Bayesian Adaptive Penalized MIDAS Regressions: Estimation, Selection, and Prediction

Clément Marsilli, Matteo Mogliani[1,*]

Banque de France - International Macroeconomics Division

Very preliminary and incomplete draft. Please do not quote.

Abstract

We propose a new approach to modeling and forecasting mixed-frequency regressions (MIDAS) that address the issues of estimation and variable selection in presence of a large number of predictors. Our approach is based on adaptive penalized regression models (Lasso, Group Lasso, and Elastic-Net) and relies on Bayesian techniques for estimation. In particular, the penalty hyper-parameters driving the model shrinkage are automatically tuned via an adaptive MCMC algorithm, which is computationally efficient compared to the standard MCEM algorithm. Simulations show that the proposed models present very good in-sample and out-of-sample performance. When applied to US GDP, the results suggest that our models produce significant out-of-sample predictive gains compared to several alternative models.

Keywords: MIDAS regressions, Variable selection, Forecasting, Bayesian estimation, Adaptive Lasso-like models

JEL: C11, C22, C53, E37

[*]Corresponding author; Banque de France, 46-1374 DGEI-DERIE-SEMSI, 31 Rue Croix des Petits Champs, 75049 Paris CEDEX 01 (France). Phone: +33(0)142929756.

Email addresses: clement.marsilli@banque-france.fr (Clément Marsilli), matteo.mogliani@banque-france.fr (Matteo Mogliani)

[1]We wish to thank participants at the 18th IWH-CIREQ-GW Macroeconometric Workshop on Mixed Frequency Data in Macroeconomics and Finance (12-13 December 2017, Halle, Germany), and the 11th International Conference on Computational and Financial Econometrics (16-18 December 2017, London, UK). The usual disclaimer applies. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Banque de France.

January 31, 2018

1. Introduction

In the macroeconomic forecasting literature, the use of mixed-data sampling (MIDAS) models has gained considerable attention in the last decade. MIDAS models (Ghysels et al., 2005) aggregate data sampled at different frequencies than the outcome variable of interest by using weighting schemes on current and past values of the high-frequency predictors that resort to either functional lag polynomials (Ghysels et al., 2007) or unrestricted linear lag polynomials (Foroni et al., 2015). Indeed, according to the empirical literature, MIDAS models implemented on high-frequency data (usually monthly or daily) have been be proved to improve forecasts of macroeconomic variables observed at lower frequency (usually quarterly), such as GDP (Clements and Galvão, 2008, 2009; Andreou et al., 2013; Mogliani et al., 2017). However, the main issue with MIDAS models relies to the selection of predictors in presence of large datasets, as the introduction of many high-frequency variables into MIDAS regressions may easily lead to overparemeterized models, overfitting issues, and poor predictive performance. In the present paper, we propose a new method that reconciles MIDAS models and variable selection in a simple regression framework.

In the recent literature, three popular model reduction strategies have been proposed to overcome the curse of dimensionality. The first one implements a model selection strategy based on U-MIDAS regressions and a general-to-specific algorithm to jointly select the relevant predictors and the relevant high-frequency lags (Castle et al., 2009; Castle and Hendry, 2010; Bec and Mogliani, 2015). The second one, known as Factor-augmented MIDAS, implies the extraction of common factors from high-frequency variables and then the estimation of standard MIDAS regression using these high-frequency factors (Marcellino and Schumacher, 2010). The third one relies on the *targeted predictors* approach developed by Bai and Ng (2008), where the pre-selection of relevant high-frequency variables is based on hard- and soft-thresholding rules carried out using the Lasso (*Least Absolute Shrinkage and Selection Operator*) technique introduced by Tibshirani (1996). Hence, this approach makes use of penalized regressions as a model reduction technique prior to factors extraction (Bessec, 2013; Bulligan et al., 2015; Girardi et al., 2017; Siliverstovs, 2017), in order to attenuate the issue raised by Boivin and Ng (2006) on the usefulness of large panels of predictors when extracting common factors for forecasting purposes.

Although very popular, the targeted predictor approach of Bai and Ng (2008) relies on a two-steps procedure, where the second step requires the extraction of static or dynamic factors (Giannone et al., 2008; Doz et al., 2011), which in turn requires the determination of the number of factors and the dynamic structure (Bai and Ng, 2002, 2007; Ahn and Horenstein, 2013). Our proposed approach is different, as it relies on a single-step procedure. The models presented in the paper take advantage of the Bayesian Lasso estimator proposed by Park and Casella (2008), and further extended by Kyung et al. (2010) and Leng et al. (2014) to other Lasso-like penalized regressions (Group Lasso, Elastic-Net) and to adaptive shrinkage (Zou, 2006). We then combine a MIDAS framework based on Almon lag polynomials, that can be cast as a linear regression model with transformed high-

frequency predictors (Pettenuzzo et al., 2016), and Bayesian penalized regressions, where variable selection and model reduction are ensured by flexible penalty hyper-parameters that are here automatically tuned via a controlled MCMC algorithm based on stochastic approximations (Atchadé et al., 2011), which is computationally efficient compared to the standard MCEM algorithm proposed by Casella (2001).

The Bayesian modeling approach offers several advantages in our setting. First, Bayesian methods exploit model inference via posterior distributions of parameters. Second, they provide a flexible way of estimating the penalty hyper-parameters, along with other parameters in the model. Lastly, they provide forecasts via predictive distributions. Such distributions can be used to evaluate a range of measures of predictive accuracy and account for parameter estimation uncertainty, which is a relevant issue in empirical applications with macroeconomic variables with many available predictors and, usually, short data samples.

The estimation and predictive accuracy of our Bayesian adaptive penalized MIDAS models (Lasso, Group Lasso, and Elastic-Net) are assessed through Monte Carlo simulations using a data generating process which involves a sparse setting of highly correlated predictors. Simulations show that the proposed penalized MIDAS models present very good in-sample and out-of-sample performance. In particular, simulation results show that variable selection, carried out using a credible interval approach, is achieved with high probability.

The paper is structured as follows. Section 2 introduces the Bayesian MIDAS penalized regressions. Section 3 investigate the estimation and predictive features of our models via Monte Carlo simulations. In Section 4, we report an empirical application on US GDP (TO BE DONE). Finally, Section 5 concludes.

2. The Bayesian MIDAS penalized regression

2.1. Basic MIDAS setup

Let us assume we want to forecast the variable y_{t+h}, which is observed at discrete times $\ldots, t-1, t, \ldots, T-h-1$, using information stemming from a set of K predictors $\mathbf{x}_t^{(m)} = (x_{1,t}^{(m)}, \ldots, x_{K,t}^{(m)})'$, which are observed m times between $t-1$ and t. The variables y_{t+h} and $x_{k,t}^{(m)}$, for $k = 1, \ldots, K$, are said to be sampled at different frequencies. For instance, quarterly and monthly frequencies, respectively, in which case $m = 3$. Let us define the high-frequency lag operator $L^{1/m}$, such that $L^{1/m} x_{k,t}^{(m)} = x_{k,t-1/m}^{(m)}$. Further, let $h \geq 0$ be an (arbitrary) forecast horizon, where $h = 0$ denotes a nowcast. The MIDAS approach plugs-in the high-frequency lagged structure of predictors $x_{k,t}^{(m)}$ in a regression model for the low-frequency response variable y_{t+h} as follows:

$$y_{t+h} = \alpha + \sum_{k=1}^{K} \mathcal{B}\left(L^{1/m}; \boldsymbol{\theta}_k\right) x_{k,t}^{(m)} + \epsilon_{t+h}, \qquad \text{for } t = 1 \ldots, T-h-1 \tag{1}$$

where ϵ_{t+h} is i.i.d. with mean zero and variance $\sigma^2 < \infty$, and $\mathcal{B}\left(L^{1/m};\boldsymbol{\theta}_k\right) = \sum_{c=0}^{C-1} B\left(c;\boldsymbol{\theta}_k\right) L^{c/m}$ is the weighting structure which depends on the weighting function $B\left(c;\boldsymbol{\theta}_k\right)$, a vector of $p+1$ parameters $\boldsymbol{\theta}_k = (\theta_{k,0}, \theta_{k,1}, \ldots, \theta_{k,p})$, and a maximum lag length C. Several functional forms have been proposed in the literature, such as the exponential Almon or the Beta lag polynomials. In this study, we consider the simple polynomial approximation of $\mathcal{B}\left(L^{1/m};\boldsymbol{\theta}_k\right)$ provided by the Almon lag polynomial, which (under the so-called "direct method") takes the unrestricted form:[1]

$$B\left(c;\boldsymbol{\theta}_k\right) = \sum_{i=0}^{p} \theta_{k,i} c^i \tag{2}$$

leading to the following MIDAS regression:

$$y_{t+h} = \alpha + \sum_{k=1}^{K} \sum_{c=0}^{C-1} \sum_{i=0}^{p} \theta_{k,i} c^i L^{c/m} x_{k,t}^{(m)} + \epsilon_{t+h} \tag{3}$$

The main advantage of using the Almon lag polynomial is that (3) is linear and can be easily reparemeterized as to depend on $K(p+1)$ coefficients, rather than $KC(p+1)$ coefficients. This can be obtained by using the transformed vector of high-frequency regressors $\mathbf{z}_{k,t}^{(m)} = \mathbf{Q}\mathbf{x}_{k,t}^{(m)}$, where $\mathbf{x}_{k,t}^{(m)} = \left(x_{k,t}^{(m)}, x_{k,t-1/m}^{(m)}, \ldots, x_{k,t-(C-1)/m}^{(m)}\right)'$ is a $(C \times 1)$ vector of high-frequency lags and

$$\mathbf{Q} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & \cdots & (C-1) \\ 0 & 1 & 2^2 & \cdots & (C-1)^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 2^p & \cdots & (C-1)^p \end{pmatrix} \tag{4}$$

is a $(p+1 \times C)$ polynomial weighting matrix. We can hence rewrite (3) as:

$$y_{t+h} = \alpha + \sum_{k=1}^{K} \sum_{i=0}^{p} \theta_{k,i} z_{k,i,t}^{(m)} + \epsilon_{t+h} \tag{5}$$

or in more compact form:

$$y_{t+h} = \alpha + \boldsymbol{\theta}' \mathbf{Z}_t^{(m)} + \epsilon_{t+h}. \tag{6}$$

[1] Linear restrictions on the value and slope of the lag polynomial $B\left(c;\boldsymbol{\theta}_k\right)$ may be placed for any $c \in (0, C-1)$, although in practice restrictions on the endpoints are usually economically meaningful. In the present framework, restrictions such as $B\left(c;\boldsymbol{\theta}_k\right) = 0$ and $\nabla_c B\left(c;\boldsymbol{\theta}_k\right) = 0$, with c evaluated at $C-1$, may be desirable, as they jointly constrain the weighting structure to tail off slowly to zero. We shall use these restrictions in Section 3.

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)'$ and $\mathbf{Z}_t^{(m)} = \left(\mathbf{z}_{1,t}^{(m)}, \ldots, \mathbf{z}_{K,t}^{(m)}\right)'$, for $k = 1, \ldots, K$. The h-step-ahead direct forecast $\hat{y}_{T+h}$ can be hence obtained using sample information known at time T as follows:

$$\hat{y}_{T+h} = \hat{\alpha} + \hat{\boldsymbol{\theta}}' \mathbf{Z}_T^{(m)}. \tag{7}$$

It is worth noting that normalized MIDAS weights (*i.e.* summing up to unity) can be computed from (2), such that h-step-ahead direct forecasts $\hat{y}_{T+h}$ can be obtained by rewriting (1) as:

$$\hat{y}_{T+h} = \hat{\alpha} + \sum_{k=1}^{K} \hat{\beta}_k \tilde{\mathcal{B}}\left(L^{1/m}; \hat{\boldsymbol{\theta}}_k\right) x_{k,T}^{(m)} \tag{8}$$

where $\tilde{\mathcal{B}}\left(L^{1/m}; \hat{\boldsymbol{\theta}}\right)$ is the normalized weighting structure, with $\tilde{B}\left(c; \hat{\boldsymbol{\theta}}_k\right) = \hat{B}\left(c; \hat{\boldsymbol{\theta}}_k\right) \Big/ \sum_{j=0}^{C-1} \hat{B}\left(j; \hat{\boldsymbol{\theta}}_k\right)$, and $\hat{\beta}_k = \sum_{c=0}^{C-1} \sum_{i=0}^{p} \hat{\theta}_{k,i} c^i$ is a slope coefficient that captures the overall impact of lagged values of $x_{k,t}^{(m)}$ on y_{t+h}.[2] Further, Equation (1) can be generalized to allow for lags of the dependent variable, as well as additional predictors sampled at the same frequency as y_{t+h} and at multiple frequencies.

2.2. MIDAS penalized regressions

Although appealing, the MIDAS regression presented above may be easily affected by over-parameterization and multicollinearity in presence of numerous and potentially highly correlated predictors.[3] To achieve variable selection and parameter estimation simultaneously, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (Lasso). For ease of exposition, let us consider a simpler version of regression (6):

$$y_{t+h} = \boldsymbol{\theta}' \mathbf{Z}_t^{(m)} + \epsilon_{t+h}. \tag{9}$$

In a nutshell, the Lasso is a penalized least squares procedure, in which the loss function $\mathcal{L}_T(\boldsymbol{\theta})$ is minimized after setting a constraint on the ℓ_1 norm of the vector of regression coefficients, where the amount of penalization is controlled by a parameter λ. The objective function of the Lasso takes the form:

$$\mathcal{Q}_{\mathrm{L}}(\boldsymbol{\theta}) = T^{-1} \mathcal{L}_T(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1, \tag{10}$$

where $\mathcal{L}_T(\boldsymbol{\theta})$ is the negative log-likelihood function, $\|\boldsymbol{\theta}\|_1 = \sum_{k=1}^{K} \sum_{i=0}^{p} |\theta_{k,i}|$ denotes the ℓ_1 norm, and $\lambda \geq 0$. In the statistical literature, the main feature of a consistent variable selection estimator

[2] An alternative approach to the Almon weighting function is provided by the U-MIDAS (Foroni et al., 2015), where the vector of transformed regressors, $\mathbf{Z}_t^{(m)}$, is replaced by the vector of untransformed regressors $\mathbf{X}_t^{(m)} = \left(\mathbf{x}_{1,t}^{(m)}, \ldots, \mathbf{x}_{K,t}^{(m)}\right)'$.

[3] The direct method used in regression (6) may be also hampered by multicollinearity in the artificial variables $\mathbf{Z}_t^{(m)}$ (Cooper, 1972).

such as the Lasso is called the *oracle property*. Let us denote $\mathcal{A} = \{k, i : \theta^*_{k,i} \neq 0\}$ the true active set of coefficients of the lag polynomials, *i.e.* those terms $(z^{(m)}_{k,i,t})$ with non-zero coefficients, $\mathcal{A}^c$ its complement set, and $\boldsymbol{\Sigma}_{\mathcal{A}}$ the covariance matrix of the elements in the active set. Further, let us denote $\hat{\mathcal{A}}$ the active set provided by an estimator, and $\hat{\boldsymbol{\theta}}$ the vector of estimated coefficients. According to Fan and Li (2001), an estimator is said to possess the oracle property if it displays the following properties:

1. It identifies the right subset model: $P(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ as $T \to \infty$, *i.e.* it selects the correct sparsity pattern with probability tending to one, leaving out all irrelevant variables and retaining all relevant variables;
2. It has the optimal estimation rate: $\sqrt{T}(\hat{\boldsymbol{\theta}}_{\hat{\mathcal{A}}} - \boldsymbol{\theta}^*_{\mathcal{A}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{A}})$ as $T \to \infty$, *i.e.* it estimates the non-zero coefficients with the same rate and asymptotic distribution as if only the relevant variables had been included in the model.

In other words, the oracle property guarantees that the estimator performs as well as if the true model had been revealed to the researcher in advance by an oracle (Callot and Kock, 2014). However, Zou (2006), Zhao and Yu (2006), and Yuan and Lin (2007) showed that the Lasso estimator possesses the oracle property if and only if the so-called *irrepresentable condition* on the design matrix is satisfied and the penalization parameter λ is chosen judiciously.[4] If this condition does not hold, the Lasso estimator chooses the wrong model with non-vanishing probability, regardless of the sample size and how λ is chosen. This happens because the Lasso estimator in (10) uses the same amount of shrinkage for each regression coefficient, leading to estimation inefficiency and selection inconsistency. To address this issue, Zou (2006) proposed the Adaptive Lasso (AL), where a different amount of shrinkage (*i.e.* a different penalty term) is used for each individual regression coefficient. The objective function of the AL takes the form:

$$\mathcal{Q}_{\text{AL}}(\boldsymbol{\theta}) = T^{-1}\mathcal{L}_T(\boldsymbol{\theta}) + \sum_{k=1}^{K}\sum_{i=0}^{p} \lambda_{k,i}|\theta_{k,i}| \tag{11}$$

Now, let us assume that $\mathcal{L}_T(\boldsymbol{\theta})$ has continuous second-order derivative with respect to $\boldsymbol{\theta}$. From a Taylor series expansion of $\mathcal{L}_T(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}$, we have that

$$T^{-1}\mathcal{L}_T(\boldsymbol{\theta}) \approx T^{-1}\mathcal{L}_T(\hat{\boldsymbol{\theta}}) + T^{-1}\dot{\mathcal{L}}_T(\hat{\boldsymbol{\theta}})'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\left[T^{-1}\ddot{\mathcal{L}}_T(\hat{\boldsymbol{\theta}})\right](\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{12}$$

where $\dot{\mathcal{L}}_T(\cdot)$ and $\ddot{\mathcal{L}}_T(\cdot)$ denote first- and second-derivatives and $\hat{\boldsymbol{\theta}} = \text{argmin}\ T^{-1}\mathcal{L}_T(\boldsymbol{\theta})$ is the maximum likelihood estimator of the unpenalized regression. Let us further assume that $\hat{\boldsymbol{\theta}}$ is $\sqrt{T}$-

[4]The irrepresentable condition states that the predictors not in the model are not representable by predictors in the true model (*i.e.* the irrelevant predictors are roughly orthogonal to the relevant ones). This represents a necessary and sufficient condition for exact recovery of the non-zero coefficients, but it can be easily violated in cases where the design matrix exhibits too strong (empirical) correlations (collinearity between predictors).

consistent and asymptotically normal. Hence, $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta}^*$ is the vector of true values of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$. By ignoring constant terms and setting $T^{-1}\ddot{\mathcal{L}}_T(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\Sigma}}^{-1}$, where $\hat{\boldsymbol{\Sigma}}^{-1}$ is an estimate of the covariance matrix $\boldsymbol{\Sigma}^{-1}$, we obtain the following *least squares approximation* (LSA) of the original loss function:

$$T^{-1}\mathcal{L}_T(\boldsymbol{\theta}) \approx \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{13}$$

leading to the approximated AL objective function (Wang and Leng, 2007):

$$\mathcal{Q}'_{\text{AL}}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \sum_{k=1}^{K}\sum_{i=0}^{p}\lambda_{k,i}|\theta_{k,i}| \tag{14}$$

with global minimizer $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}$. The approximated AL objective function (14) requires only the existence of a consistent covariance matrix estimate $\hat{\boldsymbol{\Sigma}}$ and an appropriate choice for the penalty parameters $\lambda_{k,i}$. Wang and Leng (2007) derive (under a relatively mild assumption on the asymptotic covariance matrix) the asymptotic properties of the LSA estimator of the AL and they establish its estimation and selection consistency, as well as the oracle property.[5]

In the present framework of mixed-frequency models, a natural extension to the AL in (11) is provided by the Adaptive Group Lasso (AGL) estimator proposed by Wang and Leng (2008), who extended to adaptive shrinkage the Group Lasso estimator originally proposed by Yuan and Lin (2006). This approach introduces a penalty to a group of regressors, rather than a single regressor, that may lead (if the group structure is carefully set by the researcher) to a finite sample improvement of the AL. In the present framework, it seems reasonable to define a group as each of the k lag polynomials in regression (9), with a penalization applied to the entire lag polynomial for each high-frequency variable, rather than to each term of the lag polynomials as with the AL. This is motivated by the fact that if one high-frequency predictor is irrelevant, it is reasonable to expect that zero-coefficients occur in all the coefficients of the lag polynomial. Accordingly, let us partition the parameter vector $\boldsymbol{\theta}$ into G disjoint groups, $\boldsymbol{\theta}_j$, for $j = 1, \ldots, G$, each of size g_j. Despite the change in the notation (required to avoid confusion), it is straightforward to note that in the present framework $G = K$, $\boldsymbol{\theta}_j = \boldsymbol{\theta}_k$, $g_j = p + 1$, and $\tilde{g} \equiv \sum_{j=1}^{G} g_j = K(p + 1)$. Similarly to AL, we can use the LSA of the original loss function and obtain the approximated AGL objective function:

$$\mathcal{Q}'_{\text{AGL}}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \sum_{j=1}^{G}\lambda_j\|\boldsymbol{\theta}_j\|_2 \tag{15}$$

[5] Wang and Leng (2007) point out that $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}$ is in general different from the estimator obtained by minimizing the original objective function. However, quantifying this difference, which depends on the accuracy of the approximation provided in (13), is beyond the scope of the present paper.

where $\|\boldsymbol{\theta}_j\|_2 = (\boldsymbol{\theta}_j'\boldsymbol{\theta}_j)^{1/2}$ denotes the ℓ_2 norm. As for the asymptotic properties of the LSA estimator of the AGL, it is straightforward to extend the results in Wang and Leng (2008) to LSA and to establish the consistency and the oracle property of the estimator. However, as suggested by Callot and Kock (2014), the AGL possesses a variant of the oracle property if one correctly groups the potential predictors. This happens because selection consistency concerns all groups consisting only of parameters whose true value is zero, while for those parameters whose true value is zero but are located in an active group, the oracle property states that their asymptotic distribution is equivalent to the one of least squares including all variables. Hence, the AGL only performs better than least squares including all variables if one is able to identify groups consisting of parameters whose true value is zero. In the present framework, we expect that grouping lag polynomials, rather than set of predictors, should attenuate this issue.

Finally, we also consider the Adaptive Elastic-Net (AEN) proposed by Zou and Zhang (2009), who extended to adaptive shrinkage the Elastic-Net estimator originally proposed by Zou and Hastie (2005). Numerically, the AEN can deal with multicollinearity issues better than other methods (such as AL) by encouraging grouping effect, *i.e.* either selection or omission of the correlated variables together thanks to a regularization provided by a ℓ_2 penalty term (Ghosh, 2011). Unlike the AGL, where groups are exogenously set by the researcher, the grouping effect of the AEN is here expected to apply endogenously to the k lag polynomials in regression (9). Further, Zou and Zhang (2009) showed that, under weak regularity conditions, the AEN has the oracle property. In the present work, we follow Gefang (2014) and we propose a doubly AEN (dAEN), where both the ℓ_1 and the ℓ_2 penalty parameters are allowed to adapt. Hence, the approximated dAEN objective function takes the form:

$$\mathcal{Q}'_{\text{dAEN}}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \sum_{k=1}^{K}\sum_{i=0}^{p}\left(\lambda_{1,k,i}|\theta_{k,i}| + \lambda_{2,k,i}\theta_{k,i}^2\right) \tag{16}$$

As suggested by Gefang (2014), the dAEN implies that the number of penalty parameters to tune is twice the number of regression coefficients, which could be demanding too much from the data. However, the Bayesian approach presented in the next sections provides a reasonable way to address this issue. This is also confirmed by simulation results (see Section 3).

2.3. The Bayesian MIDAS penalized regression approach

Several approaches, such as the LARS (Efron et al., 2004), Group LARS (Yuan and Lin, 2006), and LARS-EN (Zou and Hastie, 2005) algorithms (modified to account for adaptive shrinkage), have been proposed in the literature to estimate AL, AGL and AEN regressions. In this paper, we consider a Bayesian hierarchical approach, which has several advantages compared to the frequentist approach. First, Bayesian methods exploit model inference via posterior distributions of parameters. For instance, they usually provide a valid measure of standard errors based on a geometrically ergodic Markov chain. Second, they provide a flexible way of estimating the penalty parameters,

along with other parameters in the model. Lastly, they provide forecasts via predictive distributions. In what follows, we present the details on the hierarchical structure of the proposed Bayesian adaptive penalized MIDAS models.

Bayesian adaptive Lasso MIDAS. As noted by Tibshirani (1996), the Lasso estimator can be interpreted as the Bayes posterior mode using normal likelihood and independent Laplace (double-exponential) prior for the regression coefficients. Accordingly, Park and Casella (2008) propose a Bayesian Lasso where the ℓ_1 penalty corresponds to a conditional Laplace prior as:

$$\pi(\boldsymbol{\theta}|\sigma^2) = \left(\frac{\lambda}{2\sqrt{\sigma^2}}\right)^{K(p+1)} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}\|\boldsymbol{\theta}\|_1\right) \tag{17}$$

which can be represented as a scale mixture of Normals with an exponential mixing density (Andrews and Mallows, 1974). In the case of the AL, the conditional prior for $\boldsymbol{\theta}$ is modified as follows:

$$\pi(\boldsymbol{\theta}|\sigma^2) = \left(\frac{1}{2\sqrt{\sigma^2}}\right)^{K(p+1)} \prod_{k=1}^{K}\prod_{i=0}^{p} \lambda_{k,i} \exp\left(-\frac{\lambda_{k,i}|\theta_{k,i}|}{\sqrt{\sigma^2}}\right) \tag{18}$$

This motivates the following hierarchical Bayesian Adaptive Lasso MIDAS (BMIDAS-AL) model with LSA objective function (13):

$$\begin{aligned}
\pi(y_{t+h}|\boldsymbol{\theta}) &\sim \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right) \\
\pi(\boldsymbol{\theta}|\boldsymbol{\tau}) &\sim \mathcal{N}(\mathbf{0}, \mathbf{D}_\tau) \\
\pi(\tau^2_{k,i}|\lambda^2_{k,i}) &\sim \text{Gamma}\left(1, \frac{\lambda^2_{k,i}}{2}\right) \qquad k=1,\dots,K, \quad i=0,\dots,p \\
\pi(\lambda^2_{k,i}) &\sim \text{Gamma}\,(r,d)
\end{aligned}$$

where $\boldsymbol{\tau} = (\boldsymbol{\tau}_1,\dots,\boldsymbol{\tau}_K)'$, $\boldsymbol{\tau}_k = (\tau^2_{k,i},\dots,\tau^2_{k,p})'$, with $\tau^2_{k,i} > 0$, $\mathbf{D}_\tau = \text{diag}(\boldsymbol{\tau})$, and $\tau^2_{k,i}$ and $\lambda^2_{k,i}$ have, respectively, exponential and gamma priors, the latter with shape parameter r and *rate* parameter d.[6] Compared to Park and Casella (2008), the LSA estimator implies that σ^2 can be dropped from the hierarchy, as the conditional distribution of the data is provided by the (approximated)

[6]We use the equivalence $\text{Exp}(d) = \text{Gamma}\,(1,d)$, where d is the rate parameter in both the Exponential and the Gamma distributions.

likelihood function (Leng et al., 2014). The full conditional posteriors are:

$$\pi(\boldsymbol{\theta}|y_{t+h}, \boldsymbol{\tau}, \boldsymbol{\lambda}) \sim \mathcal{N}\left(\mathbf{A}^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\theta}}, \mathbf{A}^{-1}\right)$$

$$\pi(\tau_{k,i}^{-2}|y_{t+h}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \sim \text{inverse-Gaussian}\left(\frac{\lambda_{k,i}}{|\theta_{k,i}|}, \lambda_{k,i}^2\right) \qquad k = 1, \ldots, K, \quad i = 0, \ldots, p$$

$$\pi(\lambda_{k,i}^2|y_{t+h}, \boldsymbol{\theta}, \boldsymbol{\tau}) \sim \text{Gamma}\left(r + 1, d + \frac{\tau_{k,i}^2}{2}\right)$$

where $\mathbf{A} = (\hat{\boldsymbol{\Sigma}}^{-1} + \mathbf{D}_{\tau}^{-1})$, $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$, and $\boldsymbol{\lambda}_k = (\lambda_{k,0}^2, \ldots, \lambda_{k,p}^2)$.

Bayesian adaptive Group Lasso MIDAS. For the Bayesian Group Lasso, Kyung et al. (2010) consider the following conditional prior of $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta}|\sigma^2) = \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}\sum_{j=1}^{G}\|\boldsymbol{\theta}_j\|_2\right) \tag{19}$$

which can be represented as a gamma mixture of normals. For the Bayesian Adaptive Group Lasso, the conditional prior for $\boldsymbol{\theta}$ becomes:

$$\pi(\boldsymbol{\theta}|\sigma^2) = \exp\left(-\frac{1}{\sqrt{\sigma^2}}\sum_{j=1}^{G}\lambda_j\|\boldsymbol{\theta}_j\|_2\right) \tag{20}$$

The hierarchical Bayesian Adaptive Group Lasso MIDAS (BMIDAS-AGL) model with LSA objective function (13) is:

$$\pi(y_{t+h}|\boldsymbol{\theta}) \sim \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right)$$

$$\pi(\boldsymbol{\theta}_j|\boldsymbol{\tau}) \sim \mathcal{N}(\mathbf{0}, \tau_j^2\mathbf{I}_{g_j}) \qquad j = 1, \ldots, G$$

$$\pi(\tau_j^2|\boldsymbol{\lambda}) \sim \text{Gamma}\left(\frac{p+2}{2}, \frac{\lambda_j^2}{2}\right)$$

$$\pi(\lambda_j^2) \sim \text{Gamma}\,(r, d)$$

where $\boldsymbol{\tau} = (\tau_1^2, \ldots, \tau_G^2)$, $\boldsymbol{\lambda} = (\lambda_1^2, \ldots, \lambda_G^2)$, and $\mathbf{I}_{g_j}$ is the identity matrix of order g_j. It is worth noting that, compared to the Lasso, the Group Lasso requires a gamma prior on τ_j^2, for $j = 1, \ldots, G$. Let $\hat{\mathbf{S}}$ be the square-root matrix of $\hat{\boldsymbol{\Sigma}}^{-1}$, with block matrices $\hat{\mathbf{S}}_j$, each of size $\tilde{g} \times g_j$. The full

conditional posteriors are:

$$\pi(\boldsymbol{\theta}_j|y_{t+h},\boldsymbol{\theta}_{-j},\boldsymbol{\tau},\boldsymbol{\lambda}) \sim \mathcal{N}\left(\mathbf{A}_j^{-1}\hat{\mathbf{S}}_j'\left(\hat{\boldsymbol{\theta}}'\hat{\mathbf{S}} - \sum_{j'\neq j}\boldsymbol{\theta}_{j'}'\hat{\mathbf{S}}_{j'}\right), \mathbf{A}_j^{-1}\right)$$

$$\pi(\tau_j^{-2}|y_{t+h},\boldsymbol{\theta},\boldsymbol{\lambda}) \sim \text{inverse-Gaussian}\left(\frac{\lambda_j}{\|\boldsymbol{\theta}_j\|_2}, \lambda_j^2\right)$$

$$\pi(\lambda_j^2|y_{t+h},\boldsymbol{\theta},\boldsymbol{\tau}) \sim \text{Gamma}\left(r + \frac{(g_j+1)}{2}, d + \frac{\tau_j^2}{2}\right)$$

where $\boldsymbol{\theta}_{-j} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{j-1},\boldsymbol{\theta}_{j+1},\ldots,\boldsymbol{\theta}_G)'$ and $\mathbf{A}_j = \hat{\mathbf{S}}_j'\hat{\mathbf{S}}_j + \tau_j^{-2}\mathbf{I}_{g_j}$.

Bayesian doubly adaptive Elastic-Net MIDAS. For the Bayesian Elastic Net, we consider the following conditional prior of $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta}|\sigma^2) = \exp\left(-\frac{\lambda_1}{\sqrt{\sigma^2}}\sum_{k=1}^{K}\sum_{i=0}^{p}|\theta_{k,i}| - \frac{\lambda_2}{2\sigma^2}\sum_{k=1}^{K}\sum_{i=0}^{p}\theta_{k,i}^2\right) \tag{21}$$

which can be written as a normal mixture of gammas (Kyung et al., 2010). For the Bayesian doubly Adaptive Elastic-Net, the conditional prior of $\boldsymbol{\theta}$ becomes:

$$\pi(\boldsymbol{\theta}|\sigma^2) = \exp\left(-\frac{1}{\sqrt{\sigma^2}}\sum_{k=1}^{K}\sum_{i=0}^{p}\lambda_{1,k,i}|\theta_{k,i}| - \frac{1}{2\sigma^2}\sum_{k=1}^{K}\sum_{i=0}^{p}\lambda_{2,k,i}\theta_{k,i}^2\right) \tag{22}$$

The hierarchical Bayesian doubly Adaptive Elastic-Net MIDAS (BMIDAS-dAEN) model with LSA objective function (13) is:

$$\pi(y_{t+h}|\boldsymbol{\theta}) \sim \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right)$$

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau},\boldsymbol{\lambda}_2) \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_{\tau,\lambda_2}^{-1})$$

$$\pi(\tau_{k,i}^2|\boldsymbol{\lambda}_1) \sim \text{Gamma}\left(1, \frac{\lambda_{1,k,i}^2}{2}\right) \qquad k=1,\ldots,K, \quad i=0,\ldots,p$$

$$\pi(\lambda_{1,k,i}^2) \sim \text{Gamma}\,(r_1, d_1)$$

$$\pi(\lambda_{2,k,i}) \sim \text{Gamma}\,(r_2, d_2)$$

where $\mathbf{D}_{\tau,\lambda_2} = \text{diag}(\boldsymbol{\tau}^{-1}+\boldsymbol{\lambda}_2)$, $\boldsymbol{\lambda}_1 = (\boldsymbol{\lambda}_{1,1},\ldots,\boldsymbol{\lambda}_{1,K})$, and $\boldsymbol{\lambda}_{1,k} = (\lambda_{1,k,0}^2,\ldots,\lambda_{1,k,p}^2)$, $\boldsymbol{\lambda}_2 = (\boldsymbol{\lambda}_{2,1},\ldots,\boldsymbol{\lambda}_{2,K})$, and $\boldsymbol{\lambda}_{2,k} = (\lambda_{2,k,0},\ldots,\lambda_{2,k,p})$, $\tau_{k,i}^2$ has an exponential prior, and $\lambda_{1,k,i}^2$ and $\lambda_{2,k,i}$ have gamma priors.

The full conditional posteriors are:

$$\pi(\boldsymbol{\theta}|y_{t+h}, \boldsymbol{\tau}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \sim \mathcal{N}\left(\mathbf{B}^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\theta}}, \mathbf{B}^{-1}\right)$$

$$\pi(\tau_{k,i}^{-2}|y_{t+h}, \boldsymbol{\theta}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \sim \text{inverse-Gaussian}\left(\frac{\lambda_{1,k,i}}{|\theta_{k,i}|}, \lambda_{1,k,i}^2\right) \qquad k = 1, \ldots, K, \quad i = 0, \ldots, p$$

$$\pi(\lambda_{1,k,i}^2|y_{t+h}, \boldsymbol{\theta}, \boldsymbol{\tau}) \sim \text{Gamma}\left(r_1 + 1, d_1 + \frac{\tau_{k,i}^2}{2}\right)$$

$$\pi(\lambda_{2,k,i}|y_{t+h}, \boldsymbol{\theta}, \boldsymbol{\tau}) \sim \text{Gamma}\left(r_2 + 0.5, d_2 + \frac{\theta_{k,i}^2}{2}\right)$$

where $\mathbf{B} = (\hat{\boldsymbol{\Sigma}}^{-1} + \mathbf{D}_{\tau,\lambda_2})$.

The expected outcome from all these hierarchical representations of the MIDAS model is that small penalty (*i.e.* small λs) will be applied to important predictors or group of predictors, while large penalty (*i.e.* large λs) will be applied to unimportant predictors or group of predictors. In this case, it follows that estimates from AL, AGL, and dAEN are model selection consistent (Zou, 2006; Wang and Leng, 2008; Zou and Zhang, 2009).

2.4. Tuning the penalty hyper-parameters

The hierarchical models presented in Section 2.3 treat the penalty parameters as hyper-parameters, *i.e.* as random variables with gamma prior distributions $\pi(\boldsymbol{\lambda})$ and gamma posterior distributions $\pi(\boldsymbol{\lambda}|y_{t+h}, \boldsymbol{\phi})$, where $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\tau})'$. However, the main drawback of this approach is that these posterior distributions can be sensitive to the choice of the prior. An alternative approach resort to an empirical Bayes estimation of the hyper-parameters, *i.e.* using the data to propose an estimate for $\boldsymbol{\lambda}$, which can be obtained through marginal maximum likelihood. However, in the present framework, the marginal distribution $\pi(y_{t+h}|\boldsymbol{\lambda}) = \int f(y_{t+h}|\boldsymbol{\phi}, \boldsymbol{\lambda})\pi(\boldsymbol{\phi}|\boldsymbol{\lambda})d\boldsymbol{\phi}$ is not available in closed form. To deal with this issue, Park and Casella (2008) and Kyung et al. (2010) suggest to implement the Monte Carlo EM algorithm (MCEM) proposed by Casella (2001), which complements the Gibbs sampler and provides marginal maximum likelihood estimates of the hyper-parameters. The idea is to treat the parameters $\boldsymbol{\phi}$ as missing data and then use an algorithm to iteratively approximate the hyper-parameters, substituting Monte Carlo estimates for any expected values that cannot be computed explicitly. The MCEM algorithm of Casella (2001) uses N Monte Carlo iterations to maximize the marginal log-likelihood with a Q function defined as:

$$Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(n)}) = \int \log\left[f(y_{t+h}|\boldsymbol{\phi}, \boldsymbol{\lambda})\pi(\boldsymbol{\phi}|\boldsymbol{\lambda})\right] \pi(\boldsymbol{\phi}|y_{t+h}, \boldsymbol{\lambda}^{(n)})d\boldsymbol{\phi}$$

Specifically, each $n = 1, \ldots, N$ Monte Carlo iteration involves two steps. First, the E-step is solved by calculating $Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(n)})$ for a given $\boldsymbol{\lambda}^{(n)}$ (an initial value $\boldsymbol{\lambda}^{(0)}$ is used to initialize the Monte Carlo).

Then, the M-step is solved by maximizing $Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(n)})$ to give $\boldsymbol{\lambda}^{(n+1)}$. For AL, AGL, and dAEN simple analytic solutions can be used to approximate $\boldsymbol{\lambda}^{(n+1)}$ (see Park and Casella, 2008, and Kyung et al., 2010). However, since $\pi(\boldsymbol{\phi}|y_{t+h}, \boldsymbol{\lambda})$ is intractable, the algorithm requires a simulation method to approximate the quantities of interest. A run of the Gibbs sampler can then be used for this purpose.

From a computational point of view, it is straightforward that the MCEM algorithm may be extremely expensive. Indeed, each nth Monte Carlo iteration requires a fully converged Gibbs sampling from $\pi(\boldsymbol{\phi}|y_{t+h}, \boldsymbol{\lambda}^{(n)})$. Hence, a serious trade-off between accuracy of the results (S Gibbs iterations) and computational efficiency (N Monte Carlo iterations) may arise. To deal with this issue, in this work we rely on a specific class of the so-called internal adaptive MCMC algorithms, denoted controlled MCMC algorithm (see Atchadé et al., 2011). This class makes use of stochastic approximation algorithms to solve maximization problems when the likelihood function is intractable, by mimicking standard iterative methods such as the gradient algorithm. This approach is therefore computationally efficient, because it only requires a single Monte Carlo run ($N = 1$). Following Atchadé (2011), let us write the derivative of the $Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(s)})$ function with respect to $\boldsymbol{\lambda}$ as:

$$\nabla_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(s)}) = \int H(\boldsymbol{\lambda}, \boldsymbol{\phi}) \pi(\boldsymbol{\phi}|y_{t+h}, \boldsymbol{\lambda}^{(s)}) d\boldsymbol{\phi}$$

where $H(\boldsymbol{\lambda}, \boldsymbol{\phi}) = \nabla_{\boldsymbol{\lambda}} \log [f(y_{t+h}|\boldsymbol{\phi}, \boldsymbol{\lambda}) \pi(\boldsymbol{\phi}|\boldsymbol{\lambda})] = \nabla_{\boldsymbol{\lambda}} \log \pi(\boldsymbol{\phi}|\boldsymbol{\lambda})$, as the likelihood $f(y_{t+h}|\boldsymbol{\phi}, \boldsymbol{\lambda})$ does not usually depend on the hyper-parameters $\boldsymbol{\lambda}$. Note that we changed the superscript from (n) Monte Carlo iteration to (s) Gibbs sampler iteration to avoid confusion. Using a stochastic approximation to solve the maximization problem, *i.e.* replacing the maximization of the Q function by one step of the gradient algorithm, the solution to the EM algorithm takes the form:

$$\boldsymbol{\lambda}^{(s+1)} = \boldsymbol{\lambda}^{(s)} + a^{(s)} Q'(\boldsymbol{\lambda}^{(s)}|\boldsymbol{\lambda}^{(s)}) = \boldsymbol{\lambda}^{(s)} + a^{(s)} \int H(\boldsymbol{\lambda}^{(s)}, \boldsymbol{\phi}) \pi(\boldsymbol{\phi}|y_{t+h}, \boldsymbol{\lambda}^{(s)}) d\boldsymbol{\phi}$$

where $a^{(s)}$ is a step-size taking a Robbins-Monro form $a^{(s)} = 1/s^q$, with $q \in (0.5, 1)$ (Lange, 1995). If the integral $\int H(\boldsymbol{\lambda}^{(s)}, \boldsymbol{\phi}) \pi(\boldsymbol{\phi}|y_{t+h}, \boldsymbol{\lambda}^{(s)}) d\boldsymbol{\phi}$ is approximated by $H(\boldsymbol{\lambda}^{(s)}, \boldsymbol{\phi}^{(s+1)})$, we get an approximate EM algorithm, where both the E- and the M-steps are approximately implemented. Hence, marginal maximum likelihood estimates of the hyper-parameters, $\hat{\boldsymbol{\lambda}}$, and draws from the posterior distribution of the parameters, $\pi(\boldsymbol{\phi}|y_{t+h}, \hat{\boldsymbol{\lambda}})$, are both obtained using a single run of the Gibbs sampler, with $s = 1, \ldots, S$. In the present framework, making the transformation $\boldsymbol{\omega} = \frac{1}{2} \log(\boldsymbol{\lambda})$, we can build upon the analytic solutions for the EM algorithm reported in Kyung et al.

(2010) and show that:

$$H_{\text{AL}}(\boldsymbol{\omega}, \boldsymbol{\phi}) = 2 - \exp(2\boldsymbol{\omega}) \odot \boldsymbol{\tau}$$

$$H_{\text{AGL}}(\boldsymbol{\omega}, \boldsymbol{\phi}) = (\mathbf{g} + 1) - \exp(2\boldsymbol{\omega}) \odot \boldsymbol{\tau}$$

$$H_{\text{dAEN}}(\boldsymbol{\omega}_1, \boldsymbol{\phi}) = 2 - \exp(2\boldsymbol{\omega}_1) \odot \boldsymbol{\tau}$$

$$H_{\text{dAEN}}(\boldsymbol{\omega}_2, \boldsymbol{\phi}) = 1 - \exp(2\boldsymbol{\omega}_2) \odot [\text{diag}(\boldsymbol{\theta})\boldsymbol{\theta}],$$

where $\mathbf{g} = (g_1, \ldots, g_G)'$ and $\odot$ is the element-wise product. Hence, the updating rules for AL, AGL, and dAEN take the following forms, respectively:

$$\omega_{k,i}^{(s+1)} = \omega_{k,i}^{(s)} + a^{(s)} \left[2 - \exp\left(2\omega_{k,i}^{(s)}\right) \tau_{k,i}^{2,(s+1)} \right]$$

$$\omega_{j}^{(s+1)} = \omega_{j}^{(s)} + a^{(s)} \left[(g_j + 1) - \exp\left(2\omega_{j}^{(s)}\right) \tau_{j}^{2,(s+1)} \right]$$

$$\omega_{1,k,i}^{(s+1)} = \omega_{1,k,i}^{(s)} + a^{(s)} \left[2 - \exp\left(2\omega_{1,k,i}^{(s)}\right) \tau_{k,i}^{2,(s+1)} \right]$$

$$\omega_{2,k,i}^{(s+1)} = \omega_{2,k,i}^{(s)} + a^{(s)} \left[1 - \exp\left(2\omega_{2,k,i}^{(s)}\right) \theta_{k,i}^{2,(s+1)} \right]$$

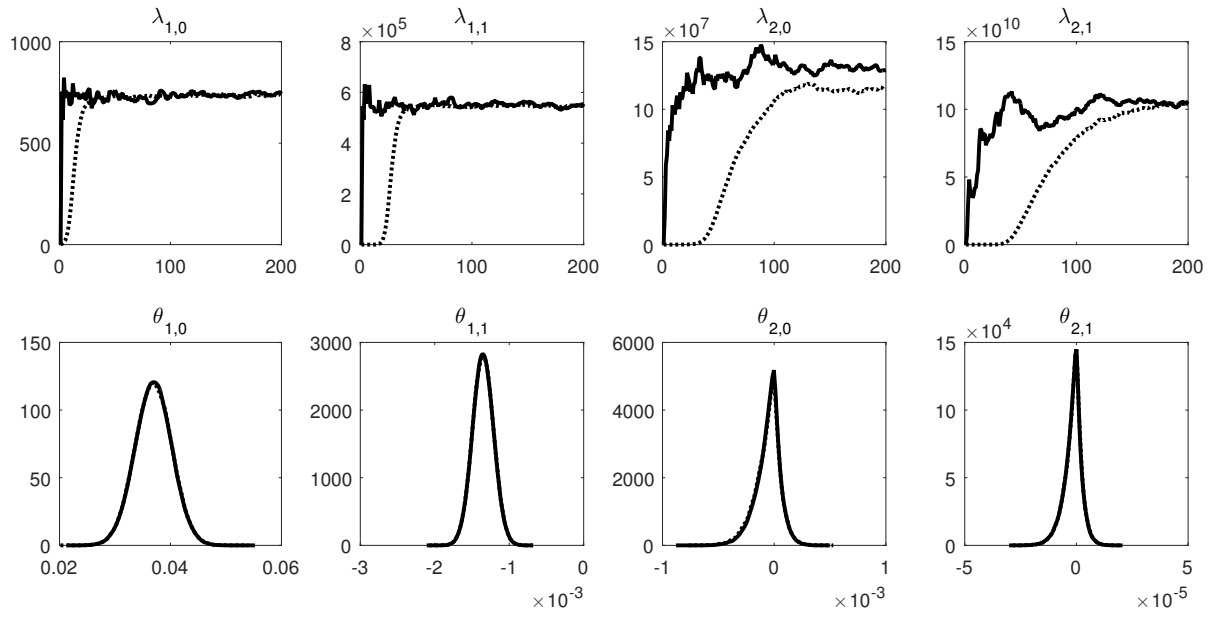
from which we get $\boldsymbol{\lambda}^{(s+1)} = \exp(2\boldsymbol{\omega}^{(s+1)})$. The algorithm can be completed by allowing for a stabilization procedure (*e.g.* truncation on random boundaries; Andrieu et al., 2005, Atchadé, 2011) ensuring the convergence of $\boldsymbol{\lambda}$ and the posterior distribution of $\boldsymbol{\phi}$ towards $\hat{\boldsymbol{\lambda}}$ and $\pi(\boldsymbol{\phi}|y_{t+h}, \hat{\boldsymbol{\lambda}})$, respectively.

We illustrate the computational advantage of the proposed methodology using simulated data. For ease of exposition, the DGP follows a simple trivariate mixed-frequency model:

$$y_t = \beta_0 + \sum_{c=0}^{C-1} \beta_1 B(c; \boldsymbol{\vartheta}) L^{c/m} x_{1,t}^{(m)} + \sum_{c=0}^{C-1} \beta_2 B(c; \boldsymbol{\vartheta}) L^{c/m} x_{2,t}^{(m)} + \epsilon_t,$$

where the regressors and the error term are iid draws from a standard normal distribution of length $T = 500$. We set the true values $(\beta_0, \beta_1, \beta_2) = (1, 2, 0)$, $C = 12$, $m = 3$, and a decreasing weighting function $B(c; \boldsymbol{\vartheta})$, such that about 90% of the weight is concentrated in the last (*i.e.* most recent) three high-frequency observations. We estimate the Bayesian Adaptive Lasso model presented in Section 2.3 using $p = 1$, such that $\boldsymbol{\theta} = (\theta_{1,0}, \theta_{1,1}, \theta_{2,0}, \theta_{2,1})'$, and we tune the penalty hyper-parameters $\boldsymbol{\lambda} = (\lambda_0, \lambda_{1,0}, \lambda_{1,1}, \lambda_{2,0}, \lambda_{2,1})'$ using either the stochastic approximation approach or the MCEM algorithm. The former updates $\boldsymbol{\lambda}$ in a single run of the Gibbs sampler by drawing $S = 200,000$ samples, while the latter uses $N = 200$ Monte Carlo runs, each one drawing $S = 50,000$ samples. The analysis is carried out using MATLAB R2015a on a workstation with a 3.40GHz Intel

Figure 1: Tuning the penalty hyper-parameters: stochastic approximation vs MCEM



Note: evolution of the penalty hyper-parameters $\boldsymbol{\lambda}$ across simulations (first panel) and posterior distributions of $\boldsymbol{\theta}$ (second panel). Solid line is the stochastic approximation approach, dotted line is the MCEM algorithm.

Core i7-4770 CPU. The evolution of $\boldsymbol{\lambda}$ across simulations is reported in the first panel of Figure 1. Each point in the plots represents the 1000th update of $\boldsymbol{\lambda}$ provided by the stochastic approximation approach (solid line) and the nth update provided by the MCEM algorithm (dotted line). Both the approaches lead to similar values of the penalty hyper-parameters, although we note that the MCEM requires a larger number of runs in order to converge. Further, the posterior densities of $\boldsymbol{\theta}$ (second panel of Figure 1), calculated dropping the first 30,000 draws for the stochastic approximation approach and using the last Monte Carlo run for the MCEM approach, are almost identical and correctly displaying largest mass at zero for $\theta_{2,0}$ and $\theta_{2,1}$. However, the computational efficiency gain of the former appears substantial: for this simple simulation experiment, the computational time was around 22 seconds using stochastic approximations, against 648 seconds required by the MCEM algorithm.

2.5. Variable selection

The penalized regression approach had originally been developed and proposed as a variable selection method. Indeed, the ℓ_1 and ℓ_2 penalty terms in Equations (14)-(16) are intended to shrink the coefficients of irrelevant predictors to zero, leading to a sparse solution. However, this attractive property vanishes in the Bayesian framework described in Section 2.3, as the Bayesian solution provides a shrinkage of the coefficients towards zero but usually not exactly to zero. Different

approaches have been proposed in the literature to achieve (*e.g.* spike-and-slab priors, Ročková and George, in press) or evaluate (*e.g.* scaled neighbourhood criterion; Li and Lin, 2010) variable selection for the models under analysis. Here we rely on the simple credible interval criterion suggested by Kyung et al. (2010). According to this criterion, a predictor k, for $k = 1, \ldots, K$, is excluded from the estimated active set if the credible interval, at say 90% level, of the posterior distribution of the slope coefficient $\beta_k = \sum_{c=0}^{C-1} \sum_{i=0}^{p} \theta_{k,i} c^i$ covers zero.

2.6. Forecasting

Forecasts are obtained from the following posterior predictive density for y_{T+h}:

$$p(y_{T+h}|\mathcal{D}) = \int p(y_{T+h}|\boldsymbol{\phi}, \boldsymbol{\lambda}, \mathcal{D}) p(\boldsymbol{\phi}, \boldsymbol{\lambda}|\mathcal{D}) d\boldsymbol{\phi} d\boldsymbol{\lambda} \tag{23}$$

where $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\tau})'$ and $p(\boldsymbol{\phi}, \boldsymbol{\lambda}|\mathcal{D})$ denotes the joint posterior distribution of the BMIDAS-AL, BMIDAS-AGL, and BMIDAS-AEN parameters conditional on past available information, $\mathcal{D}$. According to the framework described in Sections 2.3 and 2.4, draws $y_{T+h}^{(s)}$, for $s = 1, \ldots, S$, from the predictive distribution can be obtained from the Gibbs sampler.[7] This leads to a distribution of predictions that can be used for out-of-sample evaluation of the model. For instance, point forecasts can be computed by averaging over these draws, *i.e.* $\hat{y}_{T+h} = S^{-1} \sum_{s=1}^{S} y_{T+h}^{(s)}$, and evaluated using standard accuracy criteria. However, since draws from the predictive density are available, an evaluation of the entire predictive distribution can be performed using, for instance, scoring rules (Mitchell and Wallis, 2011).

3. Monte Carlo experiments

3.1. Design of the experiments

We evaluate the performance of the proposed models through Monte Carlo experiments. For this purpose, we use the following DGP involving a quarterly regressand, $K = \{10, 30, 50\}$ monthly predictors and $T = 200$ in-sample observations:

$$y_{t+h} = \alpha + \sum_{k=1}^{K} \sum_{c=0}^{C-1} \beta_k B(c; \boldsymbol{\vartheta}) L^{c/3} x_{k,t}^{(3)} + \epsilon_t$$

$$x_{k,t}^{(3)} = \mu + \rho x_{k,t-1/3}^{(3)} + \varepsilon_{k,t}$$

$$B(c; \boldsymbol{\vartheta}) = \frac{\exp(\vartheta_1 c + \vartheta_2 c^2)}{\sum_{c=0}^{C-1} \exp(\vartheta_1 c + \vartheta_2 c^2)}$$

[7] It is worth noting that we do not condition on a fixed value $\hat{\boldsymbol{\lambda}}$, such as the maximum likelihood estimate that can be obtained, for instance, by averaging over the Gibbs samples of $\boldsymbol{\lambda}$, because this would ignore the uncertainty around the estimate of the penalty parameters.

where $B(c;\boldsymbol{\vartheta})$ is parameterized as an exponential Almon lag function (Ghysels et al., 2007), with $C = 12$ and $\boldsymbol{\vartheta} = (0.10, -0.15)$, the latter implying rapidly declining weights within the last year. The lag length (C) and the parameters of the weighting function ($\boldsymbol{\vartheta}$) are chosen to be consistent with the empirical results reported, for instance, by Clements and Galvão (2008, 2009). Note that the same weighting structure applies to all the predictors entering the active set. Further, for ease of analysis we assume $h = 0$, such that the forecasting model takes the form of a nowcasting model with fully available information on predictors in the current quarter. In this specification, ϵ_t and $\boldsymbol{\varepsilon}_t$ are i.i.d. with distribution:

$$\begin{pmatrix} \epsilon_t \\ \boldsymbol{\varepsilon}_t \end{pmatrix} \sim \text{i.i.d.}\mathcal{N}\left[\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\varepsilon \end{pmatrix}\right],$$

where $\boldsymbol{\Sigma}_\varepsilon$ has elements $\sigma_\varepsilon^{|k-k'|}$, such that the diagonal elements are equal to one and the off-diagonal elements control for the correlation between $x_{k,t}^{(3)}$ and $x_{k',t}^{(3)}$, with $k \neq k'$. We set $\sigma_\varepsilon = \{0.5, 0.95\}$. As for the parameters in the DGP, we choose $\alpha = \mu = \rho = 0.5$ and $\boldsymbol{\beta} = (0.1, 0.3, 0, 0, 0.1, 0.3, 0, 0, 0.5, \mathbf{0})'$. The latter implies that only five out of K predictors are relevant. Conditional on these parameters, we set σ such that the expected noise-to-signal ratio of the mixed-frequency regression is approximately equal to 0.25.

We estimate mixed-frequency models on the data provided by the DGP above using the BMIDAS-AL, the BMIDAS-AGL, and the BMIDAS-dAEN regression approaches described in Section 2.3. As for the functional form of the weighting structure, we consider a restricted Almon lag polynomial as in (6), with $p = 3$ and endpoint restrictions (both tail and derivative; see Footnote 1). The hyper-parameters $\boldsymbol{\lambda}$ are estimated using the stochastic approximation approach described in Section 2.4, with step-size $a^{(s)} = 1/s^{0.8}$ ($q = 0.6$ for the BMIDAS-dAEN, as preliminary results suggest that a slower sequence $a^{(s)}$ is required to achieve convergence). We set the number of Monte Carlo replications at $R = 200$. The Gibbs sampler is run for $S = 60,000$ iterations, with a burn-in period of $20,000$ iterations. According to the Raftery and Lewis (1992) diagnostic, no thinning of the chains is required. Further, convergence is evaluated through the following diagnostics: i) integrated autocorrelation time, estimated using the Sokal's adaptive truncated periodogram estimator; ii) total number of draws required to achieve a given accuracy (Raftery and Lewis, 1992), with quantile = 0.025, accuracy = 0.0125, and probability = 0.95; iii) test for equality of the means of the first 10% and last 50% draws (Geweke, 1992); iv) test for stationarity of the chains (Heidelberger and Welch, 1983). Results (not reported, but available upon request from the authors) indicate that the convergence requirements are overall satisfied across the whole set of Monte Carlo experiments.

Forecasts (with $h = 0$) are computed as described in Section 2.6. Point forecasts are evaluated through the average mean squared forecast error (aMSFE) over the R Monte Carlo replications. Density forecasts (generated by the draws from the posterior predictive distribution) are evaluated by the means of the (negative) average log-score ($-$aLS), *i.e.* the average of the log of the predictive

Table 1: Monte Carlo simulations 1: estimator and selection features

K	σ_ε	aMSE	$\overline{\text{aMSE}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})$	aVAR	$\overline{\text{aVAR}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})$	aBIAS^2	$\overline{\text{aBIAS}^2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})$	$P(\hat{\mathcal{A}} = \mathcal{A})$	$P(\hat{\mathcal{A}}^c = \mathcal{A}^c)$	$P(\hat{\boldsymbol{\beta}}_{\mathcal{A}} = \boldsymbol{\beta}^*_{\mathcal{A}})$
		BMIDAS-AL								
10	0.50	2.59E-03	0.81	1.12E-03	0.81	1.47E-03	0.80	0.89	0.96	0.84
	0.95	4.70E-02	0.62	1.66E-02	0.59	3.04E-02	0.63	0.44	0.92	0.64
30	0.50	1.47E-03	0.51	0.62E-03	0.51	0.85E-03	0.50	0.89	0.95	0.84
	0.95	2.47E-02	0.43	0.89E-02	0.39	1.57E-02	0.45	0.43	0.95	0.62
50	0.50	1.23E-03	0.39	0.52E-03	0.38	0.52E-03	0.39	0.89	0.96	0.83
	0.95	2.19E-02	0.32	0.74E-02	0.29	1.45E-02	0.33	0.39	0.95	0.61
		BMIDAS-AGL								
10	0.50	2.55E-03	0.82	1.08E-03	0.84	1.47E-03	0.81	0.89	0.96	0.84
	0.95	4.59E-02	0.62	1.55E-02	0.61	3.04E-02	0.63	0.44	0.92	0.61
30	0.50	1.39E-03	0.54	0.55E-03	0.56	0.84E-03	0.52	0.89	0.95	0.84
	0.95	2.37E-02	0.46	0.79E-02	0.43	1.58E-02	0.48	0.42	0.95	0.59
50	0.50	1.13E-03	0.42	0.45E-03	0.43	0.69E-03	0.41	0.88	0.96	0.82
	0.95	2.04E-02	0.34	0.68E-02	0.31	1.35E-02	0.35	0.39	0.96	0.59
		BMIDAS-dAEN								
10	0.50	2.68E-03	0.98	0.68E-03	0.99	2.00E-03	0.98	0.82	0.99	0.66
	0.95	4.29E-02	0.67	0.54E-02	0.66	3.75E-02	0.67	0.43	0.90	0.30
30	0.50	0.90E-03	0.94	0.23E-03	0.96	0.67E-03	0.93	0.83	0.99	0.68
	0.95	1.46E-02	0.71	0.17E-02	0.67	1.29E-02	0.71	0.43	0.98	0.28
50	0.50	0.60E-03	0.92	0.14E-03	0.95	0.46E-03	0.91	0.81	0.99	0.65
	0.95	1.06E-02	0.64	0.11E-02	0.64	0.95E-02	0.64	0.42	0.98	0.26

likelihood evaluated at the out-turn of the forecast, and the average continuously ranked probability score (aCRPS), which measures the average distance between the empirical CDF of the out-of-sample observations and the empirical CDF associated with the predictive density of each model (Gneiting and Raftery, 2007).

3.2. Results

Simulation results are reported in Tables 1 and 2. In Table 1, we present the average mean squared error (aMSE), average variance (aVAR), and average squared bias (aBIAS^2) over the full set of K estimated parameters $\hat{\boldsymbol{\beta}}$ in the model, where $\hat{\beta}_k = \sum_{c=0}^{C-1} \sum_{i=0}^{p} \hat{\theta}_{k,i} c^i$ for $k = 1 \dots, K$, and over the R Monte Carlo replications. Further, we report the share of aMSE, aVAR, and aBIAS^2, denoted respectively $\overline{\text{aMSE}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})$, $\overline{\text{aVAR}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})$, and $\overline{\text{aBIAS}^2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})$, that can be attributed to the estimated parameters in the true active set ($\hat{\boldsymbol{\beta}}_{\mathcal{A}}$). Finally, we evaluate the sparsity features of our models by computing the average coverage rate of the active and inactive set, respectively $P(\hat{\mathcal{A}} = \mathcal{A})$ and $P(\hat{\mathcal{A}}^c = \mathcal{A}^c)$, as well as the average variable selection performance in the active set, $P(\hat{\boldsymbol{\beta}}_{\mathcal{A}} = \boldsymbol{\beta}^*_{\mathcal{A}})$ (see Section 2.5).

Results point to a number of interesting features. First, BMIDAS-AL and BMIDAS-AGL models provide very similar results in terms of mean squared error (including variance and bias), although the latter shows slightly better estimation performance. The BMIDAS-dAEN model shows somewhat better results compared to the other models, that can be mainly attributed to a substantially

Table 2: Monte Carlo simulations 1: predictive features

		BMIDAS-AL			BMIDAS-AGL			BMIDAS-dAEN		
K	σ_ε	aMSFE	−aLS	aCRPS	aMSFE	−aLS	aCRPS	aMSFE	−aLS	aCRPS
10	0.50	0.11	0.34	0.19	0.11	0.34	0.19	0.11	0.34	0.19
	0.95	0.30	0.83	0.31	0.31	0.83	0.31	0.30	0.82	0.31
30	0.50	0.15	0.48	0.22	0.15	0.48	0.22	0.15	0.47	0.22
	0.95	0.29	0.80	0.31	0.30	0.82	0.31	0.29	0.80	0.31
50	0.50	0.13	0.41	0.21	0.13	0.40	0.21	0.14	0.42	0.21
	0.95	0.29	0.79	0.30	0.28	0.79	0.30	0.29	0.80	0.30

lower average variance. However, we note that the share of variance attributed to the estimated parameters in the true active set is systematically higher compared to the other models, suggesting that most of the overperformance must be attributed to the ability of the BMIDAS-dAEN to correctly estimate parameters in the true inactive set. Second, average mean squared error, variance, and bias increase with both the number of variables (K) and the correlation between high-frequency variables, determined by the value of σ_ε. Third, the coverage rates are overall very high, around 80% for the of the active set and close to 100% for the inactive set when the correlation between high-frequency predictors is moderate. This suggests that the our models can select the correct sparsity pattern with a high probability even in finite samples. Further, the variable selection performance, measured as the probability of correctly estimating within the 90% credible interval the true parameters in the true active set, is also very high (around 80%).

In Table 2, we evaluate the predictive performance of our models. Results for both point and density forecasts are quite similar across models, pointing to a very good forecasting accuracy, which seems to be only marginally affected by the number of (inactive) predictors included in the DGP. However, and similarly to the in-sample results, we note a deterioration of the predictive accuracy when the correlation of high-frequency variables is very high, with average mean squared error and log score almost doubling compared to the case where correlation is moderate. This outcome is not surprising, as simulation results reported in Table 1 revealed that the models are less accurate in selecting the the true active set and correctly estimating the true coefficients in presence of very strong correlation.

Finally, we introduce more model uncertainty in our simulations by setting up a sparse recovery problem, in which the position of the active variables is more *distributed*, rather than concentrated, in the dataset. This could be a relevant, as well as more realistic, issue when the active variables are strongly correlated with adjacent inactive ones, as in our simulation experiments. To this aim, we set $K = \{30, 50\}$ and $\beta_j = \{0.1, 0.3, 0.1, 0.3, 0.5\}$, where $j = 5(i+1)$ for $K = 30$ and $j = 5(2i+1)$ for $K = 50$, with $i = 0, 1, 2, 3, 4$. Results on estimation and predictive accuracy, reported in Tables 3 and 4, are broadly in line with those reported in Tables 1 and 2. Interestingly, there are no clear signs of decrease in the models' ability to select the correct subset of variables, suggesting

Table 3: Monte Carlo simulations 2: estimator and selection features

K	σ_ε	aMSE	$\overline{\text{aMSE}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})$	aVAR	$\overline{\text{aVAR}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})$	aBIAS^2	$\overline{\text{aBIAS}^2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})$	$P(\hat{\mathcal{A}}=\mathcal{A})$	$P(\hat{\mathcal{A}}^c=\mathcal{A}^c)$	$P(\hat{\boldsymbol{\beta}}_{\mathcal{A}}=\boldsymbol{\beta}^*_{\mathcal{A}})$
		BMIDAS-AL								
30	0.50	1.25E-03	0.45	0.52E-03	0.46	0.74E-03	0.44	0.93	0.95	0.84
	0.95	2.64E-02	0.31	0.87E-02	0.29	1.77E-02	0.32	0.40	0.93	0.54
50	0.50	0.99E-03	0.33	0.40E-03	0.35	0.59E-03	0.32	0.94	0.95	0.84
	0.95	1.61E-02	0.24	0.57E-02	0.23	1.04E-02	0.25	0.47	0.94	0.61
		BMIDAS-AGL								
30	0.50	1.18E-03	0.48	0.46E-03	0.51	0.72E-03	0.46	0.93	0.95	0.84
	0.95	2.35E-02	0.33	0.76E-02	0.31	1.59E-02	0.35	0.39	0.93	0.53
50	0.50	0.93E-03	0.36	0.35E-03	0.39	0.57E-03	0.34	0.93	0.95	0.82
	0.95	1.53E-02	0.25	0.54E-02	0.23	0.99E-02	0.27	0.45	0.95	0.58
		BMIDAS-dAEN								
30	0.50	0.69E-03	0.85	0.18E-03	0.92	0.51E-03	0.82	0.89	0.98	0.75
	0.95	1.49E-02	0.49	0.15E-02	0.45	1.34E-02	0.49	0.39	0.94	0.27
50	0.50	0.39E-03	0.81	0.11E-03	0.91	0.29E-03	0.77	0.91	0.99	0.78
	0.95	0.72E-02	0.48	0.07E-02	0.43	0.65E-02	0.48	0.45	0.96	0.32

Table 4: Monte Carlo simulations 2: predictive features

		BMIDAS-AL			BMIDAS-AGL			BMIDAS-dAEN		
K	σ_ε	aMSFE	$-$aLS	aCRPS	aMSFE	$-$aLS	aCRPS	aMSFE	$-$aLS	aCRPS
30	0.50	0.13	0.41	0.20	0.12	0.40	0.20	0.13	0.40	0.20
	0.95	0.23	0.69	0.27	0.23	0.69	0.27	0.25	0.72	0.28
50	0.50	0.09	0.22	0.17	0.09	0.23	0.17	0.09	0.24	0.18
	0.95	0.21	0.65	0.26	0.21	0.64	0.26	0.22	0.66	0.26

that the introduction of additional model uncertainty does not particularly alter the in-sample and out-of-sample performance of our proposed regression approaches.

4. Empirical application

...

5. Concluding remarks

We proposed a new approach to modeling and forecasting mixed-frequency regressions (MIDAS) that address the issues of estimation and variable selection in presence of a large number of predictors. Our approach is based on a combination of MIDAS regressions based on Almon lag polynomials and adaptive penalized regression models (Lasso, Group Lasso, and Elastic-Net). The proposed models rely on Bayesian techniques for estimation. In particular, the penalty hyper-parameters driving the model shrinkage are automatically tuned via a controlled MCMC algorithm

based on stochastic approximations (Atchadé et al., 2011), which is computationally efficient compared to the standard MCEM algorithm proposed by Casella (2001). Simulations show that the proposed models present very good in-sample and out-of-sample performance. When applied to US GDP, the results suggest that our models produce significant out-of-sample predictive gains compared to several alternative models. Future research could follow recent literature and extend our models to MIDAS stochastic volatility dynamics as in Pettenuzzo et al. (2016), as well as time-variation in the parameters characterizing the lag polynomials as suggested by Carriero et al. (2015) and Schumacher (2015).

References

Ahn, S., Horenstein, A. R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81 (3), 1203–1227.

Andreou, E., Ghysels, E., Kourtellos, A., 2013. Should macroeconomic forecasters use daily financial data and how? Journal of Business & Economic Statistics 31 (2), 240–251.

Andrews, D. F., Mallows, C. L., 1974. Scale mixtures of normal distributions. Journal of the Royal Statistical Society: Series B (Methodological) 36 (1), 99–102.

Andrieu, C., Moulines, E., Priouret, P., 2005. Stability of stochastic approximation under verifiable conditions. SIAM Journal on Control and Optimization 44 (1), 283–312.

Atchadé, Y. F., 2011. A computational framework for empirical Bayes inference. Statistics and Computing 21 (4), 463–473.

Atchadé, Y. F., Fort, G., Moulines, E., Priouret, P., 2011. Adaptive markov chain monte carlo: Theory and methods. In: Barber, D., Cemgil, A. T., Chiappa, S. (Eds.), Bayesian Time Series Models. Cambridge (UK): Cambridge University Press, pp. 32–51.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70 (1), 191–221.

Bai, J., Ng, S., 2007. Determining the number of primitive shocks in factor models. Journal of Business & Economic Statistics 25 (1), 52–60.

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

Bec, F., Mogliani, M., 2015. Nowcasting French GDP in real-time with surveys and "blocked" regressions: Combining forecasts or pooling information? International Journal of Forecasting 31 (4), 1021–1042.

Bessec, M., 2013. Short-term forecasts of French GDP: A dynamic factor model with targeted predictors. Journal of Forecasting 32 (6), 500–511.

Boivin, J., Ng, S., 2006. Are more data always better for factor analysis? Journal of Econometrics 132 (1), 169–194.

Bulligan, G., Marcellino, M., F., V., 2015. Forecasting economic activity with targeted predictors. International Journal of Forecasting 31 (1), 188–206.

Callot, L. A. F., Kock, A. B., 2014. Oracle efficient estimation and forecasting with the adaptive lasso and the adaptive group lasso in vector autoregressions. In: Haldrup, N., Meitz, M., Saikkonen, P. (Eds.), Essays in Nonlinear Time Series Econometrics. Oxford (UK): Oxford University Press.

Carriero, A., Clark, T. E., Marcellino, M., 2015. Real-time nowcasting with a Bayesian mixed frequency model with stochastic volatility. Journal of the Royal Statistical Society: Series A (Statistics in Society) 178 (4), 837–862.

Casella, G., 2001. Empirical Bayes Gibbs sampling. Biostatistics 2 (4), 485–500.

Castle, J. L., Fawcett, N. W. P., Hendry, D. F., 2009. Nowcasting is not just contemporaneous forecasting. National Institute Economic Review 210 (1), 71–89.

Castle, J. L., Hendry, D. F., 2010. Nowcasting from disaggregates in the face of location shifts. Journal of Forecasting 29 (1-2), 200–214.

Clements, M. P., Galvão, A. B., 2008. Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States. Journal of Business & Economic Statistics 26 (4), 546–554.

Clements, M. P., Galvão, A. B., 2009. Forecasting US output growth using leading indicators: An appraisal using MIDAS models. Journal of Applied Econometrics 24 (7), 1187–1206.

Cooper, J. P., 1972. Two approaches to polynomial distributed lags estimation: An expository note and comment. The American Statistician 26 (3), 32–35.

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.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. The Annals of Statistics 32 (2), 407–451.

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

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 (Statistics in Society) 178 (1), 57–82.

Gefang, D., 2014. Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage. International Journal of Forecasting 30 (1), 1–11.

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M. (Eds.), Bayesian Statistics. Vol. 4. Oxford (UK): Clarendon Press, pp. 169–193.

Ghosh, S., 2011. On the grouped selection and model complexity of the adaptive elastic net. Statistics and Computing 21 (3), 451–462.

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.

Ghysels, E., Sinko, A., Valkanov, R., 2007. MIDAS regressions: Further results and new directions. Econometric Reviews 26 (1), 53–90.

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

Girardi, A., Golinelli, R., Pappalardo, C., 2017. The role of indicator selection in nowcasting Euro Area GDP in pseudo real time. Empirical Economics 53 (1), 79–99.

Gneiting, T., Raftery, A., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Heidelberger, P., Welch, P. D., 1983. Simulation run length control in the presence of an initial transient. Operations Research 31 (6), 1109–1144.

Kyung, M., Gill, J., Ghosh, M., Casella, G., 2010. Penalized regression, standard errors, and bayesian lassos. Bayesian Analysis 5 (2), 369–412.

Lange, K., 1995. A gradient algorithm locally equivalent to the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 57 (2), 425–437.

Leng, C., Tran, M. N., Nott, D., 2014. Bayesian adaptive lasso. Annals of the Institute of Statistical Mathematics 66 (2), 221–244.

Li, Q., Lin, N., 2010. The bayesian elastic net. Bayesian Analysis 5 (1), 151–170.

Marcellino, M., Schumacher, C., 2010. Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP. Oxford Bulletin of Economics and Statistics 72 (4), 518–550.

Mitchell, J., Wallis, K. F., 2011. Evaluating density forecasts: Forecast combinations,model mixtures, calibration and sharpness. Journal of Applied Econometrics 26 (6), 1023–1040.

Mogliani, M., Darné, O., Pluyaud, B., 2017. The new MIBA model: Real-time nowcasting of French GDP using the Banque de France's monthly business survey. Economic Modelling 64, 26–39.

Park, T., Casella, G., 2008. The bayesian lasso. Journal of the American Statistical Association 103 (482), 681–686.

Pettenuzzo, D., Timmermann, A., Valkanov, R., 2016. A MIDAS approach to modeling first and second moment dynamics. Journal of Econometrics 193 (2), 315–334.

Raftery, A., Lewis, S., 1992. How many iterations in the Gibbs sampler? In: Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M. (Eds.), Bayesian Statistics. Vol. 4. Oxford (UK): Clarendon Press, pp. 763–773.

Ročková, V., George, E. I., in press. The spike-and-slab LASSO. Journal of the American Statistical Association.

Schumacher, C., 2015. MIDAS regressions with time-varying parameters. Mimeo.

Siliverstovs, B., 2017. Short-term forecasting with mixed-frequency data: a MIDASSO approach. Applied Economics 49 (13), 1326–1343.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58 (1), 267–288.

Wang, H., Leng, C., 2007. Unified LASSO estimation by least squares approximation. Journal of the American Statistical Association 102 (479), 1039–1048.

Wang, H., Leng, C., 2008. A note on adaptive group lasso. Computational Statistics & Data Analysis 52 (12), 5277–5286.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1), 49–67.

Yuan, M., Lin, Y., 2007. On the non-negative garrotte estimator. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (2), 143–161.

Zhao, P., Yu, B., 2006. On model selection consistency of lasso. Journal of Machine Learning Research 7, 2541–2563.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429.

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

Zou, H., Zhang, H. H., 2009. On the adaptive elastic-net with a diverging number of parameters. The Annals of Statistics 37 (4), 1733ñ1751.