ECONOMIC STATISTICS CENTRE OF EXCELLENCE

A COLLABORATION WITH Office for National Statistics

Big Data & Macroeconomic Nowcasting: Methodological Review

George Kapetanios and Fotis Papailias

ESCoE Discussion Paper 2018-12

July 2018

ISSN 2515-4664

About the Economic Statistics Centre of Excellence (ESCoE)

The Economic Statistics Centre of Excellence provides research that addresses the challenges of measuring the modern economy, as recommended by Professor Sir Charles Bean in his Independent Review of UK Economics Statistics. ESCoE is an independent research centre sponsored by the Office for National Statistics (ONS). Key areas of investigation include: National Accounts and Beyond GDP, Productivity and the Modern economy, Regional and Labour Market statistics.

ESCoE is made up of a consortium of leading institutions led by the National Institute of Economic and Social Research (NIESR) with King's College London, innovation foundation Nesta, University of Cambridge, Warwick Business School (University of Warwick) and Strathclyde Business School.

ESCoE Discussion Papers describe research in progress by the author(s) and are published to elicit comments and to further debate. Any views expressed are solely those of the author(s) and so cannot be taken to represent those of the ESCoE, its partner institutions or the ONS.

For more information on ESCoE see www.escoe.ac.uk.

Contact Details
Economic Statistics Centre of Excellence
National Institute of Economic and Social Research
2 Dean Trench St
London SW1P 3HE
United Kingdom

T: +44 (0)20 7222 7665
E: escoeinfo@niesr.ac.uk

Big Data & Macroeconomic Nowcasting: Methodological Review

George Kapetanios[1,2] and Fotis Papailias[1,2]

[1] King's College, London
[2] Economic Statistics Centre of Excellence

Abstract

This paper is concerned with an introduction to big data which can be potentially used in nowcasting the UK GDP and other key macroeconomic variables. We discuss various big data classifications and review some indicative studies in the big data and macroeconomic nowcasting literature. A detailed discussion of big data methodologies is also provided. In particular, we focus on sparse regressions, heuristic optimisation of information criteria, factor methods and textual-data methods.

Keywords: Big Data, Machine Learning, Sparse Regressions, Factor Models

JEL classification: C32, C53, C55

Contact Details

Fotis Papailias
Data Analytics Centre for Macro and Finance
King's Business School
King's College London
Bush House
30 Aldwych
London WC2B 4BG

Email: george.kapetanios@kcl.ac.uk,fotis.papailias@kcl.ac.uk

This ESCoE paper was first published in July 2018.

Big Data & Macroeconomic Nowcasting Methodological Review

George Kapetanios Fotis Papailias[*]

King's Business School

Data Analytics Centre for Macro and Finance

Economic Statistics Centre of Excellence

July 2018

Abstract

This paper is concerned with an introduction to big data which can be potentially used in nowcasting the UK GDP and other key macroeconomic variables. We discuss various big data classifications and review some indicative studies in the big data and macroeconomic nowcasting literature. A detailed discussion of big data methodologies is also provided. In particular, we focus on sparse regressions, heuristic optimisation of information criteria, factor methods and textual-data methods.

JEL Codes: C32, C53, C55.

Keywords: Big Data, Machine Learning, Sparse Regressions, Factor Models.

[*] Corresponding Author. Kings College London, Data Analytics for Macro and Finance Research Centre. E-mail: fotis.papailias@kcl.ac.uk

Contents

1 Introduction

The real-time monitoring of economic and financial variables is now one of the main interests of policy-makers, market practitioners and other economic agents. The 2007/2008 crisis highlighted the need of central banks and other major institutions for a continuous assessment of current economic conditions. The main difficulty is the publication delay of most key macroeconomic indicators but also fiscal variables, regional/sectoral indicators and disaggregate data. For example, the main variable in the economy, the GDP (and its components), is only available on a quarterly basis. Moreover, preliminary data are often revised afterwards, in particular around turning points of the business cycle. Also, it was recently announced that the Office for National Statistics (ONS) in the UK has started producing monthly estimates for the GDP, highlighting the need of using all available data for these estimates.

On the other hand, advancements of technology now allow to organise and store a large variety of data which leads to variables that are available on a monthly, weekly, daily or even higher frequency. For example, financial market transactions, electronic payments data, internet data, etc. This has stimulated a vast amount of statistical and econometric research on how to take advantage of the large, timely and higher frequency but irregular information to provide estimates for key low frequency economic indicators. A parallel, more empirical, literature has instead focused specifically on the use of big data for nowcasting economic indicators, often using rather simple econometric techniques and specific big data sources, mainly Google Trends. Finally, a more theoretical literature has developed new, or adapted old, statistical and econometric methods to handle large sets of explanatory variables, such as those associated with big data.

In this research, we first provide a general introduction to big data classification and discuss some issues which arise when handling very large datasets. Then, we offer a detailed review of the existing literature in three areas: (i) big data and macroeconomics (in general), (ii) variable selection and dimensional reduction for big data in macroeconomics, and (iii) nowcasting in macroeconomics. Finally, drawing from the above three strands of the literature, we provide a discussion of the most widely used econometric methodologies suitable to deal with big data in macroeconomic nowcasting.

Based on the surveyed papers, we can generally say that the use of internet search data, particularly Google Trends, has been dominant in studies using big data in macroeconomics. There exist some papers based on Twitter data but they are mainly in finance. Webscrapping and collection of online prices also offer potential, especially for nowcasting inflation. However, such datasets are difficult to obtain, even more so when many countries and long enough samples are required. A similar comment applies for credit card and financial transactions data, and for data summaries resulting from textual analysis.

From the literature it also emerges that the advantages of using data like Google Trends are: (a) more timely forecasts, not subject to data revision; (b) some improvements in forecast accuracy, even though these typically emerge with respect to simple benchmarks (mostly purely autoregressive models; (c) ease of data access and collection, (d) ease of data management and treatment; (e) expected good data quality; (f) reasonable likelihood that similar data will be available on a continuous basis and without major definitional changes. There are also some disadvantages when using this data source, the main ones being: (a) a typical sole use of such data can lead to biased results (commonly known as "big data hubris"); (b) the impossibility to access the raw data, and the lack of knowledge of the precise algorithms used to pre-treat and summarise them; (c) the possibility that free access will be discontinued by the (private) data provider, or limited due to the introduction of more stringent privacy laws.

In terms of statistics and econometrics, data analysis is typically broken down into four categories: (1) pre-treatment and summarisation, (2) estimation, (3) hypothesis testing and (4) prediction. Since a large amount of data is available, penalised regressions such as LASSO, LARS, and elastic nets can be used instead of the standard linear or logistic regression. Then, the choice of the final model should come from forecasting cross-validation so that the researcher makes sure the model has good out-of-sample predictive ability. It must be highlighted that in this review we focus on univariate target variables. There are also multivariate forecasting with big data which not considered in details here.

The rest of the paper is organised as follows. Section 2 provides a general introduction to big data classification and discusses some issues which might arise when working with very large datasets. Section 3 provides a review of the existing literature and Section 4 offers a discussion of the most widely used econometric methodologies used in the academic literature. Finally, Section 5 summarises the conclusions.

2 Big Data Description

2.1 Types of Big Data

One possibility to obtain a general classification is to adopt the "4 Vs" classification, originated by the IBM, which relates to: (i) Volume (Scale of data), (ii) Velocity (Analysis of streaming data), (iii) Variety (Different forms of data) and (iv) Veracity (Uncertainty of data). However, this classification seems too general to guide empirical nowcasting applications.

Instead, we adopt Doornik and Hendry (2015) classifications who identify three main types of big data:

- "Tall" (not so many variables, N, but many observations, T, with T $\gg$ N). This is for example the case with tick by tick data on selected financial transactions or search queries. In this case T is indeed very large in the original time scale, say seconds, but it should be considered whether it is also large enough in the time scale of the target macroeconomic variable of the nowcasting exercise, say quarters.

- "Fat" (many variables, but not so many observations, N $\gg$ T). Large cross-sectional databases fall into this category. Such datasets might be useful from a nowcasting point of view if either T is large enough or the variables allow proper model estimation (e.g., by means of panel methods).

- "Huge" (many variables and many observations, i.e., very large N *and* T). This data type is ideal in the nowcasting context. The main disadvantage is that big data collection started relatively recently (in the last decade) and does not allow for a long nowcasting cross-validation. Google Trends, publicly available summaries of a huge number of specific search queries in Google, are perhaps the best example in this category, and not by chance the most commonly used indicators in economic nowcasting exercises. .

A third possibility to classify big data is to identify the data content. A particularly useful taxonomy is provided by the statistics division of the United Nations Economic Commission for Europe (UNECE), which identifies three main types of big data:

1. **Social Networks** (human-sourced information): this information is the record of human experiences, by now almost entirely digitally stored in personal computers or social networks. Data, typically, loosely structured and often ungoverned, include:

   - Social Networks: Facebook, Twitter, Tumblr etc.
   - Blogs and comments
   - Personal documents
   - Pictures: Instagram, Flickr, Picasa etc.
   - Videos: Youtube etc.
   - Internet searches
   - Mobile data content: text messages
   - User-generated maps
   - E-mail

2. **Traditional Business systems** (process-mediated data): these processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected by either private or public institutions is highly structured and includes transactions, reference tables and relationships, as well as the metadata that sets its context. Traditional business data is the vast majority of what IT managed and processed, in both operational and BI systems. Usually structured and stored in relational database systems, including also ”Administrative data”, it can be grouped into:

   - Data produced by Public Agencies: medical records, social insurance, etc.

- Data produced by businesses: commercial transactions, banking/stock records, e-commerce, credit cards, etc.

3. **Internet of Things** (machine-generated data): derived from sensors and machines used to measure and record the events and situations in the physical world. It is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches.

   - Data from sensors: Fixed sensors (Home automation, Weather/pollution sensors, Traffic sensors/webcam, etc.) or Mobile sensors (tracking: Mobile phone location, Cars, Satellite images, etc.)
   - Data from computer systems: Logs, Web logs, etc.

From an economic nowcasting point of view, all the three types of big data are potentially relevant. For example, selected internet searches and/or twits (Social Networks), credit card transactions (Traditional business systems), or number of navigating commercial vessels in a certain area (Internet of things) could all provide useful leading indicators for the GDP growth of a country.

2.2 Issues with Big Data

There is an ongoing discussion regarding the advantages and disadvantages of big data. Mainly, the main advantage is the timely nature of these sources which allow for a high frequency analysis (in nowcasting or finance context). However, the researcher must take extra care to avoid the big data hubris which states that "the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis (Lazer et al., 2014). It must be noted that we think of big data as complements rather than substitutes for more common coincident and leading indicators. In this section, we distinguish data related issues and more methodological potential problems with the use of big data.

A first issue concerns data availability. As it is clear from the data categorisation described previously, most data pass through private providers and are related to personal aspects. Hence, continuity of data provision cannot be guaranteed. For example, Google could stop providing Google Trends, or at least no longer make them available for free. In particular, recently Google stop releasing Google Trends at a weekly frequency when the sample spans more than 3 or 4 years. Another concern related to data availability is the start date, which is often quite recent for big data. For example, Google Trends data is available from 2004 onwards.

A second issue s that both the size and the quality of internet data keeps changing over time, in general much faster than for standard data collection. For example, applications such as Twitter or WhatsApp were not available just a few years ago, and the number of their users increased exponentially, in particular in the first period after their introduction. Similarly, other applications can be gradually dismissed or used for different uses. For example, the fraction of goods sold by EBay through proper auctions is progressively declining over time, being replaced by other price formation mechanisms.

A third issue is that data could not be available in a numerical format, or not in a directly usable numerical format. A similar issue emerges with standard surveys, for example on economic conditions, where discrete answers from a large number of respondents have to be somewhat summarised and transformed into a continuous index. However, the problem is more common and relevant with internet data.

A final issue, again common also with standard data but more pervasive in internet data due to their high sampling frequency and broad collection set, relates to data irregularities (outliers, working days effects, missing observations, etc.) and presence of seasonal / periodic patterns, which require properly de-noising and smoothing the data.

Overall, our suggestion is to take a pragmatic approach that balances potential gains and costs from the use of big data for nowcasting. Hence, for a specific target variable of interest, such as GDP growth or unemployment, it is worth assessing the marginal gains of big data based indicators that are rather promptly available (such as Google Trends or other variables used in previous studies and made publicly available) with respect to more standard indicators based on soft and hard data.

3 Review of Existing Literature

3.1 Big Data in Macroeconomics

In this section we briefly review academic papers which employ big data for (i) unemployment, (ii) GDP, (iii) inflation, (iv) surveys, (v) financial variables, and (iv) other studies.

3.1.1 Unemployment

Choi and Varian (2009a) and Choi and Varian (2009b) illustrate the ability of Google Trends to predict the present (*nowcasting*) using daily and weekly reports of Google Trends. In particular, they claim that people who lose their jobs search the internet for job ads. Therefore, the increasing volume of Google search queries for job-related keywords potentially has an impact on forecasting/nowcasting the initial claims.

Askitas and Zimmermann (2009) suggest an innovative method of using data on internet activity to predict economic behavior in a timely manner, which is difficult at times of structural change. They show a strong correlation between keyword searches and unemployment rates using monthly German data.

D'Amuri and Marcucci (2012) suggest the use of an index of Internet job-search intensity (the Google Index, GI) as the best leading indicator to predict the US monthly unemployment rate. They perform a deep out-of-sample forecasting comparison analyzing many models that adopt their leading indicator, the more standard initial claims or combinations of both. They find that models augmented with the GI outperform the traditional ones in predicting the unemployment rate for different out-of-sample intervals that start before, during and after the Great Recession. Google-based models also outperform standard ones in most state-level forecasts and in comparison with the Survey of Professional Forecasters. These results survive a falsification test and are also confirmed when employing different keywords.

Ross (2013) investigates the issues of identifying and extracting keywords from Google Trends relevant to economic variables. He suggests the backward induction method which identifies relevant keywords by extracting these from variable relevant websites. This backward induction method was applied to nowcast UK unemployment growth using a small set of keywords. The majority of keywords identified using the backward induction method outperformed the competing models in terms of in-sample and out-of-sample tests of predictability indicating that the backward induction method is effective in identifying relevant keywords.

3.1.2 GDP and Components

Galbraith and Tkacz (2015) assess the usefulness of a large set of electronic payments data comprising debit and credit card transactions, as well as cheques that clear through the banking system, as potential indicators of current GDP growth in Canada. These variables capture a broad range of spending activity and are available on a very timely basis, making them suitable current indicators. While every transaction made with these payment mechanisms is in principle observable, the data are aggregated for macroeconomic forecasting. Controlling for the release dates of each of a set of indicators, they generate nowcasts of GDP growth for a given quarter over a span of five months, which is the period over which interest in nowcasts would exist. They find that nowcast errors fall by about 65 per cent between the first and final nowcast. Among the payments variables considered, debit card transactions appear to produce the greatest improvements in forecast accuracy.

Schmidt and Vosen (2011) introduce an indicator for private consumption based on search query time series provided by Google Trends. The indicator is based on factors extracted from consumption-related search categories of the Google Trends application Insights for Search. The forecasting performance of this indicator is assessed relative to the two most common survey-based indicators - the University of Michigan Consumer Sentiment Index and the Conference Board Consumer Confidence Index. The results show that in almost all conducted in-sample and out-of-sample forecasting experiments the Google indicator outperforms the survey-based indicators.

Koop and Onorante (2013) suggest to nowcast using dynamic model selection (DMS) methods which allow for model switching between time-varying parameter regression models. This is potentially useful in an environment of coefficient instability and over-parameterisation which can arise when forecasting with Google variables. They allow for the model switching to be controlled by the Google variables through Google probabilities. That is, instead of using Google variables as regressors, they allow them to determine which nowcasting model should be used at each point in time. In an empirical exercise involving nine major monthly US macroeconomic variables, they find that DMS methods provide large improvements in nowcasting; the variables are: inflation, industrial production, unemployment, wage inflation, money, supply, financial conditions index (FCI), oil price inflation, commodity price inflation and the term spread. The use of Google model probabilities within DMS often performs better than conventional DMS. Also, Mitchell et al. (2013) illustrate how big qualitative survey data might be used in nowcasting.

Recently, Bok et al. (2017) present in detail the methodology underlying the New York Fed Staff Nowcast to produce early estimates of GDP growth, synthesising a wide range of macroeconomic data as they become available.

3.1.3 Inflation

Cavallo and Rigobon (2016) examines ways to deal with price data. Potential sources for micro price data include: Statistical Offices, Scanner Data (e.g. Nielsen), Online data (e.g. Billion Prices Project) etc. CPI data is useful in measuring inflation whereas Scanner and Online data can be used in marketing analytics (e.g. market shares). The Billion Prices Project is an automated web-scraping software where a robot downloads a public page, extracts the prices information and stores it in a database. A direct outcome from the papers is that online data is also useful for nowcasting inflation in the US, Latin America and Euro Area. Links between online data and CPIs are tracked using VAR models and calculating the cumulative Impulse Response Functions. The forecasting examples use predictive regressions.

Boettcher (2015) describes in detail technological, data security and legal requirements of web crawlers focusing on Austria. The paper finds that web crawling technology provides an opportunity to improve statistical data quality and reduce the overall workload for data collection. Automatic price collection methods enable statisticians to react better to the increasing amount of data sources on the internet.

Griffioen et al. (2014) discuss the usability of online apparel prices for CPI analysis. This study falls in the web scraping category and reports the findings and difficulties of online price collection during a two years period. The advantages of web scraping clothing prices are: (i) online price collection is cheaper than price collection in physical stores, (ii) given the relatively low collection costs, there is an incentive to rely on 'big data' and circumvent small sample problems (e.g. high sampling variance), (iii) the quality of online data tends to be very good and (iv) some item characteristics can be easily observed. The main disadvantages of conducting a data collection of this type are: (i) website changes can lead to data problems, (ii) the choice of web scraping strategy can affect the information collected and item representativeness, (iii) weighting information is unavailable, and (iv) the available information on characteristics may be insufficient, depending on the need for quality adjustment.

Breton et al. (2015) provide an overview of ONS research into the potential of using web scraped data for consumer price statistics. The research covers the collection, manipulation and analysis of web scraped data. As before, the main benefits of web scraped data are identified as follows: (i) reduced collection costs, (ii) increased coverage (i.e. more basket items), (iii) increased frequency, (iv) production of new or complimentary outputs/indices, and (v) improved ability to respond to new challenges. ONS use web scraped data to calculate price indices which: (i) expand the number of items used, (ii) expand the number of days considered, and (iii) expand both the number of items and days considered. The construction of this sort of indices can be useful to economists and policymakers.

3.1.4 Surveys

Nyman et al. (2014a) investigate ways to use big data in systemic risk management. News and narratives are key drivers behind economic and financial activity. Their news data consists of (i) daily comments on market events, (ii) weekly economic research reports and (iii) Reuters news. Machine Learning and Principal Components are included in the methodology in order to calculate the consensus indexes based on the above sources. Their findings include that weekly economic research reports could potentially forecast the Michigan Consumer Index and daily comments on market events could potentially forecast market volatility.

Nyman et al. (2014b) introduce the Directed Algorithmic Text Analysis and show that this methodology can improve considerably on consensus economic forecasts of the Michigan Consumer Index Survey. The approach is based upon searching for particular terms in textual data bases. In contrast to econometric approaches, their methodology is based upon a theory of human decision making under radical uncertainty. The search is directed by the theory. This direction dramatically reduces the dimensionality of the search. They look for words which convey a very limited number of emotions. As in other approaches, they also use regression analysis, but the choice of variables comes from the underlying theory of decision making.

3.1.5 Financial variables

Apart from Google Trends, economic and financial researchers have also started using Twitter posts about various economics and financial news. Cerchiello and Giudici (2014) investigate how the quality of financial tweets can be measured. They suggest that a Google Scholar 'h-index' type measure allows for improved nowcasting of financial variables using Twitter texts. The Twitter users are ranked according to their 'h-index' and confidence intervals are constructed to decide whether top Twitter users are significantly different. Twitter data are collected and R language's TwitteR package is adopted. Their methodology lies in the field of loss data modelling.

Heston and Sinha (2014), even though it is not a macroeconomics application, use a dataset of over 900,000 news stories to test whether news can predict stock returns. They find that firms with no news have distinctly different average future returns than firms with news. Confirming previous research, daily news predicts stock returns for only 1-2 days. But weekly news predicts stock returns for a quarter year. Positive news stories increase stock returns quickly, but negative stories have a long-delayed reaction.

3.1.6 Other studies

Big Data offers potential benefits for statistical modelling, but confronts problems like an excess of false positives, mistaking correlations for causes, ignoring sampling biases, and selecting by inappropriate methods. Doornik and Hendry (2015) consider the many important requirements when searching for a data-based relationship using Big Data. Paramount considerations include embedding relationships in general initial models, possibly restricting the number of variables to be selected over by non-statistical criteria (the formulation problem), using good quality data on all variables, analyzed with tight significance levels by a powerful selection procedure, retaining available theory insights (the selection problem) while testing for relationships being well specified and invariant to shifts in explanatory variables (the evaluation problem), using a viable approach that resolves the computational problem of immense numbers of possible models.

According to Bendler et al. (2014), people's tweeting behavior can be attributed to points of interest in their vicinity. This relationship can be used to identify the veracity of a social media data set. Twitter patterns are recurrent and their stability is an indicator for data certainty. Datasets with high stability estimates can be reliably used in empirical analyses.

Schubert (2015) in a recent presentation highlights the effect of big data on economic policy decision-making. In the future, macroeconomic indicators (such as unemployment), economic sentiments, house indexes and consumer price dynamics will be greatly influenced by the use of internet based data analysis.

Most of the above discussed papers are based on Google data. Lazer et al. (2014) provide a set of warnings, discussing the large errors in flu prediction using Google data, and how they could be reduced. Google Flu Trend overestimated the influenza-like illnesses during 2012-2013. Given the way Google Trends are constructed, they can be sometimes harmful in forecasting if not properly managed.

In the AAPOR (2015) report, the discussion on big data issues continues. Big Data can have positive effect on timeliness. The effects on the other quality dimensions will depend on the data source and the user needs. It is also mentioned that some National Statistical Institutes in Europe are now using internet robots to collect prices from the web or scanner data. Particularly, scanner data is used in the CPI analysis in Sweden and internet robots are used in Netherlands (also see Griffioen, de Haan and Willenborg (2015). The characteristics of big data are analysed including volume, velocity and variety. Also, there is an extensive discussion on the benefits (and the potential risks) of using big data in statistical analysis.

Baker at al. (2016) develop a new index of economic policy uncertainty (EPU) based on newspaper coverage frequency. Several types of evidence – including human readings of 12,000 newspaper articles – indicate that this index proxies for movements in policy-related economic uncertainty. The index spikes near tight presidential elections, Gulf Wars I and II, the 9/11 attacks, the failure of Lehman Brothers, the 2011 debt-ceiling dispute and other major battles over fiscal policy. Using firm-level data, they find that policy uncertainty raises stock price volatility and reduces investment and employment in policy-sensitive sectors like defense, healthcare, and infrastructure construction. At the macro level, policy uncertainty innovations foreshadow declines in investment, output, and employment in the United States and, in a panel VAR setting, for 12 major economies. Extending the index back to 1900, EPU rose dramatically in the 1930s (from late 1931) and has drifted upwards since the 1960s.

3.2 Econometric Methodologies for Big Data in Macroeconomics

As we see in Doornik and Hendry (2015), big data offers benefits for statistical modelling, but could lead to a bias result because of an excess of false positives, mistaking correlations for causes, ignoring sampling biases, and selecting by inappropriate methods. When using big data in forecasting the research must be alert that there might be embedding relationships in general initial models, poor quality data on some of the variables, not enough theoretical insights and computational problems due to the number of possible models.

Tibshirani (1996) proposes a new method for estimation in linear models. The LASSO minimises the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives more interpretable models. Simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression.

Efron, Hastie, Johnstone and Tibshirani (2004) introduced Least Angle Regression (LARS), a useful and less greedy version of traditional forward selection methods. Three main properties are derived: (1) A simple modification of the LARS algorithm implements the LASSO, an attractive version of ordinary least squares that constrains the sum of the absolute regression coefficients; the LARS modification calculates all possible Lasso estimates for a given problem. (2) A different LARS modification efficiently implements Forward Stagewise linear regression. (3) A simple approximation for the degrees of freedom of a LARS estimate is available.

LARS and its variants are computationally efficient.

Bai and Ng (2008) suggest to estimate the factors used in the forecasting equation taking into account the goal of forecasting a specific series. Hence, principal components are extracted from "targeted predictors", selected using hard and soft thresholding rules. They consider the LASSO and the Elastic Net as soft-thresholding rules, special cases as we said of the LARS algorithm developed in Efron et al. (2004).

Bai and Ng (2009) note that when it is necessary to select predictors from a large feasible set with no natural ordering, evaluating all possible combinations of predictors can be so computationally costly to become unfeasible. They suggest various forms of "boosting" to select the predictors in factor-augmented autoregressions, including a componentwise approach that treats each lag as a separate variable, and a block-wise approach that treats lags of the same variable jointly. Boosting is a procedure that estimates an unknown function, especially the conditional mean, using M stage-wise regressions.

Buhlmann and Yu (2006) propose Sparse Boosting which is a variant on boosting with the squared error loss. Sparse boosting yields sparser solutions than the previously proposed boosting by minimizing some penalized ℓ_2-loss functions, the FPE model selection criteria, through small step gradient descent. Although boosting may give already relatively sparse solutions, for example corresponding to the soft-thresholding estimator in orthogonal linear models, there is sometimes a desire for more sparseness to increase prediction accuracy and ability for better variable selection: such goals can be achieved with Sparse Boosting.

Candes and Tao (2007) suggest the Dantzig Selector to estimate β in $y = X\beta + \varepsilon$ which is the solution to the ℓ_1 regularization problem: $\min \left\| \widetilde{\beta} \right\|_{\ell_1}$ subject to $\|X^* r\|_{\ell_\infty} \leq (1 + t^{-1})\sqrt{2 \log p}\sigma$ where r is the residual vector, t is a positive scalar, p is the dimension of X and σ is the standard deviation of ε. They show that even if the number of time series observations is much less than p, the estimator achieves a loss within a logarithmic factor of the ideal mean squared error one would achieve with an oracle which would supply perfect information about which coordinates are nonzero, and which were above the noise level. However, Bickel (2009) shows that, under a sparsity scenario, the LASSO estimator and the Dantzig selector exhibit similar behavior.

Avalos, Grandvalet and Ambroise (2007) suggest a method for function estimation and variable selection, specifically designed for additive models fitted by cubic splines. This method involves regularising additive models using the ℓ_1-norm, which generalizes the lasso to the nonparametric setting. As in the linear case, it shrinks coefficients and produces some coefficients that are exactly zero. It gives parsimonious models, selects significant variables, and reveals nonlinearities in the effects of predictors.

Khan (2007) considers the problem of building a linear prediction model when the number of candidate predictors is large and the data possibly contains anomalies that are difficult to visualise and clean. The aim of the study is to predict the non-outlying cases. Therefore, a method that is robust and scalable at the same time is necessary. They consider the stepwise algorithm LARS which is computationally very efficient but sensitive to outliers. They introduce two different approaches to robustify LARS. The plug-in approach replaces the classical correlations in LARS by robust correlation estimates. The cleaning approach first transforms the dataset by shrinking the outliers toward the bulk of the data and then applies LARS to the transformed data.

Keerthi and Shevade (2007) provide an efficient algorithm for tracking the solution curve of sparse logistic regression with respect to the regularization parameter. The algorithm is based on approximating the logistic regression loss by a piecewise quadratic function and then applying a correction to get to the true path.

Haung, Ma and Zhang (2008) study the asymptotic properties of the adaptive LASSO estimators in sparse, high-dimensional, linear regression models when the number of covariates may increase with the sample size. They consider variable selection using the adaptive LASSO, where the ℓ_1-norms in the penalty are re-weighted by data dependent weights. They show that, if a reasonable initial estimator is available, under appropriate conditions, the adaptive Lasso correctly selects covariates with nonzero coefficients with probability converging to one, and that the estimators of nonzero coefficients have the same asymptotic distribution they would have if the zero coefficients were known in advance.

Van De Greer (2008) considers high-dimensional generalized linear models with Lipschitz loss functions, and proves a nonasymptotic oracle inequality for the empirical risk minimizer with LASSO penalty. The penalty is based on the coefficients in the linear predictor, after normalization with the empirical norm.

Lv and Fan (2009) study the properties of regularization methods in model selection and sparse recovery problems under the unified framework of regularized least squares with concave penalties. For model selection, they establish conditions under which a regularized least squares estimator enjoys a nonasymptotic property, called the weak oracle property, where the dimensionality can grow exponentially with sample size. For sparse recovery, they present a sufficient condition that ensures the recoverability of the sparsest solution.

Fraley and Hesterberg (2009) discuss formulations of LARS and LASSO algorithms that extend to datasets in which the number of observations could be so large that it would not be possible to access the matrix of predictors as a unit in computations. Their methods require a single pass through the data for orthogonal transformation, effectively reducing the dimension of the computations required to obtain the regression coefficients and residual sum of squares to the number of predictors, rather than the number of observations. This method could be of particular importance when dealing with big data.

Zhang (2010) proposes MC+, a fast, continuous, nearly unbiased and accurate method of penalized variable selection in high-dimensional linear regression. The MC+ has two elements: a minimax concave penalty (MCP) and a penalized linear unbiased selection (PLUS) algorithm. The MCP provides the convexity of the penalized loss in sparse regions to the greatest extent given certain thresholds for variable selection and unbiasedness. The PLUS computes multiple exact local minimizers of a possibly nonconvex penalized loss function in a certain main branch of the graph of critical points of the penalized loss. Simulation results support their claim of superior variable selection properties and demonstrate the computational efficiency of the proposed method.

Finally, Hesterberg, Choi, Meier and Fraley (2008) offer a detailed review of the least angle and ℓ_1-1 penalized regression.

From another, more econometric, perspective we also have the methods of Principal Components, Partial Least Squares and Bayesian Regression. Factor methods have been at the forefront of developments in forecasting with large data sets and in fact started this literature with the influential work of Stock and Watson (2002a). The defining characteristic of most factor methods is that relatively few summaries of the large data sets are used in forecasting equations which thereby become standard forecasting equations as they only involve a few explanatory variables. The use of principal components (PC) for the estimation of factor models is, by far, the most popular factor extraction method. It has been popularised by Stock and Watson (2002a) and Stock and Watson (2002b), in the context of large data sets, although the idea had been well established in the traditional multivariate statistical literature.

One alternative is dynamic principal components, which, as a method of factor extraction, has been suggested in a series of papers by Forni, Hallin, Lippi and Reichlin (see, e.g., Forni, Hallin, Lippi and Reichlin (2000) among others). Dynamic principal components are extracted in similar fashion to static principal components but , instead of the second moment matrix, the spectral density matrices of the data at various frequencies are used. The components are then used to construct estimates of the common component of the data set, which is a function of the unobserved factors. This method uses leads of the data and, as a result, its application to forecasting has been slow for obvious reasons. Recent work by the developers of the method has addressed this issue (see, e.g., Forni, Hallin, Lippi and Reichlin (2005)). However, overall, empirical evidence suggests that static PC are a more effective and robust technique for forecasting.

Ng (2013) reviews methods for selecting empirically relevant predictors from a set of N potentially relevant ones for the purpose of forecasting a scalar time series. The conventional case when N is smaller than the sample size T is discussed along with the opposite case. Regularisation and dimension reduction methods are reviewed in depth. Irrespective of the model size, there is an unavoidable tension between prediction accuracy and consistent model determination. Ng shows via simulations the improved forecasting performance of selected methods in a one step-ahead horizon.

Barigozzi and Brownless (2013) propose a novel network analysis techniques for multivariate time series. They define the network of a multivariate time series as a graph where nodes denote the components of the process and edges denote nonzero long run partial correlation between two components. For estimation, they introduce an algorithm called nets, based on a two step LASSO regression that allows to estimate large sparse long run partial correlation matrices. The procedure is based on a VAR approximation of the process and its spectral density. They analyze the large sample properties of the estimator and establish conditions for consistent selection and estimation of the nonzero long run partial correlations.

Kim and Swanson (2016) examine whether big data are useful for modelling low frequency macroeconomic variables such as unemployment, GDP and inflation. Using Independent Component Analysis (ICA) and Sparse Principal Component Analysis (SPCA) to reduce the dimension of the model, the authors find that various of their standard benchmarks (including AR models and model averaging) do not dominate more complicated nonlinear methods. Their findings suggest that SPCA (applied using big data) yields Mean Squared Forecast Error-best prediction models in many cases, particularly when coupled with shrinkage. This result provides strong new evidence on the usefulness of sophisticated factor based forecasting and of big data in macroeconometric forecasting.

Partial least squares (PLS) is a relatively new method for estimating regression equations, introduced in order to facilitate the estimation of multiple regressions when there is a large, but finite, amount of regressors. Herman Wold and co-workers introduced PLS regression between 1975 and 1982, see, e.g., Wold (1980). Since then it has received much attention in a variety of disciplines, especially in chemometrics, outside of economics. The basic idea is similar to principal component analysis in that factors or components, which are linear combinations of the original regression variables, are used, instead of the original variables, as regressors. A major difference between PC and PLS is that, whereas in PC regressions the factors are constructed taking into account only the values of the independent variables, in PLS the relationship between the dependent and the independent variables is considered as well in constructing the factors. PLS regression does not seem to have been explicitly considered for data sets with a very large number of series.

Kelly and Pruitt (2015) also offer a discussion of PLS and a slight generalization, labeled 3PRF, which can be easily implemented in three steps, each based on simple OLS regression. Hepenstrick and Marcellino (2016) introduce the mixed frequency version of 3PRF and show that it works quite well for nowcasting GDP growth in many countries, based on datasets of over 800 indicators.

Bayesian regression (BR) is an alternative standard tool for estimation and inference on the parameters of econometric models, and there exists a large variety of approaches for implementing it. The starting point is the specification of a prior distribution for the underlying parameters. Once this is in place, standard Bayesian analysis proceeds by incorporating the likelihood from the observed data to obtain a posterior distribution for the model parameters, which can then be used for a variety of inferential purposes and for forecasting. A special case of the BR is the shrinkage estimator, which shrinks the OLS estimators towards zero, enabling a reduction in variance at the cost of some bias. The variance reduction is a major feature of Bayesian regression that makes it useful in forecasting when large data sets are available. BR can be implemented with Ridge or LASSO estimation; see De Mol et al. (2006) and Groen and Kapetanios (2016) for a comparison of PLS and BR.

Braaksma and Zeelenberg (2015) discuss machine-learning techniques which could be used alongside more traditional methods like Bayesian techniques in the analysis of big data. Based on the experience at Statistics Netherlands they argue that National Statistics Institutes should not be afraid to use these methods, provided that their use is documented and made transparent to users.

3.3 Nowcasting in Macroeconomics

Nowcasting, which is coined by combining the terms "Now" and "Forecasting", has recently become popular in economics due to the increased demand of timely short-term analysis and forecasts of the economy. Data on key measures, such as GDP and its components, are only released after a long delay, and are then subject to subsequent revisions. There is therefore the need to use available, timely and reliable information to form preliminary estimates, i.e., nowcasts, for the key variables of interest.

There is by now a vast literature on nowcasting in macroeconomics, see, e.g., the detailed surveys by Bańbura, Giannone and Reichlin (2011), Bańbura, Giannone, Modugno and Reichlin (2013) and Foroni and Marcellino (2013, 2014). Broadly speaking, nowcasts rely either on regression based methods (Bridge, MIDAS, UMIDAS) or on the Kalman filter applied to handle mixed frequency, and other data irregularities, in VAR and factor models. Both classical and Bayesian estimation methods are available. In this Section we provide a rapid overview of the main contributions and results.

Ghysels et al. (2004) introduced the Mixed Data Sampling (MIDAS) regression models, which are now commonly used in nowcasting applications. The regressions involve time series data sampled at different frequencies. Technically speaking, MIDAS models specify conditional expectations as a distributed lag of regressors recorded at some higher sampling frequencies.

Andreou et al. (2013) introduce easy-to-implement, regression-based methods for predicting quarterly real economic activity that use daily financial data and rely on forecast combinations of MIDAS regressions. They also extract a novel small set of daily financial factors from a large panel of about 1000 daily financial assets. Their analysis is designed to elucidate the value of daily financial information and provide real-time forecast updates of the current (nowcasting) and future quarters of real GDP growth.

Foroni, Marcellino and Schumacher (2015) discuss the pros and cons of unrestricted lag polynomials in MIDAS regressions. They derive unrestricted MIDAS (U-MIDAS) regressions from linear high frequency models, discuss identification issues and show that their parameters can be estimated by ordinary least squares. In Monte Carlo experiments, they compare U-MIDAS with MIDAS with functional distributed lags estimated by non-linear least squares. It is shown that U-MIDAS performs better than MIDAS for small differences in sampling frequencies. However, with large differing sampling frequencies, distributed lag functions outperform unrestricted polynomials. The good performance of U-MIDAS for small differences in frequency is confirmed in empirical applications on nowcasting and short-term forecasting euro area and US gross domestic product growth by using monthly indicators.

Evans (2005) describes a method for calculating daily real-time estimates of the current state of the U.S. economy. The estimates are computed from data on scheduled U.S. macroeconomic announcements using an econometric model that allows for variable reporting lags, temporal aggregation, and other complications in the data. The model can be applied to find real-time estimates of GDP, inflation, unemployment or any other macroeconomic variable of interest. Daily real-time estimates of GDP are constructed that incorporate public information known on the day in question. The real-time estimates produced by the model are uniquely-suited to studying how perceived developments the macro economy are linked to asset prices over a wide range of frequencies. The estimates also provide, for the first time, daily time series that can be used in practical policy decisions.

Giannone, Reichlin and Small (2008) has shown that the process of nowcasting can be formalized in a statistical model which produces predictions without the need for informal judgement. Their method tracks the real-time flow of the type of information monitored by central banks because it can handle large data sets with staggered data-release dates. Each time new data are released, the nowcasts are updated on the basis of progressively larger data sets that, reflecting the unsynchronized data release dates, have a "jagged edge" across the most recent months.

Angelini, Bańbura and Runstler (2008) estimate and forecast growth in euro area monthly GDP and its components from a dynamic factor model which handles unbalanced data sets in an efficient way. They extend the model to integrate interpolation and forecasting together with cross-equation accounting identities and show its improved forecasting abilities.

Angelini, Camba-Mendez, Giannone and Reichlin (2011) evaluate models that exploit timely monthly releases to compute early estimates of current quarter GDP (nowcasting) in the euro area. Their method consists in bridging quarterly GDP with monthly data via a regression on factors extracted from a large panel of monthly series with different publication lags. They show that bridging via factors produces more accurate estimates than traditional bridge equations. They also show that survey data and other 'soft' information are valuable for nowcasting.

Giannone, Reichlin and Simonelli (2009) assess the role of qualitative surveys for the early estimation of GDP in the Euro Area in a model-based automated procedure which exploits the timeliness of their release. The analysis is conducted using both a historical evaluation and a real-time case study on the current conjuncture.

Altissimo, Cristadoro, Forni, Lippi and Veronese (2010) develop a method to obtain smoothing of a stationary time series by using only contemporaneous values of a large data set, so that no end-of-sample deterioration occurs. Their method is applied to the construction of New Eurocoin, an indicator of economic activity for the euro area, which is an estimate, in real time, of the medium- to long-run component of GDP growth. As their data set is monthly and most of the series are updated with a short delay, they are able to produce a monthly real-time indicator with good out-of-sample forecasting properties.

Rossiter (2010) constructs simple mixed-frequency forecasting equations for quarterly global output, imports, and inflation using the monthly global Purchasing Managers Index (PMI). When compared against two benchmark models, the results show that the PMIs are useful for forecasting developments in the global economy. As the forecasts are updated throughout the quarter with the monthly release of the PMI, forecasting performance generally improves.

Yiu and Chow (2010) apply the factor model proposed by Giannone et al. (2008) on a large data set to nowcast (i.e. current-quarter forecast) the annual growth rate of China's quarterly GDP. The identified model generates out-of-sample nowcasts for China's GDP with smaller mean-squared forecast errors than those of the random walk benchmark. Moreover, using the factor model, they find that interest rate data is the single most important block of information to improve estimates of current-quarter GDP in China. Other important blocks are consumer and retail prices data and fixed asset investment indicators

Bańbura and Runstler (2011) derive forecast weights and uncertainty measures for assessing the roles of individual series in a dynamic factor model for forecasting the euro area GDP from monthly indicators. The use of the Kalman smoother allows to deal with publication lags when calculating the above measures. They find that surveys and financial data contain important information for the GDP forecasts beyond the monthly real activity measures. However, this is discovered only if their more timely publication is taken into account properly. Differences in publication lags play a very important role and should be considered in forecast evaluation.

Mariano and Murasawa (2003) introduce a small scale mixed-frequency factor model, developed in a state space framework and estimated by means of the Kalman filer, to extend the Stock–Watson coincident index for the US economy, by combining quarterly real GDP and monthly coincident business cycle indicators.

Mariano and Murasawa (2010) estimate mixed frequency Gaussian vector autoregression (VAR) and factor models for latent monthly real GDP and other coincident indicators. For maximum likelihood estimation of the VAR model, the expectation-maximization (EM) algorithm helps in finding a good starting value for a quasi-Newton method. The smoothed estimate of latent monthly real GDP is a natural extension of the Stock-Watson coincident index.

Frale, Marcellino and Mazzi (2011) propose the EUROMIND, a new monthly indicator of the euro area economic conditions, based on tracking real gross domestic product monthly, relying on information provided in the Eurostat Euro-IND database. EUROMIND, whose underlying methodology extends to the large dataset case the model by Mariano and Murasawa (2003), has several original economic and statistical features. First, it considers both the output and the expenditure sides of the economy, as it provides a monthly estimate of the value added of the six branches of economic activity and of the main gross domestic product components by type of expenditure (final consumption, gross capital formation and net exports), and combines the estimates with optimal weights reflecting their relative precision. Second, the indicator is based on information at both the monthly and the quarterly level, modelled with a dynamic factor specification cast in state space form. Third, since estimation of the multivariate dynamic factor model with mixed frequency data can be numerically complex, computational efficiency is achieved by implementing univariate filtering and smoothing procedures. Finally, special attention is paid to chain linking and its implications, via a multistep procedure that exploits the additivity of the volume measures expressed at the prices of the previous year.

Aastveit and Trovik (2012) find that asset prices on Oslo Stock Exchange is the single most important block of data to improve estimates of current quarter GDP in Norway. They use an approximate dynamic factor model that is able to handle new information as it is released, thus the marginal impact on mean square nowcasting error can be studied for a large number of variables. The high informational content in asset prices is explained by reference to the small size of companies on Oslo Stock Exchange and the small and open nature of the Norwegian economy.

Modugno (2013) proposes a methodology for now-casting and forecasting inflation using data with a sampling frequency which is higher than monthly. The data are modeled as a trading day frequency factor model, with missing observations in a state space representation. In contrast to other existing approaches, the methodology used in this paper has the advantage of modeling all data within a single unified framework which allows one to disentangle the model-based news from each data release and subsequently to assess its impact on the forecast revision. The results show that the inclusion of high frequency data on energy and raw material prices in their data set contributes considerably to the gradual improvement of the model performance. As long as these data sources are included in their data set, the inclusion of financial variables does not make any considerable improvement to the now-casting accuracy.

Lahiri and Monokroussos (2013) study the role of the well-known monthly diffusion indices produced by the Institute for Supply Management (ISM) in nowcasting current quarter US GDP growth. They investigate their marginal impact on the nowcasts when large unbalanced (jagged edge) macroeconomic data sets are used to generate them in real time. They find evidence that the ISM indices are helpful in improving the nowcasts when new ISM information becomes available at the beginning of the month, ahead of other monthly indicators. Furthermore, while the existing literature has focused almost exclusively on manufacturing information, they establish the increasingly significant role of the recently created non-manufacturing ISM diffusion indices in such nowcasting contexts.

Foroni and Marcellino (2014) focus on the different methods which have been proposed in the literature to deal with mixed-frequency and ragged-edge datasets: bridge equations, MIDAS, and mixed-frequency VAR (MF-VAR) models. They find that MIDAS with an AR component performs quite well, and outperforms MFVAR at most horizons. Bridge equations perform well overall. Forecast pooling is superior to most of the single indicator models overall. Pooling information using factor models gives even better results. The best results are obtained for the components for which more economically related monthly indicators are available. Nowcasts of GDP components can then be combined to obtain nowcasts for the total GDP growth.

Kuzin, Marcellino and Schumacher (2013) also find that pooling nowcasts from a large set of small mixed frequency models often outperforms nowcasts from single large models when the target is GDP growth in several industrialised countries.

Aastveit, Gerdrup, Jore and Thorsrud (2014) use U.S. real-time data to produce combined density nowcasts of quarterly GDP growth, using a system of three commonly used model classes: (i) Bridge, (ii) Factor Models and (iii) Mixed-Frequency VAR. They update the density nowcast for every new data release throughout the quarter, and highlight the importance of new information for nowcasting. The results show that the logarithmic score of the predictive densities for U.S. GDP growth increase almost monotonically, as new information arrives during the quarter. While the ranking of the model classes changes during the quarter, the combined density nowcasts always perform well relative to the model classes in terms of both logarithmic scores and calibration tests. The density combination approach is superior to a simple model selection strategy and also performs better in terms of point forecast evaluation than standard point forecast combinations.

Carriero, Clark and Marcellino (2015) develop a method for producing current quarter nowcasts of GDP growth with a (possibly large) range of available within-the-quarter monthly observations of economic indicators, such as employment and industrial production, and financial indicators, such as stock prices and interest rates. In light of existing evidence of time variation in the variances of shocks to gross domestic product, they consider versions of the model with both constant variances and stochastic volatility. They use Bayesian methods to estimate the model, to facilitate providing shrinkage on the (possibly large) set of model parameters and conveniently generate predictive densities. Their method improves significantly on auto-regressive models and performs comparably with survey forecasts. In addition, it provides reliable density forecasts, for which the stochastic volatility specification is quite useful.

Bragoli, Metelli and Modugno (2014) investigate predictions updates. For example, GDP, which belongs to the Unit C1 set of variables, is a quarterly variable but many other macroeconomic indicators (which most of the times are used in the prediction of the GDP) are released with a higher frequency, and financial markets react very strongly to them. However, most of the professional forecasters, including the IMF, the OECD, and most central banks, tend to update their forecasts of economic activity only two to four times a year. The main exception is the Central Bank of Brazil which is responsible for collecting and publishing a daily survey on GDP and other variables. The authors try to evaluate the forecasting performance of the Central Bank of Brazil Survey and compare it with the mechanical forecasts based on state-of-the-art nowcasting techniques. Results indicate that institutional forecasts perform as well as model-based forecasts. The latter finding suggests that, on the one hand, judgmental forecasters do not have computational limitations and are able to incorporate very quickly new information in a way that is as efficient as a machine. On the other hand, it confirms what has been found in other studies, namely that a linear time invariant model does a good job and hence that eventual nonlinearities, time variations and soft information (such as weather conditions or government decisions) that could be incorporated by judgment, do not provide new important information.

4 Econometric Methodologies for Big Data

In this section, we provide a discussion of available econometric and statistical methods to exploit big datasets of indicators for nowcasting or forecasting one or more macroeconomic variables of interest. As it clearly emerged from the previous sections, this is a complex duty, as standard approaches for variable selection or combination are no longer computationally feasible in a big data context.

To set the scene, let y_t, $t = 1, ..., T$, be the target variable and $x_t = (x_{1t}, ..., x_{Nt})'$ be a set of potential predictors, with N very large. We do not assume a particular data generating process for y_t but simply posit the existence of a representation of the form

$$y_t = a + g(x_{1t}, ..., x_{Nt}) + u_t, \tag{1}$$

which implies that $E(u_t|x_{1t}, ..., x_{Nt}) = 0$. We consider an approximating linear representation of the form,

$$y_t = a + \sum_{i=1}^{N} \beta_i x_{it} + u_t, \tag{2}$$

with u_t denoting a martingale difference process and where the set of x_{it}s can also contain products of the original indicators in order to provide a better approximation to (1).

Our main aim is to provide estimates for current and future values of y_t. To do so, we can rely on many approaches, which can be categorised in three main strands. The first strand aims to provide estimates for $\beta = (\beta_1, ..., \beta_N)'$. While ordinary least squares (OLS) is the benchmark method for doing so, it is clear that if N is large this is not optimal or even feasible (when $N > T$). Therefore, other methods need to be used. We consider two classes of methods. The first one is sparse regression, with origins in the machine learning literature, which is discussed in Section 4.1. A main aim there is to stabilise the variability of the estimated β_i. The second class considers the use of a variety of information criteria such as AIC or BIC to select a smaller subset of all the available predictors. As the number of possible permutations of predictors is too large for all permutations to be considered, one needs to consider efficient algorithms for this selection. This is discussed in Section 4.2.

The second strand consists of reducing the dimension of x_t by producing a much smaller set of generated regressors, which can then be used to produce nowcasts and forecasts in standard ways. There are many ways to carry out dimensionality reduction, and the main ones are discussed in Section 4.3[1].

The third strand suggests the use of a (possibly very large) set of small models, one for each available indicator or small subset of them, and then the combination of the resulting many nowcasts or forecasts. Section 4.4 considers standard methods for forecast combination, focusing on those that are more promising in a big data context, and more recent proposals based on either Bayesian or classical model averaging[2].

To conclude, it is important to stress that our setup already makes a very crucial assumption, which is that the available big data are structured as time series. This is by no means a given since big data can be unstructured timewise.

4.1 Machine Learning

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. This field has pioneered many methods that are applicable to large datasets. They include penalised regression and boosting that are discussed in this section but also methods such as principal components that we discuss later (as they have also been analysed extensively outside the machine learning literature). One important issue with machine learning analysis relates to the fact that most work in this field assumes that observations are iid, thereby posing questions on the validity and applicability of the analysis in a time series context.

4.1.1 Penalised Regression

Penalised regression is one of the most popular ways for sparse regression in the literature. Various penalties have been suggested in order to effectively estimate the β_i parameters assigning zeros to the variables which should not be used in the

[1] The above categorisations are not absolute. Some methods have elements that would categorise them as, e.g., either sparse regression or dimensionality reduction ones. One such example is partial least squares, discussed in Section 4.3.2.

[2] The analysis of Bayesian methods is rather limited here, however we refer the reader to Korobilis (2017) and Koop (2018) for more information.

regression (meaning that these are not part of the true model) and consequently in the forecasting exercise. In what follows we denote $\beta_N = (\beta_1, ..., \beta_N)'$ and $x_N = (x_1, ..., x_N)'$.

Ridge Regression

Basic Concept Ridge Regression creates a linear regression model that is penalised with the L2-norm which is the sum of the squared coefficients. This has the effect of shrinking the coefficient values (and the complexity of the model) allowing some coefficients with minor contribution to the response to get close to zero (but not exactly equal to zero). The parameter estimators, $\widehat{\beta}^{Ridge}$, are then computed by solving the following optimisation problem:

$$\min_{\beta_N} \left\{ \sum_{t=1}^{T} \left(y_t - a - \beta_N' x_{t,N} \right)^2 + \lambda \sum_{i=1}^{N} \beta_i^2 \right\}, \tag{3}$$

for given values of a and λ. λ is the penalty parameter. OLS corresponds to the no penalty case, where $\widehat{\beta}^{Ridge} \to \widehat{\beta}^{OLS}$ as $\lambda \to 0$. Also, it can be easily seen that $\widehat{\beta}^{Ridge} \to 0$ as $\lambda \to \infty$. By centering the columns of x, the intercept becomes $\widehat{\alpha} = \overline{y}$. Therefore, we typically center y, x_N and do not include the intercept term.

The variance and bias of the ridge regression estimator can be shown to be

$$\begin{aligned} Var\left(\widehat{\beta}^{Ridge}\right) &= \sigma^2 W x_N' x_N W \\ Bias\left(\widehat{\beta}^{Ridge}\right) &= -\lambda W \beta \end{aligned}$$

where $W = \left(x_N' x_N + \lambda I\right)^{-1}$. It can be shown that the total variance ($\sum_j Var\left(\widehat{\beta}_j\right)$) is a monotone decreasing sequence with respect to λ, while the total squared bias ($\sum_j Bias^2\left(\widehat{\beta}_j\right)$) is a monotone increasing sequence with respect to λ.

Relationship with Bayesian Shrinkage Ridge regression shares many common elements with Bayesian shrinkage. Bayesian regression/shrinkage is a standard tool for providing inference for β in (2) and there exist a large variety of approaches for implementing Bayesian regression. We will provide a brief exposition of this method. A starting point is the specification of a prior distribution for β. Once this is in place, standard Bayesian analysis proceeds by incorporating the likelihood from the observed data to obtain a posterior distribution for β, which can then be used for a variety of inferential purposes, including, of course, forecasting.

A popular and simple implementation of Bayesian regression results in a shrinkage estimator for β in (2) given by

$$\widehat{\beta}_{BRR} = (X'X + \lambda I)^{-1} X'y \tag{4}$$

which is equivalent to that obtained in (3), where $X = (x_1, ..., x_T)'$, $y = (y_1, .., y_T)'$ and λ is a shrinkage scalar parameter, see e.g. Kapetanios, Marcellino and Venditti (2015) for details. This shrinkage estimator shrinks the OLS estimator, given by $(X'X)^{-1}X'y$ towards zero, thus enabling a reduction in the variance of the resulting estimator. This is a major feature of Bayesian regression that makes it useful in forecasting when large data sets are available. This particular implementation of Bayesian regression implies that elements of β are small but different from zero ensuring that all variables in x_t are used for forecasting. In this sense, Bayesian regression can be linked to other data-rich approaches.

It is also worth mentioning that, when N increases, the determinant of $X'X$ tends to decrease, and this is particularly true in macroeconomic applications, where the indicators tend to be correlated. In turn, this creates problems with the inversion of $X'X$ and increases the variance of the OLS estimator. Ridge and Bayesian methods address this issue by replacing $X'X$ with $(X'X + \lambda I)^{-1}$, where the added term λI regularizes the matrix, making it invertible also when $X'X$ is close to non-invertible.

Finally, it is worth noting that other implementations of Bayesian shrinkage with alternative priors lead to estimators that share common elements with other regularisation/sparse regression methods, such as Lasso, as discussed, e.g., in De Mol et al. (2006).

Main Assumptions The standard OLS assumptions are also required for Ridge regression.

Implementation Nowcasting using Ridge regression is straightforward and easy, in particular when implemented in a direct rather than iterated way (e.g., Marcellino, Stock and Watson (2006)). The algorithm can be described in three steps.

1. Replace the loss function with $\min_{\beta_{N,h}} \left\{ \sum_{t=1}^{T} \left(y_t - a - \beta_{N,h}^{'} x_{t-h,N} \right)^2 + \lambda \sum_{i=1}^{N} |\beta_{i,h}| \right\}$, where h is the forecast horizon of interest, and compute $\widehat{\beta_h}^{Ridge}$ for each of a set of values of the tuning parameter λ.
2. Use a cross-validation (CV) scheme to select the preferred tuning parameter, $\widehat{\lambda}$, by minimising the cross-validated squared error risk (or directly the MSE over a rolling window, see e.g. Kapetanios, Marcellino and Venditti (2015)).
3. Using the $\widehat{\beta_h}^{Ridge}$ associated with $\widehat{\lambda}$, produce the $h - \textit{step ahead}$ forecasts as $\widehat{\beta_h}^{Ridge} x_{T,N}$ $(+\widehat{\alpha})$.

The above procedure is then recursively repeated in order to obtain the R out-of-sample forecasts, $\widehat{y}_{T+h}, ..., \widehat{y}_{T+R+h}$.

It must be noted here that the above nowcasting implementation algorithm can be applied in all variable selection methods. Therefore, all the sparse regression methods which follow can produce nowcast estimates in the same fashion.

Since Ridge regression does not set coefficients exactly to zero (unless $\lambda \to \infty$, in which case they are all zero), ridge regression cannot perform variable selection and, even though it might perform well in terms of prediction accuracy, it does not offer a clear interpretation of the resulting forecasts.

LASSO Regression

Basic Concept Least Absolute Shrinkage and Selection Operator (LASSO) creates a regression model that is penalised with the L1-norm which is the sum of the absolute coefficients. Because of the nature of this constraint, it tends to produce some coefficients that are exactly 0 and hence gives more interpretable models. Simulation studies suggest that the LASSO enjoys some of the favourable properties of both subset selection and ridge regression. As originally noted by Tibshirani (1996), the lasso regression is better suited for predictor selection compared to the Ridge regression because the former method performs model/predictors selection keeping those variables which are more suitable for forecasting. The optimisation problem now becomes:

$$\min_{\beta_N} \left\{ \sum_{t=1}^{T} \left(y_t - a - \beta_N^{'} x_{t,N} \right)^2 + \lambda \sum_{i=1}^{N} |\beta_i| \right\}. \tag{5}$$

Although we cannot write the explicit formulas for the bias and variance of the LASSO estimator, the general trend is that the bias increases as λ increases and the variance decreases as λ increases.

Main Assumptions Following Bühlmann and van de Geer (2011), we summarise the key properties and corresponding assumptions for the LASSO. Considering the true model in Equation (2), it is:

$$\frac{1}{T} \sum_{t=1}^{T} \left(x_{t,N} \left(\widehat{\beta}^{LASSO} - \beta \right) \right)^2 = O_P \left(\sum_{i=1}^{N} |\beta_i| \sqrt{\log(N)/T} \right), \tag{6}$$

where $O_P(\cdot)$ is with respect to $N \geq T \rightarrow \infty$. This implies that we achieve consistency of prediction if $\sum_{i=1}^{N} |\beta_i| \ll \sqrt{T/\log(N)}$.

Faster convergence rate and estimation error bounds with respect to the L1- or L2-norm can be achieved using the so-called oracle optimality condition:

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} \left(x_{t,N} \left(\widehat{\beta}^{LASSO} - \beta \right) \right)^2 &= O_P \left(s_0 \phi^{-2} \log(N)/T \right), \\ \sum_{i=1}^{N} \left| \widehat{\beta}_i^{LASSO} - \beta_i \right|^q &= O_P \left(s_0^{1/q} \phi^{-2} \sqrt{\log(N)/T} \right), \; q = \{1, 2\}, \end{aligned} \tag{7}$$

where s_0 equals the true number of non-zero regression coefficients and ϕ^2 is the compatibility constant or restricted eigenvalue which is a number depending on the compatibility between the design and the L1-norm of the regression coefficient. The above rate is optimal up to the $\log(N)$ factor and the restricted eigenvalue ϕ^2.

Additionally to the oracle optimality and assuming the beta-min condition

$$\min_{i \in S_0^c} |\beta_i| \gg \phi^{-2}\sqrt{s_0 \log(N)/T},$$

we obtain the screening variable property

$$P\left[\widehat{S} \supseteq S\right] \rightarrow 1 \quad (N \geq T \rightarrow \infty), \tag{8}$$

where $\widehat{S} = \{i; \widehat{\beta}_i^{LASSO} \neq 0, i = 1, .., N\}$ and $S = \{i; \beta_i \neq 0, i = 1, .., N\}$. Consistent variable selection then means

$$P\left[\widehat{S} = S\right] \rightarrow 1 \quad (N \geq T \rightarrow \infty). \tag{9}$$

The above facts are summarized in the table below.

Property	Design Condition	Size of non-zero coef.
Slow Convergence Rate (Eq. (6))	No requirement	No requirement
Fast Convergence Rate (Eq. (7))	Compatibility	No requirement
Variable Screening (Eq. (8))	Restricted eigenvalue	beta-min condition
Variable Selection (Eq. (9))	Neighborhood Stability	beta-min condition

Adaptive LASSO

Basic Concept Zou (2006) introduces the adaptive LASSO (A-LASSO) estimator where the L1-norms in the penalty are re-weighted. He shows that, if a reasonable initial estimator is available, under appropriate conditions, the A-LASSO correctly selects covariates with nonzero coefficients with probability converging to one, and that the estimators of nonzero coefficients have the same asymptotic distribution they would have if the zero coefficients were known in advance.

The optimisation problem now is:

$$\min_{\beta_N} \left\{ \sum_{t=1}^{T} \left(y_t - a - \beta_N^{'} x_{t,N} \right)^2 + \lambda \sum_{i=1}^{N} \widehat{w}_i \, |\beta_i| \right\}, \tag{10}$$

where $\widehat{w}_i = 1/| \widehat{\beta}_{init,i}|^{\gamma}$, $\widehat{\beta}_{init}$ is an initial estimator and $\gamma > 0$. Usually, the initial estimator is the LASSO estimator with the constraint parameter tuned in the usual way with CV scheme as discussed earlier. Then, in the second stage CV is again used to select the λ parameter in Equation (10).

Main Assumptions Following Haung, Ma and Zhang (2008) we consider the following conditions to hold for the variable selection and asymptotic normality of the A-LASSO in large samples.

1. The errors are iid.
2. The initial estimators $\widehat{\beta}_{init,i}$ are r_T-consistent for the estimation of certain η_{Ti}:
$$r_T \max_{i \leq N} \left| \widehat{\beta}_{init,i} - \eta_{T,i} \right| = O_P\left(1\right), \, r_T \to \infty$$
where η_{Ti} are unknown constants depending on β_N and satisfy
$$\max_{i \notin J_{T1}} \left|\eta_{T,i}\right| \leq M_{T2}, \; \left\{ \sum_{i \in J_{T1}} \left(\frac{1}{\left|\eta_{Ti}\right|} + \frac{M_{T2}}{\left|\eta_{Ti}\right|^2} \right)^2 \right\}^{1/2} \leq M_{T1} = o\left(r_T\right).$$
3. Adaptive irrepresentable condition. For $s_{T1} = \left(\left|\eta_{Ti}\right|^{-1} sgn\left(\beta_i\right), i \in J_{T1} \right)'$ and some $\kappa < 1$
$$\frac{1}{T} \left| x_i^{'} X_1 \sum_{T11}^{-1} s_{T1} \right| \leq \frac{\kappa}{\left|\eta_{Ti}\right|}, \forall i \notin J_{T1}.$$
4. The constants $\{k_T, m_T, \lambda_T, M_{T1}, M_{T2}, b_{T1}\}$ satisfy
$$\left(\log T\right)^{I\{d=1\}} \left\{ \frac{\left(\log k_T\right)^{1/d}}{T^{1/2} b_{T1}} + \left(\log m_T\right)^{1/d} \frac{T^{1/2}}{\lambda_T} \left(M_{T2} + \frac{1}{r_T} \right) \right\} + \frac{M_{T1} \lambda_T}{b_{T1} T} \to 0.$$
5. There exists a constant $\tau_1 > 0$ such that $\tau_{T1} \geq \tau_1$ for all T.

Following Haung et al. (2008), Condition 1 is standard for variable selection in linear regression. Condition 2 assumes that the initial $\widehat{\beta}_{init,i}$ actually estimates some proxy $\eta_{T,i}$ of β_i so that the weights are not too large for $\beta_{0i} \neq 0$ and not too small for $\beta_{0i} = 0$. The adaptive irrepresentable condition becomes the strong irrepresentable condition for the sign-consistency of the Lasso if the $|\eta_{T,i}|$ are identical for all $i \leq N$. Condition 4 restricts the numbers of covariates with zero and nonzero coefficients, the penalty parameter, and the smallest non-zero coefficient. Condition 5 assumes that the eigenvalues of Σ_{T11} are bounded away from zero, which is reasonable since the number of nonzero covariates is small in a sparse model. If the above conditions hold, then $P\left[\widehat{\beta}^{A-LASSO} = \beta\right] \rightarrow 1$.

Elastic Net

Basic Concept Elastic Net (EN) creates a regression model that is penalised with both the L1-norm and L2-norm. Introduced by Zou and Hastie (2005), the elastic net has the effect of effectively shrinking coefficients (as in ridge regression) and setting some coefficients to zero (as in LASSO). The optimisation problem now is:

$$\widehat{\beta}^{naiveEN} = \min_{\beta_N} \left\{ \sum_{t=1}^{T} \left(y_t - a - \beta_N^{'} x_{t,N} \right)^2 + \lambda_1 \sum_{i=1}^{N} |\beta_i| + \lambda_2 \sum_{i=1}^{N} \beta_i^2 \right\}. \tag{11}$$

The above is called the naive elastic net. A correction which leads to the elastic net is then

$$\widehat{\beta}^{EN} = (1 + \lambda_2)\, \widehat{\beta}^{naiveEN}.$$

The correction factor $(1 + \lambda_2)$ is best motivated from the orthonormal design where $\frac{1}{T} x_N^{'} x_N = I$. The main advantage of the elastic net is its usefulness when the number of predictors is much bigger than the number of observations, which is usually the case in our big data context.

The reason for adding an additional squared L2-norm penalty is motivated by Zou and Hastie (2005) as follows. For strongly correlated covariates, the LASSO may select one but typically not both of them (and the non-selected variable can then be approximated as a linear function of the selected one). From the point of view of sparsity, this is what we would like to do. However, in terms of interpretation, we may want to have two even strongly correlated variables among the selected variables: this is motivated by the idea that we do not want to miss a “true” variable due to selection of a “non-true” which is highly correlated with the true one.

SICA

Basic Concept Smooth Integration of Counting and Absolute Deviation (SICA) was introduced by Lv and Fan (2009). The optimisation problem now is:

$$\widehat{\beta}^{SICA} = \min_{\beta_N} \left\{ \sum_{t=1}^{T} \left(y_t - a - \beta_N^{'} x_{t,N} \right)^2 + \lambda \frac{(\alpha+1) \sum_{i=1}^{N} |\beta_i|}{\left(\alpha + \sum_{i=1}^{N} |\beta_i| \right)} \right\}, \tag{12}$$

with $\alpha = 10^{-4}$. With α varying from 0 to ∞, this family provides a smooth homotopy between the L0- and L1-penalties. Each penalty function starts with slope $1 + \alpha^{-1}$ at the origin, passes through the point (1,1), and decreases its slope toward zero over the interval $\lfloor 0, \infty)$.

The above family of penalties satisfy the following condition:

$\rho(t)$ is increasing and concave in $t \in [0, \infty)$ and has a continuous derivative $\rho'(t)$ with $\rho'(0+) \in (0, \infty)$. If $\rho(t)$ is dependent on λ, $\rho'(t; \lambda)$ is increasing in $\lambda \in (0, \infty)$ and $\rho'(0+)$ is independent of λ.

The penalties which satisfy the above condition enjoy the unbiasedness, continuity and sparsity; see Lv and Fan (2009) for more information. The method is attractive to big data modelling as it avoids the single use of L0-norm which is impractical in high dimensions.

Hard Thresholding

Basic Concept Zheng, Fan and Lv (2014) consider sparse regression with a hard thresholding penalty. This approach is motivated by its close connection with in-line image-regularisation, which can be unrealistic to implement in practice but of appealing for its sampling properties and computational advantages. The function to be optimised now is:

$$Q^{TH}(\beta) = \sum_{t=1}^{T} \left(y_t - a - \beta_N^{'} x_{t,N} \right)^2 + \frac{1}{2}\lambda^2 - \left(\lambda - \sum_{i=1}^{N} |\beta_i| \right)_+^2 . \tag{13}$$

In a similar fashion to the restricted eigenvalue condition, Zheng, Fan and Lv (2014) consider the robust spark condition $s < M/2$. The robust spark $M =$

$rspark_c(X)$ of a $T \times N$ design matrix X with bound c is defined as the smallest number τ such that there exists a subgroup of τ columns from $T^{-1/2}X$ such that the corresponding submatrix has a singular value less than the given positive constant c. To ensure model identifiability and reduce the instability in the estimated model we consider the regularised estimator on the union of co-ordinate subspaces $\mathcal{S}_{M/2} = \left\{\beta \in R^N : \beta_0 \leq M/2\right\}$ (where $\beta_0 = \#\left(i|\beta_i \neq 0\right)$ denotes the number of nonzero coefficients) as:

$$\widehat{\beta}^{SICA} = \min_{\beta \in \mathcal{S}_{M/2}} Q^{TH}(\beta). \tag{14}$$

When the size of sparse models exceeds M/2 there is generally no guarantee for model identifiability. Therefore, three regularity conditions must hold:

1. $u_t \sim N\left(0, \sigma^2 I_T\right)$ for some positive σ.

2. It holds that $s < M/2$, $s = o(T)$ and $b = \min_{j \in \sup \beta_0} |\beta_{0,j}| > \{\sqrt{16/c^2} \vee 1\}c^{-1}c_2\sqrt{\{(2s+1)\log(\widetilde{N}/T)\}}$ where M is the robust spark of X with bound c (as defined above), $c_2 \geq \sigma\sqrt{10}$ for some positive constant and $\widetilde{N} = T \vee N$.

3. $\sum_{i=1}^{N} \beta_i^2$ is bounded from below by some positive constant and $\max_{\#(i|\delta_i \neq 0) < M/2, \sum_{i=1}^{N} \delta_i^2 = 1} T^{-1/2} \sum_{i=1}^{N} \left(X_i \delta_i\right)^2 \leq c_3$ for some positive constant c_3.

4.1.2 Spike and Slab Regressions

Spike and Slab regressions were originally proposed by Mitchell and Beauchamp (1988) and recently used by Scott and Varian (2013). The idea is to include an indicator variable $\gamma_i = 1$ if $\beta_i \neq 0$ (i.e. the corresponding regressor is included in the model), and $\gamma_i = 0$ if $\beta_i = 0$. Denoting the nonzero elements of β by β_γ, the spike and slab prior for β and γ can be written as

$$p(\beta, \gamma, \sigma^2) = p(\beta_\gamma|\gamma, \sigma^2) p\left(\sigma^2|\gamma\right) p\left(\gamma\right)$$

The vector of indicator variables γ is assumed to have a Bernoulli prior (independent across elements)

$$p\left(\gamma\right) = \prod_{i=1}^{N} \pi_i^{\gamma_i} \left(1 - \pi_i\right)^{(1-\gamma_i)},$$

so it represents a spike as it places positive probability mass at zero . Conditional on a particular variable being in the equation (that is, conditional on a posterior draw for γ), a standard Normal-Gamma conjugate (typically diffuse) prior for the regression parameters can be used, of the form:

$$\beta_\gamma|\sigma^2, \gamma \sim \mathcal{N}(\beta_{\gamma 0}, \sigma^2\Psi_{\gamma 0}), \quad \left(\sigma^2\right)^{-1} \sim Ga(\alpha_0/2, \delta_0/2)$$

where Ψ_γ denotes the rows and columns of Ψ for which $\gamma_i = 1$. Then, the conditional posterior of β_γ and σ^2 is also Normal-Gamma with closed form parameters

$$\beta_\gamma|\sigma^2, \gamma, y, X \sim N(\hat{\beta}_\gamma, \sigma^2\hat{\Psi}_\gamma), \quad \left(\sigma^2\right)^{-1}|y, X \sim Ga(\hat{\alpha}/2, \hat{\delta}/2) \tag{15}$$

with

$$\begin{aligned}
\hat{\Psi}_\gamma &= \left(X'X + \Psi_{\gamma 0}^{-1}\right)^{-1} \\
\hat{\beta}_\gamma &= \hat{\Psi}_\gamma\left(X'y + \Psi_{\gamma 0}^{-1}\beta_{\gamma 0}\right) \\
\hat{\alpha} &= \alpha_0 + N \\
\hat{\delta} &= \delta_0 + y'y + \beta_{\gamma 0}'\Psi_{\gamma 0}^{-1}\beta_{\gamma 0} + \hat{\beta}_\gamma'\hat{\Psi}_\gamma\hat{\beta}_\gamma.
\end{aligned}$$

Because of conjugacy, the marginal distribution of γ can be analytically derived (up to a proportionality constant):

$$p(\gamma|y, X) \propto \frac{|\Psi_{\gamma 0}|^{-\frac{1}{2}}\, p(\gamma)}{\left|\hat{\Psi}_\gamma\right|^{-\frac{1}{2}}\hat{\delta}^{N/2-1}}. \tag{16}$$

Standard Monte Carlo algorithms can be used to approximate the joint posterior density of the parameters and corresponding probabilities.

4.1.3 Boosting

As an alternative to penalised regression, a number of researchers have developed methods that focus on the predictive power of individual regressors instead of considering all the N covariates together. This approach has led to a variety of alternative specification methods sometimes referred to collectively as “greedy methods”. In this context, regressors are chosen sequentially based on their individual ability to explain the dependent variable. Perhaps the most widely known of such methods, developed in the machine learning literature, is “boosting” whose statistical properties have received considerable attention (Friedman, Hastie and Tibshirani (2000) and Friedman (2001)). Boosting constructs a regression function by considering all regressors one by one in a simple regression setting, and successively selecting the best fitting ones. More details on boosting algorithms for linear models, and their theoretical properties can be found in Bühlmann (2006).

Bühlmann (2006) proves that boosting with the squared error loss, L_2Boosting, is consistent for very high-dimensional linear models, where the number of predictor variables is allowed to grow essentially as fast as $O\left(e^T\right)$, assuming that the true underlying regression function is sparse in terms of the L1-norm of the regression coefficients. The use of an AIC-based method for tuning makes boosting computationally attractive since it is not required to run the algorithm multiple times for cross-validation. We closely follow the same algorithm as in Bühlmann (2006), which can be described as follows.

1. (Initialisation). Let $x_t = (x_{1t}, ..., x_{Nt})'$, $\mathbf{X} = (x_1, ..., x_N)$ and $\mathbf{e} = (e_1, ..., e_T)$. Define the least squares base procedure:

$$\widehat{g}_{\mathbf{X},\mathbf{e}}\left(\mathbf{x}_t\right) = \widehat{\delta}_{\widehat{s}} x_{\widehat{s}t}, \quad \widehat{\delta}_i = \frac{\mathbf{e}'\mathbf{x}_i}{\mathbf{x}_i'\mathbf{x}_i}, \quad \widehat{s} = \min_{1 \leq i \leq N} \left(\mathbf{e} - \widehat{\delta}_i \mathbf{x}_i\right)' \left(\mathbf{e} - \widehat{\delta}_i \mathbf{x}_i\right)$$

2. Given data $\mathbf{X}$ and $\mathbf{y} = (y_1, ..., y_t)'$, apply the base procedure to obtain $\widehat{g}^{(1)}_{\mathbf{X},\mathbf{y}}\left(\mathbf{x}_t\right)$. Set $\widehat{F}^{(1)}\left(\mathbf{x}_t\right) = \upsilon \widehat{g}^{(1)}_{\mathbf{X},\mathbf{y}}\left(\mathbf{x}_t\right)$, for some $\upsilon > 0$. Set $\widehat{s}^{(1)} = \widehat{s}$ and $m = 1$.

3. Compute residuals $\mathbf{e} = \mathbf{y} - \widehat{F}^{(m)}\left(\mathbf{X}\right)$ where $\widehat{F}^{(m)}\left(\mathbf{X}\right) = (\widehat{F}^{(m)}\left(\mathbf{x}_1\right), ..., \widehat{F}^{(m)}\left(\mathbf{x}_T\right))'$ and fit the base procedure to the current residuals to obtain the fit $\widehat{g}^{(m+1)}_{\mathbf{X},\mathbf{e}}\left(\mathbf{x}_t\right)$ and $\widehat{s}^{(m)}$. Update

$$\widehat{F}^{(m+1)}\left(\mathbf{x}_t\right) = \widehat{F}^{(m)}\left(\mathbf{x}_t\right) + \upsilon \widehat{g}^{(m+1)}_{\mathbf{X},\mathbf{e}}\left(\mathbf{x}_t\right).$$

4. Increase the iteration index m by one and repeat step 3 until the stopping

iteration M is achieved. The stopping iteration is given by

$$M = \min_{1 \leq m \leq m_{\max}} AIC_c\left(m\right),$$

for some predetermined large $m_{\max}$ where

$$\begin{aligned} AIC_c\left(m\right) &= \log\left(\sigma^2\right) + \frac{1 + tr\left(\mathcal{B}_m\right)/T}{1 - \left(tr\left(\mathcal{B}_m\right) + 2\right)/T} \\ \sigma^2 &= \frac{1}{T}\left(\mathbf{y} - \mathcal{B}_m\mathbf{y}\right)'\left(\mathbf{y} - \mathcal{B}_m\mathbf{y}\right) \\ \mathcal{B}_m &= I - (I - \upsilon\mathcal{H}^{(\widehat{s}_m)})(I - \upsilon\mathcal{H}^{(\widehat{s}_{m-1})})...(I - \upsilon\mathcal{H}^{(\widehat{s}_1)}) \\ \mathcal{H}^{(j)} &= \frac{\mathbf{x}_j\mathbf{x}_j^{'}}{\mathbf{x}_j^{'}\mathbf{x}_j} \end{aligned}$$

$m_{\max} = 500$ and $\upsilon = \{0.1, 1\}$ values can be used as suggested in the literature.

4.1.4 Boosting-type methods

A related approach that has a number of common elements with boosting and combines penalised regression with greedy algorithms has been put forward by Fan and Lv (2008) and analysed further by, among others, Fan and Song (2010) and Fan, Samworth and Yu (2009). This approach considers marginal correlations between each of the potential regressors and y_t, and selects either a fixed proportion of the regressors based on a ranking of the absolute correlations, or those regressors whose absolute correlation with y_t exceeds a threshold. The latter variant requires selecting a threshold and so in practice the former variant is used. As this approach is mainly an initial screening device, it selects too many regressors but enables dimension reduction in the case of ultra large datasets. As a result, a second step usually is considered, where penalised regression is applied to the regressors selected at the first stage.

4.1.5 One-Covariate at a Time, Multiple Testing Approach

A new approach that is related to those above has recently been proposed by Chudik, Kapetanios and Pesaran (2017). The main idea is to examine the net impact of each potential predictor (θ_i) on the target variable. In a second step, all covariates with statistically significant net impact are included as joint determinants of y_t in a multiple regression setting. The ideal cases are: $\beta_i = 0$ when $\theta_i = 0$ and $\beta_i \neq 0$ when $\theta_i \neq 0$. However, it might be the case when $\beta_i \neq 0$ when $\theta_i = 0$ and $\beta_i = 0$ when $\theta_i \neq 0$. In these cases, it is required to iteratively test the statistical contribution of non-selected covariates (again one at a time) to the unexplained part of y_t. While the initial regressions of this procedure is similar to the approach of Fan and Lv (2008), the multiple testing element provides additional value to the approach.

The proposed method, which is referred to as One-Covariate at a Time Multiple Testing (OCMT) approach, is computationally simple and fast even for extremely large datasets. It is based on statistical inference and is easier to interpret, relates to the classical statistical analysis, allows working under more general assumptions, and performs equally well in small and large samples.

4.1.6 Cluster Analysis

A further method in the machine learning literature which has not yet been discussed is cluster analysis. Cluster analysis is the assignment of a set of observations into subsets (i.e. clusters) so that observations within the same cluster are similar according to some predesignated criterion or criteria, while observations drawn from different clusters are dissimilar. Different clustering techniques make different assumptions on the structure of the data, often defined by some similarity metric and evaluated for example by internal compactness (similarity between members of the same cluster) and separation between different clusters. Some indicative papers in the literature include Gershenfeld, Schoner and Metois (1999) who introduced the cluster-weighted modelling for time series analysis, McCallum, Nigam and Ungar (2000) analysing the canopy clustering algorithm and Dhillon and Modha (2001) who deal with categorical series clustering in big text data. Finally, it is worth noting the work of Dablemont, Simon, Lendasse, Ruttiens, Blayo and Verleysen (2003) and Martinez Alvarez, Troncoso, Riquelme and Riquelme (2007) who provide discussions of clustering in relation to forecasting time series.

Cluster analysis, as mentioned above, can be applied to various small and big datasets. In what we are concerned, clustering could have two applications: (i) grouping the unbalanced big data into time series, and (ii) grouping the actual time series of the predictors. The goals in clustering time series are: (i) to capture global trends, (ii) to identify signals which may or may not be periodical, and (iii) to discover possibly unknown patterns. There are four major categories of time series clustering methods: (i) the relocation clustering, (ii) the Agglomerative hierarchical clustering, (iii) k-Means and fuzzy c-means and (iv) Self-organising maps. A detailed review of these methods can be found in Liao (2005).

The clustering output depends on the function used to measure the similarity between the data. These functions could be a combination of simple statistics like the minimum/maximum, the mean/median/mode, the first/third quartile, the interquartile range, the standard deviation, etc., or a distance-based measure such as the Euclidean distance, Kullback-Leibler distance, etc. Some examples of clustering with real data are: (i) clustering seasonality patterns in retail data (see Moller-Levet, Klawonn, Cho and Woklenhauer, 2003), (ii) discovery patterns from stock time series (see Fu, Chung, Ng, and Luk, 2001), and (iii) clustering personal income series (see Kalpakis, Gada and Puttagunta, 2001).

In order to forecast time series using clustering one could adopt the approach as in Hyndman, Ahmed, Athanasopoulos and Shang (2011). The researcher could forecast each time series independently and then combine the forecasts to obtain the predictions in clusters. Subsequently, the clustered forecasts are combined, or averaged using estimated regression coefficients, to estimate the predictions for the dependent variable.

4.1.7 Other machine learning methods

In this last section we briefly mention other machine learning approaches that might be useful although their usefulness is curtailed for a variety of reasons. These include bagging (see Breiman (1996)), random forest (see Breiman (2001), Shi and Horvath (2006)), logistic regression and artificial neural networks. The last one deserves special mention. Artificial neural networks (ANNs) are a family of models inspired by biological neural networks and are used to approximate functions that can depend on a large number of inputs and are unknown. They are generally presented as systems of interconnected components which exchange messages between each other.

The connections have weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning; see Blake and Kapetanios (2010) for more detailed information. While their application to econometric nowcasting has produced mixed results, we note them as they have recently given rise to methods collectively known as deep learning. Deep learning is essentially a multilayered ANN model, which has been shown to have good pattern recognition properties; see Hinton and Salakhutdinov (2006). While this set of methods might be worth further investigation, they rely on a large T and not so large N, which is not suited for the nowcasting problems under consideration. The need for a large T arises due to the fact that the multilayered ANN model has a considerable number of parameters that need to be estimated.

4.2 Heuristic Optimisation

Another approach to variable selection is the direct use of a model selection criterion (such as Akaike (1974) (AIC), Bayesian (Schwarz (1978)) (BIC) , Hannan and Quinn (1979) (HQ) etc.). For example, consider the generalised model in Equation (2). Let $\boldsymbol{\mathcal{I}} = (\boldsymbol{\mathcal{I}}_1, \ldots, \boldsymbol{\mathcal{I}}_N)'$ denote a vector of zeros and ones (which we will refer to as string). Let $\boldsymbol{\mathcal{I}}_i = 1$, if x_{it} belongs to the true model and zero otherwise. We wish to estimate $\boldsymbol{\mathcal{I}}$. We could start by selecting some of the predictors, estimate the model and calculate the criterion value. Then, we could repeat the same procedure for all possible models and select the one which optimises the selection function.

The generic form of such criteria is usually,

$$IC(\boldsymbol{\mathcal{I}}) = -2L(\boldsymbol{\mathcal{I}}) + C_T(\boldsymbol{\mathcal{I}}) \tag{17}$$

where $L(\boldsymbol{\mathcal{I}})$ is the log-likelihood of the model associated with string $\boldsymbol{\mathcal{I}}$ and $C_T(\boldsymbol{\mathcal{I}})$ is the penalty term associated with the string $\boldsymbol{\mathcal{I}}$. The three most usual penalty terms are $2\tilde{m}(\boldsymbol{\mathcal{I}})$, $ln(T)\tilde{m}(\boldsymbol{\mathcal{I}})$ and $2ln(ln(T))\tilde{m}(\boldsymbol{\mathcal{I}})$ associated with AIC, BIC and HQ information criteria. $\tilde{m}(\boldsymbol{\mathcal{I}})$ is the number of free parameters associated with the modelling of the dataset associated with $\boldsymbol{\mathcal{I}}$. Note that, in this case, $\tilde{m}(\boldsymbol{\mathcal{I}}) = \boldsymbol{\mathcal{I}}'\boldsymbol{\mathcal{I}}$. It is straightforward under relatively weak conditions on x_{jt} and u_{jt}, and using the results of, say, Sin and White (1996), to show that the string which minimises $IC(.)$ will converge to the true string with probability approaching one as $T \to \infty$ as long as (i) $C_T(\boldsymbol{\mathcal{I}}) \to \infty$ and (ii) $C_T(\boldsymbol{\mathcal{I}})/T \to 0$.

More specifically, the assumptions needed for the results of Sin and White (1996) to hold are mild and can be summarised as follows, assuming estimation of the models is undertaken in the context of Gaussian or pseudo maximum likelihood (which in the simplest case, of spherical errors, is equivalent to OLS): (i) Assumption A of Sin and White (1996) requires measurability, continuity and twice differentiability of the log-likelihood function and a standard identifiability assumption; (ii) A uniform weak law of large numbers for the log-likelihood of each observation and its second derivative; (iii) A central limit theorem for the first derivative of the log-likelihood of each observation. (ii) and (iii) above can be obtained by assuming, e.g., that x_{jt} are weakly dependent, say, near epoch dependent, processes and u_{jt} are martingale difference processes. Hence, it is clear that consistency of model selection as long as the penalty related conditions hold is straightforwardly obtained. Note that unlike BIC and HQ which consistently estimate the true model in the sense of Sin and White (1996), AIC is inconsistent, in this sense, since C_T remains bounded, as $T \to \infty$, contravening the first penalty related condition given in the preceding paragraph.

The problem is of course how to minimise the information criterion. For small dimensional x_t, evaluating the information criterion for all strings may be feasible, as, e.g., in lag order selection. In the case of lag selection the problem is made easier by the fact that there exists a natural ordering of the variables, although in many cases such an ordering may not be the optimal basis for a search algorithm. In the general variable selection case, as soon as N exceeds say 50 or 60 units, this strategy is bound to fail. Since $\boldsymbol{\mathcal{I}}$ is a binary sequence there exist 2^N strings to be evaluated. For example, when $N = 50$ and optimistically assuming that 100000 strings can be evaluated per second, we still need about 357 years for an evaluation of all strings. Clearly this is infeasible.

Although this is a maximisation problem, standard maximisation algorithms do not apply due to the discreteness of the domain over which the objective function (information criterion) needs to be optimised. To overcome this difficulty, several heuristic optimisation approaches have been suggested, including among the main ones: simulated annealing, genetic algorithm, MC^3 and sequential testing.

4.2.1 Simulated Annealing (SA)

This algorithm provides a local search for the minimum (or maximum) of a function, in our case is Equation (17). The concept is originally based on the manner in which liquids freeze or metals recrystalize in the process of annealing. In an annealing process a melt, initially at high temperature and disordered, is slowly cooled so that the system at any time is approximately in thermodynamic equilibrium. As cooling proceeds, the system becomes more ordered and approaches a 'frozen' ground state. The analogy to an optimisation problem is as follows: the current state of the thermodynamic system is analogous to the current solution to the optimisation problem, the energy equation for the thermodynamic system is analogous to the objective function, and the ground state is analogous to the global optimum. An early application of simulated annealing in econometrics is the work of Goffe, Ferrier and Rogers (1994), who suggested that simulated annealing could be used to optimise the objective function of various econometric estimators.

Below, we give a description of the algorithm together with the necessary arguments that illustrate its validity in our context. We describe the operation of the algorithm when the domain of the function (information criterion) is the set of binary strings i.e. $\{\boldsymbol{\mathcal{I}} = (\mathcal{I}_1, \ldots, \mathcal{I}_N)' | \mathcal{I}_i \in \{0, 1\}\}$.

Each step of the algorithm works as follows, starting from an initial string $\boldsymbol{\mathcal{I}}_0$.

1. Using $\boldsymbol{\mathcal{I}}_i$ choose a neighboring string at random, denoted $\boldsymbol{\mathcal{I}}^*_{i+1}$. We discuss the definition of a neighborhood below.

2. If $IC(\boldsymbol{\mathcal{I}}_i) > IC(\boldsymbol{\mathcal{I}}^*_{i+1})$, set $\boldsymbol{\mathcal{I}}_{i+1} = \boldsymbol{\mathcal{I}}^*_{i+1}$. Else, set $\boldsymbol{\mathcal{I}}_{i+1} = \boldsymbol{\mathcal{I}}^*_{i+1}$ with probability $e^{(IC(\boldsymbol{\mathcal{I}}^*_i) - IC(\boldsymbol{\mathcal{I}}_{i+1}))/T_i}$ or set $\boldsymbol{\mathcal{I}}_{i+1} = \boldsymbol{\mathcal{I}}_i$ with probability $1 - e^{(IC(\boldsymbol{\mathcal{I}}^*_i) - IC(\boldsymbol{\mathcal{I}}_{i+1}))/T_i}$.

Heuristically, the term T_i gets smaller making it more difficult, as the algorithm proceeds, to choose a point that does not decrease $IC(.)$. The issue of the neighborhood is extremely relevant. What is the neighborhood? Intuitively, the neighborhood could be the set of strings that differ from the current string by one element of the string. But this may be too restrictive. We can allow the algorithm to choose at random, up to some maximum integer (say h), the number of string elements at which the string at steps i and $i + 1$ will differ. So the neighborhood is all strings with up to h different bits from the current string. Another issue is when to stop the algorithm. There are a number of alternatives in the literature. We have chosen to stop the algorithm if it has not visited a string with lower $IC(.)$ than the current minimum for a prespecified number of steps (B_v) (Steps which stay at the same string do not count) or if the number of overall steps exceeds some other prespecified number (B_s). All strings visited by the algorithm are stored and the best chosen at the end rather than the final one.

The simulated annealing algorithm has been proven by Hajek (1998) to converge asymptotically, i.e. as $i \to \infty$, to the maximum of the function as long as $T_i = T_0/ln(i+1)$ for some T_0 for sufficiently large T_0. In particular, for almost sure convergence to the minimum it is required that $T_0 > d^*$. d^* denotes the maximum depth of all local minima of the function $IC(.)$. Heuristically, the depth of a local minimum, $\boldsymbol{\mathcal{I}}_1$, is defined as the smallest number $E > 0$ such that the function exceeds $IC(\boldsymbol{\mathcal{I}}_1) + E$ during its trajectory[3] from this minimum to any other local minimum, $\boldsymbol{\mathcal{I}}_2$, for which $IC(\boldsymbol{\mathcal{I}}_1) > IC(\boldsymbol{\mathcal{I}}_2)$.

This condition needs to be made specific for the problem at hand. We thus need to discuss possible strategies for determining d^* for model searches using information criteria. It is reasonable to assume that the space of models searched via information criteria only includes models with a prespecified maximum number of variables, otherwise problems caused by the lack of degrees of freedom will arise. Then, a possible upper limit for d^* is $2L(\boldsymbol{\mathcal{I}_B}) - 2L(\boldsymbol{\mathcal{I}_A})$ where $L(\boldsymbol{\mathcal{I}_A})$ is the likelihood associated with a regression containing just a constant term and $L(\boldsymbol{\mathcal{I}_B})$ is the likelihood associated with a regression containing the maximum allowable number of variables. Of course, there are many possible sets of variables that contain the maximum allowable number of variables. For this reason we remove the penalty terms and focus on likelihoods. This make it more likely that $-2L(\boldsymbol{\mathcal{I}_B})$, for some random $\mathcal{I}_B$ that specifies use of the maximum allowable number of variables, is a lower bound for the optimum value taken by the information criterion.

[3] A trajectory from $\boldsymbol{\mathcal{I}}_1$ to $\boldsymbol{\mathcal{I}}_2$ is a set of strings, $\boldsymbol{\mathcal{I}}_{11}, \boldsymbol{\mathcal{I}}_{12}, \ldots, \boldsymbol{\mathcal{I}}_{1p}$, such that (i) $\boldsymbol{\mathcal{I}}_{11} \in N(\boldsymbol{\mathcal{I}}_1)$, (ii) $\boldsymbol{\mathcal{I}}_{1p} \in N(\boldsymbol{\mathcal{I}}_2)$ and (iii) $\boldsymbol{\mathcal{I}}_{1i+1} \in N(\boldsymbol{\mathcal{I}}_{1i})$ for all $i = 1, \ldots, p$, where $N(\boldsymbol{\mathcal{I}})$ denotes the set of strings that make up the neighborhood of $\boldsymbol{\mathcal{I}}$.

4.2.2 Genetic Algorithms (GA)

The motivating idea of genetic algorithms is to start with a population of binary strings which then evolve and recombine to produce new populations with 'better' characteristics, i.e. lower values for the information criterion. We start with an initial population represented by an $N \times m$ matrix made up of 0's and 1's. Columns represent strings. m is the chosen size of the population. The theory of genetic algorithms suggests that the composition of the initial population does not matter. Hence, this is generated randomly. Denote this population matrix by P_0. The genetic algorithm involves defining a transition from $\mathbf{P}_i$ to $\mathbf{P}_{i+1}$. Following Kapetanios (2007), the algorithm could be described in the following steps:

1. For $\mathbf{P}_i$ create a $m \times 1$ 'fitness' vector, $\mathbf{p}_i$, by calculating for each column of $\mathbf{P}_i$ its 'fitness'. The choice of the 'fitness' function is completely open and depends on the problem. For our purposes it is the opposite of the information criterion. Normalise $\mathbf{p}_i$, such that its elements lie in $(0, 1)$ and add up to 1. Denote this vector by $\mathbf{p}_i^*$. Treat $\mathbf{p}_i^*$ as a vector of probabilities and resample m times out of $\mathbf{P}_i$ with replacement, using the vector $\mathbf{p}_i^*$ as the probabilities with which each string with be sampled. So 'fit' strings are more likely to be chosen. Denote the resampled population matrix by $\mathbf{P}_{i+1}^1$.

2. Perform cross over on $\mathbf{P}_{i+1}^1$. For cross over we do the following: Arrange all strings in $\mathbf{P}_{i+1}^1$, in pairs (assume that m is even) where the pairings are randomly drawn. Denote a generic pair by $(a_1^\alpha, a_2^\alpha, \dots, a_N^\alpha)$, $(a_1^\beta, a_2^\beta, \dots, a_N^\beta)$. Choose a random integer between 2 and $N-1$. Denote this by j. Replace the pair by the following pair: $(a_1^\alpha, a_2^\alpha, \dots, a_j^\alpha, a_{j+1}^\beta, \dots, a_N^\beta)$, $(a_1^\beta, a_2^\beta, \dots, a_j^\beta, a_{j+1}^\alpha, \dots, a_N^\alpha)$. Perform cross over on each pair with probability p_c. Denote the new population by $\mathbf{P}_{i+1}^2$. Usually p_c is set to some number around 0.5-0.6.

3. Perform mutation on $\mathbf{P}_{i+1}^2$. This amounts to flipping the bits (0 or 1) of $\mathbf{P}_{i+1}^2$ with probability p_m. p_m is usually set to a small number, say 0.01. After mutation the resulting population is $\mathbf{P}_{i+1}$.

These steps are repeated a prespecified number of times (B_g). Each set of steps is referred to as generation in the genetic literature. If a string is to be chosen this is the one with maximum fitness. For every generation, the identity of the string with maximum 'fitness' is stored. Further, this string is allowed to remain intact for that generation. So it gets chosen with probability one in step 1 of the algorithm and does not undergo neither cross-over nor mutation. At the end of the algorithm the string with the lowest information criterion value over all members of the populations and all generations is chosen. One can think of the transition from one string of maximum fitness to another as a Markov Chain. So this is a Markov Chain algorithm. In fact, the Markov chain defined over all possible strings is time invariant but not irreducible as at least the $m-1$ least fit strings will never be picked. To see this note that in any population there will be a string with more fitness than that of the $m-1$ worst strings.

There has been considerable work on the theoretical properties of genetic algorithms. Hartl and Belew (1990) have shown that with probability approaching one, the population at the n-th generation will contain the global maximum as $n \to \infty$. Perhaps the most relevant result from that work is Theorem 4.1 of Hartl and Belew (1990). This theorem states that as long as (i) the sequence of the maximum fitnesses in the population across generations is monotonically increasing, and (ii) any point in the model space is reachable from any other point by means of mutation and cross-over in a finite number of steps then the global maximum will be attained as $n \to \infty$. Both these conditions hold for the algorithm described above. The first condition holds by the requirement that the string with the maximum fitness is always kept intact in the population. The second condition holds since any string of finite length can be obtained from another by cross-over and mutation with non-zero probability in a finite number of steps. For more details on the theory of genetic algorithms see also Morinaka, Yoshikawa and Amagasa (2001).

4.2.3 MC^3

This algorithm is similar to simulated annealing for the construction of its steps. This similarity is, in fact, the main reason why we consider Bayesian methods here. The MC^3 algorithm defines a search path in the model space just like the simulated annealing algorithm we considered in the previous subsection. As a result, we refer to the setup of the previous subsection to minimise duplication for the exposition.

The difference between SA and MC^3 is the criterion used to move from one string to the other at step i. Here, the Bayes factor for string (model) $i+1$ versus string (model) i is used. This is denoted by $B_{i+1,i}$. The chain moves to the $i+1$ string with probability $min(1, B_{i+1,i})$. This is again a Metropolis-Hastings type algorithm. Following Fernandez, Ley and Steel (2001), the Bayes factor is given by:

$$B_{i+1,i} = \left(\frac{g_{0i+1}}{g_{0i+1}+1}\right)^{k_{i+1}/2} \left(\frac{g_{0i}+1}{g_{0i}}\right)^{k_i/2} \left(\frac{\frac{1}{g_{0i}+1}RSS_i + \frac{g_{0i}}{g_{0i}+1}TSS}{\frac{1}{g_{0i+1}+1}RSS_{i+1} + \frac{g_{0i+1}}{g_{0i+1}+1}TSS}\right)^{(T-1)/2}, \tag{18}$$

where RSS_i is the sum of squared residuals of the i-th model, TSS is the sum of the squared deviations from the mean for the dependent variable, k_i is the number of variables in model i and g_{0i} is a model specific constant relating to the prior relative precision. The results of Fernandez et al. (2001) suggest that for consistent model selection g_{0i} should be set to $1/T$. This is associated with prior 'a' in the terminology of subsection 4.2 of Fernandez et al. (2001), to whom we refer for more details. The chosen model is the one minimising the information criterion among all models visited by the MC^3 algorithm. This follows from the results of Appendix A.3 of Fernandez et al. (2001) concerning the asymptotic equivalence between consistent information criteria and the Bayes factor in Equation (18).

4.2.4 Sequential Testing (ST)

A general regression specification is considered and tested for misspecification using a battery of specification tests such as tests for residual autocorrelation and ARCH and tests for structural breaks. Then, a sequential testing procedure is used to remove insignificant regressors from this specification making sure that resulting specifications are acceptable using misspecification tests. This algorithm provides a tractable formalisation of the general-to-specific methodology advocated by David Hendry and his co-authors, and discussed in some detail in a number of paper such as, e.g., Hendry (1995) and Hendry (1997) (see also Brüggemann, Krolzig and Lütkepohl (2003) for an application of this methodology to model reduction in VAR processes). Also, recent work by Doornik and Hendry (2015) sheds some extra light on the use of *Autometrics*[4] in statistical model selection with big data.

A detailed description of the algorithm we use is given in steps A-H of Hoover and Perez (1999). The only modifications we suggest to this algorithm are as follows: (i) All possible search paths, rather than only 10, are considered. (ii) In step B(d) we use $CUSUM^2$ instead of Chow as a stability test. (iii) No out-of-sample evaluation is undertaken, since this would change the information set for the other algorithms

4.3 Dimensionality Reduction

Another set of methods for modelling with big data involves the adoption of dimension reduction via techniques which do not require or impose any iid assumptions. In what follows we discuss Principal Components Analysis, Partial Least Squares, and Sparse Principal Component Analysis.

4.3.1 Principal Component Analysis

The most widely used class of data-rich forecasting methods are factor methods. Factor methods have been at the forefront of developments in forecasting with large data sets and in fact started this literature with the influential work of Stock and Watson (2002a). The defining characteristic of most factor methods is that relatively few summaries of the large data sets are used in forecasting equations, which thereby become standard forecasting equations as they only involve a few explanatory variables.

The main assumption is that the co-movements across the indicator variables x_t, where $x_t = (x_{1t} \cdots x_{Nt})'$ is a vector of dimension $N \times 1$, can be captured by a $r \times 1$ vector of unobserved factors $F_t = (F_{1t} \cdots F_{rt})'$, i.e.,

$$\tilde{x}_t = \Lambda' F_t + e_t \tag{19}$$

where $\tilde{x}_t$ may be equal to x_t or may involve other variables, such as lags, leads or products of the elements of x_t, and Λ is an $r \times N$ matrix of parameters describing how the individual indicator variables relate to each of the r factors, which we denote with

[4] Autometrics is a software developed by Hendry and Doornik which makes use of sequential testing.

the terms 'loadings'. In (19) e_t is a zero-mean $I(0)$ vector of errors that represent, for each indicator variable, the fraction of dynamics unexplained by F_t, the 'idiosyncratic components'. The number of factors is assumed to be finite. So, implicitly, in (2) $\alpha' = \tilde{\alpha}'\Lambda\tilde{x}_t$, where $F_t = \Lambda\tilde{x}_t$, which means that a small, r, number of linear combinations of $\tilde{x}_t$ represent the factors and act as the predictors for y_t, the target variable. The main difference between different factor methods relates to how Λ and the factors are estimated.

The use of PCA for the estimation of factor models is, by far, the most popular factor extraction method. It has been popularised by Stock and Watson (2002a, 2002b), in the context of large data sets, although the idea had been well established in the traditional multivariate statistical literature. The method of principal components is simple. Estimates of Λ and the factors F_t are obtained by solving:

$$V(r) = \min_{\Lambda,F} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\tilde{x}_{it} - \lambda_i' F_t)^2, \tag{20}$$

where λ_i is an $r \times 1$ vector of loadings that represent the N columns of $\Lambda = (\lambda_1 \cdots \lambda_N)$. One, non-unique, solution of (20) can be found by taking the eigenvectors corresponding to the r largest eigenvalues of the second moment matrix $X'X$, which then are assumed to represent the rows in Λ, and the resulting estimate of Λ provides the forecaster with an estimate of the r factors $\hat{F}_t = \hat{\Lambda}\tilde{x}_t$. To identify the factors up to a rotation, the data are usually normalized to have zero mean and unit variance prior to the application of principal components; see Stock and Watson (2002a) and Bai (2003). We note that factor estimates obtained via PC estimation are $\min(\sqrt{N}, T)$-consistent. Further, if $\sqrt{T}/N = o(1)$, using estimated factors rather than true factors in predictive regressions produces negligible estimation errors.

PC estimation of the factor structure is essentially a static exercise as no lags or leads of x_t are considered. One alternative is dynamic principal components, which, as a method of factor extraction, has been suggested in a series of papers by Forni, Hallin, Lippi and Reichlin (see, e.g., Forni, Hallin, Lippi and Reichlin (2000) among others) and is designed to address this issue. Dynamic principal components are extracted in a similar fashion to static principal components but, instead of the second moment matrix, the spectral density matrix of the data at various frequencies is used. The dynamic PCs are then used to construct estimates of the common component of the data set, which is a function of the unobserved factors. The basic version of this method uses leads of the data, making it not suited in a forecasting context, but later work by the developers of the method has addressed this issue (see, e.g., Forni, Hallin, Lippi and Reichlin (2005)).

4.3.2 Partial Least Squares

Partial least squares (PLS) is a relatively new method for estimating regression equations, introduced in order to facilitate the estimation of multiple regressions when there is a large, but finite, amount of regressors[5]. The basic idea is similar to Principal Component Analysis (PCA) in that factors or components, which are linear combinations of the original regression variables, are used, instead of the original variables, as regressors. PLS regression does not seem to have been explicitly considered for data sets with a very large number of series, i.e., when N is assumed in the limit to converge to infinity.

There are a variety of definitions for PLS and accompanying specific PLS algorithms that inevitably have much in common. A conceptually powerful way of defining PLS is to note that the PLS factors are those linear combinations of x_t, denoted by Υx_t, that give maximum covariance between y_t and Υx_t while being orthogonal to each other. Of course, in analogy to PC factors, an identification assumption is needed, to construct PLS factors, in the usual form of a normalization.

A simple algorithm to construct k PLS factors is discussed among others, in detail, in Helland (1990). Assuming for simplicity that y_t has been demeaned and x_t have been normalized to have zero mean and unit variance, a simplified version of the algorithm is given below.

1. Set $u_t = y_t$ and $v_{i,t} = x_{i,t}$, $i = 1, ...N$. Set $j = 1$.
2. Determine the $N \times 1$ vector of indicator variable weights or loadings $w_j = (w_{1j} \cdots w_{Nj})'$ by computing individual covariances: $w_{ij} = Cov(u_t,\ v_{it})$, $i =$

[5] Herman Wold and co-workers introduced PLS regression between 1975 and 1982, see, e.g., Wold (1982). Since then it has received much attention in a variety of disciplines, especially in chemometrics, outside of economics.

$1, ..., N$. Construct the j-th PLS factor by taking the linear combination given by $w_j' v_t$ and denote this factor by $f_{j,t}$.

3. Regress u_t and $v_{i,t}$, $i = 1, ..., N$ on $f_{j,t}$. Denote the residuals of these regressions by $\tilde{u}_t$ and $\tilde{v}_{i,t}$ respectively.

4. If $j = k$ stop, else set $u_t = \tilde{u}_t$, $v_{i,t} = \tilde{v}_{i,t}$ $i = 1, .., N$ and $j = j + 1$ and go to step 2.

This algorithm makes clear that PLS is computationally tractable for very large data sets. Once PLS factors are constructed y_t can be modeled or forecast by regressing y_t on $f_{j,t}$, $j = 1, ..., k$. Helland (1990) provides a general description of the partial least squares (PLS) regression problem. Helland (1990) shows that the estimates of the coefficients α in the regression of y_t on x_t, as in Equation (2), obtained implicitly via PLS Algorithm and a regression of y_t on $f_{j,t}$ $j = 1, ..., k$, are mathematically equivalent to

$$\widehat{\beta}_{PLS} = V_k(V_k'X'XV_k)^{-1}V_k'X'y \tag{21}$$

with $V_{k_1} = (X'y \quad X'XX'y \quad \cdots \quad (X'X)^{k-1}X'y)$, $X = (x_1 \cdots x_T)'$ and $y = (y_1 \cdots y_T)'$. Thus, (21) suggests that the PLS factors that result from the PLS Algorithm span the Krylov subspace generated by $X'X$ and $X'y$, resulting in valid approximations of the covariance between y_t and x_t.

A major difference between PC and PLS is that, whereas in PC regressions the factors are constructed taking into account only the values of the x_t variables, in PLS, the relationship between y_t and x_t is considered as well in constructing the factors.

Recently, Kelly and Pruit (2015) and Groen and Kapetanios (2016) have extended and provided theoretical results on PLS, showing that it can be also applied in the large N context, while Hepenstrick and Marcellino (2016) have introduced the mixed frequency version and provided empirical evidence in favour of its use for nowcasting with very large datasets.

4.3.3 Sparse Principal Component Analysis

Empirical studies in the literature support the argument that standard PC does a good job in dimension reduction. A number of forecasting applications show that when the linear combinations of input variables is used (instead of the whole set of variables) the forecast error is reduced. However, a disadvantage of standard PC is that the principal components are combinations of all input variables. Sparse Principal Component Analysis (Sparse PC), introduced by Zou, Hastie and Tibshirani (2006), combines aspects of sparse regression and PC. In particular, the principal components are derived using linear combinations of some of the variables.

Given an integer k with $1 \leq k \leq N$ Sparse PC is aiming to maximize the variance along a vector v while constraining its cardinality:

$$\begin{gathered}\max v'\Sigma v \\ s.t. \sum_{i=1}^{N} v_i^2 = 1 \\ \#\left(i|v_i \neq 0\right) \leq k\end{gathered}$$

where Σ denotes the sample covariance matrix. The first constraint ensures that v is a unit vector and the second constraint is the L0-norm, i.e. the number of the non-zero components in v is less than k. If we take $k = N$ then the above problem reduces to the ordinary PC. After finding the optimal solution we deflate

$$S = \Sigma - (v'\Sigma v)v'v,$$

and iterate this process to obtain further principal components. Sparse PC can retain consistency even if $N \gg T$ which makes the method suitable for use with big data.

4.4 Forecast combination

Forecast combination has a long tradition, starting at least with Bates and Granger (1969), and it tends to outperform even sophisticated forecasting models, see e.g. Timmermann (2006) for a detailed overview of theoretical and empirical studies and Kuzin, Marcellino and Schumacher (2013) for a recent application in a nowcasting context. As discussed by Hendry and Clements (2004), possible reasons for the good performance of forecast pooling may be model misspecification, model uncertainty and parameter non-constancy that are attenuated by weighting. As these features are likely present when modelling with big data, forecast combination could be helpful also in this context.

A common finding in the literature is that simple weighting schemes, and even equal weighting, often perform better than more sophisticated alternatives. Computational efficiency of these simple procedures is an additional plus in a big data context.

As an alternative, weighted averaging based on past performance in terms of (inverted) mean-squared (MSE) or mean-absolute (MAE) forecast errors can be adopted. Kuzin et al. (2013) suggest to use the MSE computed over a previous rolling window, in line with Stock and Watson (2006). Finally, information-theoretic averaging can be also considered, based on information criteria such as the AIC or BIC.

4.4.1 Data-driven automated forecasting

We discuss a way to combine, in a data-dependent and, in some sense, optimal way, some of the above methods. We suggest the use of a fully automated approach of model selection and model averaging, which is similar in notion to the procedures adopted by, e.g., Kuzin et al. (2013) and Stock and Watson (2006).

The idea is simple yet intuitive: the applied researcher could adopt a *"model rotation"* strategy which chooses the best model or model(s) given a loss function or some other user-defined criterion. The algorithm can be described as follows.

- Calculate the forecast error of K candidate models, M_j for $j = 1, 2, .., K$, during the past L periods (look-back period).
- Then, rank the models using their Root Mean Squared Forecast Error (RMSFE) or another loss function of interest.
- Select the first m models with the smallest RMSFE during the examined look-back period. Then, calculate the h-step ahead forecasts for each model and compute their average. In case where $m = 1$, the forecasts of one model are computed.

This approach is based on the general principles underlying cross-validation, which a powerful and very general approach to statistical decision making. The main idea is that models, or model variants, estimated over a given set of observations, are evaluated using as a criterion their forecasting or fitting ability, in a different set of observations and then ranked according to this criterion. This approach is powerful as it has very wide applicability and can allow for data features such as structural change, that other specification approaches do not allow for. We expect this method to work well, due to its time-varying nature which means it adapts faster to the changing dynamics of the dependent series (if any).

4.5 Textual Data

In recent years, there has been a considerable attention to textual data. Once the text is transformed to numeric data and, possibly using aggregation, a structured time series indicator has been created, all the previously discussed methodologies can be employed as well. Here, we focus more on the methods which can be used to transform the textual data to numeric.

Typical examples of textual data are: (i) corporate filings (see Li, 2006, Jegadeesh and Wu, 2013 among others), (ii) media articles (see Baker et al., 2016, Garcia, 2013, King et al., 2017, Casanova et al., 2017, among others), (iii) internet postings (see Antweiler and Frank, 2004, Das and Chen, 2007, Chen et al., 2013, Levenberg et al., 2014, O'Connor et al., 2010).

The Lexicon approach is one of the most common methods used in this context (see, e.g., Tetlock, 2007, Garcia, 2013, Nyman et al., 2015, Baker et al., 2016, among others). Consider the collection of M documents $D = \{d_1, ..., d_M\}$. The researcher provides a pre-defined dictionary of V words of interest denoted by $V = \{w_1, ..., w_V\}$. Let C be the M$\times V$ document-term matrix where c_{ij} is the frequency of j word in document i. In the context of sentiment analysis, we can construct indicators which are based on the balance of positive versus negative terms according to the dictionary of the researcher. Following Nyman et al. (2015) we can construct the following indicator:

$$s_i = \frac{\sum_{j=1}^{M} c_{ij}^{+} - \sum_{j=1}^{M} c_{ij}^{-}}{M} \tag{22}$$

where c_{ij}^{+} and c_{ij}^{-} are the positive and negative terms. In the above, we assign equal weights to all terms. Obviously, the result depends on the range of dictionary as well as the weighting scheme (e.g., some terms might be more important than others).

Boolean techniques are another approach to text mining. These methods search the main body of text in the pool of documents using expressions with logical operators (OR, AND, NOT). From one hand, using boolean methods the researcher is more flexible to combine a number of terms. However, these techniques do not account for word density.

The most widely used dictionary is the Harvard-IV-dictionary. This is a good choice for general text however it might be restrictive in the context of economics or finance. Loughran and McDonald (2011) construct a more appropriate dictionary to measure the sentiment in 10-K filings published by the SEC in the US. They suggest that almost 75% of the negative words suggest by Harvard-IV are not related to negativity in financial disclosures. For example, terms such as"liability", "tax", etc.

As mentioned previously, we use an equal weighting scheme in our example of Equation (22). However, less frequent words might also be of high importance which could change the sentiment index. In the context of nowcasting, this is very important especially at the beginning of some major news or incident which has not covered extensively by the media yet. If the researcher excludes these terms decreases the timeliness of the sentiment indicator. The most usual approach is the frequency-inverse document frequency (tf.idf) where the weights are given by:

$$tf.idf_{ij} = \begin{Bmatrix} (1+\log(tf_{ij})\log(\frac{M}{df_i}), \text{ if } tf_{ij} \geq 1 \\ 0, \text{ if } tf_{ij} = 0 \end{Bmatrix}$$

where df_i is the number of documents in which the word w_i occurs, M is the number of all documents and tf_{ij} denotes the frequency of word w_i in document d_j. The first factor adds a penalty to the word which appears more frequently giving less weights, whereas the second gives higher weight to infrequent words.

5 Conclusions

This paper is concerned with an introduction to big data which can be used for UK nowcasting and a review of the machine learning and econometric techniques that could be potentially applied in macroeconomic nowcasting and forecasting using temporally structured big data.

As it has been discussed in the text, big data prevents the use of standard econometric methods. For example, when the number of regressors is larger than that of observations ($N > T$, as in FAT datasets), OLS estimation cannot be used, as well as OLS-based statistics, such as t-tests and F-tests to check the significance of regressors. Moreover, selecting regressors by means of information criteria also becomes not doable, as 2^N models should be compared, a number larger than one million already for $N = 20$ regressors. Furthermore, standard statistical theory to prove econometric properties such as unbiased and consistency of the estimators typically rely on fixed N and T diverging asymptotics (suited for TALL datasets). Instead, with big (potentially HUGE) data both N and T diverging asymptotics is needed, which is much more complex.

A first way to deal with these feature is to use machine learning methods, where the starting point is to somewhat regularise OLS estimation to make it feasible also when N is very large. This is typically achieved by adding a set of (nonlinear) constraints on the model parameters, which are thus shrunk towards pre-specified values, preferably towards zero in order to achieve a more parsimonious specification. Within this class we have considered methods such as Penalised Regression, Ridge Regression, LASSO Regression, Adaptive LASSO, Elastic Net, SICA, Hard Thresholding, Boosting and Multiple Testing.

A second class of techniques goes under the name of Heuristic Optimisation. The starting idea here is to use information criteria to reach a good balance between model fit and parsimony by assigning a penalty dependent on the number of model parameters (which is equal to that of regressors in the linear context). We have reviewed Simulated Annealing, Genetic Algorithms, and MC^3. As the methods are iterative, and sometimes simulation based, they can become computationally very demanding when N is really large. As they should be applied recursively in a macroeconomic forecasting context, not only for forecast evaluation but also for cross-validation, the computational aspect can become prohibitive.

A third class of econometric methods to properly handle big data is based on the idea of reducing the dimension of the dataset by producing a much smaller set of generated regressors, which can then be used in a second step in standard econometric models to produce nowcasts and forecasts in common ways. There are naturally many ways to carry out dimensionality reduction, and we have considered in details Principal Component Analysis, Partial Least Squares and Sparse Principal Component Analysis.

On top of the above methodologies, forecast combination should also be used by the applied researchers as the literature indicates that even simple weighting schemes, and in particular equal weighting, which can be implemented also with very large N, often perform well.

Hence, looking ahead at the empirical application which follows in the second paper of this research, we aim to evaluate the nowcasting gains using of big data, proxied by Google Trends, for the UK GDP growth. Our plan is to employ data reduction techniques, PCA and PLS, as well as sparse regression techniques. Moreover, the various resulting models will be combined using the data-driven automated forecasting methodology.

Bibliography

1. Aastveit, K. A., Gerdrup, K. R., Jore, A. S., Thorsrud, L. A. (2014). “Nowcasting GDP in Real Time: A Density Combination Approach”, *Journal of Business & Economic Statistics,* 32(1), 48-68.

2. Aastveit, K. A., Trovik, T. (2012). “Nowcasting Norwegian GDP: The Role of Asset Prices in a Small Open Economy”, *Empirical Economics,* 42, 95-119.

3. Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., Veronese, G. (2010). “New EUROCOIN: Tracking Economic Growth in Real Time”,*The Review of Economics and Statistics,* 92(4), 1024-1034.

4. American Association for Public Research Opinion (AAPOR) Big Data Task Force (2015). “AAPOR Report on Big Data”,February 12, 2015.

5. Andreou, E., Ghysels, E., Kourtellos, A. (2015). “Should Macroeconomic Forecasters Use Daily Financial Data and How?”, *Journal of Business & Economic Statistics,* 31(2), 240-251.

6. Angelini, E., Bańbura, M., Runstler, G. (2008). “Estimating and Forecasting the Euro Area Monthly National Accounts From a Dynamic Factor Model”, *ECB Working Paper Series,* 953.

7. Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L. (2011). ”Short-Term Forecasts of Euro Area GDP Growth”, *The Econometrics Journal,* 14(1), C25-C44.

8. Antweiler, W., Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance*, 59(3), 12591294.

9. Askitas, N., Zimmermann, K. F. (2009). “Google Econometrics and Unemployment Forecasting”,*Applied Economics Quarterly,* 55(2), 107-120.

10. Avalos, M., Grandvalet, Y., Ambroise, C. (2007). “Parsimonious Additive Models”, *Computational Statistics & Data Analysis*, 51, 2851-2870.

11. Bühlmann, P. (2006). “Boosting for High-Dimensional Linear Models”, *Annals of Statistics*, 34(2), 599-583.

12. Bühlmann, P., van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer.

13. Bańbura M,. Giannone D., Reichlin L. (2011). “Nowcasting”. In Oxford Handbook on Economic Forecasting, Clements MP, Hendry DF (eds). Oxford University Press: Oxford.

14. Bańbura, M., Giannone, D., Modugno, M., Reichlin, L. (2013). ”Now-Casting and the Real-Time Data Flow”, *ECB Working Paper Series,* No 1564.

15. Bańbura, M., Runstler, G. (2011). “A Look into the Factor Model Black Box: Publication Lags and the Role of Hard and Soft data in Forecasting GDP”, *International Journal of Forecasting,* 27, 333-346.

16. Bai, J. (2003). “Inferential Theory for Factor Models of Large Dimension”, *Econometrica*, 71, 135-173.

17. Bai, J., Ng, S. (2008). “Large Dimensional Factor Analysi”, *Foundations and Trends in Econometrics*, 3(2), 89-163.

18. Bai, J., Ng, S. (2009). “Boosting Diffusion Indices”, *Journal of Applied Econometrics*, 24(4), 607–629.

19. Baker, S. R., Bloom, N., Davis, S. J. (2016). “Measuring Economic Policy Uncertainty”, *The Quarterly Journal of Economics*, 131(4), 1593-1636.

20. Barigozzi, M., Brownless, C. (2013). “Nets: Network Estimation for Time Series”, *Working Paper.*

21. Bates, J. M., Granger, C. W. J. (1969). “The Combination of Forecasts”, *Operational Research*, 20(4), 451-168.

22. Bendler, J., Wagner, S., Brandt, T., Neumann, D. (2014). “Taming Uncertainty in Big Data: Evidence from Social Media in Urban Areas”, *Business & Information Systems Engineering*, 05-2014, 279-288.

23. Bickel, P. J., Ritov, Y., Tsybakov, A. B. (2009). “Simultaneous Anlysis of LASSO and the Dantzig Selector”, *Annals of Statistics*, 37(4), 1705-1732.

24. Blake, A., Kapetanios, G. (2010). “Tests of the Martingale Difference Hypothesis Using Boosting and RBF Neural Network Approximations”, *Econometric Theory*, 26(5), 1363-1397.

25. Boettcher, I. (2015). “Automatic Data Collection on the Internet (Web Scraping)”, New Techniques and Technologies for Statistics, Eurostat Conference, 9-13 March 2015.

26. Bok, B., Caratelli, D., Giannone, D., Sbordone, A., Tambalotti, A. (2017). “Macroeconomic Nowcasting and Forecasting with Big Data”, Federal Reserve Bank of New York Staff Reports, Staff Report No. 830.

27. Brüggemann, R., Krolzig, H.M., Lutkepohl, H. (2009). “Comparison of Model Reduction Methods for VAR Processes”, *Technical Report 2003-W13*, Nuffield College, University of Oxford.

28. Braaksma, B., Zeelenberg, K. (2015). ““Re-make/Re-model”: Should big data change the modelling paradigm in official statistics?”, *Statistical Journal of the IAOS*, 31, 193-202.

29. Bragoli, D., Metelli, L., Modugno, M. (2014). “The Importance of Updating: Evidence from a Brazilian Nowcasting Model”, *Federal Reserve Board Working Paper Series,* 2014-94.

30. Breiman, L. (1996). “Bagging predictors”, *Machine Learning*, 24(2), 123-140.

31. Breiman, L. (2001). “Random Forests”, *Machine Learning* , 45(1), 5-32.

32. Buhlmann, P., Yu, B. (2006). “Sparse Boosting”, *Journal of Machine Learning Research*, 7.

33. Candes, E., Tao, T. (2007). “The Dantzig Selector: Statistical Estimation when n is much larger than p”, *Annals of Statistics*, 35(6), 2313-2351.

34. Carriero, A., Clark, T. E., Marcellino, M. (2015). “Realtime Nowcasting with a Bayesian Mixed Frequency Model with Stochastic Volatility”, *Journal of the Royal Statistical Society: Series A,* 178(4), 837-862.

35. Casanova, C., Ortiz, A., Rodrigo, T., Xia, L., Iglesias, J. (2017). Tracking chinese vulnerability in real time using Big Data, Technical Report.

36. Cavallo, A., Rigobon, R. (2016). “The Billion Prices Project: Using Online Prices for Measurement and Research”. The Journal of Economic Perspectives, 30(2), 151-178.

37. Cerchiello, P., Giudici, P. (2014). “How to Measure the Quality of Financial Tweets”. Working Paper, ECB Workshop on using big data for forecasting and statistics, 07-08/04/2014, Frankfurt.

38. Chen, H., De, P., Hu, Y.J., Hu, Hwang, B.-H. (2013). Customers as advisors: The role of social media in financial markets, in 3rd Annual Behavioural Finance Conference, Queens University, Kingston, Canada. http://www. bhwang. com/customers. pdf.

39. Choi, H., Varian, H. (2009a). “Predicting the Present with Google Trends”, *Google Technical Report.*

40. Choi, H., Varian, H. (2009b. “Predicting Initial Claims for Unemployment Benefits”, *Google Technical Report.*

41. Chudik, A., Kapetanios, G., Pesaran, M.H. (2017). “A One-Covariate as a Time, Multiple Testing Approach to Variable Selection in High-Dimensional Linear Regression Models. Fothcoming, Journal of Applied Econometrics.

42. Dablemont, S., Simon, G., Lendasse, A., Ruttiens, A., Blayo, F., Verleysen, M. (2003). “Time series forecasting with SOM and local non-linear models - Application to the DAX30 index prediction”, *WSOM’2003 proceedings - Workshop on Self-Organizing Maps, Hibikino (Japan).* 11-14 September 2003, 340-345.

43. D’Amuri, F., Marcucci, J. (2012). “The Predictive Power of Google Searches in Predicting Unemployment”. Banca d’Italia Working Paper, 891.

44. Das, S. R., Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web, *Management Science*, 53, 13751388.

45. De Mol, C., Giannone, D., Reichlin, L. (2006). “Forecasting with a Large Number of Predictors: Is Bayesian Regression a Valid Alternative to Principal Components”, *Journal of Econometrics*, 146, 318-328.

46. Dhillon, I.S., Modha, D.M. (2001). “Concept decompositions for large sparse text data using clustering”, *Machine Learning* 42(1), 143-175.

47. Doornik, J. A., Hendry, D. F. (2015). “Statistical Model Selection with Big Data”. Cogent Economics & Finance, 3(1), 2015.

48. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). “Least Angle Regression”, *Annals of Statistics*, 32(2), 407-451.

49. Evans, M. D. D. (2005). “Where Are We Now? Real-Time Estimates of the Macro Economy”, *NBER Working Paper,* 11064.

50. Fan, J., Lv, J. (2008). “Sure Independence Screening for Ultra-High Dimensional Feature Space”, *Journal of Royal Statistical Society: Series B*, 70, 849-911.

51. Fan, J., Samworth, R., Wu, Y. (2009). “Ultra High Dimensional Variable Selection: Beyond the Linear Model”, *Journal of Machine Learning Research*, 10, 1829-1853.

52. Fan, J., Song, R. (2010). “Sure Independence Screening in Generalized Linear Models with NP-Dimensionality”, *Annals of Statistics*, 38, 3567-3604.

53. Fernandez, C., Ley, E., Steel, M.F.J. (2001). “Benchmark Priors for Bayesian Model Averaging”, *Journal of Econometrics*, 100, 381-427.

54. Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2000). “The Generalised Factor Model: Identification and Estimation”, *Review of Economics and Statistics*, 82, 540-554.

55. Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2005). “The Generalised Factor Model: One-Sided Estimation and Forecasting”, *Journal of the American Statistical Association*, 100(471), 830-840.

56. Foroni, C., Marcellino, M. (2013), ”A Survey of Econometric Methods for Mixed-Frequency Data”, Norges Bank, WP 2013/06.

57. Foroni, C., Marcellino, M. (2014). “A Comparison of Mixed Frequency Approaches for Nowcasting Euro Area Macroeconomic Aggregates”, *International Journal of Forecasting,* 30, 554-568.

58. Foroni, C., Marcellino, M., Schumacher, C. (2015). “Unrestricted Mixed Data Sampling (MIDAS): MIDAS Regressions with Unrestricted Lag Polynomials”, *Journal of the Royal Statistical Society: Series A,* 178(1), 57-82.

59. Frale, C., Marcellino, M., Mazzi, G. L. (2011). “EUROMIND: A Monthly Indicator of the Euro Area Economic Conditions”, *Journal of the Royal Statistical Society: Series A,* 174(2), 439-470.

60. Fraley, C., Hesterberg, T. (2009). "Least Angle Regression for Large Datasets", *Statistical Analysis and Data Mining*, 1(4), 251-259.

61. Friedman, J. (2001). "Greedy Function Approximation: a Gradient Boosting Machine", *Annals of Statistics*, 29, 1189-1232.

62. Friedman, J., Hastie, T., Tibshirani, R. (2000). "Additive Logistic Regression: a Statistical View of Boosting", *Annals of Statistics*, 28, 337-374.

63. Fu, T.-C., Chung, F.-L., Ng, V., Luk, R. (2001). "Pattern Discovery from Stock Time Series using Self-organizing Maps", *KDD 2001 Workshop on Temporal Data Mining*, August 26-29, San Francisco.

64. Galbraith, J. W., Tkacz, G. (2013). "Nowcasting GDP with Electronic Payments, Data Vintages and the Timing of Data Releases", *European Central Bank: Statistics Papers Series*, No 10.

65. Garcia, D. (2013). Sentiment during recessions, *The Journal of Finance*, 68, 1267-1300.

66. Gershenfeld, N., Schoner, B., Metois, E. (1999). "Cluster-weighted modelling for time-series analysis", *Nature*, 397(6717), 329-332.

67. Ghysels, E., Santa-Clara, P., Valkanov, R. (2004). "The MIDAS Touch: Mixed Data Sampling Regression Models", *CIRANO Working Paper,* 2004s-20.

68. Giannone, D., Reichlin, L., Simonelli (2009). "Nowcasting Euro Area Economic Activity in Real-Time: The Role of Confidence Indicators", *National Institute Economic Review,* 210, 90-97.

69. Giannone, D., Reichlin, L., Small, D. (2008). "Nowcasting: The Real-Time Informational Content of Macroeconomic Data", *Journal of Monetary Economics,* 55, 665-676.

70. Goffe, W.L., Ferrier, G.D., Rogers, J. (1994). "Global Optimisation of Statistical Functions with Simulated Annealing", *Journal of Econometrics*, 60(1), 65-99.

71. Griffioen, R., de Haan, J., Willenborg, L. (2014). “Collecting Clothing Data from the Internet”, Statistics Netherlands Technical Report.

72. Groen, J. J. J., Kapetanios, G. (2016). “Revisiting Useful Approaches to Data-Rich Macroeconomic Forecasting”, *Computational Statistics and Data Analysis,* 100, 221-239.

73. Hajek, B. (1998). “Cooling Schedules for Optimal Annealing”, *Mathematics of Operations Research*, 13(2), 311-331.

74. Hartl, H.R.F., Belew, R.K. (1990). “A Global Convergence Proof for a Class of Genetic Algorithms”, *Technical Report*, Technical University of Vienna.

75. Haung, J., Ma, S., Zhang, C.-H. (2008). “Adaptive LASSO for Sparse High Dimensional Regression Models”, *Statistica Sinica*, 18, 1603-1618.

76. Helland, I. S. (1990). “Partial Least Squares Regression and Statistical Models”, *Scandinavian Journal of Statistics*, 17(2), 97-114.

77. Hendry, D.F. (1995). Dynamic Econometrics. Oxford University Press.

78. Hendry, D.F. (1997). “On Congruent Econometric Relations: A Comment”, Carnegie-Rochester Conference Series on Public Policy, 47, 163-190.

79. Hendry, D.F., Clements, M. P. (2004). “Pooling of forecasts”, *The Econometrics Journal*, 7(1), 1-31.

80. Hepenstrick, C., Marcellino, M. (2016). “Forecasting with Large Unbalanced Datasets: The Mixed Frequency Three-Pass Regression Filter”, Working Paper, Swiss National Bank.

81. Hesterberg, T., Choi, N. H., Meier, L., Fraley, C. (2008). “Least angle and ‘ℓ_1 penalized regression: A review”, *Statistics Surveys*, 2, 61-93.

82. Heston, S. L., Sinha, N. R. (2014). “News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns”, *Working Paper.*

83. Hinton, G. E. , Salakhutdinov, R. R. (2006). “Reducing the dimensionality of data with neural networks”, *Science*, 313(5786), 504-507.

84. Hoover, K.D., Perez, S.J. (1999). "Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Set", *Econometrics Journal*, 2, 167-191.

85. Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., Shang, H.L. (2011). "Optimal Combination Forecasts for Hierarchical Time Series", *Computational Statistics and Data Analysis*, 55(9), 2579-2589.

86. Jegadeesh, N., Wu, D. (2013). Word power: A new approach for content analysis, *Journal of Financial Economics*, 110, 712729.

87. Kalpakis, K., Gada, D., Puttagunta, V. (2001). "Distance Measures for Effective Clustering of ARIMA Time-Series", *Proceedings of the 2001 IEEE International Conference on Data Mining*, San Jose, CA, November 29th - December 2nd.

88. Kapetanios, G. (2007). "Variable Selection in Regression Models using Non-Stantard Optimisation of Information Criteria", *Computational Statistics & Data Analysis*, 52(1), 4-15.

89. Kapetanios, G., Marcellino, M. and Venditti, F. (2016). "Large Time-Varying Parameter VAR: A Non-Parametric Approach", *CEPR WP 11560.*

90. Keerthi, S. S., Shevade, S. (2007). "A Fast Tracking Algorithm for Generalized LARS/LASSO", *IEEE Transactions on Neural s*, 18(6), 1826-1830.

91. Kelly, B., Pruitt, S. (2015). "The Three-Pass Regression Filter: A New Approach to Forecasting using Many Predictors", *Journal of Econometrics*, 186(2), 294-316.

92. Khan, J. A., Van Aelst, S., Zamar, R. H. (2007). "Robust Linear Model Selection Based on Least Angle Regression", *Journal of the American Statistical Association*, 102(480), 1289-1299.

93. Kim, H. H., Swanson, N. (2016). "Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods", *International Journal of Forecasting*, In Pres.

94. King, G., Schneer, B., White, A. (2017). How the news media activate public expression and influence national agendas, *Science*, 358, 776780.

95. Koop, G., Onorante, L. (2013). "Macroeconomic Nowcasting Using Google Probabilities", *Working Paper*, ECB Workshop on Big Data for Forecasting and Statistics.

96. Koop, G. (2018). "Bayesian Methods for Empirical Macroeconomics with Big Data", *Review of Economic Analysis*, 9(1), 33-56.

97. Korobilis, D. (2017). "Forecasting with Many Predictors using Message Passing Algorithms", *Essex Finance Centre Working Papers*, 19565, Essex Business School.

98. Kuzin, V., Marcellino, M., Schumacher, C., (2013). "Pooling versus model selection for nowcasting GDP with many predictors: Empirical evidence for six industrialized countries", *Journal of Applied Econometrics*, 28(3), 392-411.

99. Lahiri, K., Monokroussos, G. (2013). "Nowcasting US GDP: The Role of ISM Business Surveys", *International Journal of Forecasting*, 29, 644-658.

100. Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014). "The Parable of Google Flu: Traps in Big Data Analysis", *Science*, 143, 1203-1205.

101. Levenberg, A., Pulman, S., Moilanen, K., Simpson, E., Roberts, S. (2014). Predicting Economic Indicators from Web Text Using Sentiment Composition, *International Journal of Computer and Communication Engineering*, 3(2), 109-115.

102. Li, F. (2006). "Do stock market investors understand the risk sentiment of corporate annual reports?". Working Paper.

103. Liao, T. W. (2005). "Clustering of time series dataa survey", *Pattern Recognition*, 38(11), 1857-1874.

104. Loughran, T., McDonald, B. (2011). Barrons Red Flags: Do They Actually Work? *Journal of Behavioral Finance*, 12, 9097.

105. Lv, J., Fan, Y. (2009). "A Unified Approach to Model Selection and Sparse Recovery using Regularized Least Squares", *Annals of Statistics*, 37(6A), 3498-3528.

106. Marcellino, M., Stock, J.H., Watson, M. (2006). "A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series", *Journal of Econometrics*, 135(1-2), 499-526.

107. Mariano, R., Murasawa, Y. (2003). "A new coincident index of business cycles based on monthly and quarterly series", *Journal of Applied Econometrics*, 18(4), 427-443.

108. Mariano, R., Murasawa, Y. (2010). "A coincident index, common factors, and monthly real GDP", *Oxford Bulletin of Economics and Statistics*, 72(1), 27-46.

109. Martínez Álvarez, F., Troncoso, A., Riquelme, J.C., Riquelme, J.M. (2007). "Discovering Patterns in Electricity Price Using Clustering Techniques", *International Conference on Renewable Energies (ICREPQ'07).*

110. McCallum, A., Nigam, K., Ungar L.H. (2000). "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 169-178.

111. Mitchell, T. and Beauchamp, J. (1988). "Bayesian variable selection in linear regression", *Journal of American Statistical Association*, 83, 1023-1036.

112. Modugno, M. (2013). "Now-casting Inflation using High-Frequency Data",*International Journal of Forecasting*, 29, 664-675.

113. Moller-Levet, C.S., Klawonn, F., Cho, K.-H., Wolkenhauer, O. (2003). "Fuzzy clustering of short time series and unevenly distributed sampling points", *Advances in Intelligent Data Analysis V*, Vol. 2810 of the series Lecture Notes in Computer Science, 330-340.

114. Morinaka, Y., Yoshikawa, M., Amagasa, T. (2001). "The L-Index: An Indexing Structure for Efficient Subsequence Matching in Time Sequence Databases",

Proceedings of the Fifth Pacific-Asia Conference on Knowledge Discovery and Data Mining.

115. Ng, S. (2013). "Variable Selection in Predictive Regressions", *Handbook of Forecasting*, 2B, 753-786.

116. Nyman, R., Gregory, D., Kapadia, S., Smith, R., Tuckett, D. (2014a). "Exploiting Big Data for Systemic Risk Assessment: News and Narratives in Financial Systems", *Working Paper*, ECB Workshop on using big data for forecasting and statistics, 07-08/04/2014, Frankfurt.

117. Nyman, R., Ormerod, P., Smith, R., Tuckett, D. (2014b). "Big Data and Economic Forecasting: A Top-Down Approach Using Directed Algorithmic Text Analysis", *Working Paper*, ECB Workshop on using big data for forecasting and statistics, 07-08/04/2014, Frankfurt.

118. OConnor, B., Balasubramanyan, R., Routledge, B. R., Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series, ICWSM, 11, 12.

119. Ross, A. (2013). "Nowcasting with Google Trends: A Keyword Selection Method", *Fraser of Allander Economic Commentary*, 37(2), 54-64.

120. Rossiter, J. (2010). "Nowcasting the Global Economy", *Bank of Canada Working Paper,* 2010-12.

121. Schmidt T., Vosen, S. (2011). "Forecasting Private Consumption: Survey-based Indicators vs. Google Trends", *Journal of Forecasting*, 30(6), 565-578.

122. Schubert, A. (2015). "Data as a Core Central Banking Asset – The strategy of the ECB", *Presentation*, Big Data Workshop, Sveriges Riksbank, 09/11/2015.

123. Scott, S., Varian, H. (2013). "Predicting the Present with Bayesian Structural Time Series", *Working Paper.*

124. Shi, T., Horvath, S. (2006), "Unsupervised Learning With Random Forest Predictors", *Journal of Computational and Graphical Statistics*, 15, 118-138.

125. Sin, C.Y., White, H. (1996). “Information Criteria for Selecting Possibly Misspecifed Parametric Models”, *Journal of Econometrics*, 71(1-2), 207-225.

126. Mitchell, J., Smith, R.J., Wealer, M.R. (2013). “Efficient Aggregation of Panel Qualitative Survey Data”, *Journal of Applied Econometrics*, 28(4), 580-603.

127. Stock J.H. and M.W. Watson (2006), ”Forecasting with Many Predictors”, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.) *Handbook of Economic Forecasting.*

128. Stock, J., Watson, M. (2002a). “Forecasting Using Principal Components from a Large Number of Predictors”, Journal of the American Statistical Association, 297, 1167-1179.

129. Stock, J., Watson, M. (2002b). “Macroeconomic Forecasting using Diffusion Indexes”, Journal of Business & Economics Statistics, 20, 147-162.

130. Tibshirani, R., (1996). “Regression Shrinkage and Selection via the Lasso”, *Journal of the Royal Statistical Society: Series B*, 58(1), 267-288.

131. Timmermann, A. (2006), ”Forecast Combinations”, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.) *Handbook of Economic Forecasting.*

132. Van De Geer, S. A. (2008). “High-Dimensional Generalized Linear Models and the LASSO”, *Annals of Statistics*, 36(2), 614-645.

133. Wold, H. (1980). “Model Construction and Evaluation When Theoretical Knowledge Is Scarce”, in Kmenta, J. and Ramsey, J.B. (eds): ”Evaluation of Econometric Models”, New York: Academic Press.

134. Yiu, M. S., Chow, K. K. (2010). “Nowcasting Chinese GDP: Information Content of Economic and Financial Data”, *China Economic Journal*, 3(3), 223-240.

135. Zhang, C.-H. (2010). “Nearly Unbiased Variable Selection under Minimax Concave Penalty”, *Annals of Statistics*, 38(2), 894-942.

136. Zheng, Z., Fan, Y., Lv, J. (2014). "High Dimensional Thresholded Regression and Shrinkage Effect", *Journal of the Royal Statistical Society: Series B*, 76(3), 627-649.

137. Zou, H. (2006). "The Adaptive Lasso and Its Oracle Properties", *Journal of the American Statistical Association*, 101(746), 1418-1429.

138. Zou, H. and Hastie, T. (2005). "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society Series B*. 67(2), 301–320.

139. Zou, H., Hastie, T., Tibshirani, R. (2006). "Sparse Principal Component Analysis", *Journal of Computational and Graphical Statistics*, 15(2), 265-286.