 2019.

R. Golinelli and G. Parigi. Tracking world trade and GDP in real time. *International Journal of Forecasting*, 30(4):847–862, 2014.

V. Kuzin, M. Marcellino, and C. Schumacher. MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting*, 27(2):529–542, April 2011.

K.-C. Li. Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 14(3):1101–1112, 1986.

K.-C. Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15(3):958–975, 1987.

X. Li. Nowcasting with big data : Is Google useful in the presence of other information? *mimeo*, 2016.

H. Liu and B. Yu. Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression. *Electron. J. Statist.*, 7:3124–3169, 2013.

M. Modugno, B. Soybilgen, and E. Yazgan. Nowcasting Turkish GDP and news decomposition. *International Journal of Forecasting*, 32(4):1369–1384, 2016.

F. Narita and R. Yin. In search for information: Use of Google Trends' data to narrow information gaps for low-income developing countries. Technical Report WP/18/286, IMF Working Paper, 2018.

P. Nymand-Andersen and E. Pantelidis. Google econometrics: Nowcasting euro area car sales and big data quality requirements. Technical report, European Central Bank, 2018.

C. Schumacher. Factor forecasting using international targeted predictors: The case of German GDP. *Economics Letters*, 107(2):95–98, May 2010.

S. Scott and H. Varian. Bayesian variable selection for nowcasting economic time series. In A. Goldfarb, S. Greenstein, and C. Tucker, editors, *Economic Analysis of the Digital Economy*, pages 119–135. NBER, 2015.

H. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, 2014.

S. Vosen and T. Schmidt. Forecasting private consumption: Survey-based indicators vs. Google trends. *Journal of Forecasting*, 30(6):565–578, 2011.

A Proofs

B Proof of Theorem 3.1

Define the criterion $\widehat{Q}(\beta) := \frac{1}{T}\sum_{t=1}^{T}(y_t - \beta'_{\widehat{M}}X_t)^2 + \alpha\|\beta_{\widehat{M}}\|_2^2$. Hence by definition of the Ridge estimator after SIS selection:

$$\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_*) < 0. \tag{B.1}$$

The proof is made of two parts: in the first part we consider the event $M_* \subseteq \widehat{M}$ and in the second part we consider the complement event $M_* \supset \widehat{M}$. First, we consider the case $M_* \subseteq \widehat{M}$ with $\widehat{m} := |\widehat{M} \setminus M_*|$ and let $\widehat{\delta} := \widehat{\beta} - \beta_*$. This and (B.1) imply

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{\beta}'_{\widehat{M}}X_t)^2 - \frac{1}{T}\sum_{t=1}^{T}(y_t - \beta'_{*,\widehat{M}}X_t)^2 < \alpha\left(\|\beta_{*,\widehat{M}}\|_2^2 - \|\widehat{\beta}_{\widehat{M}}\|_2^2\right) \\
&\leq \alpha\left(\|\beta_{*,\widehat{M}}\|_2^2 - \|\beta_{*,\widehat{M}}\|_2^2 - \|\widehat{\delta}_{\widehat{M}}\|_2^2 - 2\beta'_{*,\widehat{M}}\widehat{\delta}_{\widehat{M}}\right) \leq \alpha\left(-\|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2\frac{1}{\phi(\widehat{m})} + 2\|\beta_*\|_2\,\|\widehat{\delta}_{\widehat{M}}\|_2\right)
\end{aligned}$$

by using the fact that $\|\widehat{\delta}_{\widehat{M}}\|_2^2 \geq \|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2 \min_{\|\delta_{M_*^c}\|_0 \leq \widehat{m}, \|\delta\|_0 \neq 0}\frac{\|\delta\|_2^2}{\|\delta\|_{2,T}^2} = \|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2/\phi(\widehat{m})$. Hence, since $M_* \subseteq \widehat{M}$: $(y_t - \beta'_{*,\widehat{M}}X_t) = (y_t - \beta'_*X_t) = \varepsilon_t$, and so we get

$$\|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2 - 2\frac{1}{T}\sum_{t=1}^{T}(y_t - \beta'_{*,\widehat{M}}X_t)X_t'\widehat{\delta}_{\widehat{M}} \leq \alpha\left(-\|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2\frac{1}{\phi(\widehat{m})} + 2\|\beta_*\|_2\,\|\widehat{\delta}_{\widehat{M}}\|_2\right)$$

and so,

$$\|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2\left(1+\frac{\alpha}{\phi(\widehat{m})}\right) \leq 2\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t X_t'\widehat{\delta}_{\widehat{M}} + 2\alpha\|\beta_*\|_2\,\|\widehat{\delta}_{\widehat{M}}\|_2$$

$$\leq 2\frac{1}{T}\sum_{t=1}^{T}\frac{\varepsilon_t X_t'\widehat{\delta}_{\widehat{M}}}{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}}\|\widehat{\delta}_{\widehat{M}}\|_{2,T} + 2\alpha\|\beta_*\|_2\,\|\widehat{\delta}_{\widehat{M}}\|_{2,T}\frac{1}{\kappa(\widehat{m})}$$

$$\leq 2\sup_{\|\delta_{M_*^c}\|\leq\widehat{m},\,\|\delta\|_{2,T}>0}\left|\frac{1}{T}\sum_{t=1}^{T}\frac{\varepsilon_t X_t'\delta}{\|\delta\|_{2,T}}\right|\|\widehat{\delta}_{\widehat{M}}\|_{2,T} + 2\alpha\|\beta_*\|_2\,\frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}}{\kappa(\widehat{m})}$$

$$\leq \sigma 4\sqrt{2}\frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}}{\sqrt{T}}\left(\sqrt{\log\binom{N}{\widehat{m}}} + \sqrt{(\widehat{m}+s^*)\log(D\mu(\widehat{m}))} + \sqrt{(\widehat{m}+s^*)\log(1/(\epsilon(1-1/e)e^{s^*}))}\right)$$

$$+ 2\alpha\|\beta_*\|_2\,\|\widehat{\delta}_{\widehat{M}}\|_{2,T}\frac{1}{\kappa(\widehat{m})} \quad \text{(B.2)}$$

for a universal constant $D \geq 1$, where the last inequality holds with probability at least $1-\epsilon$ and it follows by applying Belloni and Chernozhukov [2013, Lemma 5]. Moreover, by using the upper bound $\binom{N}{\widehat{m}} \leq N^{\widehat{m}}$ and by defining $\bar{D} := \max\{1, \log(D), \log(1/(\epsilon(1-1/e)e^{s^*}))\}$ we obtain

$$4\sqrt{2}\frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}}{\sqrt{T}}\left(\sqrt{\log\binom{N}{\widehat{m}}} + \sqrt{(\widehat{m}+s^*)\log(D\mu(\widehat{m}))} + \sqrt{(\widehat{m}+s^*)\log(1/(\epsilon(1-1/e)e^{s^*}))}\right)$$

$$\leq 4\sqrt{2}\frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}}{\sqrt{T}}\left(\sqrt{\widehat{m}\log(N)\bar{D}} + \sqrt{(\widehat{m}+s^*)\left(\bar{D}+\log(\mu(\widehat{m}))\right)} + \sqrt{(\widehat{m}+s^*)\bar{D}}\right)$$

$$= 4\sqrt{2}\sqrt{\bar{D}}\frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}}{\sqrt{T}}\left(\sqrt{\widehat{m}\log(N)} + \sqrt{(\widehat{m}+s^*)\log(e\mu(\widehat{m}))} + \sqrt{(\widehat{m}+s^*)}\right)$$

$$\leq 4\sqrt{6}\sqrt{\bar{D}}\frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}}{\sqrt{T}}\sqrt{\widehat{m}\log(N) + (\widehat{m}+s^*)\log(e^2\mu(\widehat{m}))} \quad \text{(B.3)}$$

where we have used the fact that $\bar{D} + \log(\mu(\widehat{m})) = \bar{D}(1+\log(\mu(\widehat{m}))) = \bar{D}\log(e\mu(\widehat{m}))$ and the Gibbs inequality for the concave function $\sqrt{\cdot}$. So, from (B.2) and (B.3) we have

$$\|\widehat{\delta}_{\widehat{M}}\|_{2,T} \leq K_\epsilon\sigma\sqrt{\frac{\widehat{m}\log(N) + (\widehat{m}+s^*)\log(e^2\mu(\widehat{m}))}{T}} + 2\alpha\|\beta_*\|_2\frac{1}{\kappa(\widehat{m})}$$

where $K_\epsilon := 4\sqrt{6\bar{D}}$. This gives the first part of the result of the theorem

Next, we consider the case $\widehat{M} \subseteq M_*$ with $\widehat{k} := |M_* \setminus \widehat{M}|$ and let $\widehat{\delta}_{\widehat{M}} := \widehat{\beta} - \beta_{*,\widehat{M}}$.

This and (B.1) imply (by using the identity $\widehat{\beta} = \widehat{\beta}_{\widehat{M}}$):

$$
\begin{aligned}
0 > \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_*) &= \frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{\beta}'_{\widehat{M}} X_t)^2 - \frac{1}{T}\sum_{t=1}^{T}(y_t - \beta'_{*,\widehat{M}} X_t)^2 + \alpha \left(\| \overbrace{\widehat{\beta}_{\widehat{M}}}^{=\widehat{\delta}_{\widehat{M}} + \beta_{*\widehat{M}}} \|^2 - \|\beta_{*,\widehat{M}}\|^2 \right) \\
&= \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2 + \frac{1}{T}\sum_{t=1}^{T}[(\beta_* - \widehat{\beta}_{\widehat{M}})' X_t]^2 + \frac{2}{T}\sum_{t=1}^{T}\varepsilon_t X_t'(\beta_* - \widehat{\beta}_{\widehat{M}}) - \frac{1}{T}\sum_{t=1}^{T}(\underbrace{y_t - \beta'_{*,\widehat{M}} X_t}_{=\varepsilon_t + (\beta_* - \beta_{*,\widehat{M}})' X_t})^2 + \alpha \|\widehat{\delta}_{\widehat{M}}\|^2 + 2\alpha \beta'_{*,\widehat{M}} \widehat{\delta}_{\widehat{M}} \\
&= \frac{1}{T}\sum_{t=1}^{T}[(\beta_* - \widehat{\beta}_{\widehat{M}})' X_t]^2 + \frac{2}{T}\sum_{t=1}^{T}\varepsilon_t X_t'(\beta_* - \widehat{\beta}_{\widehat{M}}) - \frac{1}{T}\sum_{t=1}^{T}(\beta'_{*,M_*\setminus\widehat{M}} X_t)^2 - \frac{2}{T}\sum_{t=1}^{T}\varepsilon_t X_t' \beta_{*,M_*\setminus\widehat{M}} \\
&\qquad + \alpha \|\widehat{\delta}_{\widehat{M}}\|^2 + 2\alpha \beta'_{*,\widehat{M}} \widehat{\delta}_{\widehat{M}}.
\end{aligned}
$$

Remark that $\|(\widehat{\delta}_{\widehat{M}})_{M_*^c}\|_0 \leq m$ for every $m \geq 0$, then $\|\widehat{\delta}_{\widehat{M}}\|^2 = \frac{\|\widehat{\delta}_{\widehat{M}}\|^2}{\phi(m)}\phi(m) = \frac{\|\widehat{\delta}_{\widehat{M}}\|^2}{\phi(m)} \max_{\|\delta_{M_*^c}\|_0 \leq m, \delta \neq 0} \frac{\|\delta\|_{2,T}^2}{\|\delta\|^2} \geq \frac{\|\widehat{\delta}_{\widehat{M}}\|^2}{\phi(m)} \frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2}{\|\widehat{\delta}_{\widehat{M}}\|^2} = \|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2/\phi(m)$ for every $m \geq 0$ and that $2\alpha\beta'_{*,\widehat{M}}\widehat{\delta}_{\widehat{M}} \geq -2\alpha\|\beta_{*,\widehat{M}}\|_2 \|\widehat{\delta}_{\widehat{M}}\|_2 \geq -2\alpha\|\beta_*\|_2 \|\widehat{\delta}_{\widehat{M}}\|_{2,T}/\kappa(m)$ for every $m \geq 0$ by the Cauchy-Schwartz inequality and by $\|\widehat{\delta}_{\widehat{M}}\|_2^2 = \frac{\|\widehat{\delta}_{\widehat{M}}\|^2}{\kappa(m)^2}\kappa(m)^2 = \frac{\|\widehat{\delta}_{\widehat{M}}\|^2}{\kappa(m)^2} \min_{\|\delta_{M_*^c}\|_0 \leq m, \delta \neq 0} \frac{\|\delta\|_{2,T}^2}{\|\delta\|^2} \leq \frac{\|\widehat{\delta}_{\widehat{M}}\|^2}{\kappa(m)^2} \frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2}{\|\widehat{\delta}_{\widehat{M}}\|_2^2} = \|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2/\kappa(m)^2$ for every $m \geq 0$. Therefore,

$$
\begin{aligned}
0 > \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_*) &\geq \|\widehat{\beta}_{\widehat{M}} - \beta_*\|_{2,T}^2 + \frac{2}{T}\sum_{t=1}^{T}\varepsilon_t X_t'(\beta_* - \widehat{\beta}_{\widehat{M}}) - \|\beta_{*,M_*\setminus\widehat{M}}\|_{2,T}^2 - \frac{2}{T}\sum_{t=1}^{T}\varepsilon_t X_t' \beta_{*,M_*\setminus\widehat{M}} \\
&\qquad + \frac{\alpha}{\phi(0)}\|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2 - \frac{2\alpha}{\kappa(0)}\|\beta_{*,}\|_2 \|\widehat{\delta}_{\widehat{M}}\|_{2,T} \\
&= \|\widehat{\beta}_{\widehat{M}} - \beta_*\|_{2,T}^2 - \frac{2}{T}\sum_{t=1}^{T}\varepsilon_t X_t'(\widehat{\beta}_{\widehat{M}} - \beta_{*,\widehat{M}}) - \|\beta_{*,M_*\setminus\widehat{M}}\|_{2,T}^2 + \underbrace{\frac{\alpha}{\phi(0)}\|\widehat{\delta}_{\widehat{M}}\|_{2,T}^2}_{\geq 0} - \frac{2\alpha}{\kappa(0)}\|\beta_{*,}\|_2 \|\widehat{\delta}_{\widehat{M}}\|_{2,T}.
\end{aligned}
\tag{B.4}
$$

By applying Belloni and Chernozhukov [2013, Lemma 5] (with in their notation $s = 0$, $p = s^*$, $\widetilde{T} = M_*$, $T = \widehat{M}$ and $m = \widehat{k}$) we get that with probability at least $1 - \epsilon$: $\forall \widehat{k} \leq T$

and a universal constant $D \geq 1$,

$$\begin{aligned}
\frac{2}{T}\sum_{t=1}^{T}\frac{\varepsilon_t X_t'(\widehat{\beta}_{\widehat{M}}-\beta_{*,\widehat{M}})}{\|\widehat{\beta}_{\widehat{M}}-\beta_{*,\widehat{M}}\|_{2,T}}\|\widehat{\beta}_{\widehat{M}}-\beta_{*,\widehat{M}}\|_{2,T} &\leq 2\sup_{\|\delta_{\widehat{M}^c}\|\leq\widehat{k},\,\|\delta\|_{2,T}>0}\left|\frac{1}{T}\sum_{t=1}^{T}\frac{\varepsilon_t X_t'\delta}{\|\delta\|_{2,T}}\right|\|\widehat{\beta}_{\widehat{M}}-\beta_{*,\widehat{M}}\|_{2,T}\\
&\leq \sigma 4\sqrt{2}\frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}}{\sqrt{T}}\left(\sqrt{\log\binom{s^*}{\widehat{k}}}+\sqrt{\widehat{k}\log(D\mu(0))}+\sqrt{\widehat{m}\log(1/(\epsilon(1-1/e)))}\right)\\
&\leq \sigma 4\sqrt{2}\frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}}{\sqrt{T}}\left(\sqrt{\widehat{k}\log(s^*)\bar{D}}+\sqrt{\widehat{k}\bar{D}\,(1+\log(\mu(0)))}+\sqrt{\widehat{k}\bar{D}}\right)\\
&= \sigma 4\sqrt{2\bar{D}}\frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}}{\sqrt{T}}\left(\sqrt{\widehat{k}\log(s^*)}+\sqrt{\widehat{k}\log(e\mu(0))}+\sqrt{\widehat{k}}\right)\\
&\leq \sigma 4\sqrt{6\bar{D}}\frac{\|\widehat{\delta}_{\widehat{M}}\|_{2,T}}{\sqrt{T}}\sqrt{\widehat{k}\log(s^*)+\widehat{k}\log(e^2\mu(0))}, \quad \text{(B.5)}
\end{aligned}$$

where we have used the upper bound $\binom{s^*}{\widehat{k}} \leq (s^*)^{\widehat{k}}$, the Gibbs inequality for the concave function $\sqrt{\cdot}$, and by defining $\bar{D} := \max\{1, \log(D), \log(1/(\epsilon(1-1/e)))\}$. So, by (B.4) - (B.5), by denoting $K_\epsilon := 4\sqrt{6}\bar{D}$ and by remarking that $\|\widehat{\delta}_{\widehat{M}}\|_{2,T} \leq \|\widehat{\delta}\|_{2,T} := \|\widehat{\beta}_{\widehat{M}}-\beta_*\|_{2,T}$ we obtain

$$0 > \|\widehat{\delta}\|_{2,T}^2 - \sigma K_\epsilon\frac{\|\widehat{\delta}\|_{2,T}}{\sqrt{T}}\sqrt{\widehat{k}\log(s^*)+\widehat{k}\log(e^2\mu(0))} - \|\beta_{*,M_*\setminus\widehat{M}}\|_{2,T}^2 - \frac{2\alpha}{\kappa(0)}\|\beta_*\|_2\,\|\widehat{\delta}\|_{2,T}$$

which is a second degree inequality in $\|\widehat{\delta}\|_{2,T}^2 = \|\widehat{\beta}_{\widehat{M}}-\beta_*\|_{2,T}$ and which gives

$$\|\widehat{\beta}_{\widehat{M}}-\beta_*\|_{2,T} \leq \frac{K_\epsilon\sigma}{\sqrt{T}}\sqrt{\widehat{k}\log(s^*)+\widehat{k}\log(e^2\mu(0))} + \frac{2\alpha}{\kappa(0)}\|\beta_*\|_2 + \|\beta_{*,M_*\setminus\widehat{M}}\|_{2,T}$$

which gives the second part of the result of the theorem.

C Proof of Corollary 3.1

Remark that, since $\|(\widehat{\beta}-\beta_*)_{M^{*c}}\|_0$ is equal to zero if $\widehat{M} \subset M^*$ and is upper bounded by $\widehat{m}$ if $\widehat{M} \supseteq M^*$, we have

$$\begin{aligned}
\|\widehat{\beta}-\beta_*\|_{2,T}^2 = \frac{\|\widehat{\beta}-\beta_*\|_{2,T}^2}{\|\widehat{\beta}-\beta_*\|_2^2}\|\widehat{\beta}-\beta_*\|_2^2 &\\
\geq \min_{\|\delta_{M^{*c}}\|_0\leq\widehat{m},\,\delta\neq 0}\frac{\|\delta\|_{2,T}^2}{\|\delta\|_2^2}\|\widehat{\beta}-\beta_*\|_2^2 &= \kappa(\widehat{m})^2\|\widehat{\beta}-\beta_*\|_2^2.
\end{aligned}$$

The result follows from this and the result of Theorem 3.1.

D Proof of Corollary 3.2

Let $C > 0$ be the constant as in the statement of the corollary and denote

$$\eta := \left(K_\epsilon \sqrt{\frac{\widehat{m}\log(N) + (\widehat{m}+s^*)\log(e^2\mu(\widehat{m}))}{T\kappa(\widehat{m})^2}} + 2\alpha\|\beta_*\|_2 \frac{1}{\kappa(\widehat{m})^2} \right) \mathbb{1}\{\widehat{M} \supseteq M^*\}$$
$$+ \left(\frac{K_\epsilon \sigma}{\kappa(\widehat{m})\sqrt{T}} \sqrt{\widehat{k}\log(s^*) + \widehat{k}\log(e^2\mu(0))} + \frac{2\alpha}{\kappa(\widehat{m})\kappa(0)}\|\beta_*\|_2 + \frac{1}{\kappa(\widehat{m})}\|\beta_{*,M^*\setminus\widehat{M}}\|_{2,T} \right) \mathbb{1}\{\widehat{M} \subset M^*\},$$

and $\widetilde{\eta} := \sqrt{(\widehat{m}+s^*)}\, C\eta$ where C is the constant in the statement of the theorem. Since $(\widehat{\beta}-\beta_*)_j = 0$ for every $j > \widehat{m}+s^*$, by the Cauchy-Schwartz inequality we have:

$$X_\tau'(\widehat{\beta}-\beta_*) = \sum_{j=1}^{\widehat{m}+s^*} X_{\tau,j}(\widehat{\beta}-\beta_*)_j \le \left(\sum_{j=1}^{\widehat{m}+s^*} X_{\tau,j}^2 \right)^{1/2} \|\widehat{\beta}-\beta_*\|_2$$
$$\le \sqrt{(\widehat{m}+s_*)}\, C\|\widehat{\beta}-\beta_*\|_2 \quad \text{(D.1)}$$

by using the assumption in the corollary. Therefore,

$$P_{X_\tau}\left(X_\tau'(\widehat{\beta}-\beta_*) \le \widetilde{\eta} \right) \ge P\left(\left(\sum_{j=1}^{\widehat{m}+s^*} X_{\tau,j}^2 \right)^{1/2} \|\widehat{\beta}-\beta_*\|_2 \le \widetilde{\eta} \,\middle|\, X_\tau \right)$$
$$\ge P\left(\sqrt{(\widehat{m}+s^*)}C\|\widehat{\beta}-\beta_*\|_2 \le \widetilde{\eta} \,\middle|\, X_\tau \right)$$
$$= P\left(\|\widehat{\beta}-\beta_*\|_2 \le \eta | X_\tau \right) \ge (1-\epsilon)$$

by Corollary 3.1.

Annex

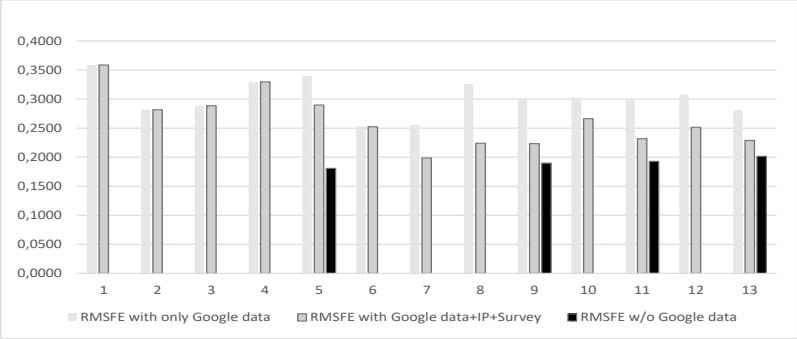

Figure 10: The importance of Google data. Pseudo-real-time analysis with pre-selection of Google data by Lasso where λ is chosen such that the number of selected Google search categories is less or equal to the number of categories selected with the SIS procedure. RMSFEs from: (i) models M1 - M13 with only variables extracted from Google data (in light gray), (ii) models M1 - M13 with all the variables (S_t, IP_t and Google data) (in gray), (iii) models with only official variables $NoGoogle_1$ - $NoGoogle_4$ (in black).

Model	Equation	Predictors
M1	$Y_t = \beta_{0,1} + \beta'_{g,1} x^{(1)}_{t,g} + \epsilon^{(1)}_t$	$x^{(1)}_{t,g} = x_{t,g,(1)}$
M2	$Y_t = \beta_{0,2} + \beta'_{g,2} x^{(2)}_{t,g} + \epsilon^{(2)}_t$	$x^{(2)}_{t,g} = \frac{\sum_{v=1}^{2} x_{t,g,(v)}}{2}$
M3	$Y_t = \beta_{0,3} + \beta'_{g,3} x^{(3)}_{t,g} + \epsilon^{(3)}_t$	$x^{(3)}_{t,g} = \frac{\sum_{v=1}^{3} x_{t,g,(v)}}{3}$
M4	$Y_t = \beta_{0,4} + \beta'_{g,4} x^{(4)}_{t,g} + \epsilon^{(4)}_t$	$x^{(4)}_{t,g} = \frac{\sum_{v=1}^{4} x_{t,g,(v)}}{4}$
M5	$Y_t = \beta_{0,5} + \beta_{s,5} x^{(5)}_{t,s} + \beta'_{g,5} x^{(5)}_{t,g} + \epsilon^{(5)}_t$	$x^{(5)}_{t,g} = \frac{\sum_{v=1}^{5} x_{t,g,(v)}}{5}$, $x^{(5)}_{t,s} = S_{t,1}$
M6	$Y_t = \beta_{0,6} + \beta_{s,6} x^{(6)}_{t,s} + \beta'_{g,6} x^{(6)}_{t,g} + \epsilon^{(6)}_t$	$x^{(6)}_{t,g} = \frac{\sum_{v=1}^{6} x_{t,g,(v)}}{6}$, $x^{(6)}_{t,s} = S_{t,1}$
M7	$Y_t = \beta_{0,7} + \beta_{s,7} x^{(7)}_{t,s} + \beta'_{g,7} x^{(7)}_{t,g} + \epsilon^{(7)}_t$	$x^{(7)}_{t,g} = \frac{\sum_{v=1}^{7} x_{t,g,(v)}}{7}$, $x^{(7)}_{t,s} = S_{t,1}$
M8	$Y_t = \beta_{0,8} + \beta_{s,8} x^{(8)}_{t,s} + \beta'_{g,8} x^{(8)}_{t,g} + \epsilon^{(8)}_t$	$x^{(8)}_{t,g} = \frac{\sum_{v=1}^{8} x_{t,g,(v)}}{8}$, $x^{(8)}_{t,s} = S_{t,1}$
M9	$Y_t = \beta_{0,9} + \beta_{s,9} x^{(9)}_{t,s} + \beta'_{g,9} x^{(9)}_{t,g} + \epsilon^{(9)}_t$	$x^{(9)}_{t,g} = \frac{\sum_{v=1}^{9} x_{t,g,(v)}}{9}$, $x^{(9)}_{t,s} = \frac{S_{t,1}+S_{t,2}}{2}$
M10	$Y_t = \beta_{0,10} + \beta_{s,10} x^{(10)}_{t,s} + \beta'_{g,10} x^{(10)}_{t,g} + \epsilon^{(10)}_t$	$x^{(10)}_{t,g} = \frac{\sum_{v=1}^{10} x_{t,g,(v)}}{10}$, $x^{(10)}_{t,s} = \frac{S_{t,1}+S_{t,2}}{2}$
M11	$Y_t = \beta_{0,11} + \beta_{s,11} x^{(11)}_{t,s} + \beta_{h,11} x^{(11)}_{t,h} + \beta'_{g,11} x^{(11)}_{t,g} + \epsilon^{(11)}_t$	$x^{(11)}_{t,g} = \frac{\sum_{v=1}^{11} x_{t,g,(v)}}{11}$, $x^{(11)}_{t,s} = \frac{S_{t,1}+S_{t,2}}{2}$, $x^{(11)}_{t,h} = IP_{t,1}$
M12	$Y_t = \beta_{0,12} + \beta_{s,12} x^{(12)}_{t,s} + \beta_{h,12} x^{(12)}_{t,h} + \beta'_{g,12} x^{(12)}_{t,g} + \epsilon^{(12)}_t$	$x^{(12)}_{t,g} = \frac{\sum_{v=1}^{12} x_{t,g,(v)}}{12}$, $x^{(12)}_{t,s} = \frac{S_{t,1}+S_{t,2}}{2}$, $x^{(12)}_{t,h} = IP_{t,1}$
M13	$Y_t = \beta_{0,13} + \beta_{s,13} x^{(13)}_{t,s} + \beta_{h,13} x^{(13)}_{t,h} + \beta'_{g,13} x^{(13)}_{t,g} + \epsilon^{(13)}_t$	$x^{(13)}_{t,g} = \frac{\sum_{v=1}^{13} x_{t,g,(v)}}{13}$, $x^{(13)}_{t,s} = \frac{\sum_{i=1}^{3} S_{t,i}}{3}$, $x^{(13)}_{t,h} = IP_{t,1}$

Table 1: Equations of the 13 models ($M1, \ldots, M13$) corresponding to (2.2) used to nowcast GDP growth over each quarter. Equations include the variables pre-selected from Google data as well as information stemming from surveys (S_t) and industrial production (IP_t). $S_{t,i}$ denotes the variable surveys S_t referring to the i-th month of the current-quarter t and $IP_{t,i}$ denotes the growth rate of the industrial production available at the 11^{th} week of the current-quarter t and referring to the i-th month of the current-quarter t.

Model	Equation	Predictors
$NoGoogle_1$	$Y_t = \beta_{0,1} + \beta_{s,1} x_{t,s}^{(1)} + \epsilon_t$	$x_{t,s}^{(1)} = S_{t,1}$
$NoGoogle_2$	$Y_t = \beta_{0,2} + \beta_{s,2} x_{t,s}^{(2)} + \epsilon_t$	$x_{t,s}^{(2)} = \frac{S_{t,1}+S_{t,2}}{2}$
$NoGoogle_3$	$Y_t = \beta_{0,3} + \beta_{s,3} x_{t,s}^{(3)} + \beta_{h,3} x_{t,h}^{(3)} + \epsilon_t$	$x_{t,s}^{(3)} = \frac{S_{t,1}+S_{t,2}}{2}$, $x_{t,h}^{(3)} = IP_{t,1}$
$NoGoogle_4$	$Y_t = \beta_{0,4} + \beta_{s,4} x_{t,s}^{(4)} + \beta_{h,4} x_{t,h}^{(3)} + \epsilon_t$	$x_{t,s}^{(4)} = \frac{S_{t,1}+...+S_{t,3}}{3}$, $x_{t,h}^{(4)} = IP_{t,1}$

Table 2: Equations of the four models used to nowcast GDP growth without the variables extracted from Google data. $x_{t,g,w}$ denotes the Google variable available at week w of period t not averaged. $S_{t,i}$ denotes the variable surveys S_t referring to the i-th month of the current-quarter t and $IP_{t,i}$ denotes the growth rate of the industrial production available at the 11^{th} week of the current-quarter t and referring to the i-th month of the current-quarter t.

2 lags between estimation period and forecasting period

Last Training Period	Nowcasting Period	1st GDP Vintage which contains the last GDP in the Training sample	Lagged GDP	week of new Vintage
2013Q3	2014Q1	08/04/2014	no	
2013Q4	2014Q2	08/04/2014	no	
		08/04/2014	no	
		15/04/2014	no	3rd week
		04/06/2014	yes	10th week
2014Q1	2014Q3	02/07/2014	no	
2014Q2	2014Q4	01/10/2014	no	
		21/10/2014	no	4th week
		14/11/2014	yes	7th week
		09/12/2014	yes	11th week
2014Q3	2015Q1	09/12/2014	no	
		17/03/2015	yes	12th week
2014Q4	2015Q2	17/03/2015	no	
		02/06/2015	yes	10th week
2015Q1	2015Q3	02/06/2015	no	
		30/07/2015	no	4th week
		09/09/2015	yes	11th week
		24/09/2015	yes	13th week
2015Q2	2015Q4	24/09/2015	no	
		13/11/2015	yes	7th week
		08/12/2015	yes	11th week
2015Q3	2016Q1	08/12/2015	no	
		12/02/2016	yes	6th week
		16/02/2016	yes	7th week
		08/03/2016	yes	10th week

Table 3: Timeline of GDP release in real-time within the quarter. The first column gives the last period used for the in-sample analysis (training sample), the second column indicates the nowcasting period, the third column indicates the date of the first vintage which contains the GDP growth in the last period of the training sample (indicated in the first column), the fourth columns indicates whether a lagged GDP growth is available to be included among the explanatory variables (the corresponding date and week of availability are given in the third and fifth columns, respectively). Finally, the fifth column gives the week, and so the model, corresponding to the date in the third column.

Pseudo real time: The importance of Google data (with SIS preselection)													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Google+S													
+IP	0.2793	0.2945	0.2985	0.2887	0.2361	0.2296	0.2362	0.2083	0.2019	0.1985	0.2127	0.2082	0.2086
Google	0.2793	0.2945	0.2985	0.2887	0.2887	0.2861	0.2993	0.2811	0.2929	0.2894	0.2658	0.2779	0.2612
No Google					0.1807				0.1897		0.1928		0.2017

Table 4: RMSFE corresponding to Figure 2. "Google+S+IP" refers to models M1 - M13 with all the variables: S_t, IP_t and Google data, "Google" refers to models M1 - M13 with only variables extracted from Google data, "No Google" refers to models $NoGoogle_1$ - $NoGoogle_4$.

Pseudo real time: The importance of Google data (with Lasso preselection)													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Google+S													
+IP	0.3702	0.3859	0.3164	0.3435	0.2498	0.2696	0.2581	0.2118	0.2211	0.2329	0.2097	0.2026	0.2038
Google	0.3702	0.3859	0.3164	0.3435	0.3093	0.4094	0.3351	0.2851	0.3352	0.3454	0.3412	0.3575	0.2739
No Google					0.1807				0.1897		0.1928		0.2017

Table 5: RMSFE corresponding to Figure 3. "Google+S+IP" refers to models M1 - M13 with all the variables: S_t, IP_t and Google data, "Google" refers to models M1 - M13 with only variables extracted from Google data, "No Google" refers to models $NoGoogle_1$ - $NoGoogle_4$.

Pseudo real time: The importance of Google data (with Lasso preselection and constrained number of Google categories)													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Google+S													
+IP	0.3586	0.2817	0.2886	0.3295	0.2898	0.2524	0.1984	0.2239	0.2232	0.2660	0.2319	0.2515	0.2289
Google	0.3586	0.2817	0.2886	0.3295	0.3396	0.2533	0.2556	0.3258	0.3002	0.3026	0.2978	0.3072	0.2807
No Google					0.1807				0.1897		0.1928		0.2017

Table 6: RMSFE corresponding to Figure 10. "Google+S+IP" refers to models M1 - M13 with all the variables: S_t, IP_t and Google data, "Google" refers to models M1 - M13 with only variables extracted from Google data, "No Google" refers to models $NoGoogle_1$ - $NoGoogle_4$.

Pseudo real time: is it worth to preselect?													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge+SIS	0.2793	0.2945	0.2985	0.2887	0.2361	0.2296	0.2362	0.2083	0.2019	0.1985	0.2127	0.2082	0.2086
Ridge+Lasso	0.3702	0.3859	0.3164	0.3435	0.2498	0.2696	0.2581	0.2118	0.2211	0.2329	0.2097	0.2026	0.2038
Ridge	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

Table 7: RMSFE corresponding to Figure 5. "Ridge+SIS" refers to model (2.2) estimated with pre-selected variables from Google data and Ridge regularization, "Ridge" refers to model (2.2) estimated with Ridge regularization without pre-selection, "Ridge+Lasso" refers to model (2.2) estimated with pre-selected variables from Google data and Ridge regularization.

True real time: is it worth to preselect?													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge + SIS	0.3496	0.3018	0.2812	0.2828	0.2636	0.2561	0.2494	0.252	0.2438	0.2375	0.2252	0.2045	0.2078
Ridge	0.4357	0.4785	0.3935	0.37	0.3628	0.4025	0.397	0.37	0.3535	0.354	0.336	0.3391	0.3372

Table 8: RMSFE corresponding to Figure 6. "Ridge + SIS" refers to model (2.2) estimated with pre-selected variables from Google data and Ridge regularization by including the lagged GDP, "Ridge" refers to model (2.2) estimated with Ridge regularization without pre-selection.

True real time: The importance of Google data (with SIS preselection)													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Google+S													
+IP+GDPlag	0.3496	0.3018	0.2812	0.2828	0.2636	0.2561	0.2494	0.252	0.2438	0.2375	0.2252	0.2045	0.2078
Google	0.3496	0.3018	0.2812	0.2828	0.2896	0.2955	0.2873	0.291	0.2836	0.2881	0.2837	0.2604	0.2544
No Google					0.2320				0.2365		0.3283		0.2576

Table 9: RMSFE corresponding to Figure 7. "Google+S+IP+GDPlag" refers to model (2.2) estimated with pre-selected variables by considering all the variables: S_t, IP_t, Google data and lagged GDP, "Google" refers to model (2.2) with only variables extracted from Google data, "No Google" refers to models $NoGoogle_1$ - $NoGoogle_4$.

Pseudo real time vs. True real time (with SIS preselection)													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Pseudo real time	0.2793	0.2945	0.2985	0.2887	0.2361	0.2296	0.2362	0.2083	0.2019	0.1985	0.2127	0.2082	0.2086
True real time (w/o GDPlag)	0.3496	0.3016	0.2813	0.2827	0.2636	0.264	0.2593	0.274	0.2525	0.2549	0.2281	0.2082	0.2146

Table 10: RMSFE corresponding to Figure 8. "Ridge + SIS" refers to model (2.2) estimated with pre-selected variables from Google data and Ridge regularization, "Ridge" refers to model (2.2) estimated with Ridge regularization without pre-selection.

Nowcasting during recession periods													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	1.0521	0.9656	0.9322	0.7299	0.7192	0.7058	0.7064	0.7122	0.6792	0.7066	0.6918	0.7720	0.8670
Ridge+SIS (Google+ S+IP)	1.7507	1.8090	1.8083	1.7993	1.7027	1.6995	1.6993	1.6994	1.6875	1.6867	1.4567	1.4265	1.4210
Ridge (Google)	1.0521	0.9656	0.9322	0.7299	0.6589	0.6204	0.7245	0.7653	0.7363	0.6982	0.6887	0.6791	0.6852
No Google					1.3459				1.2322		1.3581		1.1103

Table 11: RMSFE corresponding to Figure 9. "Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated without pre-selection, "Ridge+SIS (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated with SIS pre-selection, "Ridge (Google)" refers to model (2.2) with only Google data estimated without pre-selection, "No Google" refers to models $NoGoogle_1$ - $NoGoogle_4$.