WORKING
paper

BANQUE DE FRANCE
EUROSYSTÈME

# Bayesian MIDAS penalized regressions: estimation, selection, and prediction

Matteo Mogliani[1]

March 2019, WP #713

## ABSTRACT

We propose a new approach to mixed-frequency regressions in a high-dimensional environment that resorts to Group Lasso penalization and Bayesian techniques for estimation and inference. To improve the sparse recovery ability of the model, we also consider a Group Lasso with a spike-and-slab prior. Penalty hyper-parameters governing the model shrinkage are automatically tuned via an adaptive MCMC algorithm. Simulations show that the proposed models have good selection and forecasting performance, even when the design matrix presents high cross-correlation. When applied to U.S. GDP data, the results suggest that financial variables may have some, although limited, short-term predictive content.

Keywords: MIDAS regressions, penalized regressions, variable selection, forecasting, Bayesian estimation

JEL classification: C11 ; C22 ; C53 ; E37

[1] International Macroeconomics Division matteo.mogliani@banque-france.fr
We wish to thank Florian Pelgrin, Clement Marsilli, Daniele Siena, and participants at the 18th IWH-CIREQ-GW Macroeconometric Workshop on Mixed Frequency Data in Macroeconomics and Finance (12-13 December 2017, Halle, Germany), the 11th International Conference on Computational and Financial Econometrics (16-18 December 2017, London, UK), the 5th conference of the International Association of Applied Econometrics (26-29 June 2018, Montreal, Canada), and research seminars at Banque de France and the Institute for Mathematics and Statistics of the Vienna University (18 January 2019, Vienna, Austria). The usual disclaimer applies.

# NON-TECHNICAL SUMMARY

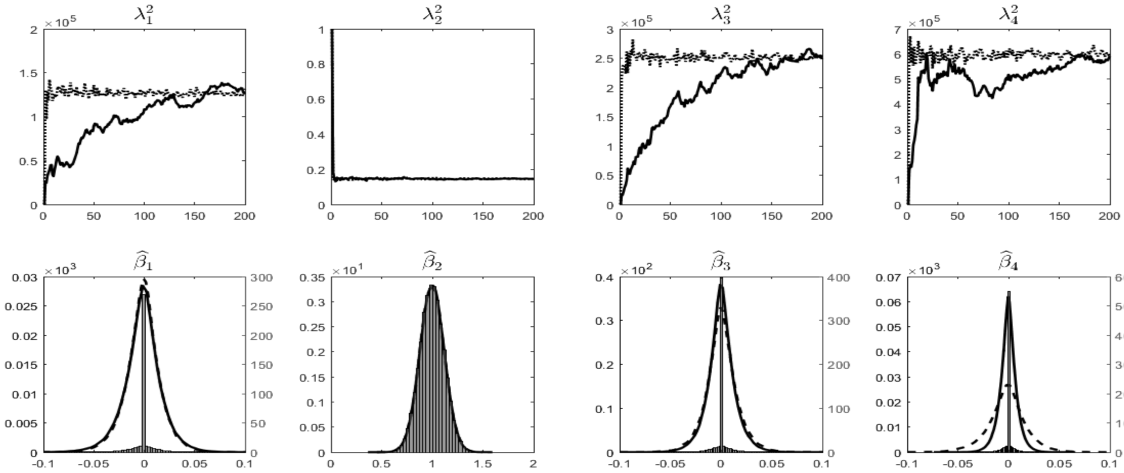
The outstanding increase in the availability of economic data has led econometricians to the development of new regression techniques based on Machine Learning algorithms, such as the family of penalized regressions. This consists in regressions with a modified objective function, such that coefficients estimated close to zero are shrunk to exactly zero, leading to simultaneous selection and estimation of coefficients associated to relevant variables only. While some of these techniques have been successfully applied to multivariate and usually highly parameterized macroeconomic models, such as VARs, only a few contributions in the literature have paid attention to mixed-frequency (MIDAS) regressions. In the classic MIDAS framework, the researcher can regress high-frequency variables (e.g. monthly variables such as surveys) directly on low-frequency variables (e.g. quarterly variables such as GDP) by matching the sampling frequency through specific aggregating (weighting) functions. The inclusion of many high-frequency variables into MIDAS regressions may nevertheless lead to overparameterized models, with poor predictive performance. This happens because the MIDAS regression approach can efficiently address the dimensionality issue arising from the number of high-frequency lags in the model, but not that arising from the number of high-frequency variables. Hence, recent literature has focused on MIDAS penalized regressions, based mainly on the so-called Lasso and Elastic-Net penalizations.

In the present paper, we propose a similar approach, but we depart from the existing literature on several points. First, we consider MIDAS regressions resorting to Almon lag polynomial weighting schemes, which depend only on a bunch of functional parameters governing the shape of the weighting function and keep linearity in the regression model. Second, we consider a Group Lasso penalty, which operates on distinct groups of regressors, and we set as many groups as the number of high-frequency predictors, allowing each group to include the entire Almon lag polynomial of each predictor. This grouping structure is motivated by the fact that if one high-frequency predictor is irrelevant, it should be expected that zero-coefficients occur in all the parameters of its lag polynomial. Third, we implement Bayesian techniques for the estimation of our penalized MIDAS regressions. The Bayesian approach offers two attractive features in our framework. The first one is the inclusion of spike-and-slab priors that, combined with the penalized likelihood approach, aim at improving the selection ability of the model by adding a probabilistic recovery layer to the hierarchy. The second one is the estimation of the penalty hyper-parameters through an automatic and data-driven approach that does not resort to extremely time consuming pilot runs. In this paper we consider an algorithm based on stochastic approximations, which consists in approximating the steps necessary to estimate the hyper-parameters in such a way that simple analytic solutions can be used. It turns out that penalty hyper-parameters can be automatically tuned with a small computational effort compared to existing and very popular alternative algorithms. We show through simple numerical experiments (see Figure below) that the suggested procedure works well in our framework: penalty hyper-parameters ($\lambda$) converge fairly quickly to their optimal values (first panel), such that the estimated coefficients ($\beta$) for irrelevant predictors are correctly centered at zero with small variance (second panel). Most importantly, the results points to substantial computational gains compared to alternative algorithms that we evaluate by a factor of 1 over 15.

We use our MIDAS models in an empirical forecasting application to U.S. GDP growth. We consider 42 real and financial indicators, sampled at monthly, weekly, and daily frequencies. We show that our models can provide superior point and density forecasts at short-term horizons (nowcasting and 1-quarter-ahead) compared to simple as well

sophisticated competing models. Further, the results suggest that high-frequency financial variables may have some, although limited, short-term predictive content for the GDP.

Figure. Convergence of penalty hyper-parameters ($\lambda$) over Gibbs sampler iterations and posterior distributions of regression parameters ($\beta$)



# Régressions pénalisées MIDAS bayésiennes : estimation, sélection et prévision

## RÉSUMÉ

Nous proposons une nouvelle méthode pour modéliser et prévoir avec des régressions à fréquences mixtes (MIDAS) en présence d'un nombre important de prédicteurs. Notre méthode s'appuie sur des régressions pénalisées telles que le Group Lasso, ainsi que sur des techniques Bayésiennes pour l'estimation des paramètres. Pour améliorer la capacité de sélection des variables du modèle, nous considérons également un Group Lasso augmenté avec des à priori de type spike-and-slab. Les hyper-paramètres de pénalisation qui gouvernent la sélection des variables sont calibrés automatiquement à partir d'un algorithme MCMC adaptatif. Des simulations Monte Carlo montrent que les modèles proposés présentent des performances en- et hors-échantillon très satisfaisantes, y compris quand les régresseurs sont très corrélés. Dans une application empirique sur le PIB des Etats-Unis, les résultats suggèrent que des variables financières à haute-fréquence (journalière) peuvent contribuer, bien que de manière limitée, à la prévision de court terme du PIB.

Mots-clés : régressions MIDAS, régressions pénalisées, sélection des variables, prévision, estimation Bayésienne

Les Documents de travail reflètent les idées personnelles de leurs auteurs et n'expriment pas nécessairement la position de la Banque de France. Ils sont disponibles sur publications.banque-france.fr

## 1. Introduction

Mixed-data sampling (MIDAS) models (Ghysels et al., 2005; Andreou et al., 2010) have been intensively used to forecast low frequency series, such as GDP, using monthly, weekly or daily predictors (Clements and Galvão, 2008, 2009; Kuzin et al., 2011; Andreou et al., 2013). This success relies mainly on the parsimonious and theoretically efficient treatment of the time-aggregation problem, compared to a traditional equal-weight approach. Indeed, data sampled at different frequencies are matched in a regression framework by using weighting schemes that resort to functional lag polynomials, where only a small number of hyperparameters that govern the shape of the aggregation function need to be estimated (Ghysels et al., 2007). However, one important issue with these models is the selection of predictors in presence of large datasets, as the inclusion of many high-frequency variables into MIDAS regressions may easily lead to overparameterized models, in-sample overfitting, and poor predictive performance. This happens because the MIDAS regression approach can efficiently address the dimensionality issue arising from the number of high-frequency lags in the model, but not that arising from the number of high-frequency variables. In the traditional MIDAS framework, a solution to this problem would consist in exploring a sub-space of possible models and selecting the best one according to some in-sample or out-of-sample criteria. However, this strategy would typically result in a prohibitive task, as enumeration could inefficiently imply too many models to explore. Hence, a number of alternative strategies have been proposed in the literature. For instance, Castle et al. (2009), Castle and Hendry (2010), and Bec and Mogliani (2015) use unrestricted MIDAS (U-MIDAS) regressions (Foroni et al., 2015) in a General-to-Specific framework (GETS) to jointly select relevant predictors and high-frequency lags through an automatic model reduction algorithm (*Autometrics*; Doornik, 2009). Marcellino and Schumacher (2010) suggest to include into MIDAS regressions common factors (static or dynamic) extracted from high-frequency variables (factor-MIDAS). Bessec (2013), Bulligan et al. (2015), and Girardi et al. (2017) pre-select instead high-frequency variables according to hard- and soft-thresholding rules (Bai and Ng, 2008) prior to factors extraction (*targeted* factor-MIDAS).

Recently, the literature has been increasingly focusing on Machine Learning and penalized regressions techniques for macroeconomic applications with large or very large variable dimension (Korobilis, 2013; Gefang, 2014; Koop et al., in press; Korobilis and Pettenuzzo, in press). Nevertheless, so far only a few contributions have paid attention to MIDAS regressions. For instance, Marsilli (2014) proposes a functional MIDAS combined with a Lasso objective function, which is solved in 1-step through a non-linear optimization algorithm. Siliverstovs (2017) proposes a 2-step targeted factor-MIDAS approach in the spirit of Bai and Ng (2008), where the soft-thresholding rule is built around U-MIDAS regressions combined with an Elastic-Net objective function. More recently, Uematsu and Tanaka (in press) propose a theoretical framework for penalized regressions with Lasso, SCAD, and MCP penalties in a general high-dimensional environment, where the number of predictors diverges sub-exponentially from the number of observations. In this framework, mixed-frequency regressions represent a special case of the general model. Indeed, Uematsu and Tanaka (in press) focus on MIDAS regressions with unconstrained lag polynomials (U-MIDAS), such that the number of parameters to estimate grows with both the number of high-frequency regressors and the length of the unconstrained lag function. Compared to Marsilli (2014), the resulting model is linear in parameters and it does not require the estimation of functional parameters for the distributed lag structure, which allows the model to deal with a large number of predictors. Further, compared to Siliverstovs (2017), estimation of model parameters and selection of relevant predictors/lags is performed in one step. However, the approach proposed by Uematsu and Tanaka (in press) may suffer from two well-known limitations of the Lasso. First, the Lasso cannot select more predictors than the number of observations. In many macroeconomic applications with monthly or daily high-frequency predictors, this constraint could be easily saturated even when the underlying U-MIDAS regression accounts only for a reasonable number of unrestricted lags. Second, and most importantly, the Lasso might not be generally suited in a mixed-frequency framework, because lags of the high-frequency predictors are by construction highly correlated, and hence the Lasso would tend to randomly select one lag and shrink the remaining unrestricted lag coefficients to zero.

In the present paper, we follow a similar strategy based on penalized regressions, but we propose to address these two issues by resorting to Almon lag polynomials and Group Lasso penalty. Distributed Almon lags allow us to keep a linear and parsimonious framework, as under this weighting scheme mixed-frequency regressions depend only on a bunch of functional parameters governing the shape of the weighting structure and can be easily cast as linear regression models (direct method). Further, linear restrictions on the lag polynomials can be placed to regularize the behavior of the weighting structure, consistently with some expected features of macroeconomic data. The Group Lasso penalty operates on distinct groups of regressors, rather than individual variables, where the grouping structure is chosen ex-ante by the researcher, usually according to some prior knowledge (common features, classification, etc.). In the present framework of distributed lags, we set as many groups as the number of high-frequency predictors, *i.e.* we let each group include one lag polynomial. This grouping structure is motivated by the fact that if one high-frequency predictor is irrelevant, it should be expected that zero-coefficients occur in all the parameters of its lag polynomial. Hence, unlike Uematsu and Tanaka (in press), selection is performed at the level of the entire lag polynomial, rather than on individual terms of the lag weighting function, overcoming the problem of extremely high correlation between lags.

A second contribution of the present paper is the implementation of Bayesian techniques for the estimation of our penalized MIDAS regressions. Following the recent literature on Bayesian penalized regressions and adaptive penalty schemes (Park and Casella, 2008; Wang and Leng, 2008; Kyung et al., 2010; Leng et al., 2014), we introduce a Bayesian MIDAS Adaptive Group Lasso, which under some conditions enjoys the oracle property by placing a different penalty term to each lag polynomial. We show that the Bayesian framework provides a simple hierarchical representation of this model, such that a Gibbs sampler can be used to draw efficiently from the posterior distribution of the parameters. Nonetheless, the Bayesian approach offers two additional attractive features in our framework. The first one is the inclusion of spike-and-slab priors that, combined with the penalized likelihood approach, aim at improving the sparsity recovery ability of the model (Zhang et al., 2014; Zhao and Sarkar, 2015; Xu and Ghosh, 2015; Ročková and George, 2018). We hence derive a Bayesian MIDAS Adaptive Group Lasso with spike-and-slab priors, which provides two shrinkage effects (zero point mass at the spike part of the prior and Group Lasso at the slab part) and it is expected to facilitate variable selection at the group level and shrinkage within the groups. The second one is the estimation of the penalty hyper-parameters through an automatic and data-driven approach that does not resort to extremely time consuming pilot runs. We depart from 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 (Park and Casella, 2008; Kyung et al., 2010; Leng et al., 2014), and we consider instead an adaptive MCMC algorithm based on stochastic approximations to solve the maximization problem (Atchadé, 2011; Atchadé et al., 2011). The algorithm consists in approximating both the E- and M-steps involved in the MCEM procedure, such that simple analytic solutions can be derived from the full posterior distribution of the unknown parameters of the Bayesian MIDAS model. Then, one step of the gradient algorithm can be used to update the penalty hyper-parameters with a small computational effort. We show through numerical experiments that substantial computational gains are obtained compared to the MCEM algorithm.

Estimation, selection and predictive accuracy are assessed through Monte Carlo simulations. Results show that the proposed models present very good in-sample and out-of-sample performance. In particular, variable selection, evaluated using a credible interval approach (Group Lasso) or a median estimator approach (Group Lasso with spike-and-slab), is achieved with high probability in a very sparse setting. Results are quite robust to the size of the design matrix (up to 50 high-frequency predictors in the Monte Carlo experiments) and to the choice of the shape of the weighting scheme in the DGP. However, the estimation and selection performance generally deteriorates with very high cross-correlation between the original high-frequency predictors. This outcome is nevertheless consistent with the theory, as the Group Lasso is not designed to handle strong collinearity between regressors. Finally, we illustrate our approach in an empirical forecasting application to U.S. GDP growth with 42 real and financial indicators sampled at monthly, weekly, and daily frequencies. We show that our models can provide superior point and density forecasts at short-term horizons (nowcasting and 1-quarter-ahead) compared to simple as well sophisticated competing models, such as Bayesian Model Averaging and optimally combined univariate Bayesian MIDAS models.

The paper is structured as follows. Sections 2 and 3 introduce the MIDAS penalized regressions and the Bayesian MIDAS framework. In Section 4 we discuss the Empirical Bayes approach used to automatically tune the penalty hyper-parameters. Section 5 investigates the estimation and predictive features of our models via Monte Carlo simulations. In Section 6, we report an empirical application to U.S. GDP. Finally, Section 7 concludes.

## 2. MIDAS penalized regressions

### *2.1. Basic MIDAS setup*

Consider the variable $y_t$, which is observed at discrete times (*i.e.* only once between $t-1$ and $t$), and suppose that we want to use information stemming from a set of $K$ predictors $\mathbf{x}_t^{(m)} = (x_{1,t}^{(m)}, \dots, x_{K,t}^{(m)})'$, which are observed $m$ times between $t-1$ and $t$, for forecasting purposes. The variables $y_t$ and $x_{k,t}^{(m)}$, for $k = 1, \dots, K$, 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 $h = 0, 1/m, 2/m, 3/m, \dots$ be an (arbitrary) forecast horizon, where $h = 0$ denotes a nowcast with high-frequency information fully matching the low-frequency sample. The MIDAS approach plugs-in the high-frequency lagged structure of predictors $x_{k,t-h}^{(m)}$ in a regression model for the low-frequency response variable $y_t$ as follows:

$$y_t = \alpha + \sum_{k=1}^{K} \mathcal{B}\left(L^{1/m}; \boldsymbol{\theta}_k\right) x_{k,t-h}^{(m)} + \epsilon_t, \tag{1}$$

where $\epsilon_t$ 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 a weighting structure which depends on the weighting function $B\left(c; \boldsymbol{\theta}_k\right)$, a vector of $p+1$ parameters $\boldsymbol{\theta}_k = (\theta_{k,0}, \theta_{k,1}, \dots, \theta_{k,p})$, and a maximum lag length $C$. Equation (1) can be also generalized to allow for lags of the dependent variable, as well as additional predictors sampled at multiple frequencies, including the same frequency as $y_t$. Several functional forms for $B\left(c; \boldsymbol{\theta}_k\right)$ have been proposed in the literature, such as the exponential Almon or the Beta lag polynomials (Ghysels et al., 2007). 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 $B\left(c; \boldsymbol{\theta}_k\right) = \sum_{i=0}^{p} \theta_{k,i} c^i$. Under the so-called "direct method" (Cooper, 1972), Equation (1) with Almon lag polynomials can be reparameterized as:

$$y_t = \alpha + \sum_{k=1}^{K} \sum_{i=0}^{p} \theta_{k,i} z_{k,i,t-h}^{(m)} + \epsilon_t \tag{2}$$

or in more compact form:

$$y_t = \alpha + \boldsymbol{\theta}' \mathbf{Z}_{t-h}^{(m)} + \epsilon_t \tag{3}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)'$, $\mathbf{Z}_t^{(m)} = \left(\mathbf{z}_{1,t}^{(m)}, \dots, \mathbf{z}_{K,t}^{(m)}\right)'$, and $\mathbf{z}_{k,t}^{(m)}$ is a vector of linear combinations of the observed high-frequency regressors, $\mathbf{z}_{k,t}^{(m)} = \mathbf{Q}\mathbf{x}_{k,t}^{(m)}$, with $\mathbf{x}_{k,t}^{(m)} = \left(x_{k,t}^{(m)}, x_{k,t-1/m}^{(m)}, \dots, x_{k,t-(C-1)/m}^{(m)}\right)'$ a

$(C \times 1)$ vector of high-frequency lags and

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

a $(p+1 \times C)$ polynomial weighting matrix. The $h$-step-ahead direct forecast $\widehat{y}_T$ can be hence obtained using (3) and sample information known at time $T-h$:

$$\widehat{y}_T = \widehat{\alpha} + \widehat{\boldsymbol{\theta}}' \mathbf{Z}^{(m)}_{T-h}. \tag{5}$$

The main advantage of the Almon lag polynomial is that (3) is linear and parsimonious, as it depends only on $K(p+1)$ parameters, and can be estimated consistently and efficiently via standard methods. However, two additional advantages make the Almon lag polynomial particularly attractive in the present framework. First, linear restrictions on the value and slope of the lag polynomial $B(c;\boldsymbol{\theta}_k)$ may be placed for any $c \in (0, C-1)$ (Smith and Giles, 1976). Restrictions such as $B(c;\boldsymbol{\theta}_k) = 0$ and $\nabla_c B(c;\boldsymbol{\theta}_k) = 0$, with $c$ evaluated at $C-1$, may be desirable and economically meaningful, as they jointly constrain the weighting structure to tail off slowly to zero. This can be obtained by modifying the $\mathbf{Q}$ matrix consistently with the form and the number of restrictions considered. As a result, the number of parameters in (2) reduces from $K(p+1)$ to $K(p-r+1)$, where $r \leq p$ is the number of restrictions. Second, a slope coefficient that captures the overall impact of lagged values of $x^{(m)}_{k,t-h}$ on $y_t$ can be easily computed as $\widehat{\beta}_k = \boldsymbol{\iota}_C \mathbf{Q}' \widehat{\boldsymbol{\theta}}'$, where $\boldsymbol{\iota}_C$ is a $(C \times 1)$ row vector of ones. This may be used to evaluate the statistical significance of each predictor in the regression and to implement model selection.

### *2.2. MIDAS penalized regressions*

Although appealing, the MIDAS regression presented above may be easily affected by overparameterization and multicollinearity in presence of a large number of potentially correlated predictors.[1] To achieve variable selection and parameter estimation simultaneously, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (Lasso). [2] 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 $\ell_1$ norm of the vector of regression coefficients, where the amount of penalization

[1] The direct method used in regression (3) may be also hampered by multicollinearity in the artificial variables $\mathbf{Z}^{(m)}_t$ (Cooper, 1972). However, if $p$ is small, the imprecision arising from multicollinearity may be compensated by the lower number of coefficients to be estimated.

[2] In the following discussion and in the next sections, we shall assume for convenience that $y_t$ is centered at 0 and regressors $\mathbf{Z}^{(m)}_{t-h}$ are standardized.

is controlled by a parameter $\lambda$. The objective function of the Lasso takes the form:

$$\mathcal{Q}_{\mathrm{L}}(\boldsymbol{\theta}) = T^{-1}\mathcal{L}_T(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_1, \tag{6}$$

where $\mathcal{L}_T(\boldsymbol{\theta})$ is the negative log-likelihood function, $\|\boldsymbol{\theta}\|_1 = \sum_{k=1}^{K}\sum_{i=0}^{p-r}|\theta_{k,i}|$ denotes the $\ell_1$ norm, and $\lambda \geq 0$.

However, it is well known (Zou, 2006; Zhao and Yu, 2006; Yuan and Lin, 2007) that the Lasso estimator does not possess the *oracle property*, which 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).[3] This can be achieved if and only if the so-called *irrepresentable condition* on the design matrix is satisfied and the penalization parameter $\lambda$ 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 $\lambda$ is chosen. This happens because the Lasso estimator in (6) uses the same amount of shrinkage for each regression coefficient, leading to estimation inefficiency and selection inconsistency. To address this issue, Zou (2006) proposes 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:

$$\mathcal{Q}_{\mathrm{AL}}(\boldsymbol{\theta}) = T^{-1}\mathcal{L}_T(\boldsymbol{\theta}) + \sum_{k=1}^{K}\sum_{i=0}^{p-r}\lambda_{k,i}|\theta_{k,i}| \tag{7}$$

However, the AL may not be suited in the present framework, as lags of high-frequency predictors are by construction highly correlated and hence the Lasso estimator would tend to select randomly only one lag and shrink the remaining polynomial coefficients to zero. The theoretical rationale for a failure in the selection ability of the AL in our mixed-frequency setting is similar to that pointed out by Efron et al. (2004) and Zou and Hastie (2005), and it is mostly related to the lack of strict convexity in the Lasso penalty. To address this issue, we propose a solution based on the Adaptive Group Lasso (AGL) estimator outlined in Wang and Leng (2008), who extend to adaptive shrinkage the Group Lasso estimator of 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 $k$ vectors of lag polynomials in the model. This grouping structure

[3] According to Fan and Li (2001), an estimator is said to possess the oracle property if *i)* it identifies the right subset model, *i.e.* $P(\widehat{\mathcal{A}} = \mathcal{A}) \to 1$ as $T \to \infty$, where $\mathcal{A}$ is the true active set of coefficients, and *ii)* it has the optimal estimation rate $\sqrt{T}(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}} - \boldsymbol{\theta}^*_{\mathcal{A}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{A}})$ 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.

[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).

is motivated by the fact that if one high-frequency predictor is irrelevant, it should be expected that zero-coefficients occur in all the parameters of its lag polynomial. Hence, unlike the AL, the AGL implies a single hyper-parameter associated to each high-frequency variable. This strategy should overcome, at least in part, the limitation of the Lasso in presence of strong correlation in the design matrix arising from the correlation among lags of the transformed high-frequency predictors. 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 notation (necessary to avoid confusion), it is straightforward that $G = K$, $\boldsymbol{\theta}_j = \boldsymbol{\theta}_k$, $g_j = p - r + 1$, and $\tilde{g} \equiv \sum_{j=1}^{G} g_j = K(p - r + 1)$. Hence, the objective function of the AGL takes the form:

$$\mathcal{Q}_{\text{AGL}}(\boldsymbol{\theta}) = T^{-1}\mathcal{L}_T(\boldsymbol{\theta}) + \sum_{j=1}^{G} \lambda_j \|\boldsymbol{\theta}_j\|_2 \tag{8}$$

where $\|\boldsymbol{\theta}_j\|_2 = (\boldsymbol{\theta}_j'\boldsymbol{\theta}_j)^{1/2}$ denotes the $\ell_2$ norm. As for the asymptotic properties, Wang and Leng (2008) establish the consistency and the oracle property of the AGL. 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 should attenuate this issue.

## 3. Bayesian MIDAS penalized regressions

Several approaches, such as the LARS (Efron et al., 2004) and the Group LARS (Yuan and Lin, 2006) algorithms (further modified to account for adaptive shrinkage), have been proposed in the literature to estimate penalized regressions. In this paper, we consider a Bayesian hierarchical approach (Park and Casella, 2008; Kyung et al., 2010), which has several advantages compared to the frequentist approach. First, Bayesian methods exploit model inference via posterior distributions of parameters, which usually provide a valid measure of standard errors based on a geometrically ergodic Markov chain (Khare and Hobert, 2013).[5] Second, they provide a flexible way of estimating the penalty parameters, along with other parameters in the model. Lastly, they provide forecasts

[5] It is nevertheless worth noting that the results in Khare and Hobert (2013) hold as long as the penalty hyper-parameters are assumed fixed, while convergence properties of the MCMC algorithm for the full Bayesian penalized regression models are still unknown (see also Roy and Chakraborty, 2017).

via predictive distributions. In what follows, we present in detail the hierarchical structure of the proposed Bayesian MIDAS penalized models.

### *3.1. Bayesian MIDAS adaptive Group Lasso*

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 $\ell_1$ penalty corresponds to a conditional Laplace prior that can be represented as a scale mixture of Normals with an exponential mixing density (Andrews and Mallows, 1974). For the Bayesian Group Lasso, Kyung et al. (2010) consider a multivariate generalization of the double exponential prior and they show that the conditional prior of $\boldsymbol{\theta}$ can be expressed as a scale mixture of Normals with Gamma hyper-priors:

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

For the Bayesian Adaptive Group Lasso, the conditional prior for $\boldsymbol{\theta}$ becomes:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\sigma^2) &\propto \prod_{j=1}^{G}\int_0^\infty\left(\frac{1}{2\pi\sigma^2\tau_j^2}\right)^{\frac{g_j+1}{2}}\exp\left(-\frac{\|\boldsymbol{\theta}_j\|_2^2}{2\sigma^2\tau_j^2}\right) f_\Gamma\left(\tau_j^2;\frac{(g_j+1)}{2},\frac{\lambda_j^2}{2}\right)\mathrm{d}\tau_j^2\\ &\propto \exp\left(-\frac{1}{\sqrt{\sigma^2}}\sum_{j=1}^{G}\lambda_j\|\boldsymbol{\theta}_j\|_2\right) \qquad (9)\end{aligned}$$

where $f_\Gamma$ denotes the pdf of a Gamma distribution, with shape $(g_j+1)/2$ and rate $\lambda_j^2/2$. This suggests the following hierarchical representation of the Bayesian MIDAS Adaptive Group Lasso model (BMIDAS-AGL):

$$\begin{aligned}y|\mathbf{Z},\boldsymbol{\theta},\sigma^2 &\sim \mathcal{N}\left(\boldsymbol{\theta}'\mathbf{Z},\sigma^2\mathbf{I}_T\right)\\ \boldsymbol{\theta}_j|\tau_j^2,\sigma^2 &\sim \mathcal{N}(\mathbf{0},\sigma^2\tau_j^2\mathbf{I}_{g_j}) \qquad j=1,\dots,G\\ \tau_j^2 &\sim \text{Gamma}\left(\frac{g_j+1}{2},\frac{\lambda_j^2}{2}\right)\\ \sigma^2 &\sim \text{iGamma}\left(a_1,b_1\right)\end{aligned}$$

where $\boldsymbol{\tau}=(\tau_1^2,\dots,\tau_G^2)$, $\boldsymbol{\lambda}=(\lambda_1^2,\dots,\lambda_G^2)$, and $\mathbf{I}_{g_j}$ is the identity matrix of order $g_j$. The full posterior distribution of all the unknown parameters conditional on the data and the penalty hyper-parameters is:

$$\begin{aligned}\pi(\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2|\boldsymbol{\lambda},y,\mathbf{Z}) \propto \left(\sigma^2\right)^{-\frac{T+\tilde{g}-1}{2}-a_1-1} &\exp\left[-\frac{1}{2\sigma^2}\|y-\boldsymbol{\theta}'\mathbf{Z}\|_2^2-\frac{b_1}{\sigma^2}\right] \\ &\times\prod_{j=1}^{G}\left(\frac{1}{2\pi\sigma^2\tau_j^2}\right)^{\frac{g_j}{2}}\exp\left(-\frac{\|\boldsymbol{\theta}_j\|_2^2}{2\sigma^2\tau_j^2}\right) \\ &\times\prod_{j=1}^{G}\left(\lambda_j^2\right)^{\frac{g_j+1}{2}}\left(\tau_j^2\right)^{\frac{g_j+1}{2}-1}\exp\left(-\frac{\lambda_j^2}{2}\tau_j^2\right) \qquad (10)\end{aligned}$$

We use an efficient block Gibbs sampler (Hobert and Geyer, 1998) for simulating from this posterior distribution. Let's denote $\boldsymbol{\theta}_{-j}=(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_{j-1},\boldsymbol{\theta}_{j+1},\dots,\boldsymbol{\theta}_G)'$ the $\boldsymbol{\theta}$ vector without the $j$th high-frequency lag polynomial, and $\mathbf{Z}_j$ and $\mathbf{Z}_{-j}$ partitions of the design matrix corresponding to $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_{-j}$, respectively. With a conjugate Gamma prior placed on the penalty hyper-parameters, $\lambda_j^2\sim\text{Gamma}\left(a_2,b_2\right)$, the full conditional posteriors are:

$$\begin{aligned}\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j},\sigma^2,\boldsymbol{\tau},\boldsymbol{\lambda},y,\mathbf{Z} &\sim \mathcal{N}\left(\mathbf{A}_j^{-1}\mathbf{C}_j,\sigma^2\mathbf{A}_j^{-1}\right) \\ \tau_j^{-2}|\boldsymbol{\theta},\sigma^2,\boldsymbol{\lambda},y,\mathbf{Z} &\sim \text{iGaussian}\left(\frac{\lambda_j\sigma}{\|\boldsymbol{\theta}_j\|_2},\lambda_j^2\right) \\ \sigma^2|\boldsymbol{\theta},\boldsymbol{\tau},\boldsymbol{\lambda},y,\mathbf{Z} &\sim \text{iGamma}\left(\frac{T+\tilde{g}-1}{2}+a_1,\frac{1}{2}\|y-\boldsymbol{\theta}'\mathbf{Z}\|_2^2+\frac{1}{2}\sum_{j=1}^{G}\frac{\|\boldsymbol{\theta}_j\|_2^2}{\tau_j^2}+b_1\right) \\ \lambda_j^2|\boldsymbol{\theta},\sigma^2,\boldsymbol{\tau},y,\mathbf{Z} &\sim \text{Gamma}\left(\frac{g_j+1}{2}+a_2,\frac{\tau_j^2}{2}+b_2\right)\end{aligned}$$

where $\mathbf{A}_j=\mathbf{Z}_j'\mathbf{Z}_j+\tau_j^{-2}\mathbf{I}_{g_j}$, and $\mathbf{C}_j=\mathbf{Z}_j'\left(y-\boldsymbol{\theta}_{-j}'\mathbf{Z}_{-j}\right)$.

### *3.2. Bayesian MIDAS adaptive Group Lasso with Spike and Slab Prior*

A typical feature of the model outlined above is that a sparse solution cannot be perfectly achieved, as the Bayesian approach provides a shrinkage of the coefficients towards zero, but usually not exactly to zero. Spike-and-slab methods (Mitchell and Beauchamp, 1988; George and McCulloch, 1993) are well known approaches for probabilistic sparse recovery, where the prior for the regression coefficients is specified as a mixture distribution taking various forms (*e.g.* two-point uniform and degenerate, or multivariate Gaussian). These approaches differ substantially from the penalized likelihood approach implemented so far in our setup, as the latter induces sparsity through penalty functions whose geometry is exerted in constrained optimization (Ročková and George, 2018). However, recent literature has increasingly focused on combining the potential advantages of the two methods by adding a point mass mixture prior to penalized regressions, and letting the slab part in the prior be a Laplace distribution (Zhang et al., 2014; Zhao and Sarkar, 2015; Ročková and George, 2018). In the present study, we follow Xu and Ghosh (2015) and we consider a Bayesian Group Lasso with spike-and-slab priors for group variable selection. Unlike the group selection method described in Section 3.1, this prior provides two shrinkage effects: the point mass at $\mathbf{0}$ (the spike part of the prior), which leads to exact zero coefficients, and the Group Lasso prior on the slab part. The combination of these two components together is expected to facilitate variable selection at the group level and to shrink coefficients in the selected groups simultaneously.

Similarly to the BMIDAS-AGL, the hierarchical Bayesian MIDAS Adaptive Group Lasso with spike-and-slab priors (BMIDAS-AGL-SS) is:

$$\begin{aligned}
y|\mathbf{Z},\boldsymbol{\theta},\sigma^2 &\sim \mathcal{N}\left(\boldsymbol{\theta}'\mathbf{Z},\sigma^2\mathbf{I}_T\right)\\
\boldsymbol{\theta}_j|\tau_j^2,\sigma^2,\pi_0 &\sim (1-\pi_0)\mathcal{N}(\mathbf{0},\sigma^2\tau_j^2\mathbf{I}_{g_j})+\pi_0\delta_0(\boldsymbol{\theta}_j) \qquad j=1,\dots,G\\
\tau_j^2 &\sim \text{Gamma}\left(\frac{g_j+1}{2},\frac{\lambda_j^2}{2}\right)\\
\sigma^2 &\sim \text{iGamma}\left(a_1,b_1\right)\\
\pi_0 &\sim \text{Beta}\left(c,d\right)
\end{aligned}$$

where $\delta_0(\boldsymbol{\theta}_j)$ denotes a point mass at $\mathbf{0}\in\mathbb{R}^{g_j}$. Note that we place a conjugate Beta prior on $\pi_0$, *i.e.* the prior probability to all sub-models, rather than a fixed value. We follow Castillo et al. (2015) and Ročková and George (2018), and we set $c=1$ and $d=G$. The full posterior distribution of all the unknown parameters conditional on the data and the penalty hyper-parameters is:

$$\begin{aligned}
\pi(\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2,\pi_0|\boldsymbol{\lambda},y,\mathbf{Z}) \propto \left(\sigma^2\right)^{-\frac{T+\tilde{g}-1}{2}-a_1-1}&\exp\left[-\frac{1}{2\sigma^2}\|y-\boldsymbol{\theta}'\mathbf{Z}\|_2^2-\frac{b_1}{\sigma^2}\right]\pi_0^{c-1}(1-\pi_0)^{d-1}\\
&\times\prod_{j=1}^{G}\left[\pi_0\left(\frac{1}{2\pi\sigma^2\tau_j^2}\right)^{\frac{g_j}{2}}\exp\left(-\frac{\|\boldsymbol{\theta}_j\|_2^2}{2\sigma^2\tau_j^2}\right)\mathbf{I}_{\boldsymbol{\theta}_j\neq 0}+(1-\pi_0)\delta_0(\boldsymbol{\theta}_j)\right]\\
&\times\prod_{j=1}^{G}\left(\lambda_j^2\right)^{\frac{g_j+1}{2}}\left(\tau_j^2\right)^{\frac{g_j+1}{2}-1}\exp\left(-\frac{\lambda_j^2}{2}\tau_j^2\right) \qquad (11)
\end{aligned}$$

The full conditional posteriors are:

$$\begin{aligned}
\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j},\sigma^2,\boldsymbol{\tau},\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\pi},\pi_0,y,\mathbf{Z} &\sim \gamma_j\ \mathcal{N}\left(\mathbf{A}_j^{-1}\mathbf{C}_j,\sigma^2\mathbf{A}_j^{-1}\right)+(1-\gamma_j)\ \delta_0(\boldsymbol{\theta}_j)\\
\tau_j^{-2}|\boldsymbol{\theta},\sigma^2,\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\pi},\pi_0,y,\mathbf{Z} &\sim \gamma_j\ \text{iGaussian}\left(\frac{\lambda_j\sigma}{\|\boldsymbol{\theta}_j\|_2},\lambda_j^2\right)+(1-\gamma_j)\ \text{Gamma}\left(\frac{g_j+1}{2},\frac{\lambda_j^2}{2}\right)
\end{aligned}$$

$$\sigma^2|\boldsymbol{\theta},\boldsymbol{\tau},\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\pi},\pi_0,y,\mathbf{Z} \sim \text{iGamma}\left(\frac{T+\widetilde{G}-1}{2}+a_1,\frac{1}{2}\|y-\boldsymbol{\theta}'\mathbf{Z}\|_2^2+\frac{1}{2}\sum_{j=1}^{G}\frac{\|\boldsymbol{\theta}_j\|_2^2}{\tau_j^2}+b_1\right)$$

$$\gamma_j|\boldsymbol{\theta},\sigma^2,\boldsymbol{\tau},\boldsymbol{\lambda},\boldsymbol{\pi},\pi_0,y,\mathbf{Z} \sim \text{Bernoulli}\left(\pi_j\right)$$

$$\lambda_j^2|\boldsymbol{\theta},\sigma^2,\boldsymbol{\tau},\boldsymbol{\gamma},\boldsymbol{\pi},\pi_0,y,\mathbf{Z} \sim \text{Gamma}\left(\frac{g_j+1}{2}+a_2,\frac{\tau_j^2}{2}+b_2\right)$$

$$\pi_0|\boldsymbol{\theta},\sigma^2,\boldsymbol{\tau},\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\pi},y,\mathbf{Z} \sim \text{Beta}\left(\sum_{j=1}^{G}\gamma_j+c,\sum_{j=1}^{G}(1-\gamma_j)+d\right)$$

where $\mathbf{A}_j = \mathbf{Z}_j'\mathbf{Z}_j + \tau_j^{-2}\mathbf{I}_{g_j}$, $\mathbf{C}_j = \mathbf{Z}_j'\left(y-\boldsymbol{\theta}_{-j}'\mathbf{Z}_{-j}\right)$, $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_G)$, $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_G)$, $\widetilde{G} = \sum_{j=1}^{G} g_j\gamma_j$, and

$$\pi_j = \pi(\boldsymbol{\theta}_j \neq \mathbf{0}|\boldsymbol{\theta}_{-j},\sigma^2,\boldsymbol{\tau},\boldsymbol{\lambda},\boldsymbol{\gamma},\pi_0,y,\mathbf{Z}) = \frac{\pi_0\left[(\tau_j^2)^{-\frac{g_j}{2}}|\mathbf{A}_j|^{-\frac{1}{2}}\exp\left(\frac{1}{2\sigma^2}\mathbf{C}_j'\mathbf{A}_j^{-1}\mathbf{C}_j\right)\right]}{1-\pi_0\left[1-(\tau_j^2)^{-\frac{g_j}{2}}|\mathbf{A}_j|^{-\frac{1}{2}}\exp\left(\frac{1}{2\sigma^2}\mathbf{C}_j'\mathbf{A}_j^{-1}\mathbf{C}_j\right)\right]}.$$

## 4. Tuning the penalty hyper-parameters

The hierarchical models presented in Section 3 treat the penalty parameters as hyper-parameters, *i.e.* as random variables with gamma prior distributions $\pi(\boldsymbol{\lambda})$ and gamma posterior distributions $\pi(\boldsymbol{\lambda}|y,\boldsymbol{\phi})$, where $\boldsymbol{\phi} = (\boldsymbol{\theta},\boldsymbol{\tau},\sigma^2)'$. However, the main drawback of this approach is that these posterior distributions can be sensitive to the choice of the prior. An alternative approach resorts to an Empirical Bayes estimation of the hyper-parameters, *i.e.* using the data to propose an estimate of $\boldsymbol{\lambda}$, which can be obtained through marginal maximum likelihood. However, in the present framework, the marginal distribution $\pi(y|\boldsymbol{\lambda}) = \int f_{\boldsymbol{\phi},\boldsymbol{\lambda}}(y)\pi(\boldsymbol{\phi}|\boldsymbol{\lambda})d\boldsymbol{\phi}$, where $f_{\boldsymbol{\phi},\boldsymbol{\lambda}}(y)$ is the likelihood, 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. More specifically, the MCEM algorithm of Casella (2001) uses $N$ Monte Carlo iterations to maximize the marginal log-likelihood $\log\pi(y|\boldsymbol{\lambda})$ and involves two steps. First (E-step), for each $n = 1,\ldots,N$, an expectation function is solved for a given $\boldsymbol{\lambda}^{(n)}$:

$$Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(n)}) = \int \log\left[f_{\boldsymbol{\phi},\boldsymbol{\lambda}}(y)\pi(\boldsymbol{\phi}|\boldsymbol{\lambda})\right]\pi(\boldsymbol{\phi}|y,\boldsymbol{\lambda}^{(n)})d\boldsymbol{\phi}$$

where $\pi(y, \boldsymbol{\phi}|\boldsymbol{\lambda}) = f_{\boldsymbol{\phi},\boldsymbol{\lambda}}(y)\pi(\boldsymbol{\phi}|\boldsymbol{\lambda})$ is the joint density of the observed and missing data, respectively, given $\boldsymbol{\lambda}$, and $\pi(\boldsymbol{\phi}|y, \boldsymbol{\lambda}^{(n)})$ is the density of the missing data conditional on the observed data and $\boldsymbol{\lambda}^{(n)}$ (an initial value $\boldsymbol{\lambda}^{(0)}$ is used to initialize the Monte Carlo). Then (M-step), $Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(n)})$ is maximized to give $\boldsymbol{\lambda}^{(n+1)}$:

$$\boldsymbol{\lambda}^{(n+1)} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\ Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(n)})$$

For the models described in Section 3, simple analytic solutions can be used to compute $\boldsymbol{\lambda}^{(n+1)}$ (Park and Casella, 2008; Kyung et al., 2010). However, since $\pi(\boldsymbol{\phi}|y, \boldsymbol{\lambda})$ 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, the MCEM algorithm may be extremely expensive, as each $n$th Monte Carlo iteration requires a fully converged Gibbs sampling from $\pi(\boldsymbol{\phi}|y, \boldsymbol{\lambda}^{(n)})$. Hence, a serious trade-off between accuracy of the results ($S$ Gibbs iterations) and computational efficiency ($N$ Monte Carlo iterations) may arise. In the present framework, careful attention must be paid to this feature, because the computational burden implied by the Group Lasso increases dramatically as the number of predictors increases (Yuan and Lin, 2006). To deal with this issue, in this work we adopt an alternative Empirical Bayes approach that relies 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 requires only a single Monte Carlo run ($N = 1$). Following Atchadé (2011), let us write the derivative of $Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(s)})$ with respect to $\boldsymbol{\lambda}$ as:

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

where $H(\boldsymbol{\lambda}, \boldsymbol{\phi}) = \nabla_{\boldsymbol{\lambda}} \log\left[f_{\boldsymbol{\phi},\boldsymbol{\lambda}}(y)\pi(\boldsymbol{\phi}|\boldsymbol{\lambda})\right] = \nabla_{\boldsymbol{\lambda}} \log \pi(\boldsymbol{\phi}|\boldsymbol{\lambda})$, as the likelihood does not usually depend on the hyper-parameters $\boldsymbol{\lambda}$. Note the change in 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 full maximization of $Q$ with one step of the gradient algorithm, the solution to the EM algorithm takes the form:

$$\boldsymbol{\lambda}^{(s+1)} = \boldsymbol{\lambda}^{(s)} + a^{(s)}\nabla_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda}^{(s)}|\boldsymbol{\lambda}^{(s)})x = \boldsymbol{\lambda}^{(s)} + a^{(s)} \int H(\boldsymbol{\lambda}^{(s)}, \boldsymbol{\phi})\pi(\boldsymbol{\phi}|y, \boldsymbol{\lambda}^{(s)})d\boldsymbol{\phi}$$

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 $\int H(\boldsymbol{\lambda}^{(s)}, \boldsymbol{\phi})\pi(\boldsymbol{\phi}|y, \boldsymbol{\lambda}^{(s)})d\boldsymbol{\phi}$ is approximated by $H(\boldsymbol{\lambda}^{(s)}, \boldsymbol{\phi}^{(s+1)})$, we get an approximate EM algorithm, where both E- and M-steps are approximately implemented. Hence, marginal maximum likelihood estimates of the hyper-parameters, $\widehat{\boldsymbol{\lambda}}$, and draws from the posterior distribution of the parameters, $\pi(\boldsymbol{\phi}|y, \widehat{\boldsymbol{\lambda}})$, are both obtained using a single run of the Gibbs sampler, with $s = 1, \dots, S$. In the present framework, taking logs of the full posterior distributions (10) and (11) and making the transformation $\boldsymbol{\omega} = \frac{1}{2}\log(\boldsymbol{\lambda})$, the function $H(\boldsymbol{\omega}, \boldsymbol{\phi}) = \nabla_{\boldsymbol{\omega}} \log \pi(\boldsymbol{\phi}|\boldsymbol{\omega})$ takes the form:

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

where $\mathbf{g} = (g_1, \dots, g_G)'$ and $\odot$ is the element-wise product. Hence, the updating rule for $\boldsymbol{\omega}$ is:

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

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 $\widehat{\boldsymbol{\lambda}}$ and $\pi(\boldsymbol{\phi}|y, \widehat{\boldsymbol{\lambda}})$, respectively. Details on the stabilization algorithm are reported in Appendix A.1.

### *4.1. Numerical illustration*

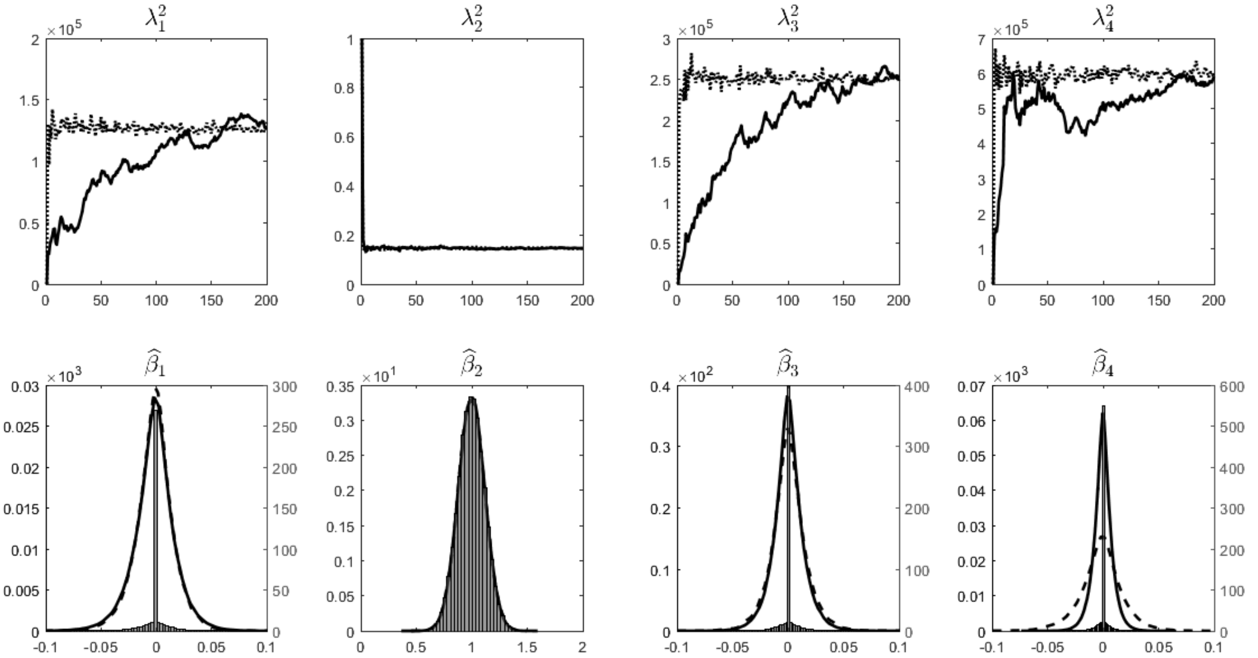
We illustrate the main features and the computational advantage of the proposed methodology using simulated data. For ease of exposition, the DGP follows a simple mixed-frequency model with four predictors:

$$y_t = \beta_0 + \sum_{k=1}^{4} \beta_k \sum_{c=0}^{C-1} B\left(c; \boldsymbol{\vartheta}\right) L^{c/m} x_{k,t}^{(m)} + \epsilon_t,$$

where the regressors and the error term are iid draws from a standard normal distribution of length $T = 500$. We set the true values $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (1, 0, 1, 0, 0)$ and $B\left(c; \boldsymbol{\vartheta}\right)$ is parameterized as an exponential Almon lag function, with $C = 12$, $m = 3$, and $\boldsymbol{\vartheta} = (0.10, -0.15)$, *i.e.* a fast-decaying weighting function where about 90% of the weight is concentrated in the three most recent high-frequency observations. We estimate the models presented in Section 3 using $p = 3$ and we tune the penalty hyper-parameters $\boldsymbol{\lambda}$ using the stochastic approximation approach. We update $\boldsymbol{\lambda}$ in a single run of the Gibbs sampler by drawing $S = 400,000$ samples. The analysis is carried out using MATLAB R2017a on a workstation with a 2.50GHz Intel Core i7-6500U CPU.

The evolution of $\boldsymbol{\lambda}$ across iterations (starting with $\lambda_j^{(0)} = 1$) is reported in the first panel of Figure 1. Each point in the plots represents the 2000th update of $\boldsymbol{\lambda}$ provided by the stochastic approximation approach for the BMIDAS-AGL model (solid line) and the BMIDAS-AGL-SS model (dotted line). For both models, the hyper-parameters converge to fairly similar values. However, while the convergence is steady and extremely fast for the active variable, the BMIDAS-AGL model displays slower convergence for the penalty terms of the inactive set compared to the BMIDAS-

Figure 1: Tuning the penalty hyper-parameters



Note: The first panel illustrates the evolution of the penalty hyper-parameters $\boldsymbol{\lambda}$ across iterations of the stochastic approximation approach for BMIDAS-AGL (solid lines) and BMIDAS-AGL-SS model (dotted lines). The second panel illustrates the posterior distributions of parameters $\boldsymbol{\beta}$ for BMIDAS-AGL (solid lines) and BMIDAS-AGL-SS model (histogram) using the stochastic approximation approach, and the BIMDAS-AGL model using the MCEM algorithm (dashed lines).

AGL-SS. Hence, it turns out that allowing for point mass at zero through the spike-and-slab prior may not only improve the sparse recovery ability of the model but also enhance the convergence of the penalty hyper-parameters. This is expected to reduce the variance of the posterior distribution around the zero-point mass when draws of the coefficients belonging to the inactive set are assigned (even with some low probability) to the slab part of the model. Results on the posterior densities of $\boldsymbol{\beta}$ seem to confirm this expected feature. Posterior densities are identical for the active set and display largest mass at zero for $\beta_1$, $\beta_3$, and $\beta_4$, but the BMIDAS-AGL-SS model displays the lowest variation around the point mass at exactly zero. Hence, by assigning small penalty (*i.e.* small $\lambda$s) to the relevant predictor and large penalty to the irrelevant predictors, both models display correct variable selection and consistent estimates of the regression coefficients (Zou, 2006; Wang and Leng, 2008; Zou and Zhang, 2009). Finally, these outcomes are compared to those obtained by tuning the penalty hyper-parameters of the BMIDAS-AGL model using the MCEM algorithm of Casella (2001) with a fairly reasonable amount of Monte Carlo runs ($N = 200$) and Gibbs draws ($S = 50,000$). A visual inspection of the second panel of Figure 1 suggests that posterior densities from the MCEM algorithm are almost indistinguishable from those obtained using the stochastic approximation approach, with the only exception of the coefficient $\beta_4$. However, the computational cost is substantially different across algorithms: for this simple simulation experiment and the settings described above, the task is executed in less than 2 minutes with stochastic approximations, against 30 minutes required by the MCEM algorithm.

## 5. Monte Carlo experiments

### *5.1. 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 $K = \{30, 50\}$ predictors sampled at frequency $m = 3$ and $T = 200$ in-sample observations:

$$
\begin{aligned}
y_t &= \alpha + \sum_{k=1}^{K} \beta_k \sum_{c=0}^{C-1} B\left(c; \boldsymbol{\vartheta}\right) L^{c/3} x_{k,t-h}^{(3)} + \epsilon_t \\
x_{k,t}^{(3)} &= \mu + \rho x_{k,t-1/3}^{(3)} + \varepsilon_{k,t} \\
B\left(c; \boldsymbol{\vartheta}\right) &= \frac{\exp(\vartheta_1 c + \vartheta_2 c^2)}{\sum_{c=0}^{C-1} \exp(\vartheta_1 c + \vartheta_2 c^2)}
\end{aligned}
$$

where $B\left(c; \boldsymbol{\vartheta}\right)$ is parameterized as an exponential Almon lag function. Following Andreou et al. (2010), we investigate three alternative weighting schemes that correspond to fast-decaying weights, $\boldsymbol{\vartheta} = (7 * 10^{-4}, -7 * 10^{-2})$, slow-decaying weights, $\boldsymbol{\vartheta} = (7 * 10^{-4}, -9 * 10^{-3})$, and near-flat weights, and $\boldsymbol{\vartheta} = (0, -5 * 10^{-4})$. In all simulations we set the lag length $C = 24$. Note that the same weighting structure applies to all the predictors entering the active set. Further, for ease of analysis we assume $h = 0$, *i.e.* a nowcasting model with fully available information on predictors in the current period. In this specification, $\epsilon_t$ 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 $\sigma_\varepsilon = \{0.50, 0.95\}$, *i.e.* from moderate to extremely high correlation structure in the design matrix $\mathbf{x}_t^{(3)}$. As for the parameters in the DGP, we choose $\alpha = 0.5$, $\mu = 0.1$, $\rho = 0.9$, and $\boldsymbol{\beta} = (0, 0.3, 0.5, 0, 0.3, 0.5, 0, 0, 0.8, \mathbf{0})'$. The latter implies that only five out of $K$ predictors are relevant. Conditional on these parameters, we set $\sigma$ such that the noise-to-signal ratio of the mixed-frequency regression is 0.20.

We estimate mixed-frequency models on the data provided by the DGP above using the regression approaches described in Section 3. As for the functional form of the weighting structure, we consider a restricted Almon lag polynomial as in (3), with $p = 3$ and $r = 2$ endpoint restrictions (both tail and derivative; see Section 2.1). The hyper-parameters $\boldsymbol{\lambda}$ are tuned using the stochastic approximation approach described in Section 4, with step-size $a^{(s)} = 1/s^{0.8}$ (preliminary results suggest that this sequence is sufficient to achieve convergence). We set the number of Monte Carlo replications at $R = 300$. The Gibbs sampler is run for $S = 250,000$ iterations, with the first $50,000$ used as a burn-in period, and every 10th draw is saved.

### *5.2. Variable selection*

The penalized regression approach had originally been proposed as a variable selection method. Indeed, the penalty terms in Equation (8) are intended to shrink the coefficients of irrelevant predictors to zero, leading to a sparse solution. However, as noted in Section 3.2, this attractive property vanishes in the Bayesian framework. Different approaches have been proposed in the literature to evaluate variable selection for the models under analysis. For instance, Li and Lin (2010) propose a scaled neighbourhood criterion, where a predictor is excluded if the posterior probability in the neighbourhood $[-\sigma_{\widehat{\beta}_k}, \sigma_{\widehat{\beta}_k}]$ of 0 exceeds a given probability threshold. Here we rely on the simple credible interval criterion suggested by Kyung et al. (2010). According to this criterion, a predictor $k$, for $k = 1, \dots, K$, is excluded from the estimated active set if the credible interval, at say 95% level, of the posterior distribution of the slope coefficient $\widehat{\beta}_k$ includes zero. For the model including a spike-and-slab prior, we resort to the posterior median estimator (Barbieri and Berger, 2004), that is, under some conditions, a soft thresholding estimator presenting model selection consistency and optimal asymptotic estimation rate (Xu and Ghosh, 2015).

### *5.3. Forecasting*

Forecasts are obtained from the following posterior predictive density for $y_T$:

$$p(y_T|\mathcal{D}) = \int p(y_T|\boldsymbol{\phi}, \boldsymbol{\lambda}, \mathcal{D}) p(\boldsymbol{\phi}, \boldsymbol{\lambda}|\mathcal{D}) d\boldsymbol{\phi} d\boldsymbol{\lambda} \tag{12}$$

where $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^2)'$ and $p(\boldsymbol{\phi}, \boldsymbol{\lambda}|\mathcal{D})$ denotes the joint posterior distribution of the BMIDAS parameters conditional on past available information, $\mathcal{D}$. According to the framework described in Sections 3 and 4, draws $y_T^{(s)}$ from the predictive distribution can be obtained from the Gibbs sampler, for $s = \bar{s}+1, \dots, S$ and $\bar{s}$ the last burn-in iteration.[6] This leads to a distribution of predictions that can be used for out-of-sample evaluation of the model. Point forecasts are computed by averaging over these draws, *i.e.* $\widehat{y}_T = (S - \bar{s} + 1)^{-1} \sum_{s=\bar{s}+1}^{S} y_T^{(s)}$, and evaluated through the average mean squared forecast error (MSFE) and the average mean absolute forecast error (MAFE) over the $R$ Monte Carlo replications. However, since draws from the predictive density are available, an evaluation of the entire predictive distribution is performed through the (negative) average log-score ($-$LS), *i.e.*

[6] It is worth noting that we do not condition on a fixed value $\widehat{\boldsymbol{\lambda}}$, 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.

the average of the log of the predictive likelihood evaluated at the out-turn of the forecast (Mitchell and Wallis, 2011), and the average continuously ranked probability score (CRPS), 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).

### *5.4. Monte Carlo results*

Simulation results for our penalized estimators are reported in Table 1. We compute the average mean squared error (MSE), the average variance (VAR), and the average squared bias ($\text{BIAS}^2$) over $R$ Monte Carlo replications and the full set of $K$ estimated parameters $\widehat{\boldsymbol{\beta}}$ in the model.[7] Further, we evaluate the selection ability of the models by computing the True Positive Rate (TPR), the False Positive Rate (FPR), and the Matthews correlation coefficient (MCC), the latter measuring the overall quality of the classification. Results point to a number of interesting features. First, the models perform overall quite similarly in terms of MSE, although the BMIDAS-AGL-SS seems to perform somewhat better across DGPs by mainly providing the smallest bias. This leads to highest TPR and lowest FPR for this model, entailing better classification of the active and inactive sets across simulations. Second, the MSE increases substantially with the degree of correlation in the design matrix (governed by the value of $\sigma_\varepsilon$), but it tends to decrease with more irrelevant predictors. To understand the latter result, it is useful to look at the breakdown of the MSE in both the active ($\mathcal{A}$) and inactive set ($\mathcal{A}^c$) reported in Figure 2. For comparison purpose, we also report results for the Oracle BMIDAS, estimated using the algorithm described in Pettenuzzo et al. (2016) on the set of relevant variables only.[8]

It turns out that while the share of variance and bias in the inactive set is broadly stable across simulations, the shares in the active set decrease substantially when $K$ increases. It follows that the performance of the models in selecting and estimating the coefficients of the relevant variables holds the same regardless the increase in the degree of sparsity, and hence the decrease in the share of variance and bias can be mainly attributed to the decrease in their relative weight (number of active predictors over $K$) in the total variance and bias. This result is confirmed by the TPR, which is relatively high and hovers around 80-90% for moderate correlation, and it's overall stable across the different values of $K$, suggesting that the models can select the correct sparsity pattern with a high probability even in finite samples. However, it is worth noting that the TPR drops

[7] For $R$ Monte Carlo replications, $K$ variables, and $S$ Gibbs draws, we have that:

$$\text{MSE} = \text{VAR} + \text{BIAS}^2 = \frac{1}{RKS}\sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{s=1}^{S}\left[\widehat{\beta}_k^{(s)} - \mathbb{E}\left(\widehat{\beta}_k\right)\right]^2 + \frac{1}{RK}\sum_{r=1}^{R}\sum_{k=1}^{K}\left[\mathbb{E}\left(\widehat{\beta}_k\right) - \beta_k\right]^2$$

where $\mathbb{E}\left(\widehat{\beta}_k\right) = \frac{1}{S}\sum_{s=1}^{S}\widehat{\beta}_k^{(s)}$ and $\widehat{\beta}_k = \boldsymbol{\iota}_C \mathbf{Q}'\boldsymbol{\theta}'_k$. Note that for the BMIDAS-AGL-SS model we use the median estimator.

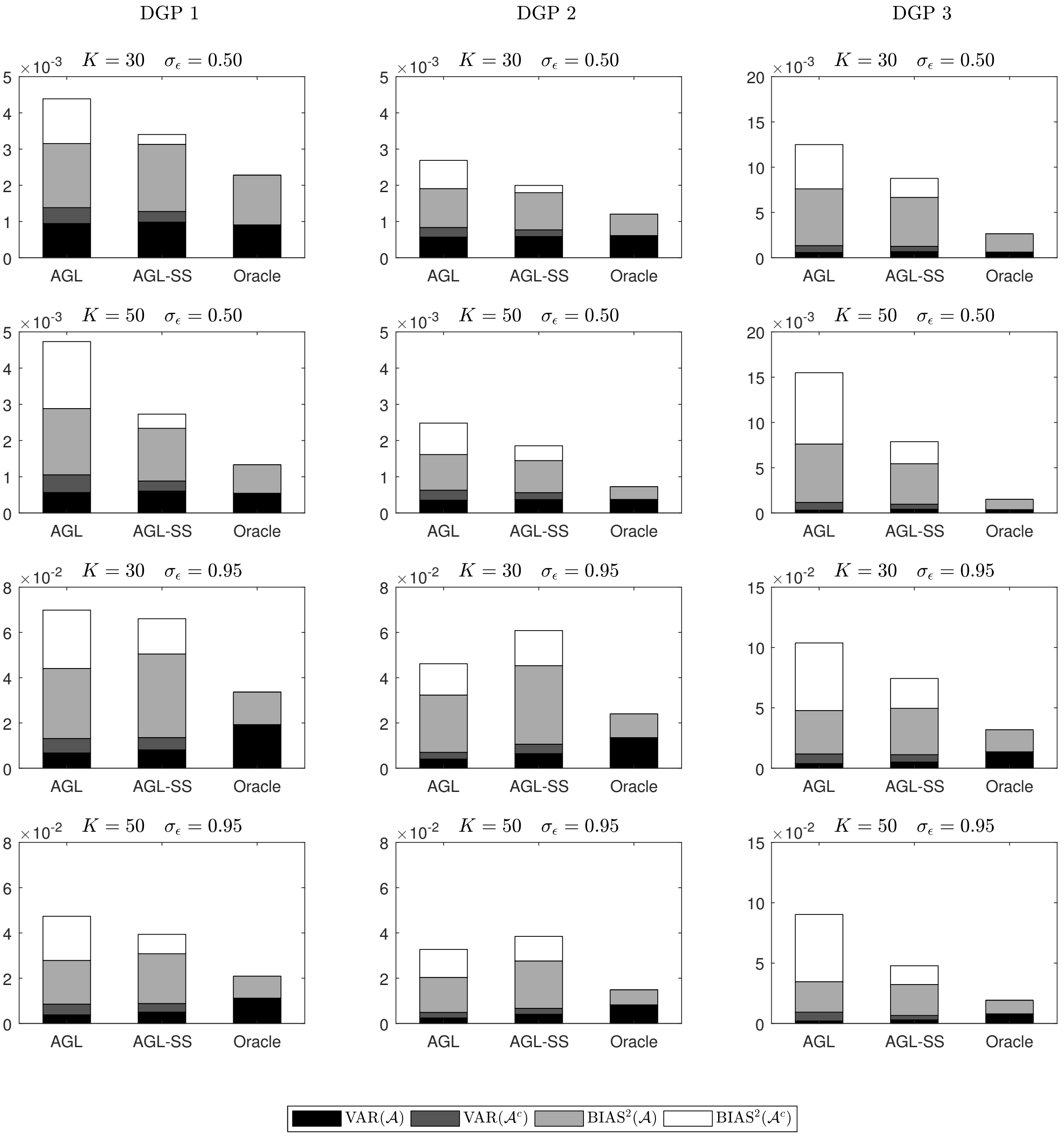
[8] We consider the same restricted Almon lag polynomial as for our models. Further, we follow Pettenuzzo et al. (2016) and we use relatively diffuse priors on both the coefficient covariance matrix and the regression variance. As for the prior mean coefficients, we set all the coefficients but the intercept to zero.

Table 1: Monte Carlo simulations

| $K$ | $\sigma_\varepsilon$ | $\mathbb{E}(\sigma)$ | MSE | VAR | BIAS$^2$ | TPR | FPR | MCC | MSFE | MAFE | $-$LS | CRPS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | DGP 1: $\boldsymbol{\vartheta} = (7 * 10^{-4}, -7 * 10^{-2})$ | | | | | | | | | |
| | | | BMIDAS-AGL | | | | | | | | | |
| 30 | 0.50 | 1.3 | 4.4E-03 | 1.4E-03 | 3.0E-03 | 0.95 | 0.03 | 0.90 | 2.12 | 1.17 | 1.79 | 0.83 |
| | 0.95 | 2.1 | 7.0E-02 | 1.3E-02 | 5.7E-02 | 0.35 | 0.04 | 0.41 | 6.20 | 1.99 | 2.32 | 1.40 |
| 50 | 0.50 | 1.3 | 4.7E-03 | 1.1E-03 | 3.7E-03 | 0.92 | 0.04 | 0.82 | 2.50 | 1.24 | 1.89 | 0.89 |
| | 0.95 | 2.1 | 4.7E-02 | 8.6E-03 | 3.9E-02 | 0.36 | 0.03 | 0.41 | 5.20 | 1.82 | 2.23 | 1.28 |
| | | | BMIDAS-AGL-SS | | | | | | | | | |
| 30 | 0.50 | 1.3 | 3.4E-03 | 1.3E-03 | 2.1E-03 | 0.95 | 0.01 | 0.94 | 2.09 | 1.14 | 1.78 | 0.81 |
| | 0.95 | 2.1 | 6.6E-02 | 1.4E-02 | 5.3E-02 | 0.37 | 0.02 | 0.47 | 6.36 | 1.95 | 2.32 | 1.40 |
| 50 | 0.50 | 1.3 | 2.7E-03 | 8.8E-04 | 1.8E-03 | 0.92 | 0.01 | 0.92 | 2.53 | 1.25 | 1.89 | 0.89 |
| | 0.95 | 2.1 | 3.9E-02 | 8.8E-03 | 3.1E-02 | 0.36 | 0.01 | 0.50 | 5.19 | 1.82 | 2.23 | 1.28 |
| | | | DGP 2: $\boldsymbol{\vartheta} = (7 * 10^{-4}, -9 * 10^{-3})$ | | | | | | | | | |
| | | | BMIDAS-AGL | | | | | | | | | |
| 30 | 0.50 | 1.1 | 2.7E-03 | 8.4E-04 | 1.9E-03 | 0.98 | 0.02 | 0.92 | 1.49 | 0.96 | 1.61 | 0.68 |
| | 0.95 | 1.8 | 4.6E-02 | 7.0E-03 | 3.9E-02 | 0.48 | 0.04 | 0.54 | 3.37 | 1.47 | 2.02 | 1.03 |
| 50 | 0.50 | 1.1 | 2.5E-03 | 6.3E-04 | 1.9E-03 | 0.97 | 0.03 | 0.87 | 1.57 | 0.99 | 1.62 | 0.70 |
| | 0.95 | 1.8 | 3.3E-02 | 4.9E-03 | 2.8E-02 | 0.44 | 0.03 | 0.48 | 3.74 | 1.49 | 2.08 | 1.08 |
| | | | BMIDAS-AGL-SS | | | | | | | | | |
| 30 | 0.50 | 1.1 | 2.0E-03 | 7.7E-04 | 1.2E-03 | 0.98 | 0.01 | 0.96 | 1.48 | 0.96 | 1.60 | 0.68 |
| | 0.95 | 1.8 | 6.1E-02 | 1.1E-02 | 5.0E-02 | 0.43 | 0.03 | 0.52 | 3.44 | 1.47 | 2.02 | 1.04 |
| 50 | 0.50 | 1.1 | 1.9E-03 | 5.6E-04 | 1.3E-03 | 0.97 | 0.01 | 0.96 | 1.47 | 0.96 | 1.59 | 0.68 |
| | 0.95 | 1.8 | 3.8E-02 | 6.7E-03 | 3.2E-02 | 0.42 | 0.01 | 0.54 | 3.54 | 1.45 | 2.05 | 1.05 |
| | | | DGP 3: $\boldsymbol{\vartheta} = (0, -5 * 10^{-4})$ | | | | | | | | | |
| | | | BMIDAS-AGL | | | | | | | | | |
| 30 | 0.50 | 1.0 | 1.2E-02 | 1.3E-03 | 1.1E-02 | 0.83 | 0.15 | 0.60 | 1.70 | 1.04 | 1.71 | 0.74 |
| | 0.95 | 1.6 | 1.0E-01 | 1.2E-02 | 9.2E-02 | 0.34 | 0.11 | 0.25 | 4.02 | 1.62 | 2.13 | 1.15 |
| 50 | 0.50 | 1.0 | 1.5E-02 | 1.2E-03 | 1.4E-02 | 0.74 | 0.16 | 0.43 | 1.81 | 1.10 | 1.74 | 0.77 |
| | 0.95 | 1.6 | 9.0E-02 | 9.4E-03 | 8.1E-02 | 0.29 | 0.10 | 0.18 | 4.29 | 1.62 | 2.17 | 1.17 |
| | | | BMIDAS-AGL-SS | | | | | | | | | |
| 30 | 0.50 | 1.0 | 8.8E-03 | 1.3E-03 | 7.5E-03 | 0.84 | 0.07 | 0.74 | 1.63 | 1.01 | 1.66 | 0.72 |
| | 0.95 | 1.6 | 7.4E-02 | 1.1E-02 | 6.3E-02 | 0.31 | 0.04 | 0.36 | 3.84 | 1.58 | 2.10 | 1.12 |
| 50 | 0.50 | 1.0 | 7.9E-03 | 9.7E-04 | 6.9E-03 | 0.79 | 0.06 | 0.67 | 1.72 | 1.05 | 1.69 | 0.74 |
| | 0.95 | 1.6 | 4.8E-02 | 6.6E-03 | 4.1E-02 | 0.28 | 0.02 | 0.36 | 3.88 | 1.55 | 2.09 | 1.11 |

Notes: BMIDAS-AGL and BMIDAS-AGL-SS refer to the models described in Section 3.

Figure 2: Breakdown of MSE by active ($\mathcal{A}$) and inactive set ($\mathcal{A}^c$)



Note: The oracle BMIDAS denotes the output from a regression including only the active set and using the algorithm described in Pettenuzzo et al. (2016) for parameters estimation.

Table 2: Monte Carlo simulations: constrained vs unconstrained weighting schemes

| | | BMIDAS-AGL | | | BMIDAS-AGL-SS | | |
|---|---|---|---|---|---|---|---|
| $K$ | $\sigma_\varepsilon$ | TPR | FPR | MCC | TPR | FPR | MCC |
| | | DGP 1: $\boldsymbol{\vartheta} = (7 * 10^{-4}, -7 * 10^{-2})$ | | | | | |
| 30 | 0.50 | 0.16 | -0.02 | 0.15 | 0.13 | -0.01 | 0.11 |
| | 0.95 | 0.06 | -0.01 | 0.08 | 0.05 | 0.00 | 0.05 |
| 50 | 0.50 | 0.21 | -0.02 | 0.20 | 0.12 | 0.00 | 0.09 |
| | 0.95 | 0.08 | 0.00 | 0.09 | 0.06 | 0.00 | 0.08 |
| | | DGP 2: $\boldsymbol{\vartheta} = (7 * 10^{-4}, -9 * 10^{-3})$ | | | | | |
| 30 | 0.50 | 0.10 | -0.02 | 0.11 | 0.07 | 0.00 | 0.07 |
| | 0.95 | 0.11 | -0.01 | 0.13 | 0.07 | 0.00 | 0.06 |
| 50 | 0.50 | 0.18 | -0.03 | 0.19 | 0.07 | 0.00 | 0.06 |
| | 0.95 | 0.10 | 0.00 | 0.11 | 0.08 | 0.00 | 0.07 |
| | | DGP 3: $\boldsymbol{\vartheta} = (0, -5 * 10^{-4})$ | | | | | |
| 30 | 0.50 | -0.10 | 0.11 | -0.26 | -0.10 | 0.05 | -0.18 |
| | 0.95 | -0.08 | 0.06 | -0.20 | -0.10 | 0.02 | -0.13 |
| 50 | 0.50 | -0.12 | 0.11 | -0.31 | -0.13 | 0.05 | -0.23 |
| | 0.95 | -0.11 | 0.07 | -0.27 | -0.12 | 0.01 | -0.16 |

Notes: See Table 1.

to 30-50% for very high correlation, while the the FPR remains overall very low. This result is nevertheless not unexpected, as the Group Lasso can address the issue of strong collinearity within the lag polynomials but is not designed to handle strong collinearity between the high-frequency regressors.

Third, not surprisingly, the in-sample results deteriorate when the DGP with near-flat weights is considered, and mostly when $\sigma_\varepsilon = 0.95$. This happens because the linear restrictions imposed on the lag polynomials incorrectly force the weighting structure to tail off to zero, while the weighting scheme under the null is almost uniform over the lag window $C$. It follows that relaxing the restrictions on the lag polynomial should lead to an improvement of the results under DGP 3. However, it is not clear how much those linear restrictions actually contribute to the selection results under DGPs 1 and 2. Table 2 provides an answer to these questions by reporting the difference in TPR, FPR, and MCC obtained with restricted ($r = 2$) and unrestricted ($r = 0$) lag polynomials. For DGPs with fast- or slow-decaying wights, the results suggest that imposing correct linear restrictions that are valid under the null seems to improve the selection ability of the models. The gain in terms of TPR ranges 10-20 percentage points for moderate correlation and 5-10 percentage points for very high correlation in the design matrix, while the gain in terms of FPR ranges 1-3 percentage point. Interestingly enough, the results reveal that the BMIDAS-AGL-SS model is relatively less affected than the BMIDAS-AGL model by the inclusion of linear restrictions. For the DGP with near-flat weights we observe an opposite outcome, as expected.

However, the magnitude of these results must be considered with care, as the number of relevant and irrelevant predictors in the true model is strongly asymmetric.

Finally, looking at the forecasting performance, the results are broadly in line with the in-sample analysis and suggest that the models perform overall quite similarly in terms of point and density forecasts, although the BMIDAS-AGL-SS model seems to perform best overall. The performance of the models deteriorates substantially with higher correlation in the design matrix, although a higher average variance in the error process, $\mathbb{E}(\sigma)$, must be discounted to explain the large differences highlighted in Table 1, but it is relatively stable with $K$ increasing.

## 6. Empirical application

We apply the proposed Bayesian MIDAS penalized regression approaches to US GDP data. Following the literature, we consider the annualized quarterly growth rate of GDP. As for the predictors, we consider a subset of 33 macroeconomic series extracted from the FRED-MD database (McCracken and Ng, 2016) and selected to provide high-frequency information on potential predictors of GDP, such as output and income, labor, housing, consumption, and orders. Further, we also consider a set of daily and weekly financial data, which have proven to improve short- to medium-term macro forecasts (Andreou et al., 2013; Pettenuzzo et al., 2016; Adrian et al., in press): the effective Federal Funds rate; the interest rate spread between the 10-year government bond rate and the Federal Funds rate; returns on the portfolio of small minus big stocks considered by Fama and French (1993); returns on the portfolio of high minus low book-to-market ratio stocks studied by Fama and French (1993); returns on a winner minus loser momentum spread portfolio; the Chicago Fed National Financial Conditions Index (NFCI), and in particular its three sub-indexes (risk, credit and leverage). Finally, we consider the Aruoba-Diebold-Scotti (ADS) daily business conditions index (Aruoba et al., 2009) to track the real business cycle at high frequency. To match the sample frequencies, we consider again a restricted Almon lag polynomial, with $p = 3$ and $r = 2$ endpoint restrictions, and twelve months of past high-frequency observations ($C = 12$). Overall, the total number of predictors entering the models is $K = 42$ (the full list of predictors is reported in Appendix A.2). The data sample starts in 1980Q1, and we set $\underline{T} = 2000\text{Q1}$ and $\overline{T} = 2017\text{Q4}$ the first and last out-of-sample observations, respectively. Estimates are carried-out recursively using an expanding window, and $h$-step-ahead posterior predictive densities are generated from (12) through a direct forecast approach. We hence dispose of $(\overline{T} - \underline{T} + 1) = 72$ out-of-sample observations. We consider three forecast horizons, namely the nowcast ($h = 0$), and 1-quarter and 4-quarters ahead forecasts ($h = 1$ and $h = 4$, respectively). For ease of analysis, we do not take into account real-time issues (ragged/jagged-edge data, revisions) and we compile the dataset using the latest vintages available at the time of writing.

Forecasts are compared to those from a benchmark model represented by a simple random-walk (RW). Point forecasts are evaluated by the means of relative RMSFE ratios:

$$\Delta\text{RMSFE} = \sqrt{\frac{\sum_{t=\underline{T}}^{\overline{T}} e_t^2}{\sum_{t=\underline{T}}^{\overline{T}} e_{\text{RW},t}^2}}$$

where $e_{\text{RW},t}$ denotes the forecast error generated by the benchmark model. Hence, values less than one suggest that our penalized mixed-frequency models outperform (in a point forecast sense) the RW. Density forecasts (generated by the draws from the posterior predictive distribution) are evaluated by the means of the average log-score differential:

$$\Delta\text{LS} = (\overline{T} - \underline{T} + 1)^{-1} \sum_{t=\underline{T}}^{\overline{T}} (\text{LS}_t - \text{LS}_{\text{RW},t})$$

Positive values of $\Delta$LS indicate that our models produce more accurate density forecasts than the RW. Further, we compute the average continuously ranked probability score (CRPS) ratio:

$$\Delta\text{CRPS} = \frac{\sum_{t=\underline{T}}^{\overline{T}} \text{CRPS}_t}{\sum_{t=\underline{T}}^{\overline{T}} \text{CRPS}_{\text{RW},t}}$$

where values less than one suggest that our models outperform (in a density forecast sense) the benchmark model. Further, to account for sample uncertainty underlying the observed forecast differences, we report results for the Diebold and Mariano (1995) and West (1996) test (DMW hereafter), which posits the null hypothesis of an unconditional equal predictive accuracy between each model and the benchmark random-walk. The resulting test statistic is computed using HAC standard errors (for $h = 4$) and a small-sample adjustment to the consistent estimate of the variance, and compared with critical values from the Student's $t$ distribution with $(\overline{T} - \underline{T})$ degrees of freedom (Harvey et al., 1997). As a robustness check, we further consider forecasts from the following competing models:

- AR(1) model.

- Combination of $K$ single-indicator Bayesian MIDAS models as in Pettenuzzo et al. (2016) (BMIDAS-comb), where the combination weights are computed using a discounted version of the optimal prediction pool proposed by Geweke and Amisano (2011). The historical performance of each individual model is hence accounted for by attaching a greater weight to recent predictive outcomes through a discount factor $\delta < 1$, that we set at 0.9 (Stock and Watson, 2004; Andreou et al., 2013). Note that $\delta = 1$ (no discounting) corresponds to the Geweke and Amisano (2011) optimal prediction pool.

Table 3: Out-of-sample forecast performance

| | $h = 0$ | | | $h = 1$ | | | $h = 4$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | ΔRMSFE | ΔLS | ΔCRPS | ΔRMSFE | ΔLS | ΔCRPS | ΔRMSFE | ΔLS | ΔCRPS |
| BMIDAS-AGL | 0.61 | 0.54 | 0.59 | 0.74 | 0.33 | 0.72 | 0.82 | 0.24 | 0.81 |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.10) | (0.07) | (0.06) |
| BMIDAS-AGL-SS | **0.57** | **0.58** | **0.56** | **0.70** | **0.39** | **0.68** | 0.81 | 0.24 | 0.77 |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.10) | (0.06) | (0.03) |
| AR(1) | 0.85 | 0.16 | 0.82 | 0.85 | 0.16 | 0.82 | 0.80 | 0.22 | 0.76 |
| BMIDAS-comb ($\delta = 0.9$) | 0.66 | 0.44 | 0.65 | 0.77 | 0.30 | 0.74 | **0.78** | **0.27** | **0.76** |
| BMS-ADMH | 1.18 | -0.67 | 1.32 | 1.12 | -0.15 | 1.15 | 0.85 | 0.18 | 0.87 |
| BMA ($g$-BRIC) | 0.61 | 0.48 | 0.59 | 0.76 | 0.30 | 0.73 | 0.84 | 0.16 | 0.80 |

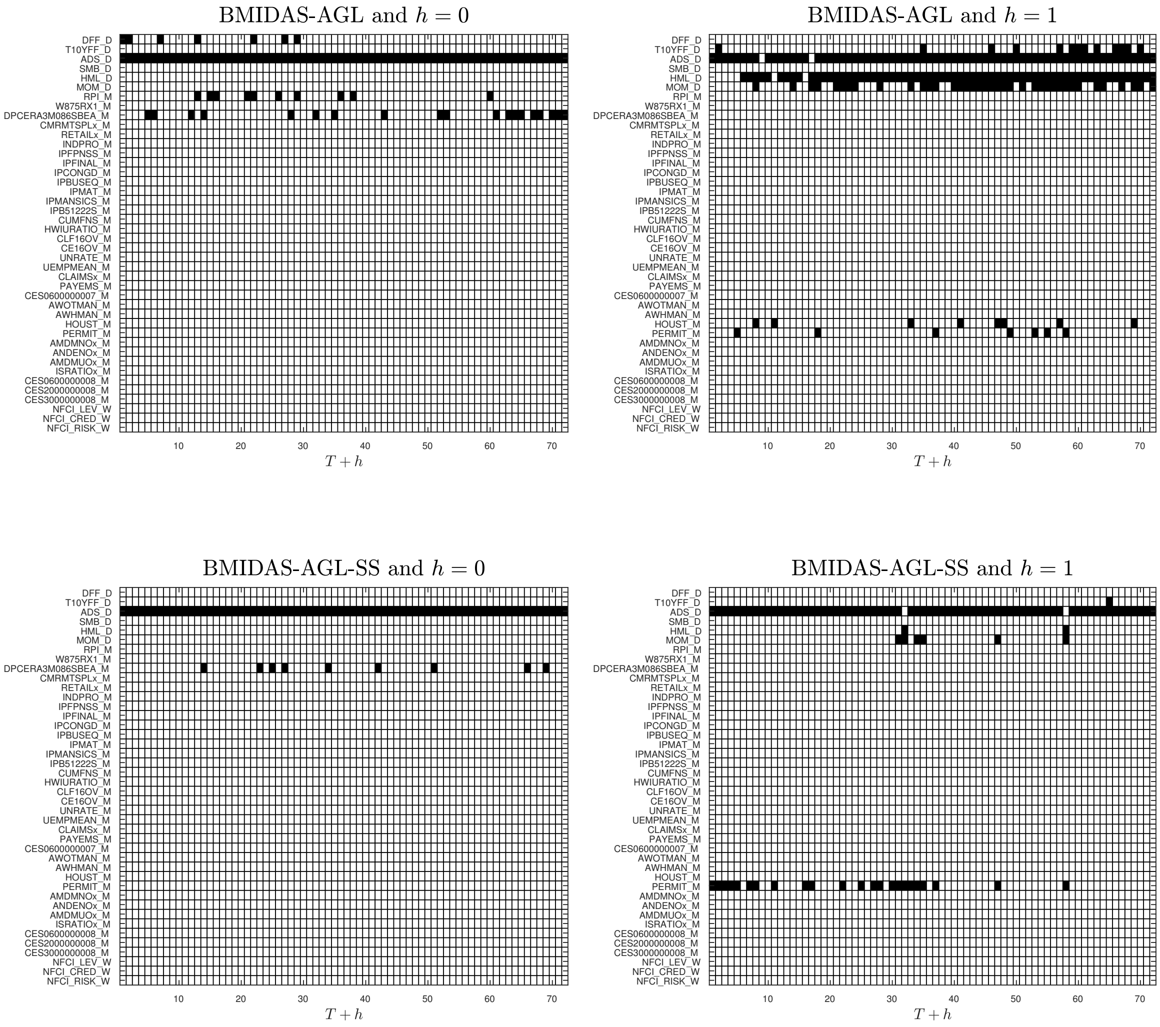
Notes: predictive performance of model $i$ compared to the random-walk benchmark. Bold values denote the best outcomes. In parentheses, $p$-values for the test of the null hypothesis of equal predictive accuracy at 10% level according to the one-sided $t$-statistic version of the DMW test.

- Bayesian model selection approach of Lamnisos et al. (2013), where an Adaptive Metropolis-Hastings algorithm is implemented to tune automatically the model proposals and achieve a targeted acceptance rate (BMS-ADMH). We use the default priors set by Lamnisos et al. (2013), and we select those variables displaying a posterior probability of inclusions greater than 50%.
- Bayesian model averaging (BMA), as in Rossi and Sekhposyan (2014), estimated using a standard $g$-BRIC prior, where $g = \max(T, K^2)$, and a reversible-jump MC$^3$ algorithm.

Both BMS-ADMH and BMA algorithms are here modified to account for groups of lag polynomials in their addition/deletion/swaption moves, to ensure that model proposals are based on selection of individual predictors rather than isolated terms of the lag polynomials. Further, all the models considered in the application always include a lag of the growth rate of GDP, which is hence excluded from the selection procedures. As for the MCMC, the Gibbs sampler is run for $S = 300,000$ iterations, with the first $100,000$ used as a burn-in period, and every 10th draw is saved. For the BMS-ADMH and BMA, we increase the number of iterations to $2,000,000$, in order to let the algorithms sufficiently explore the model space, which is fairly vast in the current application ($2^K = 4.4\text{e}{+}12$).

Results are reported in Table 3. Considering point forecasts, our findings suggest that the penalized BMIDAS models outperform the benchmark RW at all the horizons, with statistically significant (at 10% level) predictive gains hovering around 40% for $h = 0$, 30% for $h = 1$, and 20% for $h = 4$. The results for density forecasts are broadly in line with those for point forecasts, with statistically significant predictive gains decreasing almost linearly with the increase in the number of steps-ahead. When compared to the set of alternative models, our penalized BMIDAS models display predictive gains at $h = 0$ and $h = 1$. At these horizons, both point and density gains hover around 15-30% against the AR(1), but this rate decreases when more sophisticated benchmarks are considered. In particular, the BMA and the combined univariate BMIDAS models appear overall the best competitors, while the results from the BMS are disappointing. At the longer-horizon ($h = 4$), the predictive performance of most of the alternative competing models is only slightly

Figure 3: Variable selection



superior or inferior to that of our penalized regressions, with combined univariate BMIDAS models providing the best outcome.

The variables selected overtime by our models are reported in Figure 3. For ease of exposition, we consider only the short-term horizons. The selection patterns show a systematic inclusion of the ADS index and, although sporadically, a bunch of real high-frequency indicators related to the real personal consumption expenditures and the housing market. Further, selection appears more parsimonious and stable over the out-of-sample for the BMIDAS-AGL-SS model. Interestingly enough, virtually no financial indicators are selected by our models at $h = 0$. This can be attributed to the fact that the empirical analysis is carried out with information available over the whole quarter and by abstracting from real-time conditions, such that real hard- and soft-data may possibly convey enough information. However, this feature tends to attenuate for $h = 1$, where some high-frequency financial indicators, such as the the portfolio Hml and Mom indicators, are selected. All in all, this result is broadly in line with recent literature (Andreou et al., 2013) and suggests that financial variables may convey some, although limited, short-term leading information which goes beyond the predictive content of real indicators.

## 7. Concluding remarks

We proposed a new approach to modeling and forecasting mixed-frequency regressions (MIDAS) that addresses the issue of simultaneously estimating and selecting relevant high-frequency predictors in a high-dimensional environment. Our approach is based on MIDAS regressions resorting to Almon lag polynomials and an adaptive penalized regression approach, namely the Group Lasso objective function. The proposed models rely on Bayesian techniques for estimation and inference. In particular, the penalty hyper-parameters driving the model shrinkage are automatically tuned via an Empirical Bayes algorithm based on stochastic approximations. Simulations show that the proposed models present very good in-sample and out-of-sample performance. When applied to a forecasting model of U.S. GDP with high-frequency real and financial predictors, the results suggest that our models produce significant out-of-sample short-term predictive gains compared to several alternative models. Further, our findings are broadly in line with the existing literature, in the extent that high-frequency financial variables have non-zero, although limited, short-term predictive content. The models presented in the present paper could be extended in several ways. We nonetheless believe that considering a time-varying process of the parameters characterizing the lag polynomials as well as stochastic volatility error dynamics (Carriero et al., 2015; Schumacher, 2015; Pettenuzzo et al., 2016), or a quantile specification of the mixed-frequency regression, represent interesting paths for future research.

## References

Adrian, T., Boyarchenko, N., Giannone, D., in press. Vulnerable growth. American Economic Review.

Andreou, E., Ghysels, E., Kourtellos, A., 2010. Regression models with mixed sampling frequencies. Journal of Econometrics 158 (2), 246–261.

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.

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

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

Barbieri, M., Berger, J., 2004. Optimal predictive model selection. The Annals of Statistics 32 (3), 870–897.

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.

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.

Castillo, I., Schlidt-Hieber, J., Van der Vaart, A., 2015. Bayesian linear regression with sparse priors. The Annals of Statistics 43 (5), 1986–2018.

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.

Diebold, F. X., Mariano, R. S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13 (3), 253–263.

Doornik, J. A., 2009. Autometrics. In: Castle, J. L., Shephard, N. (Eds.), The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry. Oxford (UK): Oxford University Press, pp. 88–121.

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

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33 (1), 3–56.

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.

George, E. I., McCulloch, R. E., 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88 (423), 881–889.

Geweke, J., Amisano, G., 2011. Optimal prediction pools. Journal of Econometrics 164 (1), 130–141.

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.

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.

Harvey, D., Leybourne, S., Newbold, P., 1997. Testing the equality of prediction mean squared errors. International Journal of Forecasting 13 (2), 281–291.

Hobert, J. P., Geyer, C. J., 1998. Geometric ergodicity of Gibbs and Block Gibbs samplers for a hierarchical random effects model. Journal of Multivariate Analysis 67 (2), 414–430.

Khare, K., Hobert, J. P., 2013. Geometric ergodicity of the Bayesian lasso. Electronic Journal of Statistics 7, 2150–2163.

Koop, G., Korobilis, D., Pettenuzzo, D., in press. Bayesian compressed vector autoregressions. Journal of Econometrics.

Korobilis, D., 2013. Hierarchical shrinkage priors for dynamic regressions with many predictors. International Journal of Forecasting 29 (1), 43–59.

Korobilis, D., Pettenuzzo, D., in press. Adaptive hierarchical priors for high-dimensional vector autoregressions. Journal of Econometrics.

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

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

Lamnisos, D., Griffin, J. E., Steel, M. F. J., 2013. Adaptive Monte Carlo for Bayesian variable selection in regression models. Journal of Computational and Graphical Statistics 22 (3), 729–748.

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.

Marsilli, C., 2014. Variable selection in predictive MIDAS models. Working Paper 520, Banque de France.

McCracken, M. W., Ng, S., 2016. FRED-MD: A monthly database for macroeconomic research. Journal of Business & Economic Statistics 34 (4), 574–589.

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

Mitchell, T. J., Beauchamp, J. J., 1988. Bayesian variable selection in linear regression (with discussion). Journal of the American Statistical Association 83 (404), 1023–1032.

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.

Ročková, V., George, E. I., 2018. The spike-and-slab LASSO. Journal of the American Statistical Association 113 (521), 431–444.

Rossi, B., Sekhposyan, T., 2014. Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set. International Journal of Forecasting 30 (3), 662–682.

Roy, V., Chakraborty, S., 2017. Selection of tuning parameters, solution paths and standard errors for Bayesian Lassos. Bayesian Analysis 12 (3), 753–778.

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.

Smith, R. G., Giles, D. E. A., 1976. The Almon estimator: Methodology and users' guide. Discussion Paper E76/3, Reserve Bank of New Zealand.

Stock, J. H., Watson, M. W., 2004. Combination forecasts of output growth in a seven-country data set. Journal of Forecasting 23 (6), 405–430.

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

Uematsu, Y., Tanaka, S., in press. High-dimensional macroeconomic forecasting and variable selection via penalized regression. The Econometrics Journal.

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

West, K. D., 1996. Asymptotic inference about predictive ability. Econometrica 64 (5), 1067–1084.

Xu, X., Ghosh, M., 2015. Bayesian variable selection and estimation for group lasso. Bayesian Analysis 10 (4), 909–936.

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.

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., Do, K.-A., 2014. Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. Journal of the Royal Statistical Society: Series C (Applied Statistics) 63 (4), 595–620.

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

Zhao, Z., Sarkar, S. K., 2015. A Bayesian approach to constructing multiple confidence intervals of selected parameters with sparse signals. Statistica Sinica 25 (2), 725–741.

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), 1733–1751.

## Appendix

### *A.1. Stabilization algorithm*

The stabilization algorithm used in the paper is a slightly modified version of the algorithm proposed in Andrieu et al. (2005) and discussed in Atchadé (2011), which uses re-projections on randomly varying compact sets. Recall that the updating (approximate EM) algorithm described in Section 4 is:

$$\boldsymbol{\omega}^{(s+1)} = \boldsymbol{\omega}^{(s)} + a^{(s)} H(\boldsymbol{\omega}^{(s)}, \boldsymbol{\phi}^{(s+1)})$$

where we use the transformation $\boldsymbol{\omega} = 0.5\log(\boldsymbol{\lambda})$. Let $\{a^{(s)}, s \geq 0\}$ and $\{e^{(s)}, s \geq 0\}$ be two monotone non-increasing sequences of positive numbers. Here we choose $a^{(s)} = 1/s^q$, with $q = 0.8$, and $e^{(s)} = \bar{e} + (1-\bar{e})(1-\varsigma_s^{-\alpha_e})$, with $\bar{e} = 3$ and $\alpha_e = 0.1$. Let $\{\mathsf{K}^{(s)}, s \geq 0\}$ be a monotone increasing sequence of compact subsets of $\boldsymbol{\Omega}$ such that $\bigcup_{s\geq 0} \mathsf{K}^{(s)} = \boldsymbol{\Omega}$. Here we set compact subsets of the form $\mathsf{K}^{(s)} = [\max(-\kappa_s - 1, -c), \kappa_s + 1]$, where $c > 0$. To avoid unstable outcomes due to extremely small numbers in $\boldsymbol{\lambda}$, we set $c = 5$. Let $\widetilde{\boldsymbol{\Omega}} \times \widetilde{\boldsymbol{\Phi}} \subset \mathsf{K}^{(s)} \times \boldsymbol{\Phi}$ and $\Pi : \boldsymbol{\Omega} \times \boldsymbol{\Phi} \to \widetilde{\boldsymbol{\Omega}} \times \widetilde{\boldsymbol{\Phi}}$ be a re-projection function, such as $\widetilde{\boldsymbol{\Omega}} \times \widetilde{\boldsymbol{\Phi}} = (\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\omega}})$ for an arbitrary point $(\widetilde{\boldsymbol{\omega}}, \widetilde{\boldsymbol{\phi}}) \in \mathsf{K}^{(s)} \times \widetilde{\boldsymbol{\Phi}}$. Let $\varphi$ be a function such that $\varphi(w) = 1 - w$, for all $w \geq 0$.

**Algorithm 1** Stochatistic approximation with truncation on random boundaries

```
Set $\kappa_0 = 0$, $\nu_0 = 0$, $\varsigma_0 = 0$, $\boldsymbol{\omega}^{(0)} \in \boldsymbol{\Omega}$, and $\boldsymbol{\phi}^{(0)} \in \boldsymbol{\Phi}^{(0)}$.
For $s \geq 1$, compute:
(a) $\overline{\boldsymbol{\phi}} \sim P_{\boldsymbol{\omega}^{(s-1)}}(\boldsymbol{\phi}^{(s-1)}, \cdot)$
(b) $\overline{\boldsymbol{\omega}} = \boldsymbol{\omega}^{(s-1)} + a^{(\varsigma_{s-1}+1)} H(\boldsymbol{\omega}^{(s-1)}, \overline{\boldsymbol{\phi}})$,
where $P_{\boldsymbol{\omega}}$ is the Markov kernel.
if $\overline{\boldsymbol{\omega}} \in \mathsf{K}^{(\kappa_{s-1})}$ and $|\overline{\boldsymbol{\omega}} - \boldsymbol{\omega}^{(s-1)}| \leq e^{(\varsigma_{s-1})}$ then
    $(\boldsymbol{\omega}^{(s)}, \boldsymbol{\phi}^{(s)}) = (\overline{\boldsymbol{\omega}}, \overline{\boldsymbol{\phi}})$
    $\kappa_s = \kappa_{s-1}$, $\nu_s = \nu_{s-1} + 1$, $\varsigma_s = \varsigma_{s-1} + 1$
else
    $(\boldsymbol{\omega}^{(s)}, \boldsymbol{\phi}^{(s)}) = (\widetilde{\boldsymbol{\omega}}, \widetilde{\boldsymbol{\phi}}) \in \widetilde{\boldsymbol{\Omega}} \times \widetilde{\boldsymbol{\Phi}}$
    $\kappa_s = \kappa_{s-1} + 1$, $\nu_s = 0$, $\varsigma_s = \varsigma_{s-1} + \varphi(\nu_{s-1})$
end
```

With this algorithm, $\kappa_s$ is the index of the active truncation set (also equal to the number of restarts before $s$), $\nu_s$ is the number of iterations since the last restart, and $\varsigma_s$ is the current index in the step-size sequence. We set $\varphi(w) = 1$ for all $w \in \mathbb{N}$, such that $\varsigma_s = s$. Hence, if $\overline{\boldsymbol{\omega}} \notin \mathsf{K}^{(\kappa_{s-1})}$

or $|\overline{\boldsymbol{\varpi}} - \boldsymbol{\omega}^{(s-1)}| > e^{(\varsigma_{s-1})}$, we re-initialize the algorithm starting from $(\widetilde{\boldsymbol{\omega}}, \widetilde{\boldsymbol{\phi}})$, which are obtained by drawing from:

$$\begin{aligned}
\widetilde{\boldsymbol{\omega}} &\sim \text{Uniform}\left(\min(\boldsymbol{\omega}^{(s-1)}, \mathsf{K}_u^{(s-1)}), \max(\boldsymbol{\omega}^{(s-1)}, \mathsf{K}_u^{(s-1)})\right) \qquad \text{if } \overline{\boldsymbol{\varpi}} \geq \mathsf{K}_u^{(s-1)} \\
\widetilde{\boldsymbol{\omega}} &\sim \text{Uniform}\left(\min(\boldsymbol{\omega}^{(s-1)}, \mathsf{K}_l^{(s-1)}), \max(\boldsymbol{\omega}^{(s-1)}, \mathsf{K}_l^{(s-1)})\right) \qquad \text{if } \overline{\boldsymbol{\varpi}} < \mathsf{K}_l^{(s-1)}
\end{aligned}$$

where $\mathsf{K}_u^{(s-1)} = \kappa_{s-1} + 1$ and $\mathsf{K}_l^{(s-1)} = \max(-\kappa_{s-1} - 1, -c)$, and parameters $\boldsymbol{\phi}|\widetilde{\boldsymbol{\omega}}$ are drawn from the prior distributions described in Sections 3.1 and 3.2. We then iterate until the acceptance conditions stated in the algorithm are met. Finally, we set the new compact subsets to $\mathsf{K}^{(\kappa_{s-1}+1)}$ and the new sequence of step-size.

## *A.2. US Data*

| Code | Description | Frequency | Transformation |
|---|---|---|---|
| FF | Effective Federal Funds rate | Daily | $\Delta x_t$ |
| T10YFF | Spread 10-year government bond rate and Federal Funds rate | Daily | $x_t$ |
| ADS | Aruoba-Diebold-Scotti (ADS) daily business conditions index | Daily | $x_t$ |
| SMB | Returns on the portfolio of small minus big stocks | Daily | $x_t$ |
| HML | Returns on the portfolio of high minus low book-to-market ratio stocks | Daily | $x_t$ |
| MOM | Returns on a winner minus loser momentum spread portfolio | Daily | $x_t$ |
| NFCI_LEV | Chicago Fed National Financial Conditions Index - Leverage | Weekly | $x_t$ |
| NFCI_CRED | Chicago Fed National Financial Conditions Index - Credit | Weekly | $x_t$ |
| NFCI_RISK | Chicago Fed National Financial Conditions Index - Risk | Weekly | $x_t$ |
| RPI | Real Personal Income | Monthly | $\Delta \log(x_t)$ |
| W875RX1 | Real personal income ex transfer receipts | Monthly | $\Delta \log(x_t)$ |
| DPCERA3M086SBEA | Real personal consumption expenditures | Monthly | $\Delta \log(x_t)$ |
| CMRMTSPLx | Real Manu. and Trade Industries Sales | Monthly | $\Delta \log(x_t)$ |
| RETAILx | Retail and Food Services Sales | Monthly | $\Delta \log(x_t)$ |
| INDPRO | IP Index | Monthly | $\Delta \log(x_t)$ |
| IPFPNSS | IP: Final Products and Nonindustrial Supplies | Monthly | $\Delta \log(x_t)$ |
| IPFINAL | IP: Final Products (Market Group) | Monthly | $\Delta \log(x_t)$ |
| IPCONGD | IP: Consumer Goods | Monthly | $\Delta \log(x_t)$ |
| IPBUSEQ | IP: Business Equipment | Monthly | $\Delta \log(x_t)$ |
| IPMAT | IP: Materials | Monthly | $\Delta \log(x_t)$ |
| IPMANSICS | IP: Manufacturing (SIC) | Monthly | $\Delta \log(x_t)$ |
| IPB51222S | IP: Residential Utilities | Monthly | $\Delta \log(x_t)$ |
| CUMFNS | Capacity Utilization: Manufacturing | Monthly | $\Delta x_t$ |
| HWIURATIO | Ratio of Help Wanted/No. Unemployed | Monthly | $\Delta x_t$ |
| CLF16OV | Civilian Labor Force | Monthly | $\Delta \log(x_t)$ |
| CE16OV | Civilian Employment | Monthly | $\Delta \log(x_t)$ |
| UNRATE | Civilian Unemployment Rate | Monthly | $\Delta x_t$ |
| UEMPMEAN | Average Duration of Unemployment (Weeks) | Monthly | $\Delta x_t$ |
| CLAIMSx | Initial Claims | Monthly | $\Delta \log(x_t)$ |
| PAYEMS | All Employees: Total nonfarm | Monthly | $\Delta \log(x_t)$ |
| CES0600000007 | Avg Weekly Hours : Goods-Producing | Monthly | $x_t$ |
| AWOTMAN | Avg Weekly Overtime Hours : Manufacturing | Monthly | $\Delta x_t$ |
| AWHMAN | Avg Weekly Hours : Manufacturing | Monthly | $x_t$ |
| HOUST | Housing Starts: Total New Privately Owned | Monthly | $\log(x_t)$ |
| PERMIT | New Private Housing Permits (SAAR) | Monthly | $\log(x_t)$ |
| AMDMNOx | New Orders for Durable Goods | Monthly | $\Delta \log(x_t)$ |
| ANDENOx | New Orders for Nondefense Capital Goods | Monthly | $\Delta \log(x_t)$ |
| AMDMUOx | Unfilled Orders for Durable Goods | Monthly | $\Delta \log(x_t)$ |
| ISRATIOx | Total Business: Inventories to Sales Ratio | Monthly | $\Delta x_t$ |
| CES0600000008 | Avg Hourly Earnings : Goods-Producing | Monthly | $\Delta^2 \log(x_t)$ |
| CES2000000008 | Avg Hourly Earnings : Construction | Monthly | $\Delta^2 \log(x_t)$ |
| CES3000000008 | Avg Hourly Earnings : Manufacturing | Monthly | $\Delta^2 \log(x_t)$ |