Analysis of the most recent modelling techniques for big data with particular attention to Bayesian ones

GEORGE KAPETANIOS,
MASSIMILIANO MARCELLINO, KATERINA PETROVA

2018 edition

STATISTICAL
WORKING PAPERS

eurostat

Analysis of the most recent modelling techniques for big data with particular attention to Bayesian ones

GEORGE KAPETANIOS,
MASSIMILIANO MARCELLINO, KATERINA PETROVA

2018 edition

Manuscript completed in December 2017

Luxembourg: Publications Office of the European Union, 2018

PDF ISBN 978-92-79-77350-1 ISSN 2315-0807 doi: 10.2785/679114 KS-TC-18-001-EN-N

Abstract

In this report we describe various methods suited for the analysis of linear models with a very large number of explanatory variables, with a special emphasis on Bayesian approaches. We next consider some non-parametric and/or non-linear methods suited for applications with big data, such as random trees, random forests, cluster analysis, deep learning and neural networks. Finally, we survey techniques for summarizing the information in large (possibly sparse) datasets, forecast combination approaches, and techniques for the analysis of large mixed frequency datasets.

Keywords: Big Data, Machine Learning, Bayesian methods.

Authors: George Kapetanios*. Massimiliano Marcellino†, Katerina Petrova ‡

Acknowledgement:
This work has been carried out by under the supervision of the Eurostat project manager Dario Buono. Special thanks to Lisa Bosotti for taking care of the layout and to Sabine Bankwitz for the backstopping support.

*george.kapetanios@kcl.ac.uk

†massimiliano.marcellino@unibocconi.it

‡Katerina.Petrova@st-andrews.ac.uk

Table of content

List of figures

1 Introduction

One of the key issues in large data models is that the number of available economic variables is of considerable size, resulting in poor inference and forecasting performance of standard econometric techniques. Most of the large data statistical and econometric literature attempts to reduce the data dimension by 'penalising' the model for complexity. Dimension reduction is accomplished either through various penalty functions shrinking the coefficients of the large set of explanatory variables towards zero or through compressing the dimension of the explanatory variables into a much smaller set. The common idea behind the different approaches is to avoid overfitting and, as a result, considerably improve forecasting.

In this paper, we provide a survey of different models for inference with big data, focusing on the most relevant methodological improvements in the field of Bayesian econometrics. For completeness, we also include a discussion of the methods previously surveyed in Kapetanios, Marcellino and Papailias (2016).

The rest of the document is organised as follows. Section 2 describes various methods suited for the analysis of linear models with a very large number of explanatory variables. We first review various penalised regression approaches, then show how they can be given a Bayesian interpretation when choosing particular prior distributions for the model parameters, next introduce further methods based on yet other choices of prior distributions (such as spike and slab regressions and compressed regressions), and finally we consider quantile and expectile regressions. We also discuss multivariate regression methods, mostly variants of Bayesian VARs, and a set of procedures for variables selection particularly suited for the big data context (the latter were already considered in Kapetanios, Marcellino and Papailias (2016)).

In Section 3, we review some non-parametric and/or non-linear approaches suited for applications with big data, such as random trees, random forests, cluster analysis, deep learning and neural networks.

In Sections 4, 5 and 6, which are based on Kapetanios et al. (2016), for completeness, we survey, respectively, methods for summarizing the information in large datasets, forecast combination approaches, and techniques for the analysis of mixed frequency datasets. The idea of summarizing the information in large datasets by means of a few constructed series, often called factors or indexes, has a long tradition in econometrics, and can be extended to the case of big data, where sparsity is often an additional problem. Another common approach that performs well in empirical applications based on economic data is pooling a large number of forecasts from very simple models rather than using a single forecast from a big model, so that rather than selecting or summarizing the

many indicators we select or, more frequently, combine, directly the associated forecasts. Finally, approaches that can deal with mixed frequency data are relevant, as big data are typically available in higher frequency than the target indicator, as we have discussed in previous reports.

In Section 7, we compare the reviewed econometric methods for big data. Finally, in Section 8 we summarize the main results and conclude.

2 A review of parametric methods for Big Data

2.1 Penalised regressions

Penalised regressions are a simple, linear and tractable approach to dealing with large data. Let

$$y = X\beta + \varepsilon, \qquad \mathbb{E}[\varepsilon|X] = 0, \qquad \mathbb{E}[\varepsilon^2|X] = \sigma^2 \tag{1}$$

where y and ε are $T \times 1$ vectors, X is a $T \times N$ matrix containing a potentially large number of explanatory variables (N might be close or larger than T), and β is a $N \times 1$ vector of parameters. Typically the series are demeaned prior to estimation and no intercept is required. Demeaning is important as shrinking the intercept can cause bias. Penalised regressions are a wide class of models resulting from minimising the sum of squared model's residuals subject to an additional penalty

$$\hat{\beta}^{PR} = \arg\min_{\beta} \frac{1}{T}\left(\varepsilon'\varepsilon + f(\lambda, \beta)\right) \tag{2}$$

where typically the penalty $f(\lambda, \beta)$ is of the form $f(\lambda, \beta) = \lambda \| \psi\beta \|$, and ψ is a diagonal matrix consisting of penalty loadings and the notation $\|\|$ denotes a generic norm. For example, $\|\|_p$denotes the Lp-norm with $\| \beta \|_p = \sum_{j=1}^{N} |\beta_j|^p$ The penalty $\lambda \| \psi\beta \|$ 'shrinks' some of the parameters in β to zero and hence copes with the potentially large dimension of β. If λ is fixed and does not depend on T , then it is clear from (2) that $p\lim_{n\to\infty}\hat{\beta}^{PR} = \hat{\beta}^{OLS}$, the ordinary least squares estimator, as, from (2), asymptotically λ/T converges to zero and the penalty term disappears from the loss function. On the other extreme, if $\lambda \to \infty$ faster than T, then the penalty term dominates asymptotically, and $p\lim_{n\to\infty} \hat{\beta}^{PR} = 0$, so that all parameter estimators converge to zero.

In the big data context, the interest is on what are the properties of the resulting penalized estimators when the number of regressors is large and potentially diverging, $N \to \infty$. In this case, the shrinkage parameter, λ, is required to increase at some rate typically slower than T (therefore λ is now indexed by T and $\lambda T \to \infty$). In this setup, Knight and Fu (2000) study the asymptotic properties of penalised regression estimators of the general form

$$\hat{\beta}^{PR} = \arg\min_{\beta} Z_n := \frac{1}{T}\left(\sum_{t=1}^{T} \varepsilon_t^2 + \lambda_T \sum_{i=1}^{N} |\beta_i|^{\gamma}\right). \tag{3}$$

Note that this specification includes a wide range of estimators, such as the Ridge, Lasso and *Lp*-norm estimators discussed individually below, and ψ is taken to be the $N \times N$ identity matrix. The two main assumptions that they make are:

1. $C_N := \frac{1}{T}\sum_{t=1}^{T} x_t x_t' \rightarrow C$ as $T \rightarrow \infty$, where C is a nonnegative definite matrix

2. $\frac{1}{T}\max x_t' x_t \rightarrow 0$ as $T \rightarrow \infty$.

The first result is that under assumptions 1 and 2, with C nonsingular, the sufficient condition for consistency is $\lambda T = o(T)$. That is, when $\frac{\lambda T}{T} \rightarrow \lambda_0 \geq 0$ the penalised regression estimators of the general form in (3) are consistent: $\hat{\beta}^{PR} \rightarrow_p \beta$ as $T \rightarrow \infty$. More interestingly, in order to obtain standard asymptotic normality results, λT is required to grow at a rate slower than T, and the exact rate depends on the penalty function and, in particular, on whether $\gamma \geq 1$ or $\gamma < 1$. In their Theorem 2, Knight and Fu (2000) show that whenever $\gamma \geq 1, \lambda_T = O(\sqrt{T})$ ' is sufficient for asymptotic normality:

$$\sqrt{T}\left(\hat{\beta}^{PR} - \beta\right) \rightarrow_d N(0, \sigma^2 C^{-1}).$$

For the case when $\gamma < 1$, the sufficient condition is actually weaker: $\lambda_T = O(T^{\gamma/2})$ (note that $\lambda_T = O(\sqrt{T})$ suffices for $\gamma < 1$).
There are several important and interesting implications of the asymptotic results presented in Knight and Fu (2000). If $\gamma < 1$, the nonzero regression parameters can be estimated at the usual $\sqrt{T}$ rate without asymptotic bias while shrinking the estimates of zero regression parameters to zero with positive probability. In contrast, when $\gamma \geq 1$, their results indicate that nonzero parameters are estimated with some asymptotic bias if $\lambda_0 > 0$.
In practice, determining how much shrinkage to use in applications is an important question, as it affects the bias-variance tradeoff of the penalised regression estimators, with large values of λ generally resulting in smaller variance and larger bias and vice versa. K-fold crossvalidation methods are often used to get an optimal value for λ; these are popular and work well in practice but lack theoretical justification. Theoretical results for alternative procedures are often available for special cases, for instance, in the case of Lasso, theoretical results are presented by Belloni and Chernozhukov (2013).
The penalised regression model is a general framework and different norms and penalty functions give rise to different estimators. Below we discuss some examples, and explain how all these estimators can be given a Bayesian interpretation. In fact, they are typically the mode of the posterior distributions of the parameters in Bayesian regression models, using different prior distributions.

2.1.1 RIDGE REGRESSION

The L2-norm in (2) gives rise to the Ridge estimator, first introduced by Hoerl and Kennard (1970):

$$\begin{aligned} \hat{\beta}^{RIDGE} &= arg\min_{\beta} \frac{1}{T}(\varepsilon'\varepsilon + \lambda \parallel \psi\beta \parallel_2) \\ &= (X'X + \lambda\psi)^{-1}X'_y \end{aligned} \tag{4}$$

where $\parallel \beta \parallel_2 = \sum_{j=1}^{N} \beta_j^2$

In most applications, Ψ is taken to be a $N \times N$ identity matrix. The shrinkage is typically imposed on the slope parameters β and not on the intercept term, since constraining the intercept might cause serious bias and therefore have important negative implications on forecasting. For this reason, the data are demeaned before estimation. Moreover, the ridge regression is not invariant to scaling of X and y, so sometimes data are standardised before estimation.

Asymptotic results for Ridge regressions are available, although not at the level of generality of the

results of Knight and Fu (2000). In the case of very large datasets, when $N \gg T$, under some regularity conditions, Dobriban and Wager (2017) fully characterize the first order asymptotic behavior of the predictive risk of ridge regression. They also find that whenever λ is optimally chosen in their setup, there might be a tradeoff between the asymptotic predictive and estimation risks of the ridge regression. This inverse relationship is an interesting and surprising result. The intuition behind it is that when the regressors x_i are highly correlated, prediction can easily be achieved because y lies close to the column space of the matrix X, but estimation of the parameters can be difficult as a result of multicollinearity. On the other hand, as correlation among the regressors x_i decreases, estimation gets easier but prediction becomes problematic.

2.1.2 LASSO REGRESSION

Tibshirani (1996) suggested using the L1-norm in (2), which produces the Lasso (Least Absolute Shrinkage and Selection Operator) estimator:

$$\hat{\beta}^{LASSO} = arg\min_{\beta} \frac{1}{t}(\varepsilon'\varepsilon + \lambda \parallel \psi\beta \parallel_1) \tag{5}$$

where $\parallel \beta \parallel_1 = \sum_{j=1}^{N} |\beta_j|$. Belloni, Chen, Chernozhukov and Hansen (2012) extend the estimation problem to errors ε_i that may be non-Gaussian or heteroskedastic. The resulting estimator shares similarity with the Ridge estimator. However, the choice of L1-norm makes the problem nonlinear and no closed form expression is available for $\hat{\beta}^{LASSO}$. In addition, the L1 norm results into some coefficients in β to be exactly equal to zero, which can be convenient in economic applications to enhance interpretability of results. This is the main difference between the two penalised methods: lasso is more appropriate in cases where the model is sparse (i.e. there are many irrelevant regressors in X), while ridge performs better in situations where the model is approximately sparse (i.e. there are many very small but not zero elements in β).

Under regularity conditions, Tibshirani (2013) proves that the minimisation problem in (5) has a unique solution. Several algorithms have been proposed in the literature in order to obtain the solution numerically. For example, Wang, Gordon and Zhu (2006) develop an efficient linear programming algorithm which can solve the entire regularization path in one pass. Wu and Lange (2008) compare the $l2$ algorithm based on cyclic coordinate descent and propose a new $l1$ algorithm based on greedy coordinate descent and Edgeworth's algorithm.

There are also various variations of the basic Lasso estimator. For example, the post-Lasso estimator was introduced and analysed in Belloni and Chernozhukov (2013). Moreover, adaptive Lasso (A-Lasso) extension was proposed by Zou (2006), 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 for the adaptive Lasso is:

$$\min_{\beta_n}\left\{\sum_{t=1}^{T}\left(y_t - a - \beta_N' x_{t,N}\right)^2 + \lambda\sum_{i=1}^{N}\widehat{\omega}_i|\beta_i|\right\}, \tag{6}$$

where $\widehat{\omega}_i = 1/|\hat{\beta}_{init,i}|^{\gamma}$, $\hat{\beta}_{inil}$ 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 (6).

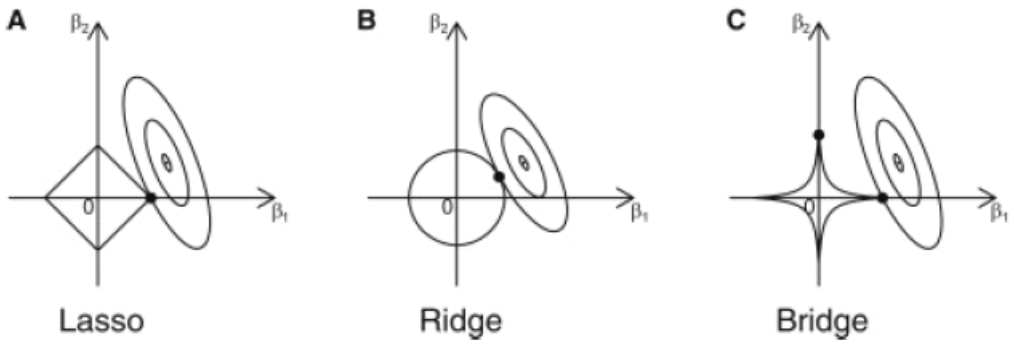
2.1.3 THE LP NORM

It is straightforward to generalise the penalty function $\lambda \parallel \psi\beta \parallel$ to an Lp norm

$$\hat{\beta}^{Pen} = arg\min_{\beta}\frac{1}{T}\left(\varepsilon'\varepsilon + \lambda \parallel \psi\beta \parallel_p\right)$$

where $\parallel \beta \parallel_p = \sum_{j=1}^{N}|\beta_j|^p$ and, as discussed earlier, the asymptotic results in Knight and Fu (2000) are available for this class of estimators. The Lp modification gives rise to a large class of penalty functions. For example, values $p \in (1,2)$ provide a combination between ridge and lasso. In particular, when $p > 1$, the penalty is differentiable at 0, and the resulting estimator, while it does not have the property of setting elements in β exactly to zero (as is the case of Lasso), is more computationally tractable.

Figure 1: Graphical illustration of the properties of $L1-$, $L2-$ and $Lp-$norm for the case when $N = 2$. The ellipses represent the contours of the unconstrained objective function of sum of squared residuals, the shapes represent the constraints in the case of $L1-$, $L2-$ and $Lp-$norms respectively. The dots represent the resulting regularised estimators, where the contours are tangent to the constraints.



Source: Based on author's calculations

2.1.4 ELASTIC NET

The advantage of the Ridge estimator is that it shrinks all parameters towards zero (without setting them exactly to zero), which is expected to perform well in approximately sparse models (models in which the parameters of all explanatory variables are small but different from 0). On the other hand, the $L1$ norm used in Lasso has the property of setting some elements in β exactly equal to zero, which is desirable whenever the true model is exactly sparse (that is when some of the variables in X do not appear in the true model). When working with data, researchers typically do not know the data generating process, so allowing for the possibility of both exact and approximate sparsity is desirable. This is the motivation behind using the Lp norm with $p \in (1,2)$.

An alternative way to combine ridge and lasso regressions is the elastic net, proposed by Zou and Hastie (2005), whose penalty function is a weighted average of those for Lasso and Ridge:

$$\hat{\beta}^{NET} = arg\min_{\beta} \frac{1}{T}\left(\varepsilon'\varepsilon + \lambda \sum_{j=1}^{N} \left(\alpha\beta_j^2 + (1-\alpha)|\beta_j|\right)\right). \alpha \in (0,1)$$

The advantage of penalised regression methods is that they are linear (hence easily interpretable) and overall computationally fast to estimate.

2.1.5 BAYESIAN INTERPRETATION OF PENALISED REGRESSION MODELS

The penalized estimators considered above have a Bayesian interpretation, as they can be considered as the mean or more of the posterior distributions of the parameters of the linear regression model for specific choices of priors, where the prior is incorporated in the penalty function to impose a 'prior' belief on the model parameters.

In particular, if we assume that $\varepsilon \sim N(0,\sigma^2)$, then $y \sim N(X\beta,\sigma^2)$ and the likelihood function of the linear regression model is:

$$f(y|\beta,\sigma^2) = (2\pi\sigma^2)^{-\frac{T}{2}} e^{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)} \quad (7)$$

Bayesian analysis requires specifying a prior density, $p(\beta,\sigma^2)$, for the model parameters β and σ^2. The prior is then combined with the likelihood $f(y|\beta,\sigma^2)$ in (7) to produce a posterior distribution for the model parameters, $p(\beta,\sigma^2|y)$. Using Bayes formula, it is:

$$p(\beta,\sigma^2|y) = \frac{p(y|\beta,\sigma^2)p(\beta,\sigma^2)}{p(y)} \quad (8)$$

where $p(y)$is the marginal data density.

The Ridge estimator (4) for the parameter vector β coincides with the mode of the posterior $p(\beta,\sigma^2|y)$ if a prior of the form $\beta \sim N(0,(\lambda\Psi)^{-1})$ is specified. Usually, σ^2, the variance of the residual also receives Bayesian treatment. In particular, whenever a Normal-inverse-Gamma prior is specified for β and σ^2 of the form

$$\beta|\sigma^2 \sim N(\beta_0,\sigma^2\Psi_0),\ (\sigma^2)^{-1} \sim Ga(\alpha_0/2,\delta_0/2) \quad (9)$$

the joint posterior of β and σ^2 is also Normal-Gamma with closed form parameters of the form

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

with

$$\hat{\Psi} = (X'X + \Psi_0^{-1})^{-1}$$
$$\hat{\beta} = \hat{\Psi}(X'y + \Psi_0^{-1}\beta_0)$$
$$\hat{\alpha} = \alpha_0 + N$$
$$\hat{\delta} = \delta_0 + y'y + \beta_0'\Psi_0^{-1}\beta_0 - \hat{\beta}'\hat{\Psi}\hat{\beta}.$$

The Lasso estimator is also a special case of the posterior mode of a Bayesian estimator in (8) and results from combining the likelihood in (7) with a Laplace prior density $\beta \sim La(0, (\lambda\Psi)^{-1})$ (see for example Park and Casella (2008)). Similarly, a generalised Laplace prior distribution gives rise to a posterior mode which coincides with the Lp norm estimators. The elastic net model is another special case of the Bayesian estimators in (8), where the prior for β is a mixture density between the Normal and Laplace priors with mixing weights α and $(1-\alpha)$.

Hierarchical priors The hyperparameter λ is crucial also in a Bayesian setting, as it controls the overall shrinkage of the prior. Some papers propose considering an $N \times 1$ vector λ (note that in the frequentist setup λ is a scalar) with elements controlling the shrinkage of each corresponding element of the vector β. In order to provide a Bayesian treatment for λ, a prior distribution is specified for it, $p(\lambda)$, resulting into hierarchical Bayesian priors of the form

$$p(\beta, \sigma^2, \lambda^2) = p(\beta, \sigma^2|\lambda)p(\lambda^2).$$

Different combinations between the choice of the mixing density $p(\lambda^2)$.and the prior $p(\beta, \sigma^2|\lambda^2)$ give rise to different estimators. Examples include:

Normal prior with Jeffreys mixing. This is the standard Normal prior considered in (9), with an uninformative Jeffreys prior for λ^2 of the form $p(\lambda^2) = \frac{1}{\lambda_i^2},\ i = 1, \ldots, N$ (see Figueiredo, (2003) and Bae and Mallick (2004))

Horseshoe prior. This results a special case of a normal/inverted-beta class distribution with $p(\lambda^2) = IB(a, b)$. Carvalho, Polson and Scott (2010) study the model when $a = b = 0.5$ and Polson and Scott (2009) generalise to a wider class of models.

Student-t prior for β with an inverse-gamma mixing density: (see Tipping (2001)).

Double-exponential prior with an exponential mixing density (e.g. West (1987), Carlin and Polson (1991), Pericchi and Smith (1992), Tibshirani (1996), Park and Casella (2008), and Hans (2009)).

Normal/exponential-gamma prior with an exponential mixing density. The mixing density is exponential of the form $p(\lambda_i^2|k) = Exp(k)$, and a second level prior for the exponential parameter k,

assumed to have a Gamma prior $p(k) = Ga(c, 1)$ as in Griffin and Brown (2005), which leads to a prior for λ of the form $p(\lambda_i^2) \propto (1 + \lambda^2)^{1-c}$.

Normal/gamma and normal/inverse-Gaussian, respectively characterized by gamma and inverse-Gaussian mixing densities (Caron and Doucet (2008), Griffin and Brown, (2010)

As long as the prior densities for λ^2 and β are proper densities and do not depend on the sample size T, standard Least Square asymptotic results apply as the priors of the different models become negligible and vanish asymptotically as T increases. For the case when $\lambda \to \infty$ as a slowly varying function, some asymptotic results are presented for a general class of priors in Polson and Scott (2010).

2.2 Spike and Slab regressions

Spike and Slab regressions were originally proposed by Mitchell and Beauchamp (1988) and recently popularised by Scott and Varian (2013). They are another case resulting from a special choice of prior in the penalised regression model in (3). 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 equation), and $\gamma_i = 0$ if $\beta_i = 0$. Denoting the nonzero elements of β by β_y, the spike and slab prior for β and γ can be written as

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

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

$$p(\gamma) = \prod_i^N \pi_i^{\gamma_i} (1 - \pi_i)^{(1-\gamma_i)},$$

so it represents a spike as it places positive probability mass at zero[1]. 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 used, of the form:

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

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

$$\beta_\gamma \left| \sigma^2, \gamma, y, X \sim N\left(\hat{\beta}_\gamma, \sigma^2 \hat{\Psi}_\gamma\right), (\sigma^2)^{-1} \right| y, X \sim Ga\left(\frac{\hat{\alpha}}{2}, \frac{\hat{\delta}}{2}\right) \quad (10)$$

with

$$\hat{\Psi}_\gamma = (X'X + \Psi_{\gamma 0}^{-1})^{-1}$$
$$\hat{\beta}_\gamma = \hat{\Psi}_\gamma (X'y + \Psi_{\gamma 0}^{-1} \beta_{\gamma 0})$$

[1] Note that, in applications, the Bernoulli prior can be simplified by setting $\pi_i = \pi$ for all $i = 1, \ldots, N$ and π can be set to be m/N, where m is the prior belief about the number of nonzero predictors.

$$\hat{\alpha} = \alpha_0 + N$$
$$\hat{\delta} = \delta_0 + y'y + \beta'_{\gamma 0}\Psi^{-1}_{\gamma 0}\beta_{\gamma 0} - \hat{\beta}'_{\gamma}\hat{\Psi}_{\gamma}\hat{\beta}_{\gamma}.$$

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

$$p(\gamma|y,X) \propto \frac{|\Psi_{\gamma 0}|^{-\frac{1}{2}}p(\gamma)}{|\hat{\Psi}_{\gamma}|^{-\frac{1}{2}}\hat{\delta}^{\frac{N}{2}-1}} \quad (11)$$

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

Metropolis within Gibbs Algorithm

1. Initialise the algorithm with guesses for $\gamma^0, \beta^0, (\sigma^2)^0$.

   For $i = 1, \ldots, N^{sim}$ iterate between the following steps:
2. Draw the vector of indicator variables γ^i from the posterior (11).
3. Conditional on γ^i, draw $\beta_{\gamma}|\sigma^2, \gamma, y, X$ and $(\sigma^2)^{-1}|\, y, X$ from the (10) posterior.

The Spike and Slab prior diff from Lasso, Ridge and other penalised regression in that it gives a nonzero prior probability mass to the parameters of being exactly equal to zero (note that the $L1$-norm of lasso places positive density (not mass) of coefficients equal to zero). It is that feature that Scott and Varian (2013) claim can help in reducing considerably the dimension and complexity of a large sparse model, where sparsity can be achieved for the full posterior rather than just the mode as with lasso.

2.3 Compressed regressions

Random compressed regressions were first introduced by Raftery, Madigan, and Hoeting (1997). Early work on the theory of compressed regressions include Donoho (2006), Candes and Tao (2006) and Zhou, Lafferty and Wasserman (2007). More recently, Guhaniyogi and Dunson (2015) extend the analysis by introducing a Bayesian version of the model, which adds shrinkage to the model's parameters and also per- forms Bayesian Model Averaging. The compression regression is an alternative way of reducing the dimension of the regressors in the linear regression model in (1). The idea is to calculate the standard conjugate Gaussian-inverse gamma posterior for the regression coefficients conditional on a subset of regressors (compressed) many times for various random projections with differing subspace dimensions. Then model averaging, introduced by Raftery, Madigan, and Hoeting (1997), is employed to obtain a single posterior. The method requires no MCMC and has the advantage of being computationally fast and cheap.

Let

$$y = X\Phi\beta + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where X is $T \times N$, N is very large, possibly larger than T; Φ is a $N \times m$ matrix projection matrix and β is

a vector of coefficients of the compressed regressors, with dimension $m \times 1$. The matrix Φ is drawn independently of the data, and Guhaniyogi and Dunson (2015) suggest setting the ij^{th} element of Φ as

$$\Phi_{ij} = \begin{cases} -\sqrt{\frac{1}{\phi}} \text{ with probability } \phi^2 \\ 0 \text{ with probability } 2(1-\phi)\phi \\ \sqrt{\frac{1}{\phi}} \text{ with probability } (1-\phi)^2. \end{cases}$$

Finally, the rows of Φ are normalised by Gram-Schmidt orthonormalisation. Moreover, Guhaniyogi and Dunson (2015) estimate ϕ, assuming a uniform prior with values in (0, 1).

Conditional on Φ, the model has standard Normal-inverse Gamma posteriors as in (10), with regressors $X\Phi$ instead of X. In order to limit the sensitivity of the estimates to the choice of ϕ and m, Guhaniyogi and Dunson (2015) propose generating s random projection matrices Φ with different values for (ϕ,m), resulting into s regression models with different posteriors for β and σ^2. These are averaged across using Bayesian Model Averaging (BMA). Denote the different models by $\{M_1, \ldots, M_s\}$, then:

$$p(M_j|y,X) = \frac{p(y,X|M_j)p(M_j)}{\sum_{i}^{s}{}_1 p(y,X|M_i)p(M_i)}, \quad (12)$$

where the prior $p(M_j)$ is usually assumed to be 1, giving equal prior probability to each model. The marginal likelihood $p(y,X|M_j)$ can also be computed in closed form as

$$p(y,X|M_j) = \iint p(y,X|M_j,\beta^j(\sigma^2))p(\beta^j,(\sigma^2)^j)d\beta^j d(\sigma^2)^j,$$

where β^j and $(\sigma^2)^j$ denote $\beta$and σ^2 in model *j.* Forecast densities can thereafter be generated as

$$p(y_{N+h}|y_{1:N}, x_{1:N}, x_{N+h}) = \sum_{i=1}^{s} p(y_{N+h}|y_{1:N}, x_{1:N}, x_{N+h}, M_i)p(M_i|y,X).$$

An extension of the compressed regression approach to VAR models has been proposed by Koop, Korobilis and Pettenuzzo (2016).

From the computational point of view, the following algorithm can be used to implement compressed regression.

Algorithm

1. For each model $Mj, j = 1,\ldots,s$, and for each $i = 1,\ldots,N^{sim}$ iterate between Steps a and b:

a) Draw ϕ^{ji} and the resulting Φ^{ji}

b) Conditional on Φ^{ji}, draw β^{ji} and $(\sigma^2)^{ji}$ from (11).with regressors $X\Phi^{ij}$instead of X.

2. Compute $p(M_j|y,X)$ using (12) and average desired posterior quantities across all s models

$\{M1, \ldots, Ms\}$.

2.4 Bayesian VARs

Bayesian VARs (BVARs) can be considered as a multivariate extension of penalised regression models, where the right hand side variables (that is, the predictors) are lagged values of the left hand side variables. Suppose that we have an $N \times 1$ dimensional vector y_t generated by a VAR model of lag order p. Then, y_t can be written as:

$$y_t = B_0 + \sum_{i=1}^{p} B_i y_{t-i} + \varepsilon_t \tag{13}$$

where B_0 is an $N \times 1$ vector of intercepts, and B_i is an $N \times N$ matrix of autoregressive coefficients for lag $i = 1, \ldots, p$. The error term, ε_t, is an $N \times 1$ vector of normally distributed zero mean random variables, with a positive definite symmetric $N \times N$ contemporaneous covariance matrix R^{-1}. Then, ε_t can be written as: $\varepsilon_t = R^{-1/2}\eta_t$ where $\eta_t \sim NID(0_N, I_N)$.

VAR models are richly parameterised, containing $N + pN^2 + N(N+1)/2$ parameters. Therefore, a Bayesian estimation procedure is convenient, in particular for medium-large N, in order to avoid overfitting. Denote by $x_t := (1, y'_{t-1}, \ldots, y'_{t-p})$ a $1 \times (N_p + 1)$ vector and by $B := (B_0, B_1, \ldots, B_p)$ an $N \times (N_p + 1)$ matrix. Then, the model (13) can be written as

$$y_t = Bx'_t + \varepsilon_t$$

After vectorising, one can write

$$y_t = (I_N \otimes x_t)\beta + R^{-\frac{1}{2}}\eta_t, \tag{14}$$

where $\beta := vec(B')$ is an $N(N_p + 1) \times 1$ vector.

The likelihood of the sample $(y_1, \ldots, y_T)$ for the VAR(p) model (13) is given by

$$L(y|\beta, R, X) = (2\pi)^{-\frac{NT}{2}} |R|^{\frac{T}{2}} e^{-\frac{1}{2}\sum_{t=1}^{T}(y_t - (I_N \otimes x_t)\beta)' R (y_t - (I_N \otimes x_t)\beta)} \tag{15}$$

Denote by $Y = (y_1, \ldots, y_T)'$ a $T \times N$ matrix of stacked vectors $y'_1, \ldots, y'_T$ and define $y = vec(Y)$ as a $TN \times 1$ vector. Similarly, define $E = (\varepsilon_1, \ldots, \varepsilon_T)$ and the $TN \times 1$ vector $\varepsilon = vec(E)$. Let X be a $T \times N_p + 1$ matrix defined as $X = (x'_1, \ldots, x'_T)'$. Then, the likelihood (15) can be written in a more compact form as:

$$L(y|\beta, R, X) \propto |R|^{\frac{T}{2}} exp\left\{-\frac{1}{2}(y - (I_N \otimes X)\beta)'(R \otimes I_T)(y - (I_N \otimes X)\beta)\right\}. \tag{16}$$

Let us assume that β and R have Normal-Wishart prior distributions of the form

$$\beta|R \sim \mathcal{N}(\beta_0, (R \otimes \kappa_0)^{-1}), R \sim W(\alpha_0, \gamma_0), \tag{17}$$

where β_0 is a $(N_p + 1)N \times 1$ vector of prior means, κ_0 is a $(N_p + 1) \times (N_p + 1)$ positive definite symmetric matrix, α_0 is a scalar scale parameter of the Wishart distribution and γ_0 is a $N \times N$ positive definite symmetric matrix. Combining the likelihood L in (16) with the prior in (17), β and R have Normal-Wishart quasi-posterior distributions of the form

$$\begin{aligned} \beta\big|R, X, Y &\sim \mathcal{N}(\tilde{\beta}, (R \otimes \tilde{\kappa})^{-1}), \\ R &\sim W(\tilde{\alpha}, \tilde{\gamma}), \end{aligned} \tag{18}$$

with posterior parameters:

$$\begin{aligned} &\tilde{\beta} = (I_N \otimes \kappa^{-1})\left[(I_N \otimes X'X)\hat{\beta} + (I_M \otimes \kappa_0)\beta_0\right], \\ &\tilde{\kappa} = \kappa_0 + X'X, \qquad \tilde{\alpha} = \alpha_0 + T, \\ &\tilde{\gamma} = \gamma_0 + Y'Y + B_0\kappa_0 B_0' - \tilde{B}\tilde{\kappa}_0\tilde{B}', \end{aligned} \tag{19}$$

where

$$\hat{\beta} = (I_N \otimes (X'X)^{-1}X')y \tag{20}$$

is the OLS estimator for β.

The Gaussian assumption together with the independent Normal-Wishart prior above delivers posteriors in closed form, which alleviates the need for MCMC simulations and makes estimation fast. The prior parameters are set by the researcher; however, in a large dimension context, this could be a difficult task. For example,κ_0 is a $(N_p + 1) \times (N_p + 1)$ shrinkage matrix, so even for small models it requires specifying a large number of prior parameters (e.g. for a VAR with 10 variables and 4 lags, 1681 prior parameters are required). Therefore, it is desirable to have an automatic way to set the priors and ideally a much smaller number of tuning parameters should be used. This is the idea behind the so-called Minnesota prior of Doan, Litterman and Sims (1984) and Litterman (1986), where all prior means β_0 are set to zero and

$$\kappa_{0,ij} = \begin{cases} \lambda \dfrac{1}{k^2} \; for \; j = i, \forall k = 1, \dots, p \\ \lambda \dfrac{1}{k^2} \dfrac{\sigma^2}{\sigma_j^2} \; for \; j \neq i, \forall k = 1, \dots, p \end{cases}$$

where σ^2 denotes the i^{th}variable variance and is usually estimated from the data prior to estimation.

The ratio σ_i^2/σ_j^2 accounts for the different scale and variability of the data. The factor $1/k^2$ is the rate at which the prior variance decreases with increasing lag length.

Kadiyala and Karlsson (1997) propose a way to automatically set the prior parameters for the volatility α_0 and γ_0. Another issue that has received attention in the literature is optimal shrinkage, determining the optimal value of the hyperparameter λ. For example, Carriero, Clark and Marcellino (2015) use a grid search to find the value of λ that delivers best forecasting performance, while Giannone, Lenza and Primiceri (2015) use a hierarchical prior for λ, showing that maximising the posterior of λ corresponds to maximising the marginal likelihood and hence the one step ahead forecasts.

A number of extensions to the model in (13) have been proposed. We discuss some of them below.

2.4.1 REDUCED RANK VARS

Reduced Rank Regressions (RR) have a long history in the time series literature but have been mainly applied in small models, see e.g. Velu, Reinsel, Wichern (1986), Reinsel (1983), Reinsel and Velu (1998), Camba-Mendez, Kapetanios, Smith and Weale (2003). Carriero, Kapetanios and Marcellino (2011, 2015) show that RR are also well suited for medium to large datasets of the dimension typically of interest for central banks, i.e. about 50-60 variables. In theory, the methods can deal with larger datasets, but this poses serious computational burdens. In particular, as the number of regressors grows, classical RR can suffer numerical problems in the estimation of the covariance matrix of the unrestricted residuals while Bayesian RR requires simulations involving in each step the inversion of larger matrices. For this reason, a careful pre-selection or aggregation of the proper big dataset of indicators should be performed prior to the application of the RR or BVAR approaches (and this can be done with the methods discussed so far).

It is often the case that estimation of VAR(p) models results in a large number of insignificant coefficients. Therefore, in order to obtain a more parsimonious model, one might impose rank reduction, i.e., assume that $rk(B') = r < N$. This is equivalent to the parametric specification:

$$Y_t = \alpha\left(\sum_{i=1}^{p}\beta_i' Y_{t-i}\right) + e_t = \alpha\beta' X_t + e_t, \tag{21}$$

where α and $\beta = (\beta_1', \ldots, \beta_p')'$ are, respectively, $N \times$r and $M \times r$ matrices. The model (21) was studied by Velu et al. (1986). Ahn and Reinsel (1988) suggested a more general specification where the rank of the coefficient matrix on each lagged vector of the explanatory variables may differ. However, this generalization creates computational problems in the large N case. Therefore, we focus on (21).

In equation (21), it is assumed that the true rank of the matrices α and β are identical and equal to r which is thus referred to as the rank of the system (21). However, note that the ranks of $\beta_i, i = 1, \ldots, p$, need not equal r; in particular, it can be $rk(\beta_i) \leq r, i = 1, \ldots, p$.

An interesting special case of the RRVAR model (21), which resembles the autoregressive index model of Reinsel (1983), results if $\beta_i = \beta_* K_i$ with $rk(\beta_*) = r$ for some (r, r) matrix K_i which need not be full rank, $i = 1, \ldots, p$, although $K = (K_1', \ldots, K_p')'$ is. Hence, $\beta = (I_p \otimes \beta_*)K$ and $\beta_i' = \alpha_i \beta_*'$, where $\alpha_i = \alpha K_i'$, in which case $\beta_*' y_{t-i}, i = 1, \ldots, p$, may be interpreted as dynamic factors for y_t.

Given the assumed system rank r, Velu et al. (1986) suggested an estimation method for the parameters α and β that may be shown to be quasi-maximum likelihood (see also Reinsel and Velu, 1998). Denote the sample second moment matrices by $S_{YY} = T^{-1}Y'Y, S_{YX} = T^{-1}Y'X, S_{YX} = S_{XY}'$, and $S_{XX} = T^{-1}X'X$. Hence, the covariance matrix of the unrestricted LS residuals,$S_{YY,X} = S_{YY} - S_{YX}S_{XX}^{-1}S_{XY}$

is the unrestricted quasi-ML estimator of the error process variance matrix. Additionally, let $\{\lambda\}_t^T{}_1, \lambda_1^2 \geq, \lambda_2^2 \geq \ldots \geq, \lambda_N^2 \geq 0$ denote the ordered squared eigenvalues of the $N \times N$ matrix $S_{YY,X}^{-1/2} S_{YX} S_{XX}^{-1} S_{XY}\, S_{YY,X}^{-1/2}$ with associated eigenvectors $\{v_i\}_{t=1}^T$ subject to the normalization $v_i' v_j = 1$ if $i = j$ and 0 otherwise, and let $\hat{V} = (v_1, v_2, \ldots, v_r)$. The quasi-ML estimators for α and β are given by $\hat{\alpha} = S_{YY-X}^{1/2} \hat{V}$ and $\hat{\beta} = S_{XX}^{-1} S_{XY}\, S_{YY,X}^{-1/2} \hat{V}$, so that $\hat{B}' = S_{YY,X}^{1/2} \hat{V} \hat{V}' S_{YY,X}^{-1/2} S_{XX}^{-1}\, S_{XY}$.

2.4.2 BAYESIAN REDUCED RANK VAR MODELS

It is possible to impose both rank reduction and shrinkage on the VAR. Bayesian analysis of reduced rank regression has been introduced by Geweke (1996). As for the reduced rank case, the $M \times N$ matrix of coefficients B is assumed to have rank r, where $r < N$. This rank reduction assumption is equivalent to the parametric specification

$$Y = X\Psi\Phi + E \tag{22}$$

with Ψ and Φ being respectively $M \times r$ and $r \times N$ matrices. To identify these matrices Geweke (1996) proposes the following normalization:

$$\Phi = [I_r | \Phi^*]. \tag{23}$$

Given this normalization, a proper prior is:

$$|\Sigma|^{-(N+v_0+1)} exp\left[-\frac{1}{2} tr S_0 \Sigma^{-1}\right] exp\left[-\frac{r^2}{2}(tr\Phi^{*\prime}\Phi^* + tr\Psi'\Psi)\right], \tag{24}$$

namely, the product of an independent Wishart distribution for Σ with v_0 degrees of freedom and matrix parameter S0, and independent $N(0, \tau^{-2})$ shrinkage priors for each element of the coefficient matrices Φ^* and Ψ. The conditional posterior distribution of Σ is:

$$|\Sigma|(\Phi^*, \Psi, X, Y) \sim IW[T + v_0, S_0 + (Y - XB)'(Y - XB)]. \tag{25}$$

The conditional posterior distributions of the coefficients Φ^*, Ψ, are multivariate normals.

In particular, the conditional posterior distribution of Φ^* is:

$$vec(\Phi^*)|(\Psi, \Sigma, X, Y) \sim N\left[\Pi_\Phi * vec\left(\widehat{\Phi}^*\right), \Pi_\Phi\right], \tag{26}$$

where:

$$\widehat{\Phi}^* = (\Psi' X' X \Psi)^{-1} \Psi' X' Y_1 \Sigma^{12} (\Sigma^{22})^{-1} - \Sigma^{12} (\Sigma^{22})^{-1} \tag{27}$$

$$+ (\Psi' X' X \Psi)^{-1} \Psi' X' Y_2,$$
$$\Pi_\Phi = [(\Sigma^{22})^{-1} \otimes (\Psi' X' X \Psi)^{-1} + \tau^2 I_{r(N-r)}]^{-1}, \quad (28)$$

and where $Y = [Y_1|Y_2]$ is a partitioning of Y into its first r and last $N-r$ columns, while Σ^{ij} denotes the partitioning of Σ^{-1} into its first and last $N-r$ rows and columns.

The conditional posterior distribution of Ψ is:

$$vec(\Psi)|(\Phi, \Sigma, X, Y) \sim N\left[\Pi_\Psi * vec(\widehat{\Psi}), \Pi_\Psi\right], \quad (29)$$

where:

$$\widehat{\Psi} = \hat{B}\left[\Phi^+ + \Phi^0 \tilde{\Sigma}^{21}\left(\tilde{\Sigma}^{11}\right)^{-1}\right], \quad (30)$$
$$\Pi_\Psi = [\tilde{\Sigma}^{11} \otimes X'X + \tau^2 I_{Mr}]^{-1}, \quad (31)$$

and where $\hat{B}$ is the OLS estimator, Φ^+ is the generalized inverse of Φ, Φ^+ is column-wise orthogonal to Φ^+, and where $\tilde{\Sigma}^{ij}$ denotes the partitioning of $\tilde{\Sigma}^{-1} = ([\Phi + \Phi^0]' \Sigma [\Phi^+ \Phi^0])^{-1}$ into its first r and last $N-r$ rows and columns.

Unconditional posterior distributions can be simulated by using a Gibbs sampling algorithm which draws in turn from (26), (29), and (25). See Geweke (1996) for details.

The BRR has the shortcoming of being computationally challenging when the assumed rank is high, as the estimation of this model requires simulation involving inversion of Mr -dimensional matrices. As a computationally quicker way to impose both rank reduction and shrinkage, Carriero, Kapetanios and Marcellino (2011) suggest to simply impose rank reduction on the posterior estimates of a BVAR.

The implementation of the method is straightforward. First, the system is estimated under the prior distribution described by equation (18), then a rank reduction is imposed as follows. Let $\hat{B}$ be the posterior mean of B and let $\hat{B} = U \Lambda V$ be its singular value decomposition. Collecting the largest r singular values and associated vectors in the matrices $\Lambda^* = diag(\lambda_1, \lambda_2, \ldots, \lambda_r)$. $U^* = (u_1, u_2, \ldots, u_r)$ and $V^* = (v_1, v_2, \ldots, v_r)$ a reduced rank approximation (of rank r) of the posterior mean is given by:

$$\hat{B}_r^* = U^* \Lambda^* V^*, \quad (32)$$

which is the RRP estimator.

2.4.3 TIME VARYING PARAMETER VAR MODEL

Most of the methods that we have considered so far assume that the model parameters are stable over time, a common hypothesis in the statistical literature which is unfortunately often violated in empirical economic analyses. Hence, we now briefly discuss how to allow for time variation in a VAR model, with similar techniques applicable also in several other models and estimation methods.

In order to accommodate the possibility of structural change in the parameters and/or the volatility of the VAR model in (13), we can write it as:

$$y_t = B_{0,t} + \sum_{i\ 1}^{p} B_{i,tyt-i} + \varepsilon_t, \quad \varepsilon_t \sim NID(0_N, R_t^{-1}),$$

where now all parameters are indexed by the time index t. The standard estimation approach of time varying parameter (TVP) VAR models employs state space methods. In a seminal paper, Cogley and Sargent (2002) study the changing dynamics of macroeconomic variables in the U.S. using a TVP VAR with autoregressive coefficients modelled as random walk processes. To proceed, stack the drifting parameters in a $N(Np+1)\times 1$ vector $\beta_t = vec(B_{0,t}, B_{1,t}, \ldots, B_{p,t})'$ and specify a process of the form

$$\beta_t = \beta_{t-1} + v_t, \quad v_t \sim N(0, Q).$$

Cogley and Sargent (2002) further assume that

$$\mathrm{E}\begin{bmatrix}\varepsilon_t \\ v_t\end{bmatrix}[\varepsilon_t' v_t'] = V = \begin{bmatrix}R^{-1} & 0 \\ 0 & Q\end{bmatrix}$$

and specify an inverse-Wishart prior for V of the form:

$$p(V) = IW(V_0^{-1}, T_0)$$

and the initial condition for the Kalman filter is

$$p(\beta_0) \sim N(b, P). \tag{33}$$

Below, we outline the estimation algorithm from Cogley and Sargent (2002).

Gibbs Algorithm

1. Initialise the algorithm with a guess for V. Then, for $j = 1,..,N^{sims}$ iterate between steps 2 and 3 below.
2. Conditional on the data y_t and the prior hyperparameters, the transition law of motion for β_t is linear and Gaussian, so the Kalman filter can be used with the initial condition in (33) to draw the history of time varying parameters $\beta_t, t = 1, \ldots, T$.
3. Conditional on y_t and a draw from the parameters' history β_t, the innovations ε_t and v_t are observable. Draw the hyperparameters V from their conjugate posterior

$$p(V|y_{1:T}, \beta_{1:T}) = IW(\tilde{V}^{-1}, \tilde{T})$$

where

$$\tilde{V} = V_0 + \frac{1}{T}\sum_{l=1}^{T}\begin{bmatrix}\varepsilon_t \\ v_t\end{bmatrix}[\varepsilon_t' v_t']$$

and

$$\tilde{T} = T_0 + T.$$

2.4.4. STOCHASTIC VOLATILITY VAR MODELS

A different strand in the literature has insisted on the importance of stochastic volatilty in the VAR model above (e.g Sims (1980), Bernanke and Mihov (1998), Kim and Nelson (1999), McConnell and Perez Quiros (2000), Sims and Zha (2006) and Primiceri (2005). Primiceri (2005) and Cogley and Sargent (2005) allow for drifting volatility R_t^{-1} of the VAR model, the former by employing a procedure suggested by Kim, Shephard and Chib (1998) and the latter by utilising a Metropolis within Gibbs algorithm with a technique from Jacquier, Polson and Rossi (1994).

Primiceri (2005) estimates the model

$$y_t = B_{0,t} + \sum_{i=1}^{p} B_{i,t} y_{t-i} + R_t^{-1/2}\varepsilon_t, \qquad \varepsilon_t \sim NID(0_N, I_N)$$

Now the volatility R_t^{-1} is changing over time, with

$$R_t^{-1} = A_t \Omega_t A_t'$$

where A_t is a lower triangular matrix with ones on its main diagonal and nonzero elements stacked into a vector at, and Ω_t is a diagonal matrix with diagonal elements $[\sigma_{1,t}, \ldots, \sigma_{N,t}]$. In addition to the assumption made in Cogley and Sargent (2002) that

$$\beta_t = \beta_{t-1} + v_t, \qquad v_t \sim N(0, Q),$$

Primiceri (2005) also assumes that

$$a_t = a_{t-1} + \zeta_t$$

$$\log \sigma_t = \log \sigma_{t-1} + \eta_t. \tag{34}$$

Moreover, the disturbances $[\varepsilon_t, v_t, \zeta_t, \eta_t]$ are assumed to have a diagonal covariance matrix of the form

$$V = \begin{bmatrix} I_M & 0 & 0 & 0 \\ 0 & Q & 0 & 0 \\ 0 & 0 & S & 0 \\ 0 & 0 & 0 & W \end{bmatrix}.$$

Recall that

$$y_t = (I_M \otimes x_t)\beta_t + R_t^{-1/2}\varepsilon_t$$

where $x_t := (1, y'_{t-1}, \ldots, y'_{t-k})a1 \times (Np + 1)$ vector and by $\beta_t = vec(B_0, B_1, \ldots, B_p)'$ an $N(Np + 1)$ vector. the key idea of the algorithm is to note that, conditional on V and the history for $\beta_{1:T}$ and $a_{1:T}$, the model is linear in $\log \sigma_t$ because

$$\log\ (A_t^{-1}y_t - A_t^{-1}(I_N \otimes x_t)\beta_t)^2 = \log \sigma^2 + \log \varepsilon_t^2. \tag{35}$$

However, the model is no longer Gaussian as $\log \varepsilon_t^2 \sim \chi^2(1)$. However, as suggested in Kim, Shephard and Chib (1998), $\log \varepsilon_t^2$ can be approximated with a mixture of normal distributions (more details on the means and variances of the seven Normal distributions can be found in Kim, Shephard and Chib (1998)). Conditional on the mixture of normals $s_{1:T}$, and on V , $\beta_{1:T}$ and $a_{1:T}$, equations (35) and (34) are a linear and Gaussian, so the Kalman filter can be applied.

An estimation algorithm for the model with SV proceeds as follows:

Gibbs Sampling Algorithm

1. Initialise the algorithm with a guess for V, and the history for $\beta_{1:T}$ and $a_{1:T}$.
   Then, for $j = 1, .., N^{sims}$ iterate between steps 2 through 5 below.
2. Draw $\log \sigma_t$, conditional on $s_{1:T}$, and on V , $\beta_{1:T}$ and $a_{1:T}$.
3. Draw $a_{1:T}$. conditional on V, and the history for $\beta_{1:T}$ and $\sigma_{1:T}$.
4. Draw (β_t, s_t) conditional on V, $\sigma_{1:T}$ and $a_{1:T}$ in two steps
   (a) Draw β_t conditional on V, $\sigma_{1:T}$ and $a_{1:T}$.
   (b) Draw a_t conditional on $\beta_{1:T}$, V, σ1:T and $a_{1:T}$
5. Draw V from its conjugate inverted Wishart posterior, conditional on $\sigma_{1:T}$, $s_{1:T}$, $\beta_{1:T}$ and $a_{1:T}$.

More recently, Cogley, Primiceri and Sargent (2010) propose a VAR model, which in addition to drifts in the parameters and volatilities, also features time varying volatility in the state equations of the autoregressive parameters.

The estimation procedure outlined above is very computationally demanding, so that VAR-SV models are typically only applied for very small values of , say N =3 or 4. However, Carriero, Clark and Marcellino (2016) develop a particular estimation algorithm that allows for very large values of N, they present an application where $N > 100$, even though it cannot still be directly applied to the case of big data.

2.4.5 NONPARAMETRIC TIME VARYING PARAMETERS

An alternative to the parametric state space approach to handle parameter time variation is presented in Giraitis, Kapetanios and Yates (2014). They propose a nonparametric method for the estimation of the coefficient and variance processes in a time varying linear regression setting and establish the theoretical properties of their kernel-type estimator. Their method accommodates consistent and asymptotically normal extremum estimation and has been extended to a general local likelihood framework. In addition, Petrova (2017) extends the approach to provide a Bayesian treatment for it, while Kapetanios, Marcellino and Venditti (2016) introduce a con- strained version of the model. Both extensions allow to handle large dimensional models, which was not feasible in the original proposal.

Here, we provide an overview of the frequentist estimators in Giraitis et al. (2014), while we refer to the papers mentioned above for more details on the large N case.

Let y_t be an observed time series with log-density $l_t(y_t|y^{t-1}, \theta_t)$, conditional on history $y^{t-1} = \{y_1, \dots, y_{t-1}\}$, and depending on a time varying finite-dimensional vector of parameters θ_t, satisfying one of the conditions (i)-(ii) presented below.

(i) For each $t \in \{1, \dots, T\}, \theta_t$ is a deterministic function of time given by

$$\theta_t = \theta \frac{t}{T} \tag{36}$$

where $\theta(.)$ is a piecewise differentiable function.

(ii) θ_t is a vector-valued stochastic process satisfying: for $1 \leq h \leq t$ as $h \to \infty$

$$\sup_{j:\,|j-t| \leq h} ||\theta_t - \theta_j||^2 = O_p\left(\frac{h}{l}\right). \tag{37}$$

Both (36) and (37) imply that the sequence of parameters drifts slowly with time, a property that is important for consistent estimation of θ_t. An extremum estimator $\hat{\theta}_j = \arg\max_\theta \ell_j(\theta_j)$ for θ_j is derived by maximising an objective function given by

$$\ell_j(\theta_j) := \sum_{t=1}^{T} w_{jt} l_t\,(y_t|y^{t-1})\;\; j \in \{1, \dots, T\} \tag{38}$$

where $l_t(y_t|y^{t-1})$ is the conditional log-density for observation t and the weights w_{jt} are computed using a kernel function and normalised to sum to one:

$$w_{jt} = \widetilde{w}_{jt} / \textstyle\sum_{t-1}^{T} \widetilde{w}_{jt},\; \widetilde{w}_{jt} = K\left(\frac{j-l}{H}\right) \text{ for } j, t \in \{1, \dots, T\}. \tag{39}$$

The kernel function K is assumed to be non-negative, continuous and bounded function with domain

$\mathbb{R}$. The bandwidth parameter H satisfies $H \to \infty$ and $H = o(T/\log T)$. For example, the widely used Normal kernel weights are given by

$$\widetilde{w}_{jt} = (1/\sqrt{2\pi})\exp((-1/2)((j-t)/H)^2) \text{ for } t, j = 1, .., T, \tag{40}$$

while the rolling window procedure results as a special case of the choice of a flat kernel weights: $w_{jt} = \mathbb{I}(|t-j| \leq H)$. For further discussion of the advantages of exponential kernels over the flat kernel for introducing time variation, refer to Girairis et al. (2014) and Giraitis, Kapetanios and Price (2013) In this setup, Giraitis, Kapetanios, Wetherilt and Zikes (2016) show that, under regularity conditions, $\hat{\theta}_j$ is an $H^{1/2} + (T/H)^{1/2}$ – consistent estimator of θ_j for all $j = [\tau T], 0 < \tau < 1$. Furthermore, defining

$$\hat{\Sigma}_{Tj} := \frac{1}{\varkappa_{Tj}}\left(-\frac{\partial^2 \ell_j(\hat{\theta}_j)}{\partial\theta\,\partial\theta'}\right), \quad \varkappa_{Tj} := \left(\sum\nolimits_{t=1}^{T} w_{jt}^2\right)^{-1} \tag{41}$$

a.s. positive definiteness of $\hat{\Sigma}_j$ and the bandwidth rate $H = o(T^{1/2})$ are sufficient for asymptotic normality of $\hat{\theta}_j$:

$$\hat{\Sigma}_{Tj}^{-\frac{1}{2}}\left(\hat{\theta}_j - \theta_j^0\right) \to_d \mathcal{N}(0, I) \text{ as T } \to \infty$$

for all $j = [\tau T], 0 < \tau < 1$, with $\mathcal{N}(0, I)$ denoting the multivariate standard normal distribution.

2.5 Quantile and expectile regressions

The starting point of quantile regressions (QR) models in the linear regression model in (1) but the emphasis is on modelling the quantiles of the dependant variable y conditional on the set of explanatory variables X: $Q\tau(y|X)$, where $0 < \tau < 1$ is the τ^{th} quantile of y. The τ^{th} quantile can be estimated via minimisation of a 'check' objective function

$$\begin{aligned} \hat{\beta}_\tau &= \min_{\beta} \sum_{i=1}^{T} \left(\rho_\tau(y_i - x_i'\beta)\right) \\ &= \sum_{i=1}^{T} \left((1-\tau)(y_i - x_i'\beta)_- + \tau(y_i - x_i'\beta)_+\right), \end{aligned} \tag{42}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is known as a 'check' function (see, e.g., Koenker and Bassett (1978) for more details). The asymmetric weights τ and $1-\tau$ basically weight observations below and above the quantile of interest differently to shift the estimate to upper or lower parts of the sample. Computationally, the minimisation in (42) is solved by linear programming which can be computationally costly. When $\tau = 0.5$, the check function is symmetric and minimisation results in the ordinary least squares. Moreover, the conditional quantile function is linear and given by

$$Q_\tau(y|X) = X\hat{\beta}_\tau.$$

Poiraud-Casanova and Thomas-Agnan (2000) and Komunjer (2005) show that the minimisation problem in (42) is equivalent to doing Quasi-Maximum Likelihood (QML) with tick-exponential family of densities – family whose role in Quantile Regression is analogous to the role of the linear-exponential family in mean regression estimation; with the most well-known member of the tick-exponential family being the asymmetric Laplace density function.

Adopting a Bayesian approach for the estimation of the QR model is useful when the number of explanatory variables to condition on is very large and can help avoid overfitting. Sampling from the posterior of the parameters is done through standard Normal-inverse Gamma conjugacy results.

The QR model looks at the entire conditional distribution of y, rather than just at the conditional expectation of y given X (which is the case with most methods surveyed earlier). QR models can therefore be useful for modelling the tail behaviour of time series in the presence of large data and can help in forecasting the probability of time series exceeding some threshold. While this might be useful for financial time series, where emphasis is placed on studying the tail behaviour, it should not have a major effect when forecasting macroeconomic variables. Moreover, while the sample size for financial time series is typically very long, due to high frequency sampling, the sample size for macroeconomic time series is not so long, reducing the degrees of freedom for estimation of low and high percentiles.

Expectile regression is an extension of quantile regression. Just as the idea of the median may be extended to quantiles, so the concept of the mean may be extended to percentiles (or expectiles). The population expectiles are similar to quantiles but they are determined by tail expectations rather than tail probabilities. The ω^{th} expectile of y determines the point at which $\omega\%$ of the mean distance between the expectile and y comes from the mass below it. Expectile regression requires minimizing an asymmetrically weighted least-squares criterion:

$$\begin{aligned}\hat{\beta}_{\omega} &= \min_{\beta} \sum_{i\ 1}^{T} (\rho_{\omega}(y_i - x_i'\beta)) \\ &= \sum_{i=1}^{T} ((1-\omega)((y_i - x_i'\beta)_{-})^2 + \tau((y_i - x_i'\beta)_{+})^2),\end{aligned} \tag{43}$$

Similarly to the Bayesian quantile regression, the asymmetric quasi-likelihood can be considered as a likelihood of the data, which can be combined with a prior (see for example Yue and Rue (2011)). The expectile regression is very similar to quantile regression but, due to the square term in the loss function, it is less robust to outliers, which can be problematic in particular when working with big data.

2.6 Other methods covered in Kapetanios, Marcellino and Papailias (2016)

2.6.1 SICA

Smooth Integration of Counting and Absolute Deviation (SICA) is another penalised regression, introduced by Lv and Fan (2009). It results from the optimisation problem:

$$\hat{\beta}^{SICA} = \min_{\beta} \frac{1}{n} \left\{ \varepsilon'\varepsilon + \lambda \frac{(\alpha+1) \parallel \beta \parallel_1}{(\alpha + \parallel \beta \parallel_1)} \right\}, \tag{44}$$

with α varying from 0 to ∞, providing a smooth homotopy between the L_0 and L_1 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 $\lfloor l0, \infty)$.

The above family of penalties satisfies the following conditions:

$\rho(t)$ 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 unbiasedness, continuity and sparsity; see Lv and Fan (2009) for more information. The method is attractive for big data modelling as it avoids the single use of the L_0 norm, which is impractical in high dimensions.

2.6.2 HARD THRESHOLDING

Zheng, Fan and Lv (2014) consider sparse regression with a hard thresholding penalty, with objective function:

$$\hat{\beta}^{TH} = \min_{\beta} \frac{1}{n} (\varepsilon'\varepsilon + \frac{1}{2}\lambda^2 - (\lambda - \parallel \beta \parallel_1)^2) \tag{45}$$

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 $S_{M/2} = \{\beta \in R^N : \beta_0 \leq M/2\}$(where $\beta_0 = \#(i|\beta_i \neq 0)$ denotes the number of non-zero coefficients) as:

$$\hat{\beta}^{SICA} = \min_{\beta \in S_{M/2}} Q^{TH}(\beta). \tag{46}$$

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

1. $\varepsilon_t \sim N(0, \sigma^2 I_T)$ for some positive σ.
2. It holds that $s < M/2, s = o(T)$ and $b = \min_{j \in sup\beta_0} |\beta_{0,j}| > \{\sqrt{16/c^2} \vee 1\} c^{-1} c_2 \sqrt{\{(2s+1)\log(\tilde{N}/T)\}}$ 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 $\tilde{N} = T \vee N$.
3. $\sum_{i=1}^{N} \beta_i^2$ is bounded from below by some positive constant and

$$\max_{\#(i|\delta_i \neq 0) < M/2,), \sum_{i=1}^{N} \perp \ \delta_i^2 = 1 N \delta 2 = 1} T^{-1/2} \sum_{i=1}^{N} (X_i \delta_i)^2 \leq c_3$$

for some positive constant c_3.

2.6.3 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.). The idea is to start by selecting some of the regressors, estimate the model and calculate the criterion value. Then, 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,

$$IC(I) = -2L(I) + CT(I) \tag{47}$$

where $L(I)$ is the log-likelihood of the model associated with string I and C_T (I) is the penalty term associated with the string I. The three most usual penalty terms are $2\tilde{m}$ (I), $ln(T)\tilde{m}$ (I) and $2ln(ln(T))\tilde{m}$(I) associated with AIC, BIC and HQ information criteria. $\tilde{m}$ (I) is the number of free parameters associated with the modelling of the dataset associated with I. Note that, in this case, $\tilde{m}$ (I) = III. 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) $C_T(I) \to \infty$ and (ii) $C_T(I)/T \to 0$.

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. The drawback with this approach in the case of large k is that the number of possible models resulting from diff t combination of regressors increases exponentially fast: the total number of models to be compared is 2^k, making the approach computationally infeasible.

To overcome this difficulty, several heuristic optimisation approaches have been suggested, including among the main ones: simulated annealing (Goffe, Ferrier and Rogers (1994)), genetic algorithm (Kapetanios (2006)), MC^3 (Fernandez et al. (2001)) and sequential testing (Hendry (1995)), which we review in the next subsections.

2.6.4 SIMULATED ANNEALING

This algorithm provides a local search for the minimum (or maximum) of a function, in our case is Equation (47). 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. $\{I = (I_1, \ldots, I_N)' | I_i \in \{0,1\}\}$.

Each step of the algorithm works as follows, starting from an initial string I_0.

1. Using I_i choose a neighboring string at random, denoted I_{i+1}^*. We discuss the definition of a neighborhood below.
2. If IC(I_i) > IC(I_{i+1}^*), set $I_{i+1} = I_{i+1}^*$. Else, set $I_{i+1} = I_{i+1}^*$ with probability $e^{(\mathrm{IC}(I_i) > \mathrm{IC}(I_{i+1}^*))/T_i}$ or set $I_{i+1} = I_i$ with probability $1 - e^{(\mathrm{IC}(I_i) > \mathrm{IC}(I_{i+1}^*))/T_i}$.

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 neighbor- hood 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, I_1, is defined as the smallest number $E > 0$ such that the function exceeds $IC(I_1) + E$ during its trajectory[2] from this minimum to any other local minimum, I_2, for which $IC(I_1) > IC(I_2)$.

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 $2L(I_B) - 2L(I_A)$ where $L(I_A)$ is the likelihood associated with a regression containing just a constant term and $L(I_B)$ is the likelihood associated with a regression containing the maximum allowable number of variables. Of course, there are many

[2] A trajectory from I_1 to I_2 is a set of strings, $I_{11}, I_{12}, \ldots, I_{1p}$, such that (i) $I_{11} \in N(I_1)$, (ii) $I_{1p} \in N(I_2)$,) and (iii) $I_{1i_{+1}} \in N(I_{1i})$, for all $i = 1, \ldots, p$, where $N(I)$ denotes the set of strings that make up the neighborhood of I.

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 $-2L(I_B)$, for some random 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.

2.6.5 GENETIC ALGORITHMS

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 (2006), 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 $(a_1^\alpha, a_2^\alpha, \ldots, a_N^\alpha)$, $(a_1^\beta, a_2^\beta, \ldots, a_N^\beta)$. Choose a random integer between 2 and $N-1$. Denote this by j. Replace the pair by the following pair: $(a_1^\alpha, a_2^\alpha, \ldots, a_j^\alpha, a_{j+1}^\beta, \ldots, a_N^\beta)$, $(a_1^\beta, a_2^\beta, \ldots, a_j^\beta, a_{j\,|1}^\alpha, \ldots, a_N^\alpha)$. Perform cross over on each pair with probability p_c. Denote the new population. 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 pre-specified 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).

2.6.6 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:

$$B_{i+1,i} = \left(\frac{g_{0i+1}}{g_{0i+1}+1}\right)^{\frac{k_{i+1}}{2}} \left(\frac{g_{0i}+1}{g_{0i}}\right)^{\frac{k_i}{2}} \left(\frac{\frac{1}{g_{0i}+1}RSS_i + \frac{g_{0i}}{g_{0i}+1}TSS}{\frac{1}{g_{0i+1}\,|1}RSS_{i+1} + \frac{g_{0i+1}}{g_{0i+1}\,|1}TSS}\right)^{\frac{T-1}{2}}, \qquad (48)$$

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 (48).

2.6.7 SEQUENTIAL TESTING

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 (2014) sheds some extra light on the use of *Autometrics*[3] in statistical model selection with big data.

A detailed description of the algorithm 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

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

only 10, should be considered. (ii) In step B(d) $CUSUM^2$ should be used instead of Chow as a stability test. (iii) No out-of-sample evaluation should be undertaken, since this would change the information set for the other algorithms.

3 A review of non-parametric methods for Big Data

In the previous section we have reviewed methods based on the specification of a parametric model, typically a linear regression, which links the target variable y with a, possibly big, number of explanatory variables X. In this section we consider other methods that do not require an explicit parametric formulation of the relationship between y and X, focusing on those cases where X can be big.

3.1 Regression trees

Regression trees are based on a partition of the space of the dependent variable y into M subsets R_m, with y allocated to each subset according to a given rule and modelled as a different constant c_m in each subset. This is a powerful idea, since it can fit various functional relationship between y and a set of explanatory variables X, say $y = \boldsymbol{f}(\boldsymbol{X})$, without imposing linearity or additivity, which are commonly assumed in standard linear regression models. Let

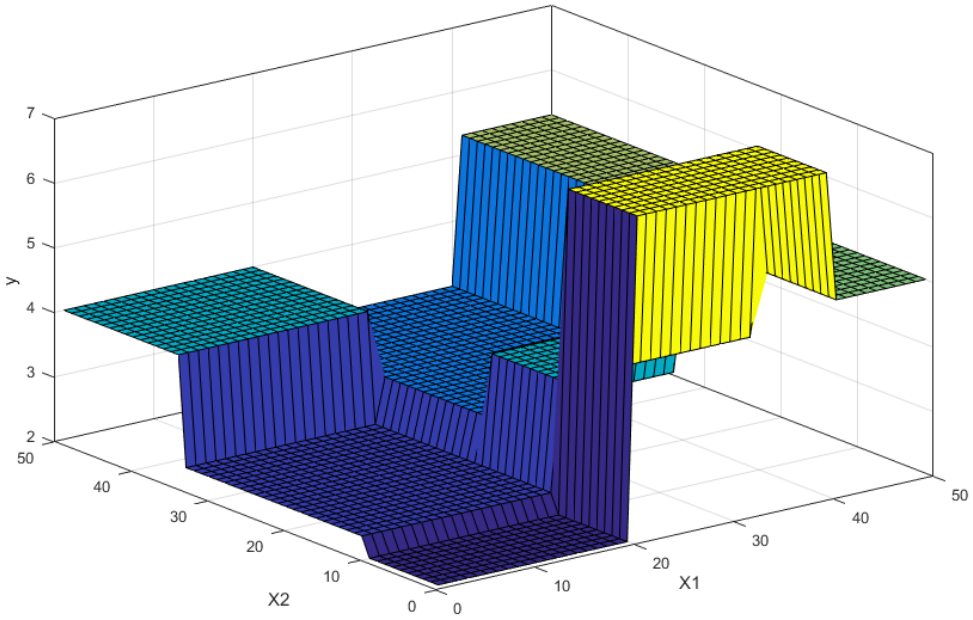
$$y = f(X) = \sum_{m=1}^{M} c_m 1\,(X \in R_m),$$

where 1 denotes the indicator variable taking value 1 if the condition is satisfied, 0 otherwise. Then, given a partition, minimizing

$$\| y - f(y) \|_2 = \sum_{i=1}^{N} \left(y_i - f(y_i)\right)^2, \tag{49}$$

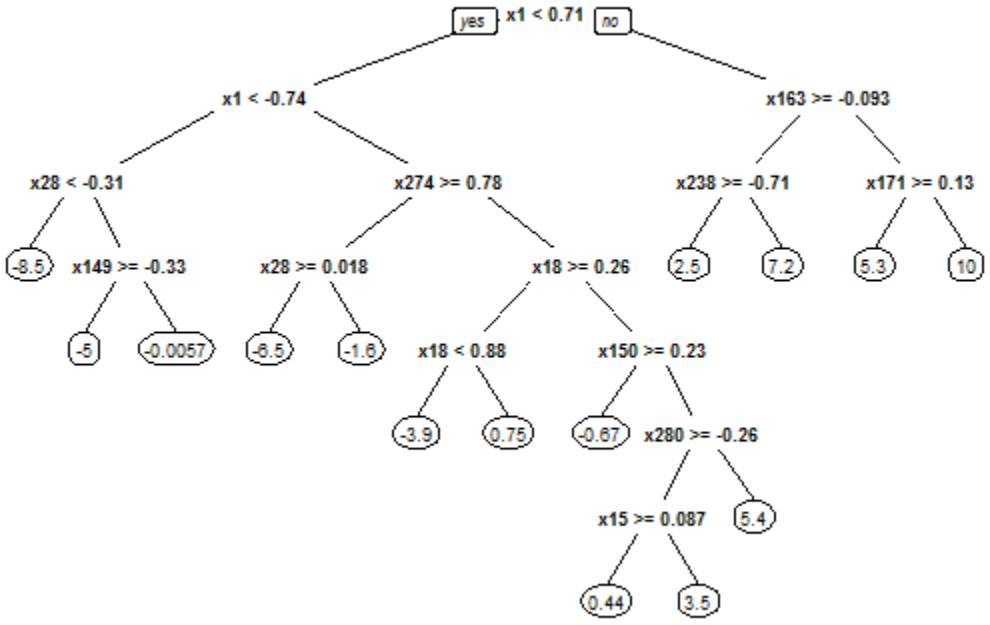
with respect to c_m yields $\hat{c}_m = \bar{y}_m$, where $\bar{y}_m$ denotes the sample mean of y over each region R_m.

Figure 2: Example of a Regression tree with two explanatory variables X_1 and X_2



Source: Based on author's calculations

Figure 3: Example of a Regression tree with many explanatory variables.



Source: Based on author's calculations

A much more difficult problem is to find the best partition in terms of minimum sum of squares (49). Even in the two dimensional case, i.e when $k = 2$ so that $X = [x1, x2]$, finding the best binary partition to minimise (49) is not computationally feasible. Instead, greedy algorithms are commonly used. The idea is to do one split at a time. Consider a splitting variable j (where $j = 1, .., k$) and a splitpoint s such that a region $R_1(j, s)$ is defined as

$$R_1(j, s) = \{X | X_j \leq s\} \text{ and } R_2(j, s) = \{X | X_j > s\}$$

Then, (49) is minimised wrt j and s. For each splitting variable, the split point s can be found and hence by scanning through all of the variables X_j, determination of the best pair (j, s) is feasible. Having found the best split, the data are partitioned into **two** resulting regions and the same splitting exercise is repeated on each of the two regions. Then this process is repeated on all of the resulting regions and so on. How many rounds of the algorithm are done determines how deep the resulting tree is. On one hand, shallow trees might fail to capture the structure of the data. On the other hand, however, deeper trees might overfit the data and hence do poorly in prediction.

A common way to proceed requires to grow a very large tree T_0, which is then pruned using a penalty function. Define a subtree $T \subset T_0$ to be any tree that can be obtained by collapsing any number of its non-terminal nodes. Recall that T_0 partitions the space of y into M regions Rm, $m = 1, \ldots, M$, and hence contains M terminal nodes; and define $|T|$ to be the number of terminal nodes of a subtree T. Define N_m to be the cardinality of R_m, i.e. $N_m = |x_i \in R_m|$. Recall that

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_{i \in R_m}} y_i,$$

and denote the function

$$Qm(T) = \frac{1}{N_m} \sum_{x_{i \in R_m}} (y_i - \hat{c}_m)^2.$$

Then a complexity criterion function can be specified in the following way

$$CC_\alpha(T) = \sum_{m=1}^{M} N_m Q_m(T) + \alpha M.$$

The idea is to find (for a given α) a subtree $T_\alpha \subset T$ such that $CC_\alpha(T)$ is minimised. The tuning parameter $\alpha \geq 0$ governs how much large trees are penalised, so whenever $\alpha = 0$, the solution is the full tree T_0, while large values of α result into smaller trees. It turns out that, for a given α, a unique smallest tree $T_\alpha^* \subset T$ exists that minimises $CC_\alpha(T)$. To find T_α^* an algorithm called 'weak link pruning' is used. The idea is to successively collapse subnodes that produce the smallest per-node increase in $\sum_{m=1}^{M} N_m Q_m(T)$, until a single root tree is obtained. Breiman et al. (1984) show that this results into a finite sequence of subtrees that contains T_α^*.

3.2 Bootstrap, bagging, boosting

There are ways to further improve the performance of regression trees. Bootstrap requires choosing with replacement a subsample and re-estimating the tree in order to get a sampling distribution of various statistics. Bagging is a general method that generates multiple versions of a predictor and uses these to get an aggregated predictor. Breiman (1996) offers an overview. In the context of regression trees, bagging averages across trees estimated with different bootstrapped samples. Boosting focuses on the predictive power of individual predictors one at a time. In an economic context, boosting has been applied by Bai and Ng (2009) and Ng (2014). For example, Ng (2014) uses boosting in order to screen a large number of potentially relevant predictors and their lags and give warning signals of recessions. We review boosting in more detail in the next subsection.

3.2.1 BOOSTING

As an alternative to including many regressors simultaneously in a penalised regression setup, a number of papers have developed methods that focus on the predictive power of individual regressors instead of considering all 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 R. Tibshirani (2000) and Friedman (2001)).

Boosting is an iterative procedure where misclassified observations are given increasing cost in each estimation repetition. The idea is to consider regressors one by one in a simple regression setting, and successively selecting the best fitting ones, giving rise to 'greedy' algorithms. 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 (e^T), assuming that the true underlying regression function is sparse in terms of the L_1-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 $x_l = (x_{1t}, \ldots, x_{Nt})'$, $\mathbf{X} = (x_1, \ldots, x_N)$ and $\mathbf{e} = (e_1, \ldots, e_T)$.

   Define the least squares base procedure:

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

2. Given data $\mathbf{X}$ and $\mathbf{y} = (y_1, \ldots, y_t)'$, apply the base procedure to obtain $\hat{g}^{(1)}_{\mathbf{x},\mathbf{y}}(\mathbf{x}_t)$.

   Set $\hat{F}^{(1)}(\mathbf{x}_t) = \upsilon\hat{g}^{(1)}_{\mathbf{x},\mathbf{y}}$, for some $\upsilon > 0$. Set $\hat{s}^{(1)} = \hat{s}$ and $m = 1$.

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

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

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

$$M = \min_{1 \le m \le m_{max}} AIC_c(m),$$

for some predetermined large m_{max} where

$$AIC_c(m) = \log(\sigma^2) + \frac{1 + \frac{tr(B_m)}{T}}{1 + \frac{lr(B_m) + 2}{T}}$$

$$\sigma^2 = \frac{1}{T}(y - B_m y)'(y - B_m y)$$

$$B_m = I - \left(I - \upsilon \mathcal{H}^{(\hat{s}_m)}\right)\left(I - \upsilon \mathcal{H}^{(\hat{s}_{m-1})}\right) \dots \left(I - \upsilon \mathcal{H}^{(\hat{s}_1)}\right)$$

$$\mathcal{H}^{(j)} = \frac{\mathbf{x}_j \mathbf{x}_j'}{\mathbf{x}_j' \mathbf{x}_j}$$

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

3.2.2 RELATED 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.

A new approach that is related to those above has recently been proposed by Chudik, Kapetanios and Pesaran (2016). The main idea is to test the statistical significance of the net contribution of each potential covariate to y_t separately, whilst taking full and rigorous account of the multiple testing nature of the problem under consideration. In a second step, all statistically significant covariates are included as joint determinants of y_t in a multiple regression setting. In some exceptional cases it might also be required to iterate on this process by testing the statistical contribution of non-selected covariates (again one at a time) to the unexplained part of y_t. But it can be shown that asymptotically the number of such additional iterations will be less than the number of true variables explaining y_t. Whilst the initial regressions are common to boosting and the screening approach of Fan and Lv (2008), the multiple testing element provides a powerful stopping rule without needing to resort to model selection or penalised regression in the subsequent steps.

The proposed method, which is referred to as multiple testing (MT) procedure, is computationally simple and fast even for extremely large datasets, unlike penalised regression which presents some computational challenges in such cases. The method is extremely effective in selecting regressors that are correlated with the true unknown conditional mean of the target variable and, as a result, it also has good estimation properties for the unknown coefficient vector, which should turn into improved forecasts. Like penalised regressions, the method is applicable when the underlying regression model is sparse, but unlike penalised regressions it does not require that predictors have a sparse covariance matrix.

3.2.3 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 diff t 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 in this paper, 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, e.g., 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.

3.3 Random forests

Random forests were introduced by Breiman (2001). The idea is exactly as bagging applied on regression trees: to grow a large collection of de-correlated trees (hence the name forest) and then average them. This is achieved by bootstrapping a random sample at each node of every tree. In order to induce "decorrelation" of trees, when growing trees, before each split, select a subset of the input variables at random as candidates for splitting. This prevents the "strong" predictors imposing too much structure on the trunk of the tree. Although their asymptotic properties are not well understood yet, forests can deliver good out-of-sample performance, documented for instance in Howard and Bowles (2012). Random forests, however, are not suited for time series data. The could perform better compared to logit for iid binary data and therefore could be considered in Early Warning Signals or other indicator variables forecasting; for an example on random forests and logit comparison see Bhattacharjee (2016).

3.4 Deep learning and neural networks

Artificial neural networks (ANNs) are a family of models inspired by biological neural networks and are used to approximate functions and recognise patterns 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, e.g., Blake and Kapetanios (2010) for more detailed information.

While the application of ANNs 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 , which is not suited for the nowcasting problems under consideration. The need for a large T arises as the multilayered ANN model has a considerable number of parameters that need to be estimated. Let

$$y_t = \delta + \alpha' \mu(x_l) + \epsilon_l \tag{50}$$

Neural networks provide an approximation of the unknown function $\mu(.)$ and their approximation properties have been established formally in the literature (see, e.g., Hornik, Stinchcombe and White (1989)). An algorithm for estimation of the regression function $\mu(x_l)$ proceeds as follows.

Algorithm 1 *(RBF) MDH Boosting algorithm*

1. *Let σ_T be some sequence such that $\sigma_T = o(1)$. We construct an initial set of T RBF nodes given by: $\Psi^{(1,\ldots,T)} = \{\psi(x, x_1, \sigma_T), \psi(x, x_2, \sigma_T), \ldots, \psi(x, x_T, \sigma_T)\}$.*

2. *These are ranked according to their ability to reduce the residual variance, when each $\psi(x_t, x_i, \sigma_T), i = 1, \ldots, T$, is entered individually in (50).*

3. *The node that minimises the residual variance becomes the first node in the ranking of the nodes. Denote this node by $\psi(x, x_{S_1}, \sigma_T)$. Denote the residual from the regression of y_t on $\psi(x, x_{S_1}, \sigma_T)$, by $y_t^{(1)}$. Let $\tilde{S}_1 = \{S_1\}$. Let $\Psi^{(1,\ldots,T)/\tilde{S}_1}$ be the set of nodes in $\Psi^{(1,\ldots,T)}$ apart from the nodes indexed by the elements of $\tilde{S}_1$.*

4. *Set* $i = 1$.

5. *The nodes in* $\Psi^{(1,\dots,T)/\tilde{S}_1}$ *are ranked according to their ability to reduce the residual variance of* $y_t^{(i)}$, *when* $y_t^{(i)}$ *is regressed on each* $\psi(x_t, x_i, \sigma_T).)$, $i \in \tilde{S}_1$.

6. *The node that minimises the residual variance becomes the* $i + 1 - th$ *node in the ranking of the nodes. Denote this node by* $\psi(x, x_{S_{i+1}}, \sigma_T)$. *Denote the residual from the regression of* $y_t^{(1)}$ *on* $\psi(x_t, x_{S_{i+1}}, \sigma_T \psi(x, x_{S_{i\,|1}}, \sigma_T)$, *by* $y_t^{(i+1)}$. *Let* $\tilde{S}_{i+1} = \tilde{S}_i \cup \{\tilde{S}_{i+1}\}$. *Let* $\Psi^{(1,\dots,T)/\tilde{S}_{i+1}}$ *be the set of nodes in* $\Psi^{(1,\dots,T)}$ *apart from the nodes indexed by the elements of* $\tilde{S}_{i+1}$.

7. *If* $i = m$ *for some* $m = m_T \to \infty$ *stop, else set* $i = i + 1$ *and go to Step 5.*

Theorem 1 in Kapetanios and Blake (2010) states that the estimate of the regression function $\mu(x_t)$ obtained using the iterative boosting algorithm above, denoted $\hat{\mu}(x_t)$, satisfies

$$\hat{\mu}(x_t) - \mu(x_t) = o_p(m^{-1/C_1}) ,$$

for all $C_1 > 6$, if $m \equiv m_T \leq \log_a T$, for all a that satisfy $\log_a e < \frac{In(5/2)}{4}$. Further, $m \leq \log_a T$, for all a that satisfy $\log_a e < \frac{In(5/2)}{2}$. This suggests that the maximum possible rate for m is logarithmic in T and the choice for $\sigma_T = O((\ln \ln T)^{-1})$ is acceptable.

In practice, it is common to split the dataset into three subsets: training, validation, and testing sets. The training set is used to adjust the weights of the network; the validation set is used to minimize the overfitting through choosing values of hyperparameters and selecting the appropriate model. Finally, the testing set is used to confirm the actual out-of-sample predictive power of the model. Deep learning has been applied in financial applications: for example, Sirignano, Sadhwani and Giesecke (2016) use neural networks to analyze mortgage risk using a dataset of over 120 million prime and subprime US mortgages between 1995 and 2014. Heaton, Polson, and Witte (2016a, 2016b) also employ neural networks in the context of portfolio theory.

4 Summarizing the information in Big Data

The methods we have considered so far include all the, possibly many, explanatory variables in the regression model for the variable of interest, and then induce shrinkage by either penalizing over-parameterization, or using an appropriate prior distribution, or selecting the most relevant regressors, based on either parametric or non-parametric approaches. An alternative procedure requires instead to summarize the relevant information in a first step, producing a (much) smaller set of generated regressors, and then use the generated regressors in a second step in a regression model for the target variable. We now review methods for the first step, i.e., to efficiently summarize the information in (possibly very) large datasets.

4.1 Principal component analysis and factor models

Factor models are commonly used data-rich forecasting 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) and Forni, Hallin, Lippi and Reichlin (2000). The defining characteristic of most factor methods is that relatively few summaries of the many available variables 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 (weakly stationary and standardized) 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 $F_t = (F_{1l} \cdots F_{rl})'$,

i.e.,

$$x_t = \Lambda' e_t + e_t, \tag{51}$$

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 the terms 'loadings'. In (51) 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 (1) $\alpha' = \tilde{\alpha}'\Lambda\tilde{x}_t$, 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 principal component analysis (PCA) for the estimation of factor models is, by far, the most popular 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:

$$V(r) = \min_{\Lambda,F} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\tilde{x}_{it} - \lambda_i' F_t)^2, \tag{52}$$

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 (52) 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)).

An alternative is to use filters such as the Kalman filter which requires additional assumptions on the stochastic law of motion of the factors F_t. Another alternative way of extracting factors is Partial Least Squares (PLS), introduced in order to facilitate the estimation of multiple regressions when there is a large, but finite, amount of regressors, and to target the estimated factors towards the specific variable of interest, y. PLS will be described in a later subsection.

4.2 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 Prin- cipal 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 υ while constraining its cardinality:

$$\max \upsilon' \Sigma \upsilon$$

$$s.t. \sum_{i=1}^{N} \upsilon_i^2 = 1$$

$$\#(i|\upsilon_i \neq 0) \leq k,$$

where Σ denotes the sample covariance matrix. The first constraint ensures that υ is a unit vector and the second constraint is the L_0-norm, i.e. the number of the non-zero components in υ 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

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

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

4.3 Partial least squares

Partial least squares (PLS) is a relatively new method with a very similar idea to Principal Component Analysis (PCA) in that a number of factors or components, which are linear combinations of the original regression variables, are extracted and used as regressors instead of the original typically much larger set of variables.

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 $v_{i,t} = x_{i,t}, i = 1, \ldots N$. 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: $w_{ij} = Cov(u_t, v_{it}), i = 1, \ldots, N$. 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 $v_{i,t}, i = 1,\dots,N$ 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 forecasted by regressing y_t on $f_{j,t}, j = 1,\dots,k$. 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 (1), obtained implicitly via PLS Algorithm and a regression of y_t on $f_{j,t}, j = 1,\dots,k$, are mathematically equivalent to

$$\hat{\beta}_{PLS} = V_k(V_k'X'XV_k)^{-1}V_k'X'y \tag{53}$$

with $V_{k_1} = (X'y \;\; X'XX'y \cdots (X'X)^{k-1}X'y), X = (x_1 \cdots x_T)'$ and $y = (y_1 \cdots y_T)'$. Thus, (53) 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.

The main difference between PC and PLS is that, PLS takes into account the relationship between y_t and x_t when constructing the factors, while factors extracted by PC are constructed taking into account only the values of the x_t variables.

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.

5 Forecast combination

Pooling the forecasts from different models can improve forecasting performance (see for instance Hendry and Clements (2004)). Possible reasons for this result are model misspecification or model uncertainty. In addition, the effect of changing importance of various explanatory variables (documented in Stock and Watson (2006) for example) can be reduced by weighting over models containing different regressors. In the context of big data where little is known ex-ante about the relevance and predictive power of variables in the large dimensional dataset, forecast combination could be very useful.

The most simple and computationally efficient weighting schemes is equal weighting, whereby the different models $\{M_1, \ldots, M_s\}$, receive an equal weight of $\frac{1}{s}$. The resulting forecast density of the combination is then given by

$$p(y_{N+h} \mid y_{1:N}, x_{1:N}, x_{N+h}) = \frac{1}{s}\sum_{i=1}^{s} p(y_{N+h} \mid y_{1:N}, x_{1:N}, x_{N+h}, M_i).$$

As an alternative, weighted averaging is usually based on past forecast performance in terms of mean-squared (MSE) forecast errors. For example, Kuzin, Marcellino and Schumacher (2013) suggest using the MSE computed over a previous rolling window.

5.1 Bayesian model averaging

The Bayesian approach to combination forecasting requires computing the posterior probability of each model M_j through the Bayes theorem:

$$p(M_j \mid y, X) = \frac{p(y, X \mid M_j)p(M_j)}{\sum_{i=1}^{s} p(y, X \mid M_i)p(M_i)}, \tag{54}$$

where $p(M_j)$ is the prior probability given to each model. For discussion, see for example Hoeting, Madigan, Raftery and Volinsky (1999). Forecast densities can thereafter be computed as a weighted sum, with weights determined by the posterior probabilities

$$p(y_{N+h} \mid y_{1:N}, x_{1:N}, x_{N+h}) = \sum_{i=1}^{s} p(y_{N+h} \mid y_{1:N}, x_{1:N}, x_{N+h}, M_i)p(M_i \mid y, X).$$

Recent papers on Bayesian model averaging (see, for example, Koop and Korobilis (2012)) have used a dynamic model averaging, emphasising the importance of allowing the models' weights to be time varying, in order to capture the changing relevance of different models over time.

5.2 Frequentist model averaging

As an alternative to Bayesian model averaging, there is a sizable literature, competently summarised by Burnham and Anderson (1998), on a frequentist information theoretic approach in an analogous vein. In this context, information theory suggests ways of constructing model confidence sets. Given the existence of a set of models, relative model likelihood can be defined. Model weights within this framework have been suggested by Akaike in a series of papers (see Akaike (1978)) and expounded further by Burnham and Anderson (1998).

In practical terms such weights are easy to construct using standard information criteria such as Akaike's information criterion. Kapetanios, Labhard and Price (2008) have considered this way of model averaging as an alternative to Bayesian model averaging for forecasting. Similarly to the work of Eklund and Karlsson (2007), Kapetanios, Labhard and Price (2008) use an out-of-sample measure of fit in standard information criteria when constructing weights for forecast combination in an information theoretic approach. They find that the proposed method performs well and, in some respects, outperforms other averaging methods considered.

5.3 Forecast combination with Big Data

While there are no studies focusing specifically on the use of forecast pooling in the presence of big data, one could apply some of the methods described above also in this context. Specifically, equal based or information criterion based forecast averaging remains doable even in the presence of a big set of competing forecasts or nowcasts, each based on a single variable (or a small set of them). Penalized or shrinked regression methods could be also used to determine optimally the pooling weights, by regressing the actual values on the many competing forecasts over an evaluation sample. And the standard or sparse principal components could be also used to summarize the many competing forecasts into a single pooled one.

6 Modelling mixed frequency data

A key feature of economic time series data is that they are released by statistical agencies in differing frequency. For example, variables such as GDP are typically released at quarterly frequency, while various prices are measured monthly and many financial variables are available daily or even hourly. This problem can be solved by using mixed frequency methods, which allow for all possible information to be taken into account when forecasting with large data. Different methods are available in the econometrics literature for nowcasting with mixed frequency data, Bańbura Giannone and Reichlin (2011), Bańbura, Giannone, Modugno, Reichlin (2012) and Foroni and Marcellino (2013, 2014) provide overviews. The simplest approaches are bridge modelling (e.g., Baffigi, Golinelli and Parigi (2004), Diron (2008)) and unrestricted mixed data sampling (U-MIDAS, see Foroni, Marcellino and Schumacher (2015)). Both methods are based on univariate linear regressions, and are therefore suited for direct application of most of the big data specific model specification and estimation methods that we have discussed in the previous sections.

Another approach is mixed-data sampling (MIDAS) models, which rely on lag polynomials to aggregate higher frequency data, which can deal with data sampled at diff t frequencies and provide a direct forecast of the low-frequency variable (see e.g. Ghysels, Santa-Clara and Valkanov (2004), Clements and Galvao (2008)). MIDAS models have an advantage of being able to handle different frequency mismatches (e.g., daily/quarterly/monthly) but the nonlinearity introduced through the lag polynomials requires numerical optimisation and therefore complicates estimation substantially.

An even more sophisticated, and optimal in a linear context, approach is based on the use of the Kalman filter applied to models cast in state-space form and properly extended to include aggregation constraints. Mixed-frequency VAR (MF-VAR) and factor models belong to this class. Both are system approaches that jointly describe the dynamics of the variable to be explained and of the indicators, where the use of the Kalman filter provides not only predictions of the future observations but also estimates of the current latent state (see, e.g., Mariano and Murasawa (2003, 2010)). MF-VARs are suited to handle few variables while MF-factor models can also handle large datasets, but the computational costs increase very fast with the dimension. Hence, state-space approaches do not appear very suited to use with big data, unless the big data are substantially summarized in the pre-treatment phase, for example, reduced to a single Google Trend.

Factors can be also directly included in bridge, MIDAS and U-MIDAS models, as suggested by Marcellino and Schumacher (2010). Factor-Bridge and Factor-UMIDAS can be used in the context of big data summarized by means of some type of factor models. Factor-MIDAS is also feasible in the presence of very few factors.

6.1 Bridge models

One of the early econometric approaches in the presence of mixed-frequency data relies on the use of bridge equations, see e.g. Baffigi et al. (2004) and Diron (2008). Bridge equations are linear regressions that link (”bridge”) high frequency variables, such as industrial production or retail sales or Google Trends, to low frequency ones, e.g. the quarterly real GDP growth, providing some estimates of current and short- term developments in advance of the release. The ”Bridge model” technique allows computing early estimates of the low-frequency variables by using high frequency indicators. They are not standard macroeconometric models, since the inclusion of specific indicators is not based on causal relations, but on the statistical fact that they contain timely updated information. In principle, bridge models require that the whole set of regressors should be known over the projection period, allowing for an estimate only of the current period. In practice, anyway, this is not the case, even though the forecasting horizon of the bridge models is quite short, one or two quarters ahead at most.

Taking forecasting GDP as an example, since the monthly indicators are usually only partially available over the projection period, the predictions of quarterly GDP growth are obtained in two steps. First, monthly indicators are forecasted over the remainder of the quarter, usually on the basis of univariate time series models (in some cases VAR have been implemented in order to obtain better forecasts of the monthly indicators), and then aggregated to obtain their quarterly correspondent values. Second, the aggregated values are used as regressors in the bridge equation which allows obtaining forecasts of GDP growth.

Therefore, the bridge model to be estimated is:

$$y_{t_q} = \alpha + \sum_{i=1}^{j} \beta_i (L) x_{it_q} + u_{t_q}$$

where $\beta_i(L)$ is a lag polynomial of length k, and x_{it_q} are the monthly indicators aggregated at quarterly frequency.

The selection of the monthly indicators included in the bridge model is usually based on a general-to-specific methodology and relies on different in-sample or out- of-sample criteria, like information criteria or RMSE performance. In our big data context, we can use any of the model selection and estimation techniques discussed in the previous sections.

6.2 MIDAS

Distributed lag (DL) models have been typically employed in the literature to describe the distribution over time of the lagged effects of a change in the explanatory variable. In general, a stylized distributed lag model is given by

$$y_{l_q} = \alpha + B(L) x_{l_q} + \varepsilon_{l_q} \tag{56}$$

where $B\,(L)$ is some finite or infinite lag polynomial operator. This kind of models underlies the construction of the bridge equations, once all the high frequency values are aggregated to the

corresponding low-frequency values.

In order to take into account mixed-frequency data, Ghysels et al. (2004) introduce the Mixed-Data Sampling (MIDAS) approach, which is closely related to the distributed lag model, but in this case the dependent variable y_{l_q}, sampled at a lower-frequency, is regressed on a distributed lag of x_{t_m}, which is sampled at a higher-frequency. In what follows, we present the basic features of the model, as presented by Ghysels et al. (2004).

In terms of notation, $t_q = 1, \dots T_q$ indexes the basic time unit (e.g. quarters), and m is the number of times the higher sampling frequency appears in the same basic time unit. For example, for quarterly GDP growth and monthly indicators as explanatory variables, $m = 3.w$ is the number of monthly values of the indicators that are earlier available than the lower-frequency variable to be estimated. The lower-frequency variable can be expressed at the high frequency by setting $y_{t_m} = y_{t_q}, \forall t_m = mt_q$, where t_m is the time index at the high frequency.

As mentioned, the response to the higher-frequency explanatory variable is modelled using highly parsimonious distributed lag polynomials, to prevent the proliferation of parameters that might otherwise result, as well as the issues related to lag-order selection. Specifically, the basic MIDAS model for a single explanatory variable, and h_q -step-ahead forecasting, with $h_q = h_m/m$, is given by:

$$y_{t_q} + mh_q = y_{t_m} + h_m = \beta_0 + \beta_1 b(L_m; 0) x_{t_m+w}^{(m)} + \varepsilon_{t_m+h_m} \tag{57}$$

where $b(L^{1/m}; \theta) = \sum_{k=0}^{K} c(k; \theta)\, L_m^k$, and $L_m^x x_{t_m}^{(m)} = x_{t_m-x}^{(m)}. x_{t_m\,|w}^{(m)}$ is skip-sampled from the high frequency indicator x_{t_m}.

The parameterization of the lagged coefficients of $c(k; \theta)$ in a parsimonious way is one of the key MIDAS features. One of the most used parameterizations is the one known as “Exponential Almon Lag”, since it is closely related to the smooth polynomial Almon lag functions that are used to reduce multicollinearity in the Distributed Lag literature. It is often expressed as

$$c(k; \theta) = \frac{\exp\left(\theta_1 k + \cdots + \theta_Q k^Q\right)}{\sum_{k=1}^{K} \exp\left(\theta_1 k + \cdots + \theta_Q k^Q\right)} \tag{58}$$

This function is known to be quite flexible and can take various shapes with only a few parameters. These include decreasing, increasing or hump-shaped patterns. Ghysels, Santa-Clara and Valkanov (2005) use the functional form with two parameters, which allows a great flexibility and determines how many lags are included in the regression. Notice that the standard practice in bridge equations of calculating a quarterly series from the monthly indicators corresponds to imposing restrictions on this parameterization function. To be concrete, in the case of the quarterly-monthly example, taking the last month in the quarter to produce a quarterly series amounts to setting $c(2; \theta) = c(3; \theta) = c(5; \theta) = c(6; \theta) = \dots = c(11; \theta) = c(12; \theta) = 0$.

Another possible parameterization, also with only two parameters, is the so-called “Beta Lag”, because it is based on the Beta function:

$$c(k;0_1,0_2)=\frac{f\left(\frac{k}{K},\theta 1;\theta 2\right)}{\sum_{k=1}^{K} f\left(\frac{k}{K},\theta 1;\theta 2\right)} \tag{59}$$

where $c(x,a,b)=\frac{x^{a-1}(1-x)^{b-1}\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ and $\Gamma(a)=\int_0^{\infty} e^{-x}x^{a-1}dx$. Ghysels, Rubia and Valkanov (2009) propose also three other different parameterizations of the lag coefficients: a linear scheme, with c(k;θ)=1, where there are no parameters to estimate in the lagged weight function; an hyperbolic scheme, with $c(k;\theta)=\frac{g(\frac{k}{K},\theta)}{\sum_{k=1}^{K} g(\frac{k}{K},\theta)}, g(k,\theta)=\frac{\Gamma(k+\theta)}{\Gamma(k+1)\Gamma(\theta)}$ where the gamma function has only one parameter to estimate, but it's not as flexible as the Beta specification; a geometric scheme, with $c(k;\theta)=\frac{\theta^k}{\sum_{k=1}^{\infty}\theta^k}, |\theta|\leq 1$ and $c(k;\theta)$ are normalized so that they sum up to one.

The parameterizations described above are all quite flexible. For different values of the parameters, they can take various shapes: weights attached to the different lags can decline slowly or fast, or even have a hump shape. Therefore, estimating the parameters from the data automatically determines the shape of the weights and, accordingly, the number of lags to be included in the regression.

The MIDAS model can be estimated using nonlinear least squares (NLS) in a regression of y_t onto $x_{t-h}^{(m)}$. Ghysels et al. (2004) show that MIDAS regressions always lead to more efficient estimation than the typical approach of aggregating all series to the least frequent sampling. Moreover, they also show that discretization biases are the same for MIDAS and distributed lag models and vanish when regressors are sampled more frequently. However, as mentioned, NLS can be computationally challenging, in particular with many regressors, and require frequent user interventions in case of convergence to local minima.

The forecast (or nowcast) is given by

$$y_{T_m^y+h_m\,|T_m^x}=\hat{\beta}_0+\hat{\beta}_1 b(L_m;\theta)x_{T_m^x}^{(m)}. \tag{60}$$

Note that MIDAS is h−dependent, and thus needs to be re-estimated for each forecast horizon.

6.3 U — MIDAS

Foroni, Marcellino and Schumacher (2015) study the performance of a variant of MIDAS which does not resort to functional distributed lag polynomials. In the paper, the authors discuss how unrestricted MIDAS (U-MIDAS) regressions can be derived in a general linear dynamic framework, and under which conditions the parameters of the underlying high-frequency model can be identified.

The U-MIDAS model is based on a linear lag polynomial such as

$$c(L^m)\omega(L)y_{t_m}=\delta_1(L)x_{1t_m}-1+\ldots+\delta_N(L)x_{Nt_{m^{-1}}}-1+\epsilon_{t_m}, \tag{61}$$
$$t=1,2,3\ldots$$

where $c(L^m) = (1 - c_1 L^m - \ldots - c_c L^{mc}), \delta_j(L) = (\delta_{j,0} + \delta_{j,1} L + \ldots + \delta_{j,v} L^v), j = 1, \ldots, N$.

Note that if we assume that the lag orders c and v are large enough to make the error term $\in_{t_m}$uncorrelated, then, all the parameters in the U-MIDAS model (61) can be estimated by simple OLS (while the aggregation scheme $\omega(L)$ is supposed known). From a practical point of view, the lag order v could diff across variables, and v_i and c could be selected by an information criterion such as BIC. As the model is linear, all the specification methods for big data we have discussed in the previous sections can still be applied.

A simple approach to forecasting is to use a form of direct estimation and construct the forecast as

$$\tilde{y}_{T_m^x + m \mid T_m^x} = \tilde{c}(L^k) y_{T_m^x} + \tilde{\delta}_1(L) x_{1T_m^x} + \ldots + \tilde{\delta}_N(L) x_{NT_m^x}, \quad (62)$$

where the polynomials $\tilde{c}(Z) = \tilde{c}_1 L^m - \ldots - \tilde{c}_c L^{mc}$ and $\tilde{\delta}_i(L)$ are obtained by projecting y_{t_m} on information dated $mt_m - m$ or earlier, for $t = 1,2,\ldots,T_m^x$. In general, the direct approach of (62) can also be extended to construct h_m-step ahead forecasts given information in T_m^x :

$$\bar{y}_{T_m^x + h_m \mid T_m^x} = \bar{c}(L^k) y_{T_m^x} + \bar{\delta}_1(L) x_{1T_m^x} + \cdots + \bar{\delta}_N(L) x_{NT_m^x}, \quad (63)$$

where the polynomials $\bar{c}(Z)$ and $\bar{\delta}_1(L)$ are obtained by projecting y_{t_m} on information dated $mt - h_m$ or earlier, for $t = 1,2,\ldots,T_m^x$.

In the case of U-MIDAS, an autoregressive term can be also included easily, without any common factor restriction as instead often needed in a MIDAS context, see Clements and Galvao (2009).

Carriero, Clark and Marcellino (2013) use Bayesian techniques to estimate specifications similar to U-MIDAS models with several regressors and stochastic volatility, which can easily produce not only point but also interval and density forecasts. We refer to their paper for the technical details. Due to the computationally complexity, it does not seem suited for application with big data.

Finally, as for the case of bridge models, linearity of the UMIDAS specification allows for model selection and estimation in the presence of big data using any of the methods discussed in the previous sections, while this does not hold for MIDAS, due to its non-linearity.

6.4 Mixed frequency VAR

While so far, we have seen models which take into account mixed-frequency data in a univariate approach, we now focus on multivariate methods which jointly specify the dynamics of the indicators and of the variable to be explained. To exploit the information available in series released at diff t frequencies and jointly analyze them, there is a growing literature which looks at mixed-frequency VARs, which aim to characterize the co-movements in the series and summarize the information contained in the mixed-frequency data.

Nowadays, in the literature, there are both classical and Bayesian approaches to estimate MF-VAR models. In what follows, we describe the main features of these two classes of estimation, following two of the most representative studies in the literature, Mariano and Murasawa (2010) for the classical approach and Schorfheide and Song (2011) for the Bayesian approach.

One of the most compelling approaches in the literature to deal with mixed- frequency time series at the moment is the one proposed by Zadrozny (1988) for directly estimating a VARMA model sampled at diff t frequencies; see also Harvey (1989). The approach treats all the series as

generated at the highest frequency, but some of them are not observed. Those variables that are observed only at the low frequency are therefore considered as periodically missing.

Following the notation of Mariano and Murasawa (2010), we consider the state- space representation of a VAR model in a classical framework, treating quarterly series as monthly series with missing observations and taking GDP growth as an example. The disaggregation of the quarterly GDP growth, y_{t_m}, observed every $t_m = 3,6,9,\dots,T_m$, into the month-on-month GDP growth, $y^*_{t_m}$, never observed, is based on the following aggregation equation:

$$\begin{aligned} y_{t_m} &= \frac{1}{3}\left(y^*_{t_m} + y^*_{t_{m-1}} + y^*_{t_{m-2}}\right) + \frac{1}{3}\left(y^*_{t_{m-1}} + y^*_{t_{m-2}} + y^*_{t_{m-3}}\right) + \\ &+ \frac{1}{3}\left(y^*_{t_{m-2}} + y^*_{t_{m-3}} + y^*_{t_{m-4}}\right) \\ &= \tfrac{1}{3}y^*_{t_m} + \tfrac{2}{3}y^*_{t_{m-1}} + y^*_{t_{m-2}} + \tfrac{2}{3}y^*_{t_{m-3}} + \tfrac{1}{3}y^*_{t_{m-4}} \end{aligned} \tag{64}$$

This aggregation equation comes from the assumption that the quarterly GDP series (in log levels), Y_{t_m}, is the geometric mean of the latent monthly random sequence $Y^*_{t_m}, Y^*_{t_{m-1}}, Y^*_{t_{m-2}}$. Taking the three-period differences and defining $y_{t_m} = \Delta_3 Y_{t_m}$ and $y^*_{t_m} = \Delta Y^*_{t_m}$, we obtain eq. (64).

Let for all t_m the latent month-on-month GDP growth $y^*_{t_m}$ and the corresponding monthly indicator x_{t_m} follow a bivariate VAR(p) process

$$\phi(L_m)\begin{pmatrix} y^*_{t_m} - \mu^*_y \\ x_{t_m} - \mu_x \end{pmatrix} = u_{t_m}, \tag{65}$$

where $u_{t_m} \sim N(0, \Sigma)$.

The VAR(p) process in eq. (65) together with the aggregation equation (64) is then cast in a state-space representation.

Assuming $p \leq 4$ [4]and defining

$$s_{t_m} = \begin{pmatrix} z_{t_m} \\ \vdots \\ z_{t_m-4} \end{pmatrix}, \qquad z_{t_m} = \begin{pmatrix} y^*_{t_m} - \mu^*_y \\ x_{t_m} - \mu_x \end{pmatrix}$$

a state-space representation of the MF-VAR is

$$s_{t_m} = F s_{t_{m-1}} + G v_{t_m} \tag{66}$$

[4] For the sake of conciseness, we do not report the state-space representation for p > 4. Details for this case can be found in Mariano and Murasawa (2010).

$$\begin{pmatrix} y_{t_m} - \mu_y \\ x_{t_m} - \mu_x \end{pmatrix} = Hs_{t_m} \tag{67}$$

with $\mu_y = 3\mu_y^*$ that holds, and $v_{t_m} \sim N(0, I_2)$ The matrices are defined as:

$$F = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix};\ F_1 = [\phi_1\ \dots\ \phi_p\ 0_{2x2(5-p)}];\ F_2 = [I_8\ 0_{8x2}], \tag{68}$$

$$G = \begin{bmatrix} \Sigma^{1/2} \\ 0_{8x2} \end{bmatrix};\ H = [H_0\ \dots\ H_4\ ] \tag{69}$$

where H contains the lag polynomial

$$H(L_m) = \begin{bmatrix} 1/3 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 2/3 & 0 \\ 0 & 0 \end{bmatrix} L_m + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} L_m^2 + \begin{bmatrix} 2/3 & 0 \\ 0 & 0 \end{bmatrix} L_m^3 + \begin{bmatrix} 1/3 & 0 \\ 0 & 0 \end{bmatrix} L_m^4 \tag{70}$$

The state-space model consisting of equations (66) and (67) can be estimated with maximum-likelihood techniques or the expectation-maximization algorithm.

6.4.1 BAYESIAN MIXED FREQUENCY VAR

The estimation of MF-VAR model with Bayesian techniques has been recently considered as an alternative framework in the literature. One of the earliest studies on this is the paper by Eraker, Chiu, Foerster, Kim and Seoane (2015). In this paper, the authors develop a Gibbs sampling approach to estimate a VAR with mixed and irregularly sampled data. The algorithm they develop is a Gibbs sampler which iterates over the draws from the missing data and from the unknown parameters in the model. Under the assumption of a normally distributed error term, the algorithm allows for draws from Gaussian conditional distributions for estimating the missing data, and for draws from Gaussian and inverse Wishart conditional posterior distributions for the parameters in the model.

As an example for the Bayesian estimation of a MF-VAR, we present the algorithm developed by Schorfheide and Song (2011). The authors represent the MF-VAR as a state-space model, and use MCMC methods to conduct Bayesian inference for model parameters and unobserved monthly variables.

The state equation of the model is represented by the VAR(p) model written in the companion form:

$$z_{t_m} = F_1(\Phi) z_{t_m-1} + F_c(\Phi) + v_{t_m}, \qquad v_{t_m} \sim iidN\big(0, \Omega(\Sigma)\big). \tag{71}$$

To write the measurement equation, the authors need to write the aggregation equation, which is in this case diff t from the one considered by Mariano and Murasawa (2010). In this case, the quarterly variable is seen as the three-month average of the monthly process, which in the previous notation is:

$$y_{t_m} = \frac{1}{3}\left(y_{t_m}^* + y_{t_m-1}^* + y_{t_m-2}^*\right) = \Lambda_{mz} z_{t_m}. \tag{72}$$

However, since y_{t_m} is observed only every third month, there is a need of a selection matrix that equals the identity matrix if t_m corresponds to the last month of the quarter and is empty otherwise. Therefore, the measurement equation can be written as

$$\begin{pmatrix} y_{t_m} \\ x_{t_m} \end{pmatrix} = M_{t_m} \Lambda_z z_{t_m}, \tag{73}$$

where M_{t_m} is the selection matrix. A Minnesota prior that shrinks the VAR coefficients toward univariate random walk representations is introduced to cope with the issue of dimensionality.

Both classical and Bayesian estimation methods are, unfortunately, computation- ally intensive. In addition, only few variables can be included in the MF-VARs, even more so in the classical ones, due to the curse of dimensionality, thus making them not so suited to handle big data, unless they are substantially summarized in the pre-treatment phase.

6.5 Mixed frequency factor models

Closely related to the MF-VAR for their state-space representation, factor models have also been employed in the literature to handle data with different frequencies. These models have been utilized to extract an unobserved state of the economy and create a new coincident indicator, but also to exploit more information and obtain more precise forecasts. In what follows, we discuss the Mariano and Murasawa (2003) small scale mixed-frequency factor model, developed to extend the Stock–Watson coincident index for the US economy by combining quarterly real GDP and monthly coincident business cycle indicators. An extension of this model with big data based indicators, such as Google Trends, could be interesting by itself, though beyond the scope of the current research. Interesting applications of a similar approach can be found in Frale, Marcellino, Mazzi and Proietti (2010, 2011).

Then, we present an example of large scale mixed-frequency factor model, as proposed by Giannone, Reichlin and Small (2008), whose aim is to bridge the information in a large monthly dataset with the forecast of a quarterly variable. As an extension to it, we present the mixed-frequency state-space framework as developed by Bańbura and Rünstler (2011).

Finally, based on Marcellino and Schumacher (2010), we analyze the approach that merges factor models and the MIDAS / UMIDAS framework presented above. Factor models have a long tradition in econometrics and they are also appealing from an economic point of view. In fact, they decompose each time series under analysis into a common component, driven by few factors that represent the key economic driving forces, and an idiosyncratic component.

Mariano and Murasawa (2003) set up a static one-factor model for a small set of observable monthly and quarterly series, and derive its state-space representation.

Following their notation, consider a one-factor model for y_t^*, such that for all t_m,

$$y_{t_m}^* = \mu^* + \Lambda f_{t_m} + u_{t_m} \tag{74}$$

$$\Phi_f(L) f_{t_m} = v_{t_m} \tag{75}$$

$$\Phi_u(L)u_{t_m} = w_{t_m} \tag{76}$$

$$\begin{pmatrix} v_{t_m} \\ w_{t_m} \end{pmatrix} \sim N\left(0, \begin{bmatrix} \Sigma_{vv} & 0 \\ 0 & \Sigma_{ww} \end{bmatrix}\right) \tag{77}$$

where $\Phi_f(.)$ is a pth-order polynomial on $\mathbb{R}$ and $\Phi_u(.)$ is a qth-order polynomial on $\mathbb{R}^{NxN}$. In order to have identification, we assume $\Lambda := [I, \Lambda_2']'$ and $\Phi_u(.)$ and Σ_{ww} diagonal.

Assuming $p, q \leq 4$, for all t_m, and defining

$$s_t = \begin{pmatrix} f_{t_m} \\ \vdots \\ f_{t_m-4} \\ u_{t_m} \\ \vdots \\ u_{t_m-4} \\ \end{pmatrix},$$

the state-space representation of the factor model is

$$s_{t_m+1} = Fs_{t_m} + Gv_{t_m} \tag{78}$$

$$y_{t_m} = \mu + IIs_{t_m} \tag{79}$$

with $v_{t_m} \sim N(0, I_3)$, where

$$F = \begin{bmatrix} F_1 & F_2 \\ F_3 & F_4 \end{bmatrix};\ F_1 = \begin{bmatrix} \Phi_{f,1} \dots \Phi_{f,p} & 0_{1x(5-p)} \\ I_4 & 0_{4x1} \end{bmatrix};\ F_2 = 0_{5x10}; \tag{80}$$

$$F_3 = 0_{10x5};\ F_4 = \begin{bmatrix} \Phi_{u,1} \dots \Phi_{u,q} & 0_{1x(5-q)} \\ I_8 & 0_{8x2} \end{bmatrix}$$

$$G = \begin{bmatrix} \Sigma_{vv}^{1/2} & 0_{1x2} \\ 0_{4x1} & 0_{4x2} \\ 0_{2x1} & \Sigma_{ww}^{1/2} \\ 0_{8x1} & 0_{8x2} \end{bmatrix};\ H = [H_0\Lambda \ \dots \ H_4\Lambda \ H_0 \ \dots \ H_4] \tag{81}$$

where $II(L_m)$ is defined as in equation (70).

In the estimation, Mariano and Murasawa (2003) cannot use the standard EM algorithm, since the measurement equation has unknown parameters. The procedure they followed is similar to the one described for the MF-VAR.

The dynamic factor model as extended by Mariano and Murasawa (2003) is also used in Frale et al. (2011) to handle mixed frequency data, in order to obtain estimates of the monthly Euro area GDP components from the output and expenditure sides, to be later aggregated into a single indicator,

called EUROMIND. Broadly speaking, GDP is disaggregated by supply sectors and demand components. For each of these sectors and components, timely and economically sensible observable monthly indicators are then selected and represented with a dynamic factor model, as described above. The single models are then linked together based on the composition of GDP.

6.5.1 BRIDGE FACTOR MODELS

We now discuss a large mixed frequency factor model as proposed by Giannone, Reichlin and Small (2008), which exploits a large number of series that are released at different times and with different lags. The methodology the authors propose relies on the two-step estimator by Doz, Giannone and Reichlin (2011). This framework combines principal components with the Kalman filter. First, the parameters of the model are estimated by OLS regression on the estimated factors, where the latter are obtained through principal components calculated on a balanced version of the dataset. Then, the Kalman smoother is used to update the estimate of the signal variable on the basis of the entire unbalanced panel.

The dynamic factor model of Doz et al. (2011) is given by

$$x_{t_m} = \Lambda f_{t_m} + \xi_{t_m} \quad \xi_{t_m} \sim N(0, \Sigma_\xi) \tag{82}$$

$$f_{t_m} = \sum_{i=1}^{p} A_i f_{t_m - i} + B\eta_{t_m} \quad \eta_{t_m} \sim N(0, I_q) \tag{83}$$

Equation (82) relates the N monthly series x_{t_m} to a $r \times 1$ vector of latent factors f_{t_m}, through a matrix of factor loadings Λ, plus an idiosyncratic component ξ_{t_m}, assumed to be a multivariate white noise with diagonal covariance matrix Σ_ξ. Equation (83) describes the law of motion of the latent factors, which are driven by a $q-$dimensional standardized white noise η_{t_m}, where B is a $r \times q$ matrix $(r \leq q)$. Hence, $\zeta_{t_m} \sim N(0, BB')$.

To deal with missing observations at the end of the sample, the authors use a two-step estimator. In the fi step, the parameters of the model are estimated consistently through principal components on a balanced panel, created by truncating the data set at the date of the least timely release. In the second step, the Kalman smoother is applied to update the estimates of the factor and the forecast on the basis of the entire unbalanced data set.

The model is then complemented by a forecast equation for mean-adjusted quarterly GDP. The forecast is defined as the projection of the quarterly GDP growth on the quarterly aggregated estimated common factors:

$$\hat{y}_{t_q} = \alpha + \beta \hat{f}_{t_q}, \tag{84}$$

where $\hat{f}_{t_q}$ is the quarterly aggregated correspondent of $\hat{f}_{t_m}$.

If we look at eq. (84), we see that this is exactly in line with bridge modelling. In fact, the framework can be interpreted as a large bridge model that uses a large number of variables and bridges monthly data releases with the forecast of the quarterly variable.

6.5.2 FACTOR MODELS IN A MIXED FREQUENCY STATE SPACE

Bańbura and Rünstler (2011) extend the model of Giannone et al. (2008), by integrating a forecast equation for quarterly GDP. More specifically, they introduce the forecast of monthly GDP growth y_{t_m} as a latent variable, related to the common factors by the static equation

$$y_{t_m} = \beta' f_{t_m} + \varepsilon_{t_m}, \;\; \varepsilon_{t_m} \sim N(0, \sigma_\varepsilon^2). \tag{85}$$

The quarterly GDP growth, y_{t_m}, is assumed to be the quarterly average of the monthly series:

$$y_{t_m} = \frac{1}{3}\left(y^*_{t_m} + y^*_{t_{m^{-1}}} + y^*_{t_{m^{-2}}}\right). \tag{86}$$

The innovations $\varepsilon_{t_m}, \eta_{t_m}, \xi_{t_m}$ are assumed to be mutually independent at all leads and lags.

Equations (82) to (86) can be cast in state-space form. y_{t_m} is constructed in such a way that it contains the quarterly GDP growth in the third month of each quarter, while the other observations are treated as missing. The state-space representation, when $p = 1$, is:

$$\begin{bmatrix} x_{t_m} \\ y_{t_m} \end{bmatrix} = \begin{bmatrix} \Lambda & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_{t_m} \\ y^*_{t_m} \\ y^c_{t_m} \end{bmatrix} + \begin{bmatrix} \xi_{t_m} \\ \varepsilon_{t_m} \end{bmatrix} \tag{87}$$

$$\begin{bmatrix} \mathrm{I}_r & 0 & 0 \\ -\beta' & 1 & 0 \\ 0 & -1/3 & 1 \end{bmatrix} \begin{bmatrix} f_{t_{m+1}} \\ y^*_{t_{m+1}} \\ y^C_{t_{m+1}} \end{bmatrix} = \begin{bmatrix} \Lambda_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Xi_{t_{m\ |1}} \end{bmatrix} \begin{bmatrix} f_{t_m} \\ y^*_{t_m} \\ y^C_{t_m} \end{bmatrix} + \begin{bmatrix} B\eta_{t_{m+1}} \\ 0 \\ 0 \end{bmatrix} \tag{88}$$

The aggregation rule (86) is implemented in a recursive way, by introducing a latent cumulator variable $y^C_{t_m} = \Xi_{t_m} y^C_{t_{m-1}} + \frac{1}{3} y^*_{t_m}$, where $\Xi_{t_m} = 0$ for t_m corresponding to the first month of the quarter and $\Xi_{t_m} = 1$ otherwise. The estimation of the model parameters follows Giannone, Reichlin and Small (2008).

6.6 Factor MIDAS models

It is possible to augment the MIDAS regressions with the factors extracted from a large dataset to obtain a richer family of models that exploit a large high-frequency dataset to predict a low-frequency variable.

While the basic MIDAS framework consists of a regression of a low-frequency variable on a set of high-frequency indicators, the Factor-MIDAS approach exploits estimated factors rather than single or small groups of economic indicators as regressors.

Marcellino and Schumacher (2010) propose alternative factor-MIDAS regressions. In the standard MIDAS case, they follow Clements and Galvao (2008), while as a modification they evaluate a more general regression approach, labeled unrestricted Factor-MIDAS, where the dynamic relationship between the low-frequency variables and the high-frequency indicators is unrestricted, in contrast to the distributed lag functions as proposed by Ghysels, Sinko and Valkanov (2007). As a third alternative, they consider a regression scheme proposed by Altissimo et al. (2010), which considers only correlation at certain frequencies between variables sampled at high- and low- frequencies. This approach is called smoothed MIDAS, since the regression essentially eliminates high-frequency correlations.

The information set consists of a large set of stationary monthly indicators, X_{t_m}.

The last observation is at time $T_m + w, w > 0$, allowing for at most $w > 0$ monthly values of the indicators that are earlier available than the lower-frequency variable to be estimated. X_{t_m} is modeled using a factor representation, where r factors F_{t_m} are estimated in order to summarize the information in X_{t_m} The estimated factors, F_{t_m} , are then used in the various MIDAS based projections for the quarterly-frequency variable.

In the basic Factor-MIDAS approach the explanatory variables used as regressors are estimated factors. Assume for simplicity r = 1, so that there is only one factor f_{tm}. The Factor-MIDAS model for forecast horizon h_q quarters with $h_q = h_m/3$ is

$$y_{t_q\ |h_q} = y_{t_m\ |h_m} = \beta_0 + \beta_1 b(L_m;\theta) f^{(3)}_{t_{m+w}} + \varepsilon_{tm\ |h_m} \tag{89}$$

where $b(L_m;\theta) = \sum_{k=0}^{K} c\ (k;\theta) L_m^k$ and $c(k;\theta) = \frac{\exp(\theta_1 k + \theta_2 k^2)}{\sum_{k=0}^{K} \exp(\theta_1 k + \theta_2 k^2}$.

$\hat{f}^{(3)}_{t_m}$ is skip-sampled from the monthly factor $\hat{f}_{t_m}$. Every third observation starting from the final one is

included in the regressor $\hat{f}_{t_m}^{(3)}$, i.e.$\hat{f}_{t_m|w}^{(3)} = f_{t_{m+w}}, \forall l_m + w = \ldots, T_m + w - 6, T_m + w - 3, T_m + w$. As described above in the MIDAS models, the exponential lag function provides a parsimonious way to consider monthly lags of the factors.

The model can be estimated using nonlinear least squares in a regression of y_{t_m} onto the factors $\hat{f}_{t_{m+w-h}}^{(3)}$. The forecast is given by

$$y_{T_{m+h_m|T_{m+w}}} = \hat{\beta}_0 + \hat{\beta}_1 b\left(L_m; \hat{\theta}\right)\hat{f}_{t_{m+w}}^{(3)}. \tag{90}$$

The projection is based on the final values of estimated factors.

Factor-MIDAS regression can be generalized to more than one factor and extended with the addition of autoregressive dynamics. But the presence on non- linearity complicates estimation. Hence, as a simpler alternative, a U-MIDAS approach can be adopted. Specifically,

$$y_{T_{m+h_m}} = \beta_0 + D(L_m)\hat{f}_{t_{m+w}}^{(3)} + \varepsilon_{tm+h_m}, \tag{91}$$

where $D(L_m) = \sum_{k\ 0}^{K} D_k L_m^k$is an unrestricted lag polynomial of order K. $D(L_m)$ and β_0 are estimated by OLS. To specify the lag order in the empirical application, Marcellino and Schumacher (2010) consider a fixed scheme with $k = 0$ and an automatic lag length selection using the BIC.

Factor U-MIDAS provides a nice way to summarize (possibly very) large sets of indicators and use them in a nowcasting and forecasting framework.

7 A comparison of the reviewed methods

The performance of the big data modelling methods reviewed depends on what the underlying data generating process (DGP) is. The penalised regression models are linear so that, under correct specification, they all deliver consistent and asymptotically normal estimators, as shown in Knight and Wu (2000).

When the number of regressors is fixed standard $\sqrt{T}$ asymptotics applies whereas, if the number of regressors N is increasing as $T \to \infty$, as Knight and Fu (2000) show, $\lambda_T = o(T)$ is sufficient for consistency and $\lambda_T = O\sqrt{T}$ is sufficient for asymptotic normality for the case of the general penalised regression model. Some work has been done in the special case of Lasso (Bühlmann and van de Geer (2011)) and, when $N \to \infty$, the rate of convergence of the estimator is reduced but nonetheless consistency is achieved. Ridge-regularized methods have also been shown to work even when $N \gg T$ for a range of problem settings (Sutton and McCallum, 2006; Toutanova et al., 2003). Bernau et al. (2014) find that ridge regression performs better than lasso regression and boosting. The advantages of penalised regression methods is that they are linear (hence easily interpretable) and computationally cheap to estimate (especially in cases when the estimator is available in closed form). Moreover, Lasso, Slab and Spike regression and Compressed regression are more appropriate in cases where the model is sparse (i.e. there are many irrelevant regressors in X), as an inherent property of these methods is that the dimension of the regressors X is reduced, while ridge and Lp norm penalised regression models perform better in situations where the model is approximately sparse (i.e. there are many very small but not zero elements in β). Elastic net and Lp norm with $p \in (1,2)$ provide a compromise between sparsity and approximate sparsity; in that respect, they will be more robust under model uncertainty.

VAR and BVAR model have been documented to perform well in forecasting real data, but the underlying assumption made is that the data are dynamic and generated using a linear autoregressive process; so these models will perform poorly in cases when the data are not dependent, for example i.i.d. Moreover, allowing for time variation in the parameters of VAR models can help alleviate forecast bias and improve density forecasts when the true parameters are subject to structural change. On the other hand, allowing for time variation when the parameters are constant, will result in higher variance and hence worsen forecasting. Moreover, in general introducing time variation further increases the number of model parameters, worsening the "curse of dimensionality". The latter is typically addressed by introducing an amount of shrinkage, more so the large the number of variables under analysis. Hence, the various shrinkage methods discussed in the univariate case can be also applied in the vector context.

Choosing between a univariate or a multivariate specification is not obvious. The latter can be preferable either when multiple variables are of interest, so y is a vector, or to produce iterated rather than direct forecasts, where the former are typically more efficient than the latter. However, if one of the equations of the VAR is misspecified, this can affect the entire system, and in this sense the

univariate approach can be more robust, besides being computationally faster.

A common feature of univariate and multivariate regression models is to impose linearity in the formulation of the expectation of y conditional on (the many) elements of X. If, however, the DGP is nonlinear, then the linear specification, and its related extensions such as penalised or shrinked regressions, will generally not perform well. A non- or semi-parametric specification can be preferable in this case, at the cost of losing substantially in terms of efficiency if instead a linear model would suffice.

Regression trees and cluster analysis are popular nonlinear semi-parametric methods in the machine learning literature, and their theoretical properties have been established in cases when the data are i.i.d. In the time series context, their theoretical properties are unknown for dependent data and, since both methods approximate a nonlinear function of the regressors X, they require reshuffling and reordering of the observations, which will not respect the time series structure of the data. Hence, we do not expect in general these methods to perform well with dependent data.

Quantile and expectile regressions could instead be useful whenever tail behavior (fat tails, asymmetry) is of interest. However, they require very long time series for reliable estimation of the tail behavior, and they are computationally intensive for consideration in a big data context. Further, there is no conclusive evidence in the macroeconomic literature that they provide any added value in terms of forecasting compared to other methods in this context.

Factor based methods, and other approaches that summarize the information prior to its use in regression models, typically perform well for empirical macroeconomic forecasting. Hence, they could be reasonable competitors for penalized and shrunk regressions when the focus is on forecasting a target variable based on a (very) large dataset. While many methods are available for factor extraction, static principal component analysis and PLS seem particularly suited, due to their computational simplicity, possibly after modification to allow for sparsity in the information set.

As discussed, there are no studies on the performance of forecast pooling when the set of available forecasts is very large, but proper methods for averaging exist also in this context, and an evaluation of their (possibly relative) performance would be quite interesting.

It is worth mentioning that some of the methods we have considered could be also jointly or sequentially applied. For example, heuristic optimization or principal component analysis could be used to reduce the dimensionality of the big data, and then penalized or shrinkage regression could be applied on the subset of regressors obtained in the first step.

Finally, it should be emphasized that in the context of economic forecasting it is also important that the adopted model has an adequate economic interpretation and justification. This criterion can be quite relevant in the model specification process, to eliminate models that are clearly capturing spurious correlation with the target variable.

8 Conclusions

In this paper we provide a detailed survey of various methods for estimation and inference in the presence of large data, with a particular focus on the most recent methodological improvements in the field of Bayesian econometrics.

We surveyed diff t penalised regression estimators such as Ridge, Lasso, and Lp-norm penalised regression, and showed how they can be equivalently interpreted as posterior modes of a linear Bayesian regression model with an appropriate choice of prior distributions for the model's parameters. We also reviewed additional Bayesian methods, resulting from the choice of other prior distributions (such as spike and slab regressions and compressed regressions). The advantages of penalised regression methods is that they are simple, linear (and hence tractable) and have known asymptotic and good finite sample properties even when N is large. Finally, we considered methods that go beyond mean estimation, such as quantile and expectile regressions. We also discussed multivariate dynamic regression methods such as Bayesian VARs and some extensions, such as allowing for parameter time variation and stochastic volatility.

We also discussed some non-parameteric and non-linear approaches suited for applications with big data, such as random trees, random forests, cluster analysis, deep learning and neural networks. The relative advantage of these approaches is robustness: they do not require a parametric model for the variable of interest and can handle nonlinearities. The disadvantage is that they can be computationally intensive and their properties for time series data are not well studied.

Finally, and for completeness, based on Kapetanios, Marcellino and Papailias (2016) we also reviewed methods for summarizing the information in large datasets, such as principal components, partial least squares, and their sparse versions; procedures for forecast combination, based on the idea of combining a big set of simple models based on single indicators; and techniques to handle mixed frequency indicators, which is relevant as big data based variables are typically available at higher frequency than common macroeconomic indicators.

9 Bibliography

1. Ahn, S.K., Reinsel, G.C. (1988). Nested Reduced-Rank Autoregressive Models for Time Series, *Journal of the American Statistical Association*, 83, 849-856.
2. Akaike, H. (1974). A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, 19, 716-723.
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. Bae, K. and Mallick, B. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18):3423-3430.
5. Baffigi, A., Golinelli, R, Parigi, G. (2004). Bridge Models to Forecast the Euro Area GDP, *International Journal of Forecasting*, 20(3), 447-460.
6. Bai, J. (2003). Inferential Theory for Factor Models of Large Dimension, *Econometrica*, 71, 135-173.
7. Bai, J. and Ng, S. (2014) Boosting diff indices, Journal of Applied Econometrics, 24(4): 607–629.
8. Bańbura M., Giannone D., Reichlin L. (2011). Nowcasting. In Oxford Handbook on Economic Forecasting, Clements MP, Hendry DF (eds). Oxford University Press: Oxford.
9. Bańbura, M., Giannone, D., Modugno, M., Reichlin, L. (2013). Now-Casting and the Real-Time Data Flow, *ECB Working Paper Series*, No 1564.
10. 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.

11. Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012) Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica*, 80, 2369-2429.
12. Belloni, A., and Chernozhukov, V. (2013) Least Squares After Model Selection in High-dimensional Sparse Models, *Bernoulli*, 19(2), 521-547.
13. Bernanke, B. and Mihov, I. (1998). Measuring monetary policy, *Quarterly Journal of Economics* 113: 869–902.
14. Bernau, C., Riester, M., Boulesteix, A., Parmigiani, G., Huttenhower, C., Waldron, L. and Trippa, L. (2014) Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, 30(12):105-112.
15. Bhattacharjee, S. (2016). Big Data Seminar Presentation, Eurostat, European Commission.
16. Blake, A., Kapetanios, G. (2010). Tests of the Martingale Difference Hypothesis Using Boosting and RBF Neural Network Approximations, *Econometric Theory*, 26(5), 1363–1397.
17. Breiman, L. (1996) Bagging predictors, Machine Learning, 24(2):123–140. Breiman, L (2001). Random Forests. Machine Learning 45(1): 5-32.
18. Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and Regression Trees, *Wadsworth*, New York
19. 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.
20. Bühlmann, P. (2006). "Boosting for High-Dimensional Linear Models", Annals of Statistics, 34(2), 599-583.
21. Bühlmann, P., van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer.
22. Burnham, K.P., Anderson, D.R. (1998). Model Selection and Inference: a Practical Information-Theoretic Approach, Springer-Verlag, New York, USA.
23. Camba-Mendez, G., Kapetanios, G., Smith, R.J., Weale, M.R. (2003). Tests of Rank in Reduced Rank Regression Models, *Journal of Business and Economic Statistics*, 21, 145-155.
24. Candes, E. and Tao, T. Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? *IEEE Transactions on Information Theory* 52(2), 5406 - 5425.

25. Carlin, B. and Polson, G. (1991) Inference for nonconjugate Bayesian models using the Gibbs sampler. The Canadian Journal of Statistics, 19(4):399-405.

26. Caron, F. and Doucet, A. (2008) Sparse Bayesian nonparametric regression. In ICML '08: Proceedings of the 25th international conference on Machine learning, 88-95.

27. Carriero, A., Clark, T. and Marcellino, M. (2013) Bayesian VARs: Specification Choices and Forecast Accuracy. Journal of Applied Econometrics. 30(1), 46-73.

28. Carriero, A., Clark, T., 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.

29. Carriero, C., Kapetanios, G., Marcellino, M. (2011). Forecasting Large Datasets with Bayesian Reduced Rank Multivariate Models, Journal of Applied Econometrics, 26, 736-761.

30. Carriero, C., Kapetanios, G., Marcellino, M. (2015). Structural Analysis with Multivariate Autoregressive Index Models, Journal of Econometrics, 192, (2) 332- 348.

31. Carriero, Andrea, Todd E. Clark, and Massimiliano Marcellino, (2016) Large Vector Autoregressions with Asymmetric Priors and Time-Varying Volatilities, Federal Reserve Bank of Cleveland Working Paper, no. 16-17

32. Carvalho, C., Polson, N. and Scott, J. (2010) The horseshoe estimator for sparse signals. Biometrika, 97 (2): 465-480.

33. Chudik, A., Kapetanios, G., Pesaran, M.H. (2015). A Multiple Testing Approach to Variable Selection in Linear Regression models with a Large Number of Covariates, Working Paper.

34. Clements, M.P., Galvao, A.B. (2008). Macroeconomic Forecasting with Mixed- Frequency Data: Forecasting US Output Growth, Journal of Business and Economic Statistics, 26, 546-554.

35. Clements, M.P., Galvao, A.B. (2009). Forecasting US Output Growth using Leading Indicators: An Appraisal using MIDAS Models, Journal of Applied Econometrics, 24(7), 1057-1217.

36. Cogley, T., Primiceri, G. E. and Sargent, T. J. (2010). Inflation-gap persistence in the US, American Economic Journal: Macroeconomics 2(1): 43-69.

37. Cogley, T. and Sargent, T. J. (2002). Evolving post-World War II U.S. inflation dynamics, in B. S. Bernanke and K. Rogoff (eds), NBER Macroeconomics Annual, MIT Press: Cambridge, pp. 331-88.

38. Cogley, T. and Sargent, T. J. (2005). Drifts and volatilities: Monetary policies and outcomes in the post World War II US, *Review of Economic Dynamics*. 8: 262-302.

39. 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.

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

41. Diron, M. (2008). Short-term Forecasts of Euro Area Real GDP Growth. An Assessment of Real-Time Performance Based on Vintage Data, *Journal of Forecasting*, 27(5), 371-390.

42. Doan, T., Litterman, R., Sims, C.A. (1984). Forecasting and Conditional Projection Using Realistic Prior Distributions, Econometric Reviews, 3, 1-100.

43. Dobriban, E. and Wager, S. (2017) High-Dimensional Asymptotics of Prediction: Ridge Regression and Classification. Annals of Statistics, Forthcoming

44. Donoho, D. (2006). Compressed sensing. IEEE Transactions on Information Theory, 52(4), 1289-1306.

45. Doornik, J. A., Hendry, D. F. (2014). Statistical Model Selection with Big Data, Cogent Economics & Finance, 3(1), 2015.

46. Doz, C., Giannone, D., Reichlin, L. (2011). A Two-step Estimator for Large Approximate Dynamic Factor Models Based on Kalman Filtering, Journal of Econometrics, 164(1), 188-205.

47. Eklund, J., Karlsson, S. (2007). Forecast Combination and Model Averaging Using Predictive Measures, Econometric Reviews, 26(2), 329-363.

48. Eraker, B., Chiu, C.W., Foerster, A.T., Kim, T.B., Seoane, H.D. (2015). Bayesian Mixed Frequency VARs, Journal of Financial Econometrics, 13(3), 698-721.

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

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

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

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

53. Figueiredo, M. (2003). Adaptive sparseness for supervised learning. IEEE Trans- actions on Pattern Analysis and Machine Intelligence, 25(9):11501159.

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, Working Paper 2013/06.

57. Foroni, C., Marcellino, M. (2014). A Comparison of Mixed Frequency Approaches for Nowcasting Euro Area Macroeconomic Aggregates, International Journal of Fore- casting, 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., Proietti, T. (2010). Survey Data as Coinci- dent or Leading indicators, Journal of Forecasting, 29(1-2), 109-131.

60. Frale, C., Marcellino, M., Mazzi, G., Proietti, T. (2011). EUROMIND: A Monthly Indicator of the Euro Area Economic Conditions, Journal of the Royal Statistical Society: Series A, 174, 439-470.

61. Friedman, J. (2001). Greedy Function Approximation: a Gradient Boosting Ma- chine, 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. Gershenfeld, N., Schoner, B., Metois, E. (1999). Cluster-weighted modelling for time-series analysis, Nature, 397(6717), 329-332.

65. Geweke, J. (1996). Bayesian Reduced Rank Regression in Econometrics, Journal of Econometrics, 75, 121-146.

66. Giraitis, L., Kapetanios, G. and Price, S. (2013). Adaptive forecasting in the presence of recent and ongoing structural change, Journal of Econometrics 177: 153- 170.

67. Giraitis, L., Kapetanios, G., Wetherilt, A. and Zikes, F. (2016). Estimating the dynamics and persistence of financial networks, with an application to the Sterling money market, Journal of Applied Econometrics 31(1): 58-84.

68. Giraitis, L., Kapetanios, G. and Yates, T. (2014). Inference on stochastic time- varying coefficient models, Journal of Econometrics 179(1): 46-65.

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

70. Ghysels, E., Santa-Clara, P., Valkanov, R. (2005). There is a Risk-Return Trade- off After All, Journal of Financial Economics, 76(3), 509-548.

71. Ghysels, E., Sinko, E., Valkanov, R. (2007). MIDAS Regressions: Further Results and New Directions, Econometric Reviews, 26(1), 53-90.

72. Giannone, D., Reichlin, L., Small, D. (2008). Nowcasting GDP and Inflation: The Real-Time Informational Content of Macroeconomic Data Releases, Journal of Monetary Economics, 55, 665-676.

73. Giannone, D., Lenza, M. and Primiceri, G. (2015) Prior Selection for Vector Autoregressions. Review of Economics and Statistics. 97(2), 436-451.

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

75. Griffin, J. and Brown, P. (2005) Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick.

76. Griffin, J. and Brown, P. (2010). Inference with normal-gamma prior distributions in regression problems. Bayesian Analysis, 5(1):171-188.

77. Groen, J., Kapetanios, G. (2015). Revisiting Useful Approaches to Data-Rich Macroeconomic Forecasting, Computational Statistics & Data Analysis.

78. Guhaniyogi, R. and Dunson, D. (2015). Bayesian compressed regression. Journal of the American Statistical Association, 110, 1500-1514.

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

80. Hans, C. (2009) Bayesian lasso regression. Biometrika, 96(4):835-845.

81. Harvey, A. (1989). Forecasting: Structural Time Series Models and the Kalman Filter. Cambridge University Press.

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

83. Heaton, J. B., Polson, N. G., Witte, J. H. (2016a) Deep Portfolio Theory, Working Paper

84. Heaton, J. B., Polson, N. G., Witte, J. H. (2016b) Deep Learning in Finance, Working Paper

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

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

87. Hepenstrick, C., Marcellino, M. (2015). Forecasting with Large Unbalanced Datasets: The Mixed Frequency Three-Pass Regression Filter, Working Paper, Swiss National Bank.

88. Hinton, G. E., Salakhutdinov, R.R. (2006). Reducing the Dimensionality of Data with Neural Networks, Science, 313, 504-507.

89. Hoerl, A. E. and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12: 55–67.

90. Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T. (1999). Bayesian Model Averaging: a Tutorial, Statistical Science, 14, 382-417

91. 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.

92. Hornik, K., M. Stinchcombe, and H. White (1989): Multi-Layer Feedforward Networks and Universal Approximators, Neural Network, 2, 359–366.

93. 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.

94. Jacquier, E., Polson, N. G. and Rossi, P. (1994). Bayesian analysis of stochastic volatility models, Journal of Business and Economic Statistics 12: 371-418.

95. Kadiyala, K. R., Karlsson, S. (1997). Numerical Methods for Estimation and Inference in Bayesian VAR-Models, Journal of Applied Econometrics, 12(2), 99-132.

96. 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.

97. Kapetanios, G. (2006). Variable Selection in Regression Models using Non- Standard Optimisation of Information Criteria, Computational Statistics & Data Analysis, 52(1), 4-15.

98. Kapetanios, G., Labhard, V., Price, S. (2008). Forecasting using Bayesian and Information Theoretic Model Averaging: An application to UK inflation Journal of Business and Economic Statistics, 26(1), 33-41.

99. Kapetanios, G. and Blake, A. (2017) Tests of the Martingale Difference Hypothesis using Boosting and RBF Neural Network Approximations. Working Paper

100. Kapetanios, G., Marcellino, M. and Papailias, F. (2016) Big Data and Macroeconomic Nowcasting Task 3: Big data and modelling, Eurostat Report.

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

102. 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.

103. Kim, C. and Nelson, C. R. (1999). Has the U.S. Economy become more stable? A Bayesian Approach based on a Markov-Switching Model of the Business Cycle, Review of Economics and Statistics 81(4): 608–618.

104. Kim, S., Shephard, N. and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models, Review of Economic Studies, 65: 361- 393.

105. Knight, K. and Fu, W. (2000) Asymptotics for Lasso-Type Estimators, The Annals of Statistics, 28(5):1356-1378.

106. Koenker, R. and Bassett, G. (1978) Regression Quantiles. Econometrica, 46(1), 33-50.

107. Koop, G. and Korobilis, D. (2012) Forecasting Inflation using Dynamic Model Averaging, International Economic Review. 53(3), 867-886.

108. Koop, G., Korobilis, D. and Pettenuzzo, D. (2016) Bayesian Compressed VARs, Working Paper.

109. Komunjer (2005) Quasi-Maximum Likelihood Estimation for Conditional Quantiles, Journal of Econometrics, 128, 137–164.

110. Kuzin, V., Marcellino, M. and C. Schumacher (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.

111. Litterman, R. (1986). Forecasting With Bayesian Vector Autoregressions - Five Years of Experience, Journal of Business and Economic Statistics, 4, 25-38.

112. 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.

113. Marcellino, M., Schumacher, C. (2010). Factor-MIDAS for Now- and Forecasting with Ragged-Edge Data: A Model Comparison for German GDP, Oxford Bulletin of Economics and Statistics, 72, 518-550.

114. 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.

115. Mariano, R., Murasawa, Y. (2010). A Coincident Index, Common Factors, and Monthly Real GDP, Oxford Bulletin of Economics and Statistics, 72(1), 27-46.

116. Mart´ınez Á lvarez, 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).

117. 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.

118. McConnell, M. and Perez Quiros, G. (2000). Output fluctuations in the U.S.: what has changed since the early 1980s?, American Economic Review 90: 1464–1476.

119. Mitchell,T. and Beauchamp, J. (1988) Bayesian variable selection in linear regression. Journal of American Statistical Association, 83:1023–1036.

120. 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.

121. 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.

122. Ng, S. (2014) Viewpoint: Boosting Recessions, Canadian Journal of Economics, 47(1), 1-34.

123. Park, T. and Casella, G. (2008) The Bayesian Lasso. Journal of American Statistical Association. 103: 681-686.

124. Pericchi, L. and Smith, A. Exact and approximate posterior moments for a normal location parameter. Journal of the Royal Statistical Society (Series B), 54(3):793-804.

125. Petrova, K. (2017). A quasi-bayesian local likelihood approach to time varying parameter VAR models, Working Paper.

126. Poiraud-Casanova, S. and Thomas-Agnan, C. (2000). About monotone regression quantiles, Statistics and Probability Letters 48, 101–104.

127. Polson, N. and Scott, J. (2009) Alternative global local shrinkage rules using hypergeometric beta mixtures. Technical Report 14, Duke University Department of Statistical Science.

128. Polson, N. and Scott, J. (2010) Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction, in J. M. Bernardo, M. J. Bayarri, J. O. Berger, A.

129. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (eds.). Bayesian Statistics, Oxford University Press

130. Raftery, A., Madigan, D. and Hoeting, J (1997), Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association, 92: 179–191.

131. Reinsel, G. (1983). Some Results on Multivariate Autoregressive Index Models, Biometrika, 70, 145-156.

132. Reinsel, G.C., Velu, R.P. (1998). Multivariate Reduced Rank Regression, Lecture Notes in Statistics 136. New York: Springer-Verlag.

133. Schorfheide, F., Song, D. (2011). Real-time Forecasting with a Mixed Frequency VAR, NBER Working Paper No. 19712.

134. Scott, S. and Varian, H. (2013) Predicting the Present with Bayesian Structural Time Series, Working Paper.

135. Sims, C. A. (1980). Macroeconomics and reality, Econometrica 48: 1–48.

136. Sims, C. A. and Zha, T. (2006). Were there regime switches in U.S. monetary policy?, American Economic Review 96: 1193–1224.

137. Sin, C.Y., White, H. (1996). Information Criteria for Selecting Possibly Misspecified Parametric Models, Journal of Econometrics, 71(1-2), 207-225.

138. Sirignano, J., Sadhwani, A. and Giesecke, K. (2016) Deep Learning for Mortgage Risk, Working Paper

139. Stock, J.H., Watson, M.W. (2002a). Forecasting using Principal Components from a Large Number of Predictors, Journal of the American Statistical Association, 97, 147–162.

140. Stock, J.H., Watson, M.W. (2002b). Macroeconomic Forecasting using Diffusion Indexes, Journal of Business and Economic Statistics, 20, 147–162.

141. 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.

142. Sutton, C. and McCallum, (2006) A. An introduction to conditional random fields for relational learning. Introduction to statistical relational learning, 93–128.

143. Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Series B, 58, 267–288.

144. Tibshirani, R. (2013), The Lasso problem and Uniqueness, Electronic Journal of Statistics, 7, 1456-1490.

145. Tipping, M. (2001) Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research, 1:211-244, .

146. Toutanova, D. Klein, C. D. Manning, and Y. Singer (2003), Feature-rich part-of- speech tagging with a cyclic dependency network. NAACL.

147. Velu, R.P., Reinsel, G.C., Wichern, D.W. (1986). Reduced Rank Models for Multiple Time Series, Biometrika, 73, 105-118.

148. Wang, L., Gordon, M. D. and Zhu, J. (2006). Regularized least absolute deviation regression and an efficient algorithm for parameter tuning. Proceedings of the Sixth International Conference on Data Mining, IEE Computer Society, 690-700.

149. West, M. (1987) On scale mixtures of normal distributions. Biometrika, 74(3):646-8.

150. Wu, T.T. and Lange, K. (2008). Coordinate Descent Algorithms for Lasso Penalised Regression, The Annals of Applied Statistics, 2(1), 224-244.

151. Yue Y and Rue H (2011) Bayesian inference for additive mixed quantile regression models. Computational Statistics & Data Analysis, 55, 84–96.

152. Zadrozny, P (1988). Gaussian Likelihood of Continuous-Time ARMAX Models When Data Are Stocks and Flows at Different Frequencies, Econometric Theory, 4(1), 108-124.

153. Zheng, Z., Fan, Y., Lv, J. (2013). High Dimensional Thresholded Regression and Shrinkage Effect, Journal of the Royal Statistical Society: Series B, 76(3), 627-649.

154. Zhou, S., Lafferty, J. and Wasserman, L. (2007) Compressed Regression. IEEE Transactions on Information Theory, 55(2)

155. 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.

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

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

Getting in touch with the EU

In person
All over the European Union there are hundreds of Europe Direct Information Centres. You can find the address of the centre nearest you at: http://europa.eu/contact

On the phone or by e-mail
Europe Direct is a service that answers your questions about the European Union. You can contact this service
– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
– at the following standard number: +32 22999696 or
– by electronic mail via: http://europa.eu/contact

Finding information about the EU

Online
Information about the European Union in all the official languages of the EU is available on the Europa website at: http://europa.eu

EU Publications
You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see http://europa.eu/contact)

EU law and related documents
For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU
The EU Open Data Portal (http://data.europa.eu/euodp/en/data) provides access to datasets from the EU. Data can be downloaded and reused for free, both for commercial and non-commercial purposes.

KS-TC-18-001-EN-N

Analysis of the most recent modelling techniques for big data with particular attention to Bayesian ones

In this report we describe various methods suited for the analysis of linear models with a very large number of explanatory variables, with a special emphasis on Bayesian approaches. We next consider some non-parametric and/or non-linear methods suited for applications with big data, such as random trees, random forests, cluster analysis, deep learning and neural networks. Finally, we survey techniques for summarizing the information in large (possibly sparse) datasets, forecast combination approaches, and techniques for the analysis of large mixed frequency datasets.

For more information
http://ec.europa.eu/eurostat/

Publications Office

ISBN 978-92-79-77350-1