When are Google data useful to nowcast GDP?

An approach via preselection and shrinkage[1]

Laurent Ferrara* Anna Simoni[2]

SKEMA Business School CREST, CNRS

Abstract

Alternative data sets are nowadays widely used for macroeconomic nowcasting together with new Machine Learning-based tools which often are applied without having a complete picture of their theoretical nowcasting properties. Against this background, this paper proposes a theoretically-funded nowcasting methodology allowing to incorporate alternative Google Search Data (GSD) among the predictors and combining targeted preselection, Ridge regularization and Generalized Cross Validation. Breaking with most of the existing literature that focuses on asymptotic in-sample theoretical properties, we establish the theoretical out-of-sample properties of our methodology, that are supported by Monte-Carlo simulations. We apply our methodology to GSD in order to nowcast GDP growth rate of different countries during various economic periods. Our empirical findings support the idea that GSD tend to increase nowcasting accuracy, even after controlling for official variables, but that the gain differs between periods of recessions and of macroeconomic stability.

Keywords: Nowcasting, Big data, Google search data, Targeted preselection, 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*, EDHEC, Brunel University, 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 CAMA, Australian National University, e-mail: laurent.ferrara@skema.edu

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

1 Introduction

Large sets of alternative data – such as web scraped data, Google data (Trends, Correlate or Search), scanner data or satellite data – are now widely used by practitioners for short-term macroeconomic forecasting and nowcasting purposes, see *e.g.* Ng [2017]. The main research questions related to alternative datasets are: (A) when such data improve nowcasting accuracy, and (B) whether they are useful even after controlling for official variables, such as opinion surveys or production, generally used by forecasters. Answering these questions requires the use of appropriate Machine Learning/econometrics methods. Recent macroeconomic empirical literature has seen an explosion of various methods to account for the specificity of alternative data, but for most of them we do not know their out-of-sample (OOS hereafter) theoretical properties, which matter the most for forecasting purposes. This paper puts forward a new methodology to deal with Google Search data (GSD in the remaining) for nowcasting purposes and establishes OOS large-sample properties for the proposed method. The challenging feature of GSD as a whole is their high dimension. GSD differ from Google Trends mainly because GSD are volume variations of Google queries with respect to the first value while Google Trends provides the ratio between the search shares for a particular keyword/category over a given sub-period and the maximum search share for the same keyword/category over a chosen larger period. Macroeconomic nowcast based on GSD have been proposed in *e.g.* Goetz and Knetsch [2019] while Google Trends have been used *e.g.* in Choi and Varian [2009, 2012], Scott and Varian [2015], Vosen and Schmidt [2011], D'Amuri and Marcucci [2017], Niesert et al. [2020] and references therein. Contrarily to ours, all these papers do not include theoretical contributions.

Our proposed nowcasting method, which we call *Ridge after Model Selection*, is a two-step approach: (i) first, GSD variables are preselected, conditionally on the official variables, by targeting the macroeconomic aggregate to be nowcast, and (ii) second, a Ridge regularization is applied to those preselected GSD and official variables. The

Ridge tuning parameter is chosen by Generalized Cross Validation (GCV, in the remaining). The main theoretical contributions of this paper are threefold. First, we prove that our targeted preselection retains all the variables in the true model with probability approaching one. Second, we establish an upper bound for both the in-sample and OOS prediction error associated with the Ridge after model selection estimator. Third, we evaluate the OOS performance of GCV. To the best of our knowledge, previous literature has established in-sample optimality of the GCV in the setting of Ridge regularization but not OOS optimality, see *e.g.* Li [1986] and Carrasco and Rossi [2016].

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, which come out with a delay.[1]

To clarify the presentation of our nowcasting methodology, suppose the variable one wants to nowcast is the quarterly GDP growth rate. The predictors are made of two subsets: a set of official variables and a set of the weekly GSD variables. Our nowcasting model is based on linear regression models that incorporate predictors sampled over different frequencies (*e.g.* monthly and weekly) and released with various reporting lags so that the relevant information set for calculating the nowcast evolves within the quarter. To explicitly account for the information set available at a specific time within the quarter a forecaster will consider a different set of predictors for each week which is the frequency of the higher-frequency variable considered in our example.

To estimate the models we use our *Ridge after Model Selection* procedure described

[1]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. See the websites https://www.newyorkfed.org/research/policy/nowcast and https://www.frbatlanta.org/cqer/research/gdpnow.aspx. 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. There is a large literature dealing with nowcasting GDP growth for different countries, see *e.g.* Aastveit and Trovik [2012], Aruoba et al. [2009], Doz et al. [2011] and Ferrara and Marsilli [2018].

above. Preselection in the first step is based on the t-statistics associated with each GSD variable in a regression that includes the official predictors as well, see *e.g.* Bai and Ng [2008] and Barut et al. [2016]. In the second step, a Ridge regularization is applied to the linear regression model incorporating the official variables and the GSD variables that have been preselected. The regularization parameter α is set equal to the minimum of the GCV criterion. Past literature has proposed forecasting approaches based on Ridge regression to deal with dense models with a large number of predictors, *e.g.* De Mol et al. [2008] and Carrasco and Rossi [2016]. We go beyond this literature by considering models where the dimension can be ultra-high and that can be either sparse or dense.

For our two-step procedure we show three types of theoretical properties. First, we establish the sure screening property of our preselection procedure. Second, we provide an upper bound for both the in-sample and OOS prediction error associated with the Ridge after model selection estimator. This upper bound is a function of the number of predictors N, the number of time-series observations T and the Ridge regularization parameter α. Third, we evaluate optimality of GCV to choose the regularization parameter α for OOS prediction. We know from the previous literature that GCV has optimality properties for in-sample prediction, see *e.g.* Li [1986], Andrews [1991] and Carrasco and Rossi [2016]. We complete this result by showing that the minimiser of the GCV is as good as the minimiser of the conditional mean squared prediction error for OOS prediction. The latter is the objective of nowcasting for central bankers.

We study finite sample properties of our procedure through a Monte Carlo exercise. Our study analyzes how the dimension of the problem, N and T, the degree of sparsity s in the model and the correlation among the predictors affect the performance of our method compared with other widely used methods in macroeconomic nowcast, like Lasso, Ridge without preselection and Principal Component Analysis estimators. We show that when the true data generating process is sparse with a large number of active predictors our *Ridge after Model Selection* procedure outperforms all the considered

competitors for OOS prediction.

Finally, we conduct an empirical study to answer questions (A) and (B) stated above for GSD with respect to GDP growth nowcast for three countries/areas: the euro area, the U.S. and Germany. Usual GDP nowcasting tools integrate standard official macroeconomic information stemming, for instance, from national statistical institutes, central banks and international organizations. Typically, two sources of official data are considered: (i) hard data (production, sales, employment ...) and (ii) opinion surveys (households or companies are asked about their view on current and future economic conditions). Sometimes, financial markets information, generally available on high frequency basis, is also integrated into the information set. However, Adrian et al. [2019] have recently shown that the macro-financial relationship is highly non-linear. In our study we also consider financial market information for robustness check. In addition to these official data, we include the alternative GSD into our information set.

We analyze three different periods: a period of cyclical stability ($2014q1 - 2016q1$), a period that exhibits a sharp downturn in GDP ($2017q1 - 2018q4$) and a period of recession (the *Great Recession* period from $2008q1$ to $2009q2$). Overall, empirical results show that GSD are useful when trying to nowcast GDP growth. At the beginning of the quarter, when there is no official information available about the current state of the economy, we show that using only Google data leads to very reasonable Mean Squared Forecasting Errors (MSFEs), sometimes only slightly higher than those obtained at the end of the quarter when the information set is complete. As soon as we integrate official macroeconomic information, starting from the fifth week of the quarter, MSFEs decrease reflecting the importance of this type of data in nowcasting. Overall, combining macroeconomic variables and GSD variables in the same model appears to be generally fruitful.

A striking result coming out from our empirical analysis is that, on the one hand, the preselection step is crucial in the first two periods considered as it generates better

outcomes compared to nowcasting procedures without any preselection. This result confirms previous findings from the nowcasting literature, see *e.g.* Bai and Ng [2008] and Boivin and Ng [2006] for dynamic factor models. On the other hand, recession periods present specific patterns as a model that only contains GSD, without any preselection step, tends to be preferred in terms of nowcasting accuracy. This result is quite robust over the three countries/areas that we consider in the study.

The rest of the paper is organized as follows. In Section 2 we describe the nowcasting model and our *Ridge after Model Selection* procedure for which theoretical results are provided in Section 3. The Monte Carlo exercise is in Section 4 while the empirical application is in Section 5. Section 6 concludes. Additional material and proofs are in the Supplementary Appendix.

2 Methodology

2.1 The nowcasting equation

To nowcast any series of interest Y_t we focus on linear bridge equation models which allow to construct Y_t nowcasts by using predictors available at different frequencies. To fix ideas, suppose the frequency of Y_t is quarterly. We include three types of predictors: *soft* variables, such as opinion surveys, *hard* variables, such as industrial production or sales, and variables stemming from GSD. GSD are available on a weekly basis while the other predictors are available monthly. 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 bridge equation model to nowcast 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 GSD, and ε_t is an unobservable shock. Because variables $x_{t,s}$, $x_{t,h}$ and $x_{t,g}$ are sampled over different frequencies – monthly and weekly, respectively – and are released with various reporting lags the relevant information set for calculating the nowcasts evolves within the quarter. We assume in the remaining of this paper that a given quarter is made up of thirteen weeks. Thus, by denoting with $x_{t,j,i}^{(w)}$, $j \in \{s,h,g\}$, the i-th series in vector $x_{t,j}$ released at week less or equal than $w = 1,\ldots,13$ of quarter t, we define the relevant information set at week w of a quarter t as $\Omega_t^{(w)} := \{x_{t,j,i}^{(w)}, i = 1,\ldots,N_j,\ j \in \{s,h,g\}\}$. For simplicity, we keep in $\Omega_t^{(w)}$ only the observations relative to the current quarter t and do not consider past observations. While the series in $x_{t,g}$ are 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 (or after) their release and so the dataset is unbalanced.

To explicitly account for the relevant information set at each week of the quarter, we replace model (2.1) by a model for each week w denoted by $M_{(w)}$ and defined as: $\forall t = 1,\ldots,T$, $\forall w = 1,\ldots,13$,

$$M_{(w)}: \qquad \mathbf{E}[Y_t|\Omega_t^{(w)}] \;=\; \beta_{0,w} + \beta_{s,w}' x_{t,s}^{(w)} + \beta_{h,w}' x_{t,h}^{(w)} + \beta_{g,w}' x_{t,g}^{(w)}, \tag{2.2}$$

where $\beta_{j,w,i} = 0$ if $x_{t,j,i}^{(w)} \notin \Omega_t^{(w)}$. For instance, as the first observation of euro area-industrial production relative to the current quarter t is released in week 9, then we set the corresponding $\beta_{h,w} = 0$ for every $w = 1,\ldots,8$. The bridge equations (2.2) exploit weekly information to obtain more accurate nowcasts of Y_t. The idea of having thirteen models is that a researcher aiming at nowcasting the current-quarter values of Y_t will use the model corresponding to the current week of the quarter. For instance, to nowcast the current-quarter value of Y_t at the end of week 2, model $M_{(2)}$ will be used.

2.2 Step 1: preselection of Google Search 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 macroeconomic variable such as GDP, at least in terms of nowcasting accuracy, see for instance Boivin and Ng [2006], Barhoumi et al. [2010]. The problem arises because we have too many variables and using all the variables would only add noise in the estimation process. As shown in Bai and Ng [2008], an empirical way to circumvent this issue is to target more accurately the choice of predictors.

As we will explain in Section 5, the GSD have a very high dimension N_g compared with T and in our empirical analysis we have experienced that using all the variables in the GSD is not always a good strategy because one would pay the price of dealing with ultra-high dimensionality without increasing the nowcasting accuracy as measured by the MSFE. In fact, it might be that the series Y_t to be nowcast is highly predictable by a subset of the GSD and that this subset is specific to Y_t. For this reason, before estimating model (2.2) we preselect GSD by retaining the most relevant variables for Y_t nowcasting, capturing much of the variability in it. We refer to this approach as a targeted preselection.

Preselection is based on the procedure proposed in Bai and Ng [2008] and Barut et al. [2016] which, in our framework, works as follows. Let us start from the standard linear regression equation (2.1). Then, by denoting with $x_{t,g,j}$ the j-th GSD variable, we apply the following approach:

(1) for each $j = 1, \ldots, N_g$, regress Y_t on a constant, $x_{t,s}$, $x_{t,h}$ and $x_{t,g,j}$, and compute the corresponding t-statistics t_j associated with $x_{t,g,j}$;

(2) select the Google variables that have the absolute value $|t_j|$ largest than a given threshold $\lambda > 0$: $\widehat{M}_g := \widehat{M}_g(\lambda) := \{1 \leq j \leq N_g : |t_j| > \lambda\}$.

The basic idea of this approach is that a GSD variable is retained depending on its contribution for predicting Y_t after controlling for a set of official variables. Associated with each λ there is a selected submodel $\widehat{M}_g := \widehat{M}_g(\lambda)$. In practice, we select λ as the $(1-\tau)$-quantile of a $\mathcal{N}(0,1)$ distribution with $\tau \in \{20\%, 10\%, 5\%, 2.5\%, 1\%, 0.5\%\}$. The parameter τ must be interpreted as the percentage of false positives that can be tolerated. Compared to Barut et al. [2016], we do not assume that the conditional variance of Y_t is known and we estimate it to construct the t-statistics. This modify in part the proof of the sure screening property in Theorem 2.1 below. This property is not established in Bai and Ng [2008].

To take into account the problem of frequency mismatch among the different sets of variables and make each monthly/weekly series comparable to the quarterly Y_t series in terms of frequency, one can replace each predictor with either its value available at a given week during the quarter t or with the average of its observations over the quarter t. In the second case, $x_{t,s}$, $x_{t,h}$ and $x_{t,g}$ are replaced by $x_{t,s}^{(13)} := \sum_{m=1}^{3} x_{t,g,(m)}/3$, $x_{t,h}^{(13)} := \sum_{m=1}^{3} x_{t,g,(m)}/3$ and $x_{t,g}^{(13)} := \sum_{w=1}^{13} x_{t,g,(w)}/13$, respectively, with $x_{t,s,(m)}$, $x_{t,h,(m)}$ and $x_{t,g,(v)}$ denoting the vectors of soft, hard and Google variables released at month m and week v of quarter t, respectively.

Let $N_1 := 1 + N_s + N_h$, $N := N_1 + N_g$, $X_t := (1, x_{t,s}', x_{t,h}', x_{t,g}')'$, $X_{t,O} := (1, x_{t,s}', x_{t,h}')'$, $X_{t,O,j} := (X_{t,O}', x_{t,g,j})'$ for every $j \in \{1, \ldots, N_g\}$, and $\beta := (\beta_0, \beta_s', \beta_h', \beta_g')'$. We introduce the following assumption.

Assumption A.1. *Assume that : (i) $Y_t = \beta_*' X_t + \varepsilon_t$, $t = 1, \ldots, T$, with β_* the true value of β, $\mathbf{E}[\varepsilon | X_t] = 0$ and $\mathbf{E}[\varepsilon\varepsilon' | X_t] = \sigma^2 I$, where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)'$; (ii) $\beta_{*j} \neq 0$, $\forall j \leq N_1$ and for $1 \leq s_g^* \leq N_g$, $\beta_{*g} = (\beta_{*g,1}, \ldots, \beta_{*g,s_g^*}, \mathbf{0}')'$, where $\mathbf{0}$ is a $(N_g - s_g^*)$-vector of zeros and $\beta_{*g,j} \neq 0$ for all $j = 1, \ldots, s_g^*$; (iii) $\varepsilon_t | X_t$, $t \geq 1$ are independent zero-mean sub-Gaussian random variables.*

Assumption A.1 *(i)* states that the true model is linear. Assumption A.1 *(ii)* states that the subvector of the true β_{*g} corresponding to the Google variables is s_g^*-sparse

and that the true sparse model is $M^* := \{1, \ldots, N_1\} \cup \{N_1 + j; j \in M_g^*\}$, where $M_g^* := \{1 \leq j \leq N_g : \beta_{*g,j} \neq 0\}$ is the subset of the true sparse model containing only the indices of the active Google variables with size $s_g^* = |M_g^*|$. Assumption A.1 *(iii)* assumes sub-Gaussianity of the errors conditional on the covariates X_t. This assumption is more general than assuming Gaussianity of $\varepsilon_t|X_t$ and allows for distributions whose tails are dominated by the tails of a Gaussian distribution.

For every $j \in \{1, \ldots, N_g\}$, define $\widetilde{\beta}_{O,j} := \arg\min_{\beta_O, \beta_{g,j}} \mathbf{E}(Y_t - X_{t,O}'\beta_{O,j}^1 - x_{t,g,j}\beta_{g,j})^2$, which is the pseudo-true value of $\beta_{O,j} := (\beta_{O,j}^{1\prime}, \beta_{g,j})' \in \mathbb{R}^{N_1+1}$ in the j-th misspecified model and define $\sigma_{O,j}^2 := \mathbf{E}[(Y_t - X_{t,O,j}'\widetilde{\beta}_{O,j})^2]$. Misspecification arises because, in general, $\widetilde{\beta}_{O,j}$ differs from the corresponding coefficients in β_*. Finally, $\mathcal{B} := \{\beta_{O,j}, j = 1, \ldots, N_g; |\beta_{O,j,1}| \leq B, \ldots, |\beta_{O,j,N_1}| \leq B, |\beta_{O,j,N_1+1}| \leq B\}$ for a large positive constant B is the set over which the Least Squares estimates in (1) are searched. The next assumption allows us to control for the estimated $\sigma_{O,j}^2$ in the construction of the t-statistics and is not present in Barut et al. [2016].

ASSUMPTION A.2. *Assume that: (i)* $\forall j, k, k' \in \{1, \ldots, N_g\}$ *with* $k, k' \neq j$, $cov(x_{t,g,k}x_{t,g,k'}|X_{t,O,j}) = c_{k,k',j}$ *where* $c_{k,k',j}$ *is a bounded constant that may depend on* (k, k', j) *but is independent of* t*; (ii)* $\{(x_{t,s}', x_{t,h}', x_{t,g}')'\}_{t\geq 1}$ *are i.i.d. random vectors in* $\mathbb{R}^{N-1}$*; (iii) for every* $j \in \{1, \ldots, N_g\}$, $\mathbf{E}[\widetilde{\varepsilon}_{t,j}^4|X_{t,O,j}]$ *is bounded, where* $\widetilde{\varepsilon}_{t,j} := (y_t - X_{t,O,j}'\widetilde{\beta}_{O,j})$*; (iv) there exist two constants* $0 < \underline{C}_x^2 < \overline{C}_x^2 < \infty$ *such that* $\underline{C}_x^2 \leq \min_{1\leq j\leq N_g} \lambda_{\min}\left(\mathbf{E}[X_{t,O,j}X_{t,O,j}']\right) \leq \max_{1\leq j\leq N_g} \lambda_{\max}\left(\mathbf{E}[X_{t,O,j}X_{t,O,j}']\right) \leq \overline{C}_x^2$*, where* $\lambda_{\min}(\cdot)$ *and* $\lambda_{\max}(\cdot)$ *denote the minimum and maximum eigenvalues of a matrix; (v) there exist two constants* $0 < \underline{\sigma}_O^2 < \overline{\sigma}_O^2 < \infty$ *such that* $\underline{\sigma}_O^2 \leq \min_{1\leq j\leq N_g} \sigma_{O,j}^2 \leq \max_{1\leq j\leq N_g} \sigma_{O,j}^2 \leq \overline{\sigma}_O^2$.

The next assumption is the same as Conditions 1 and 2 in Barut et al. [2016] made specific to our framework and we refer to that paper for comments about it.

ASSUMPTION A.3. *Assume that: (i) for* $j \in M_g^*$*, there exist two positive constants* $c_1, c_2 > 0$ *and* $0 < \kappa < 1/2$ *such that* $|cov(Y_t, x_{t,g,j}|X_{t,O})| \geq c_1 T^{-\kappa}$*, and uniformly in* $j \in \{1, \ldots, N_g\}$*:* $\mathbf{E}[x_{t,g,j}^2] \leq c_2$*; (ii) the Fisher information* $\mathbf{E}[X_{t,O,j}X_{t,O,j}']/\sigma_{O,j}^2$ *is finite*

and positive definite, and the operator norm $\|\mathbf{E}[X_{t,O,j}X'_{t,O,j}]\|_{op}$ *is bounded from above; (iii) there exists a* $\epsilon_T > 0$ *such that for all* $j \in \{N_1+1,\ldots,N\}$*:*

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

where K_T *is an arbitrarily large constant; (iv) there exist positive constants* m_0, m_1, s_0, s_1 *and* ρ *such that for sufficiently large* τ*,* $P(|X_{t,j}| > \tau) \leq m_1 \exp\{-m_0\tau^\rho\}$ *for all* $j \in \{2,\ldots,N\}$*, and* $\mathbf{E}[\exp\{2\beta'_*X_ts_0\} + \exp\{-2\beta'_*X_ts_0\}] \leq s_1$*; (v) for all* $\beta_{O,j} \in \mathcal{B}$*,*

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

for some positive constant V *bounded from below uniformly over* $j = N_1+1,\ldots,N$*.*

The next theorem establishes the sure screening property of our selection procedure.

THEOREM 2.1. *Suppose that Assumptions A.1-A.3 hold. Let* $\kappa_T := K_TB(N_1+1) + m_0K_T^\rho/s_0$*, with* K_T *given in Assumption A.3 (iii). Assume that* $T^{1-2\kappa}/(\kappa_TK_T)^2 \to \infty$ *and that* $T^{-\kappa/2}K_T^{\rho/2} = \mathcal{O}(1)$ *with* $\kappa < 1/2$ *given in Assumption A.3 (i). Then, by taking* $\lambda = c_7T^{1/2-\kappa}$ *for some constant* $c_7 > 0$ *it holds*

$$\begin{aligned} P\left(M^*_g \subset \widehat{M}_g(\lambda)\right) &\geq 1 - 4s^*_g(N_1+1)\exp\left\{-\frac{\min\{c_4, b_1/4, 1/4\}}{\kappa_T^2K_T^2}T^{1-2\kappa}\right\} \\ &- 2s^*_gTm_2e^{-m_0K_T^\rho} - 2s^*_gT\exp\left\{-\frac{K_T^\rho}{4CK_2^2}\right\} - 2s^*_g\exp\left\{-C_3T^{1-2\kappa}\min\left\{\frac{c_\epsilon^2}{K_1^2}, \frac{c_\epsilon}{K_1}\right\}\right\} \end{aligned}$$

where c_4, C, c_ϵ *are positive constants,* $m_2 := (N_1m_1 + \sqrt{s_1}\sqrt{\mathbf{E}[\exp\{4C_ms_0^2\|\varepsilon_t\|^2_{\psi_2}\}]})$ *for some positive constant* $C_m > 0$ *and with* $\|\varepsilon_t\|_{\psi_2} := \sup_{p\geq1} p^{-1/2}\left(\mathbf{E}\left[|\varepsilon_t|^p \mid X_t\right]\right)^{1/p}$*,* $K_1 := \max_{j\in M^*_g}\max_t\|(\widetilde{\varepsilon}^2_{t,j} - \sigma^2_{O,j})\|_{\psi_1}$*, and* $K_2 := \max_{j\in M^*_g}\max_t\|\widetilde{\varepsilon}_{t,j}\|_{\psi_2}$ *with* $\|\cdot\|_{\psi_1}$ *(resp.* $\|\cdot\|_{\psi_2}$*) denoting the sub-exponential (resp. sub-Gaussian) norm.*

This theorem is proved in Appendix B.5. The result of the theorem is similar to Barut

et al. [2016, Theorem 3]. The only difference, in the result and in the proof, is the presence of additional terms in the lower bound of the probability. These terms are due to the variance estimation in our approach (to construct the t-statistic t_j and $\widehat{M}_g$) that instead is assumed to be known in Barut et al. [2016]. If $\log(s_g^* N_1)\frac{(\kappa_T K_T)^2}{T^{1-2\kappa}} < \min\{c_4, b_1/4, 1/4\}$, $\log(s_g^* T m_2)/K_T^\rho < m_0$, $\log(s_g^*) = o\left(T^{1-2\kappa}/\max(K_1^2, K_1)\right)$ then Theorem 2.1 establishes that the selected submodel includes the true model M_g^* with probability approaching one. More precisely, if $0 < c < \min\left\{\frac{c_\epsilon^2}{K_1^2}, \frac{c_\epsilon}{K_1}\right\} < C$ and $0 < c < K_2 < C$, for two positive constants c, C, then the last two terms in the rate are negligible with respect to the other ones, and if we take the optimal order $K_T \asymp T^{(1-2\kappa)/A}$ where $A := \max\{4+\rho, 2+3\rho\}$, then $P\left(M_g^* \subset \widehat{M}_g(\lambda)\right) \gtrsim 1 - s_g^* m_2 \exp\{-CT^{(1-2\kappa)\rho/A}\}$. In this case it follows from Lemma A.1 in the Supplementary Appendix that with our methodology we can deal with an N_g such that $\log(N_g) = o(T^{(1-2\kappa)\rho/A})$. Similarly, we can deal with a m_2 such that $\log(m_2) = o(T^{(1-2\kappa)\rho/A})$ which means that N_1 and $\|\varepsilon_t\|_{\psi_2}$ are allowed to increase at a certain rate. If $X_{t,j}$ are sub-Gaussian then Assumption A.2 *(iv)* is satisfied with $\rho = 2$ which gives $\log(N_g) = o(T^{(1-2\kappa)/4})$.

2.3 Step 2: Ridge regression

Because GSD have a very high dimension, with the number of variables much larger than the number of observations, even after the preselection in Step 1 the number of selected Google variables may still be large compared to the time dimension T. To deal with this large number of preselected covariates, in Step 2 we use Ridge regularization. Let $\widehat{M} = \widehat{M}(\lambda) := \{1, \ldots, N_1\} \cup \widehat{M}_g$ and denote $X_{t,\widehat{M}}^{(w)} := (1, x_{t,s}^{(w)\prime}, x_{t,h}^{(w)\prime}, x_{t,g,\widehat{M}_g}^{(w)\prime})'$, where $x_{t,g,\widehat{M}_g}^{(w)} = \{x_{t,g,j}^{(w)}; j \in \widehat{M}_g\}$ is the vector containing only the preselected Google variables. We estimate the parameter $\beta_w := (\beta_{0,w}, \beta_{s,w}', \beta_{h,w}', \beta_{g,w}')'$ in equation (2.2) by minimizing a penalized residuals sum of squares where the penalty is given by the Euclidean squared norm $\|\cdot\|_2$. By using model (2.2) for each week $w \in \{1, \ldots, 13, \}$ we define the *Ridge*

after Model Selection estimator as: $\widehat{\beta}^{(w)} := \widehat{\beta}^{(w)}(\alpha)$ where

$$\widehat{\beta}^{(w)}(\alpha) := \underset{\beta\in\mathbb{R}^N;\beta_{g,j}=0,j\in\widehat{M}_g^c}{\operatorname{argmin}} \left\{ \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \beta_0 - \beta_s' x_{t,s}^{(w)} - \beta_h' x_{t,h}^{(w)} - \beta_g' x_{t,g}^{(w)}\right)^2 + \alpha\|\beta\|_2^2 \right\} \tag{2.3}$$

and $\alpha > 0$ 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. Let $\mathbf{0}$ be the $(N_g - |\widehat{M}|)$-dimensional column vector of zeros. Then, we can write $\widehat{\beta}^{(w)}$ as $\widehat{\beta}^{(w)} := (\widehat{\beta}^{(w)'}_{1:|\widehat{M}|}, \mathbf{0}')'$, where

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

and I is the $|\widehat{M}|$-dimensional identity matrix with $|\widehat{M}| = N_1 + |\widehat{M}_g|$.

Empirical choice of the parameter α is based on the Generalized cross-validation (GCV in the following) 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 an in-sample forecasting setting. To complement their study, we show in Section 3.2 that GCV performs well also for out-of-sample prediction. GCV selects, for each model w, the value of α that minimizes the following quantity:

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

where T denotes the last quarter of the training sample and $tr(\cdot)$ denotes the trace operator. We denote by $\widehat{\alpha}^{(w)}$ the value of α that minimizes $GCV^{(w)}(\alpha)$ in model w. Theoretical properties of $\widehat{\beta}^{(w)}$ and $\widehat{\alpha}^{(w)}$ are established in the next section.

3 Theoretical Properties

3.1 In-sample and Out-of-sample Prediction Error

In this section we study the convergence to zero of in-sample and out-of-sample prediction error associated with the *Ridge after Model Selection estimator*. Asymptotic properties for the out-of-sample prediction error associated with the Ridge estimator without model selection have been analysed in De Mol 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 [2012, 2016] for a Bayesian interpretation of the Ridge estimator.

To the best of our knowledge theoretical properties of the Ridge estimator coupled with a targeted selection have not been established in the previous literature. Here, we fill this gap and establish an upper bound for both the in-sample and out-of-sample prediction error for sparse models. This upper bound gives the rate of convergence as $N, T \to \infty$. For dense models, previous results in the literature, together with our result in Theorem 2.1, give the rate.

In the following, for simplicity, we leave implicit the dependence of each model on the week w and write $\widehat{\beta}$ instead of $\widehat{\beta}^{(w)}$. Let $X := (X_1, \ldots, X_T)'$ be a $(T \times N)$ matrix. Let M^{*c} denote the complementary set of $M^* := \{1 \leq j \leq N : \beta_{*j} \neq 0\}$ in $\{1, \ldots, N\}$ with $s^* := |M^*|$. For a vector $\beta \in \mathbb{R}^N$ and an index set $M \subset \{1, \ldots, N\}$, denote $\beta_M := (\beta_{M,j})_{j=1}^N$ with $\beta_{M,j} := \beta_j \mathbb{1}\{j \in M\}$, 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 $X_{t,M}$ denote the transpose of the t-th row of X_M. Thus, β_M has zero outside the set M. We denote by P_X the conditional probability given the covariates X. 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 $\|\delta\|_2 := \sqrt{\delta'\delta}$.

Next, we introduce an assumption which is known in the literature as a restricted sparse eigenvalue condition on the empirical Gram matrix $X'X/T$, see *e.g.* Belloni and Chernozhukov [2013], and it is an extension of the restricted isometry condition, see Candes and Tao [2007]. The quantity m in the assumption restricts the number of nonzero components outside the set M^* of the vectors $\delta \in \mathbb{R}^N$ considered.

ASSUMPTION A.4. *For a given $m < T$, for a $\delta \in \mathbb{R}^N$, with probability $1-o(1)$, $\underline{\varphi}(m)^2 := \min_{\|\delta_{M^{*c}}\|_0 \leq m,\, \delta \neq 0} \frac{\|\delta\|_{2,T}^2}{\|\delta\|_2^2} > 0$ and $\overline{\varphi}(m) := \max_{\|\delta_{M^{*c}}\|_0 \leq m,\, \delta \neq 0} \frac{\|\delta\|_{2,T}^2}{\|\delta\|_2^2} > 0$.*

We also define the condition number associated with the empirical Gram matrix $(X'_{\widehat{M}} X_{\widehat{M}})/T$: $\mu(\widehat{m}) = \frac{\sqrt{\overline{\varphi}(\widehat{m})}}{\underline{\varphi}(\widehat{m})}$, where $\widehat{m} := |\widehat{M} \setminus M_*| \, \mathbb{1}\{\widehat{M} \supseteq M_*\}$ is the number of incorrect covariates selected. Similarly, define $\widehat{k} := |M_* \setminus \widehat{M}| \, \mathbb{1}\{M^* \nsubseteq \widehat{M}\}$. We start by establishing an upper bound on the in-sample prediction error. Its proof is provided in Appendix B.1.

THEOREM 3.1 (In-sample prediction error). *Suppose that Assumptions A.1 (i)-(ii) and A.4 are satisfied and that $\varepsilon_t | X_t$ is Gaussian. Let $\widehat{M}$ be the model selected in the first step. Let $\widehat{\beta}$ be the Ridge estimator defined in (2.3). Then, for every $\epsilon > 0$, there is a constant K_ϵ independent of T such that with P_X-probability at least $1 - \epsilon$,*

$$\begin{aligned}\|\widehat{\beta} - \beta_*\|_{2,T} \leq & \left(K_\epsilon \sqrt{\frac{\widehat{m} \log(N) + (\widehat{m} + s^*) \log(e^2 \mu(\widehat{m}))}{T}} + 2\alpha \|\beta_*\|_2 \frac{1}{\underline{\varphi}(\widehat{m})} \right) \mathbb{1}\{M^* \subseteq \widehat{M}\} \\ & + \left(\frac{K_\epsilon \sigma}{\sqrt{T}} \sqrt{\widehat{k} \log(s^*) + \widehat{k} \log(e^2 \mu(0))} + \frac{2\alpha}{\underline{\varphi}(0)} \|\beta_*\|_2 + \|\beta_{*, M^* \setminus \widehat{M}}\|_{2,T} \right) \mathbb{1}\{M^* \nsubseteq \widehat{M}\}.\end{aligned}$$

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. As discussed below Theorem 2.1, $P(M^* \subset \widehat{M}) \to 1$ under some conditions (and so is the probability of $\{M^* \subseteq \widehat{M}\}$). We remark that on the event $\{M^* \nsubseteq \widehat{M}\}$, instead, we get a bias term given by $\|\beta_{*, M^* \setminus \widehat{M}}\|_{2,T}$. 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 (i)-(ii) and A.4*

are satisfied and that $\varepsilon_t|X_t$ *is Gaussian. Let* $\widehat{M}$ *be the model selected in the first step. Let* $\widehat{\beta}$ *be the Ridge estimator defined in (2.3). Then, for every* $\epsilon > 0$*, there is a constant* K_ϵ *independent of* T *such that with* P_X*-probability at least* $1-\epsilon$*,*

$$\|\widehat{\beta} - \beta_*\|_2 \leq \left(K_\epsilon \sqrt{\frac{\widehat{m}\log(N) + (\widehat{m}+s^*)\log(e^2\mu(\widehat{m}))}{T\underline{\varphi}(\widehat{m})^2}} + 2\alpha\|\beta_*\|_2 \frac{1}{\underline{\varphi}(\widehat{m})^2} \right) \mathbb{1}\{M^* \subseteq \widehat{M}\}$$
$$+ \left(\frac{K_\epsilon \sigma}{\underline{\varphi}(0)\sqrt{T}} \sqrt{\widehat{k}\log(s^*) + \widehat{k}\log(e^2\mu(0))} + \frac{2\alpha}{\underline{\varphi}(0)^2}\|\beta_*\|_2 + \frac{\|\beta_{*,M^*\setminus\widehat{M}}\|_{2,T}}{\underline{\varphi}(0)} \right) \mathbb{1}\{M^* \not\subseteq \widehat{M}\}.$$

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 either $1/\underline{\varphi}(\widehat{m})$ or $1/\underline{\varphi}(0)$. As seen in Assumption A.4, $\underline{\varphi}(\widehat{m})$ 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. For this let $(Y_\tau, X_\tau')'$, $\tau > T$, be a new copy of $(Y_t, X_t')'$ that satisfies Assumption A.1 and that is independent of (Y, X) with $Y := (Y_1, \ldots, Y_T)'$. We denote by P_{X_τ} the conditional probability given the covariates X and X_τ.

Corollary 3.2 (Out-of-sample prediction error)**.** *Suppose that Assumptions A.1 (i)-(ii) and A.4 are satisfied and that* $\varepsilon_t|X_t$ *is Gaussian. Let* $\widehat{M}$ *be the model selected in the first step. Let* $\widehat{\beta}$ *be the Ridge estimator defined in (2.3). 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* $(1-\epsilon)$*,*

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

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. Preselection allows to reduce this dimension 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].

3.2 Out-of-sample evaluation of the selection of α

In this section we evaluate the performance of $\widehat{\alpha}^{(w)}$, the minimiser of the GCV as described in Section 2.3, for out-of-sample prediction which is the objective of nowcasting for central bankers. For simplicity, in this section we continue to leave implicit the dependence of each model on the week w. Optimality of GCV minimisation for in-sample prediction in the setting of Ridge regularization has been established in Li [1986] and in Carrasco and Rossi [2016] but they do not consider optimality for out-of-sample prediction. The latter has been considered in Leeb [2008] in a setting different from Ridge regularization to evaluate the performance of model selection. Our proofs departs entirely from their proofs.

Consider a new copy $(Y_{T+1}, X'_{T+1})'$ of $(Y_t, X'_t)'$ that satisfies Assumption A.1 *(i)-(ii)* and that is independent of (Y, X), where $Y^{(T)} := (Y_1, \ldots, Y_T)'$ and $X^{(T)} := (X_1, \ldots, X_T)'$ that is: $Y_{T+1} = \sum_{j=1}^{s^*} X_{T+1,j}\beta_{*,j} + \varepsilon_{T+1} = \sum_{j=1}^{\widehat{m}+s^*} X_{T+1,j}\beta_{*,j} + \varepsilon_{T+1}$ with $\beta_{*,j} = 0$ for every $j \in \{s^*+1, \ldots, \widehat{m}+s^*\}$, $\mathbf{E}[\varepsilon_{T+1}|X_{T+1}] = 0$ and $Var(\varepsilon_{T+1}|X_{T+1}) = \sigma^2$. With the purpose of forecast, one would like that the $\widehat{\alpha}$ selected by GCV based on the sample (Y, X) be good for out-of-sample prediction. That is, given a selected value $\widehat{\alpha}$, its out-of-sample performance is evaluated by considering the conditional mean squared prediction error given by

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

where $\widehat{\beta}(\alpha)' X_{T+1} = \sum_{j=1}^{\widehat{m}+s^*} X_{T+1,j}\widehat{\beta}_j(\alpha)$, $\widehat{\beta}(\alpha)$ is the *Ridge after Model Selection* estimator defined in (2.3) and $\mathcal{A} := \{M_* \subseteq \widehat{M}\}$. Theorem 3.2 below provides the rate at which the GCV criterion converges to $\rho^2(\alpha; Y^{(T)}, X^{(T)})$ uniformly over α in a

given region as $T \to \infty$. For this we introduce the following assumption and notation: $\widehat{\Sigma}_{\widehat{M}} := X'_{\widehat{M}} X_{\widehat{M}}/T$, $\Sigma_{\widehat{M}} := \mathbf{E}[X_{t,\widehat{M}} X'_{t,\widehat{M}}]$, $X_{\widehat{M}} := (X_{1,\widehat{M}}, \ldots, X_{T,\widehat{M}})'$ and for a matrix A, $\|A\|_{op}$ denotes its operator norm. Moreover, $\mathcal{B} := \left\{ X^{(T)}; \left\| \widehat{\Sigma}_{\widehat{M}} - \Sigma_{\widehat{M}} \right\|_{op} \leq C \sqrt{\frac{\widehat{m}+s^*}{T}} \right\}$ with $0 < C < \infty$ a universal constant.

ASSUMPTION A.5. *Assume that: (i)* $\mathbf{E}[\varepsilon_t^4] < C_1$ *for some constant* $0 < C_1 < \infty$*; (ii)* $\{X_t\}_{t \geq 1}$ *are i.i.d. random vectors in* $\mathbb{R}^N$ *with finite second moment; (iii)* $P(M_* \nsubseteq \widehat{M}) = o(r_{\mathcal{A},T})$ *where* $r_{\mathcal{A},T}$ *is a non-stochastic sequence independent of* α *that converges to zero as* $T \to \infty$*; (iv)* $P(\mathcal{B}^c) = o(r_{\mathcal{B},T})$ *where* $r_{\mathcal{B},T}$ *is a non-stochastic sequence independent of* α *that converges to zero as* $T \to \infty$*; (v) there exists* $\gamma > 0$ *such that* $\|\Sigma_{\widehat{M}}^{-\gamma/2} \beta_*\|_2 < \infty$ *(source condition).*

Assumption A.5 *(iii)* is satisfied by our preselection method in Step 1 under Assumptions A.1-A.3. Our Theorem 2.1 gives an explicit expression for $r_{\mathcal{A},T}$ under those assumptions. Assumption A.5 *(iv)* is for instance satisfied if $X_{t,\widehat{M}}$ are sub-Gaussian random vectors with sub-Gaussian norm bounded by a constant K. In this case C in the definition of $\mathcal{B}$ is equal to $4K^2$ and $r_{\mathcal{B},T} = 2\exp\{(\widehat{m}+s^*)\log(9) - c\min\{\sqrt{\frac{\widehat{m}+s^*}{T}}, \frac{\widehat{m}+s^*}{T}\}T\}$ for a numerical constant $c > 0$.

THEOREM 3.2. *Let Assumptions A.1 (i)-(ii) and A.5 hold and* $\widetilde{\gamma} := \min\{\gamma, 2\}$*. Then, for every* $\alpha > 0$*:*

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

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

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

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

This theorem establishes the rate at which the conditional mean squared prediction error converges to the *GCV*-criterion. If $\max\{\frac{(\widehat{m}+s^*)}{T^{1/2+\underline{u}}}, \frac{(\widehat{m}+s^*)}{T^{2\underline{u}}T^{\widetilde{\gamma}(1/2-\underline{u})}}\} \to 0$ then the convergence of the two criteria is uniform over a shrinking interval. The next theorem establishes out-of-sample optimality of the GCV-minimiser $\widehat{\alpha}$

THEOREM 3.3. *Consider the minimizers of* $GCV(\alpha)$ *and* $\rho^2(\alpha; Y^{(T)}, X^{(T)})$*:*

$$\widehat{\alpha} = \arg \min_{\alpha \in [\underline{\alpha} T^{-1/2+\underline{u}}, \overline{\alpha}_T]} GCV(\alpha)$$

and $\alpha^* = \arg\min_{\alpha \in [\underline{\alpha} T^{-1/2+\underline{u}}, \overline{\alpha}_T]} \rho^2(\alpha; Y^{(T)}, X^{(T)})$*. Then, in the setting of Theorem 3.2, (i)* $\widehat{\alpha}$ *is as good as* α^* *for out-of-sample prediction in the sense that*

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

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

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

4 Monte-Carlo exercise

This section presents the results of a simulation exercise. We are interested in understanding how the dimension of the problem, N and T, the degree of sparsity s^* in the model and the correlation among the predictors affect the nowcasting performance of our estimation method in finite sample. In this respect, we conduct two exercises. The first one consists in comparing our *Ridge after Model Selection* procedure with the mostly used methods in the macroeconometric nowcasting/forecasting literature. The description and results of this exercise are postponed to Section C in the Supplementary

Appendix. Here, we show the results of the second exercise which aims at looking at the effect of varying N, T, s^* on the in-sample and out-of-sample prediction error to confirm the theoretical results in Section 3.1.

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

$$\begin{aligned} y_t &= \gamma' z_t + \beta' x_t + v_t, \qquad z_t = (z_{1,t}, z_{2,t})' \sim \mathcal{N}_2 \left(0, \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix} \right), \\ \underset{(N\times 1)}{x_t} &= \delta' z_t + u_t, \end{aligned} \tag{4.1}$$

where $\gamma = (1, 2)'$, $u_t \sim \mathcal{N}_N(0, \Psi)$, Ψ is an $(N \times N)$-full rank covariance matrix and $v_t \sim \mathcal{N}(0, 1)$. In the DGP for y_t we have two sets of covariates: z_t, which we are sure are in the model, and x_t, which has to be preselected. The possible sparsity of the model only affects x_t. Specification of β governs the sparsity of the model, δ determines the correlation between x_t and z_t and Ψ affects the correlation among covariates in x_t that are included in the model and others that are not. We consider a sparse structure: $\beta_j \sim \mathcal{N}(0, 1)$ for $j \leq s^*$ and $\beta_j = 0$ for $j > s^*$.

For the parameter vector δ we use the specifications: $\delta = 0.2\iota$ and $\delta = 0.8\iota$ with ι a $(N \times 2)$-matrix of ones. For the covariance matrix Ψ we consider two cases: (I) *uncorrelated:* $\Psi = I_N$ and (II) *decreasing correlation*: $\Psi = \left(|0.5|^{j-k}\right)_{j,k}$. For N, T, s we consider: $(N = 150, T = 100, s = 105)$, $(N = 200, T = 150, s = 105)$, $(N = 200, T = 150, s = 110)$ and $(N = 200, T = 100, s = 110)$. The split between in-sample and out-of-sample is adjusted to keep the same fraction of observations in the two samples.

The results are in Tables 1-2 in Appendix A and are presented as ratios with respect to the case $N = 150, T = 100, s = 105$. That is, we show the in-sample MSE of each triplet (N, T, s) relative to the in-sample MSE for $N = 150, T = 100, s = 105$, labeled "MSER", and the out-of-sample MSE of each triplet (N, T, s) relative to the out-of-sample MSE for $N = 150, T = 100, s = 105$, labeled "MSFER". The threshold parameter λ is set equal to the $\{80\%, 90\%, 95\%, 97.5\%, 99\%, 99.5\%\}$-quantiles of a

$\mathcal{N}(0,1)$ distribution and is indicated in the first column of each table below. This choice corresponds to a false discovery rate (FDR) equal to $20\%, 10\%, 5\%, 2.5\%, 1\%, 0.5\%$, respectively.

We see that when T increases from 100 to 150 both the MSE (in-sample error) and the MSFE (out-of-sample error) decrease even if N increases as well (first two columns in the Table). When s is also increasing (going from $s = 105$ to $s = 110$) there is a reduction in the MSE and MSFE even if it is smaller (third and fourth columns in the Table) and for the values of λ corresponding to a FDR of 20% (and also 10% in the case $\delta = 0.2\iota$) only the MSFE decreases. The conclusions are similar for the two structures of Ψ. Instead, when N and s increase but T remains fixed (last two columns in the Table), the MSE is decreasing in most of the cases but the MSFE is increasing. This illustrates our extra term $\sqrt{\widehat{m} + s^*}$ in the upper bound for the out-of-sample prediction error in Corollary 3.2.

5 Empirical Study

In this section, we present the empirical results obtained by applying our *Ridge after Model selection* procedure to nowcast GDP growth rate with weekly GSD for three countries/areas: the euro area (EA, hereafter), the U.S. and Germany. We will present our out-of-sample results for three various phases of the business cycle: a calm period (2014-16), a period with a sudden downward shift in GDP growth (2017-18) and a recession period with large negative growth rates (2008-09). As a calm period we focus on 2014-16 as this period does not show any strong GDP movements, excepting a decline in oil prices starting from mid-2014. Then the period 2017-18 is interesting as the global economy was recovering in 2017 at a faster pace than expected by economists but the trade war initiated by the Trump's administration came by surprise, leading to a sharp slowdown in global GDP amidst rising uncertainties around global trade. Last, all the

considered countries/areas experienced a large drop in the level of GDP during the Great Recession of 2008-09, in the wake of the global financial crisis.

The empirical results are computed against a background of *pseudo* real-time analysis, that is, we account for the release dates of official variables but we do not use vintages of data. A robustness check for a *true* real-time nowcasting is carried out on EA data in subsection 5.3.

5.1 Design of the empirical analysis

The objective of this empirical application is to nowcast on a high-frequency basis quarterly GDP growth rate, which is variable Y_t in model (2.1)-(2.2), for three countries/areas[2]. The official monthly macroeconomic series that we use as regressors $x_{t,h}^{(w)}$ and $x_{t,s}^{(w)}$ are, respectively, the growth rate of the industrial production index, which is the most used measure of hard data by practitioners, denoted by IP_t, and a composite index of opinion surveys from various sectors as a proxy for soft variables, denoted by S_t. As regards Germany and EA, we use for S_t sentiment indexes computed by the European Commission, while we use the well known ISM index for the U.S. economy.

GSD are weekly data which are received and made available by the European Central Bank every Tuesday and are not the same as Google Trends data. GSD are data related to queries performed with Google search machines. The queries are assigned by Google to particular categories using natural language processing methods. The data are indexes of weekly volume changes of Google queries grouped by category and by country. Data are normalized at 1 at the first week of January 2004 which is the first week of availability of these data. Then, the following values indicate the deviation from the first value. The GSD we use for our study cover weekly Google searches for the six main euro area countries: Belgium, France, Germany, Italy, Netherlands and Spain, as well as for the U.S., ranging from the January 1st, 2004, to end of December 2018. For each country,

[2]Official GDP data are stemming from Eurostat for EA as a whole, from Destatis for Germany and from the BEA for the U.S.

we have at disposal a total of $N_g = 296$ variables, grouped into 26 categories. When dealing with the EA as the whole, we account for information conveyed by the six main countries, meaning that we have access to a total of $N_g = 1776$ GSD variables. 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. Original data are not seasonally adjusted, thus we take the growth rate over 52 weeks to eliminate the seasonality within the data, a standard procedure when dealing with weekly data (see Lewis et al. [2020]). In order to account for the variation over a quarter, we add a second differentiation over 13 weeks. Consequently, models are estimated on a recursive basis starting the last week of March 2005.

In addition to frequency mismatch in the data, another challenge arises because data on official series and GSD are released with various reporting lags, leading thus to an unbalanced information set 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]) and, as discussed in Section 2.1, we deal with it by considering a different model for every week of the quarter (*i.e.* the thirteen models $M_{(w)}$ given by equation (2.2)). As regards macroeconomic series, we mimic the exact release dates as published by statistical offices. Within the EA, for instance, 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. The IP_t for the first month of the quarter is 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 (see Figure 1 in the Supplementary Appendix). As regards the U.S. economy, survey data are released at the same dates as for the EA while industrial production data are released about four weeks earlier.

To construct the vector $x_{t,g}^{(w)}$ in equation (2.2) containing GSD variables, we take for each Google variable the sample average of its observations over the period week 1 to

week w of quarter t. That is, by denoting with $x_{t,g,(v)}$ the N_g-vector of Google variables released at week v of period t (not averaged) we construct $x_{t,g}^{(w)} = \sum_{v=1}^{w} x_{t,g,(v)}/w$. 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]

As regards the survey S_t in $x_{t,s}^{(w)}$ and the industrial production IP_t in $x_{t,h}^{(w)}$, we impose the following structure which mimics the data release explained above in the case of the EA. The variable $x_{t,s}^{(w)}$ is not present in models 1 to 4 because the current S_t is not available in the first four weeks of the quarter, so that $\beta_{s,1} = \beta_{s,2} = \beta_{s,3} = \beta_{s,4} = 0$. Then, for models $w \in \{5, \ldots, 8\}$, $x_{t,s}^{(w)}$ is the value of S_t 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 $w \in \{9, \ldots, 12\}$, $x_{t,s}^{(w)}$ is equal to the average of the survey data available at the end of the first and second month of the quarter: $x_{t,s}^{(w)} = (S_{t,1} + S_{t,2})/2$. Lastly, 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 $w \in \{1, \ldots, 10\}$ (so that $\beta_{h,1} = \ldots = \beta_{h,10} = 0$), and in models $w \in \{11, \ldots, 13\}$, $x_{t,h}^{(w)}$ is the value of IP_t for the first month of quarter t. For Germany and U.S. we use a similar structure that mimics the specific data release for these countries.

We split our dataset in two non-overlapping subsamples: the training set and the out-of-sample set. The latter starts at $2014q1$, $2017q1$ or $2008q1$ depending on the period we are considering, and the training sample always finishes two quarters before the beginning of the out-of-sample period to take into account the delay due to the release of GDP figures, see Section D.3 in the Supplementary Appendix. We use a recursive scheme method, that is, the parameters are re-estimated at each new nowcasting quarter using all the past information available until the penultimate quarter before the nowcasting one.

[3] As robustness check we have considered models that do not use the average over weeks of GSD as explanatory variables, but instead construct $x_{t,g}^{(w)} = x_{t,g,(w)}$. Our findings clearly point that averaging GSD reduces the MSFE.

5.2 Overall evaluation of Google Search Data

This subsection presents our empirical results for the three countries/areas and the three different economic periods as explained before in order to answer questions (A) and (B) in the Introduction. We estimate the following models: (a) the nowcasting models $M_{(1)}, \ldots, M_{(13)}$ accounting for the full set of information (GSD, hard and soft data) without preselection, (b) the nowcasting models $M_{(1)}, \ldots, M_{(13)}$ accounting for the full set of information (GSD, hard and soft data) with the preselection of Step 1 (i.e. Ridge after Model selection approach), (c) the nowcasting models $M_{(1)}, \ldots, M_{(13)}$ by using only GSD (i.e. $\beta_{s,w} = \beta_{h,w} = 0$ for every $w = 1, \ldots, 13$ in Equation (2.2)), and (d) the models that only account for hard and soft data (i.e. without GSD, $\beta_{g,w} = 0$ for every $w = 1, \ldots, 13$ in Equation (2.2)). Root MSFE (RMSFE) results are presented in Tables 3 to 11 in Appendix A, each row corresponding to those four models.

By looking at rows 2-4 of Tables 3-11, one can compare nowcasts obtained with and without GSD in a pseudo real-time exercise to assess: (i) if GSD are informative when there is no official data available for the forecaster and (ii) to what extent GSD remain informative when official data become available. The first stylized fact that we observe in our results is the downward sloping evolution of RMSFEs throughout the quarter stemming from models $M_{(1)}, \ldots, M_{(13)}$ with full information (second row). 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]). Second, when using only Google information (third row), we still observe a decline of RMSFEs throughout the quarter but to a much lower extent. Third, when focusing on the beginning of the quarter, models that only integrate Google information provide very reasonable RMSFEs, which are slightly higher than those obtained at the end of the quarter. This means that using only Google data at the beginning of the quarter, when there is no other available official information about the current state of the economy, is a pertinent strategy for economists aiming at tracking GDP.

Let us now focus on the gain from using our *Ridge after Model Selection* strategy. During both calm and *sudden shift* periods (namely, 2014-16 and 2017-18), this estimation strategy applied to the whole dataset (hard data, soft data and GSD) generally tends to provide the lowest RMSFEs (second row of Tables 3-11 in the Appendix, lowest values in bold). This result, which is robust over the different countries/areas, does not hold during the recession period (2008-09). Thus, we have here few important results. First, by comparing the first and second rows of Tables 3-11, we can conclude that, outside recession periods, our *Ridge after Model Selection* strategy outperforms a strategy that would skip the data preselection step. Second, by comparing the second, third and fourth rows of Tables 3-11, we point out that, outside recession periods, combining information (i.e. macroeconomic and Google data) generally leads to more accurate nowcasts than those only based on either pure macroeconomic information or pure Google information. Third, recession periods possess a very specific pattern as preselecting data does not necessarily generate lower RMSFEs during those phases of the business cycle. Indeed, for almost all weeks within the quarter, the Ridge model that only integrates Google data without preselection outperforms other models (third rows of Tables 5, 8, 11). Hence, during a recession, (i) we do not have to preselect data and (ii) GSD provide more accurate information than official macroeconomic data. This latter result has been also observed during the Covid-19 recession for which it has been shown that using new sources of high-frequency data is a way to get more reactive and accurate nowcasts of the economic activity (Lewis et al. [2020]).

5.3 Robustness checks

This section shows the results of two robustness checks we have carried out for the euro area. The first one concerns a true real-time analysis executed by using vintages of GDP and industrial production. The second robustness check consists in controlling for additional official macroeconomic series other than industrial production and opinion

surveys.

5.3.1 A true real-time analysis

The analysis is done over the period 2014-16 for which we use vintages of data for EA GDP and industrial production[4] and account for the observed timeline of data release as provided by Eurostat. When available, we also include lagged GDP growth among the explanatory variables of the nowcasting models[5].

The RMSFEs values that compare the same four models (a)-(d) as in the previous pseudo real-time analysis are reported in Tables 34-36 in the Supplementary Appendix. Overall, the results that we have illustrated above for the pseudo real-time exercise still hold in true real-time. In particular, we get that (i) nowcasting accuracy improves throughout the quarter, (ii) GSD provide valuable information at the beginning of the quarter when there is no official information, (iii) combining macroeconomic information with Google information improves the results and (iv) our Ridge after Model Selection strategy outperforms the other approaches in terms of nowcasting accuracy. Those results are reassuring about the reliability of the real-time use of GSD when nowcasting EA GDP.

5.3.2 Controlling for additional macroeconomic variables

So far, we have only controlled for economic surveys, S_t, and industrial production, IP_t, as official series. This makes sense to us as both series are considered by practitioners as the two most important variables to asses the EA economic state. Here, we aim at checking the robustness of our evaluation about GSD to a richer macroeconomic information set. For this purpose we include additional macroeconomic series among the covariates $x_{t,s}^{(w)}$ and $x_{t,h}^{(w)}$ in model (2.2). Those macroeconomic series – such as sales,

[4]Survey data are generally not revised.

[5]Table 23 in the Supplementary Appendix gives the exact weeks in the out-of-sample period 2014q1-2016q1 where the lagged GDP is included in the real-time analysis.

exports or unemployment rate – are commonly used in the nowcasting literature and are continuously monitored by policymakers and market participants[6]. We refer to this richer set of variables as *Big Official Set* and to the set made of the previously considered variables, IP_t and S_t, as *Small Official Set*. The robustness check is made for two periods: the period $2014q1 - 2016q4$ of cyclical stability and the period $2017q1 - 2018q4$ where there is a downward shift in the EA-GDP series.

As performance measure, we look at the ratios between the RMSFEs obtained by using: GSD together with the *Small Official Set* of data (resp. the *Big Official Set* of data) in the numerator (resp. in the denominator). A ratio larger than one indicates that including additional official series improves nowcasting accuracy, and conversely.

Results are in Table 43 of the Supplementary Appendix. They highlight that in the period of cyclical stability $2014q1 - 2016q1$ the inclusion of additional macroeconomic variables when using our Ridge after Model Selection strategy does not generally improve the nowcasting accuracy except for week 4, and week 2 at a lower extent. This result still holds when there is no preselection of data and when we only account for macroeconomic variables. This result is quite strong as it means that using only industrial production and a survey does a pretty good job when trying to nowcast EA GDP. However, when there is a downward shift in the GDP, as in the period $2017q1 - 2018q4$, it seems worth including a larger set of macroeconomic variables, especially in the middle of the quarter (rows four and six of Table 43). For the first three weeks and last three weeks, the *Small Official* dataset is preferred. On the other hand if we compare the results obtained without preselection of GSD (row five) inclusion of additional macroeconomic variables does not improve the nowcasting accuracy. It seems that when economic uncertainty is high, it is useful to keep all the GSD variables into the models. A possible explanation is that this uncertainty generated by the trade war does not have a strong common impact on all macroeconomic variables and does not adversely affect economic activity across

[6]A description of the 22 macroeconomic variables used is in Table 24 of the Supplementary Appendix.

the board.

6 Conclusions

Large data sets arising from alternative sources of information have gained in popularity among macroeconomists when trying to assess the current state of the economy on a high-frequency basis. The design of an appropriate econometric methodology for macroeconomic nowcasting has to take into account the specific structure of the alternative data used. This paper deals with the particular structure of Google Search Data (GSD) and proposes an econometric approach for nowcasting macroeconomic quantities based on GSD together with official data when available. Our proposed *Ridge after Model Selection* approach is two-step where in the first step GSD variables are preselected, conditionally on the official variables, by targeting the macroeconomic aggregate to be nowcast. In the second step, a Ridge regularization is applied to those preselected GSD and official variables. The Ridge tuning parameter is chosen by Generalized Cross Validation.

Our theoretical contribution consists in showing different types of optimality properties of our proposed procedure. First, we prove that our targeted preselection retains all the active variables in the true model with probability approaching one. Second, we show convergence to zero of both the in-sample and OOS prediction error associated with the *Ridge after Model Selection* estimator as well as consistency of parameter estimation. Finally, we are the first to demonstrate optimality of GCV for OOS prediction in the setting of Ridge regularization.

We illustrate our procedure through numerical studies and an empirical application where we nowcast GDP growth rates for the EA as whole, the U.S., and Germany by combining standard macroeconomic variables and alternative GSD. Empirical results show that GSD contain valuable information about the current economic state and that

combining standard macroeconomic variables with GSD variables is generally fruitful. They also suggest that the preselection step is crucial as it consistently leads to better outcomes. On the other hand, recession periods present specific patterns as, during those phases of the business cycle, models that only contain non preselected GSD tend to outperform in terms of nowcasting accuracy.

Supplementary Appendix: it contains all the proofs of the results in the paper, additional simulations, figures, tables and the discussion of macroeconomic series used in the empirical study.

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.

T. Adrian, N. Boyarchenko, and D. Giannone. Vulnerable growth. *American Economic Review*, 109(4):1236–1289, 2019.

D. W. Andrews. Asymptotic optimality of generalized CL, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, 47(2):359–377, 1991.

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.

S. B. Aruoba, F. X. Diebold, and C. Scotti. Real-time measurement of business conditions. *Journal of Business & Economic Statistics*, 27(4):417–427, 2009.

J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304 – 317, 2008.

K. Barhoumi, O. Darne, and L. Ferrara. Are disaggregate data useful for forecasting French GDP with dynamic factor models ? *Journal of Forecasting*, 29:132–144, 2010.

E. Barut, J. Fan, and A. Verhasselt. Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277, 2016.

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.

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.

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.

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

C. De Mol, D. Giannone, and L. Reichlin. Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318–328, 2008.

C. Doz, D. Giannone, and L. Reichlin. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 164(1): 188–205, 2011.

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. Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized posterior. *Journal of Econometrics*, 170(2):458–475, 2012.

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

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

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

H. Leeb. Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli*, 14(3):661 – 690, 2008.

D. J. Lewis, K. Mertens, and J. H. Stock. Monitoring Real Activity in Real Time: The Weekly Economic Index. Liberty street economics, Federal Reserve Bank of New York, 2020.

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

Table 1: Effect of N, T, s on the in-sample and out-of-sample prediction error. In-sample (MSER) and out-of-sample MSE (MSFER) expressed as ratios with respect to the case $N = 150, T = 100, s = 105$. FDR denotes the percentage of false positive that can be tolerated.

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.

S. Ng. *Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data*, volume 2 of *Econometric Society Monographs*, page 1–34. 2017.

R. F. Niesert, J. A. Oorschot, C. P. Veldhuisen, K. Brons, and R.-J. Lange. Can Google search data help predict macroeconomic series? *International Journal of Forecasting*, 36(3):1163–1172, 2020.

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.

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

A Appendix

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

Table 2: Effect of N, T, s on the in-sample and out-of-sample prediction error. In-sample (MSER) and out-of-sample MSE (MSFER) expressed as ratios with respect to the case $N = 150, T = 100, s = 105$. FDR denotes the percentage of false positive that can be tolerated.

EA – Nowcasting during $2014q1 - 2016q1$													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.4467	0.4816	0.3897	0.3659	0.3239	0.3829	0.3901	0.3609	0.3427	0.3422	0.3103	0.3142	0.3111
Sel + Ridge (Google+ S+IP)	**0.2889**	**0.2607**	**0.2400**	**0.2493**	**0.1747**	**0.1706**	**0.1695**	**0.1608**	**0.1641**	**0.1668**	0.2222	**0.2178**	0.2082
Sel + Ridge (Google)	0.3026	0.2769	0.2841	0.3008	0.3052	0.3107	0.3001	0.2974	0.2984	0.2975	0.2964	0.2867	0.2880
No Google					0.1807				0.1897		**0.1928**		**0.2017**

Table 3: RMSFEs corresponding to the nowcasting period $2014q1 - 2016q1$. "Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated without pre-selection, "Sel + Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated with our *Ridge after Model Selection* procedure, "Sel + Ridge (Google)" refers to model (2.2) with only Google data (preselected), "No Google" refers to models without Google data ($NoGoogle_1$ - $NoGoogle_4$ in Table 21 in the Supplementary Appendix).

EA – Nowcasting during $2017q1 - 2018q4$													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.5592	0.5956	0.5713	0.5642	0.3604	0.3605	0.3186	0.3064	0.5100	0.4836	0.4181	0.4527	0.4788
Sel + Ridge (Google+ S+IP)	**0.3505**	**0.3306**	0.3341	**0.3227**	**0.2330**	**0.2664**	**0.2501**	**0.2415**	**0.2720**	**0.2451**	**0.1316**	**0.1340**	**0.1314**
Sel + Ridge (Google)	0.3760	0.3431	**0.3262**	0.3276	0.3230	0.3167	0.3051	0.2875	0.2894	0.2856	0.2795	0.2763	0.2700
No Google					0.4340				0.4841		0.2871		0.3177

Table 4: RMSFEs corresponding to the nowcasting period $2017q1 - 2018q4$. "Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated without preselection, "Sel + Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated with our *Ridge after Model Selection* procedure, "Sel + Ridge (Google)" refers to model (2.2) with only Google data (preselected), "No Google" refers to models without Google data ($NoGoogle_1$ - $NoGoogle_4$ in Table 21 in the Supplementary Appendix).

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

Table 5: RMSFEs corresponding to the nowcasting period $2008q1 - 2009q2$. "Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated without preselection, "Sel + Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated with our *Ridge after Model Selection* procedure, "Ridge (Google)" refers to model (2.2) with only Google data estimated without preselection, "No Google" refers to models without Google data ($NoGoogle_1$ - $NoGoogle_4$ in Table 21 in the Supplementary Appendix).

U.S. – Nowcasting during $2014q1-2016q1$													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.5843	0.5177	0.5771	0.5769	0.4588	0.4735	0.4207	0.4213	0.4315	0.4313	0.4479	0.4488	0.4786
Sel + Ridge (Google+ S+IP)	0.4889	**0.4792**	**0.4647**	**0.4670**	0.4101	**0.4277**	**0.3957**	**0.3922**	**0.3948**	**0.3933**	**0.4233**	**0.4273**	0.4509
Sel + Ridge (Google)	**0.4873**	0.4833	0.4829	0.4816	0.4777	0.4740	0.4750	0.4751	0.4745	0.4746	0.4753	0.4749	0.4703
No Google					**0.4062**		0.4061		0.4156		0.4260		**0.4466**

Table 6: RMSFEs corresponding to the nowcasting period $2014q1-2016q1$. "Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated without preselection, "Sel + Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated with our *Ridge after Model Selection* procedure, "Sel + Ridge (Google)" refers to model (2.2) with only Google data (preselected), "No Google" refers to models without Google data ($NoGoogle_1$ - $NoGoogle_4$ in Table 21 in the Supplementary Appendix).

U.S. – Nowcasting during $2017q1-2018q4$													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.4746	0.4090	0.4625	0.4549	0.2762	0.3021	0.2401	0.2407	0.2963	0.2506	0.2555	0.1791	0.2081
Sel + Ridge (Google+ S+IP)	0.3639	0.3601	**0.3092**	**0.3181**	**0.1735**	**0.1685**	**0.1347**	**0.1330**	**0.1042**	**0.0991**	**0.1187**	**0.1081**	**0.1320**
Sel + Ridge (Google)	**0.3482**	**0.3335**	0.3177	0.3270	0.3168	0.3103	0.3061	0.3055	0.2949	0.2833	0.2770	0.2816	0.2757
No Google					0.2598		0.2255		0.3604		0.3510		0.2979

Table 7: RMSFEs corresponding to the nowcasting period $2017q1-2018q4$. "Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated without preselection, "Sel + Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated with our *Ridge after Model Selection* procedure, "Sel + Ridge (Google)" refers to model (2.2) with only Google data (preselected), "No Google" refers to models without Google data ($NoGoogle_1$ - $NoGoogle_4$ in Table 21 in the Supplementary Appendix).

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

Table 8: RMSFE corresponding to the nowcasting period $2008q1-2009q2$. "Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated without preselection, "Sel + Ridge (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 without Google data ($NoGoogle_1$ - $NoGoogle_4$ in Table 21 in the Supplementary Appendix).

Germany – Nowcasting during $2014q1 - 2016q1$													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.3316	0.3225	0.3296	0.3365	0.3137	0.2905	0.2687	0.2690	0.2695	0.2684	0.2713	0.2754	0.2698
Sel + Ridge (Google+ S+IP)	0.2619	0.2454	**0.2219**	**0.2306**	**0.2378**	0.2406	0.2373	0.2382	0.2429	0.2460	0.2717	0.2794	0.2754
Sel + Ridge (Google)	**0.2265**	**0.2266**	0.2484	0.2436	0.2433	**0.2341**	**0.2310**	**0.2296**	**0.2362**	**0.2372**	**0.2380**	**0.2470**	**0.2453**
No Google					0.3977				0.4208		0.2914		0.4325

Table 9: RMSFEs corresponding to the nowcasting period $2014q1 - 2016q1$. "Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated without preselection, "Sel + Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated with our *Ridge after Model Selection* procedure, "Sel + Ridge (Google)" refers to model (2.2) with only Google data (preselected), "No Google" refers to models without Google data ($NoGoogle_1$ - $NoGoogle_4$ in Table 21 in the Supplementary Appendix).

Germany – Nowcasting during $2017q1 - 2018q4$													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
Ridge (Google+ S+IP)	0.4889	0.4894	0.4837	0.4777	0.4352	0.4800	0.4649	0.4441	0.5038	0.4927	0.4382	0.4059	0.4181
Sel + Ridge (Google+ S+IP)	0.3814	**0.3794**	**0.3751**	0.3769	0.3831	0.3826	0.3876	0.3898	0.3238	0.3181	0.3141	**0.3088**	**0.2917**
Sel + Ridge (Google)	**0.3812**	0.3800	0.3788	**0.3757**	**0.3711**	**0.3680**	**0.3712**	**0.3584**	**0.3184**	**0.3097**	**0.3141**	0.3187	0.3140
No Google					0.6532				0.6241		0.3632		0.6433

Table 10: RMSFEs corresponding to the nowcasting period $2017q1 - 2018q4$. "Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated without preselection, "Sel + Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated with our *Ridge after Model Selection* procedure, "Sel + Ridge (Google)" refers to model (2.2) with only Google data (preselected), "No Google" refers to models without Google data ($NoGoogle_1$ - $NoGoogle_4$ in Table 21 in the Supplementary Appendix).

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

Table 11: RMSFEs corresponding to the nowcasting period $2008q1 - 2009q2$. "Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated without preselection, "Sel + Ridge (Google+S+IP)" refers to model (2.2) with Google data, Survey, and IPI estimated with our *Ridge after Model Selection* procedure, "Ridge (Google)" refers to model (2.2) with only Google data estimated without preselection, "No Google" refers to models without Google data ($NoGoogle_1$ - $NoGoogle_4$ in Table 21 in the Supplementary Appendix).