Lasso variable selection in predictive mixed-frequency model

Clément Marsilli*

Preliminary draft, February 19, 2018

Abstract

In short-term forecasting, it is essential to take into account all available information on the current state of the economic activity. Yet, the fact that various time series are sampled at different frequencies prevents an efficient use of available data. In this respect, the *Mixed-Data Sampling* (MIDAS) model has proved to outperform existing tools by combining data series of different frequencies. However, a major issue remain regarding the choice of explanatory variables. The paper addresses this point by developing MIDAS based dimension reduction techniques and by introducing a novel approach based on a method of penalized variable selection, the Lasso. This feature integrates a cross-validation procedure that allows automatic in-sample selection based on forecasting performances. Then the developed technique is assessed with regards to its forecasting power of US economic growth during the period 1990-2015 using an augmented version of the Stock and Watson database jointly involving daily, weekly, monthly, and quarterly data as the real-time economic data-flow. Our model succeeds in identifying leading indicators and constructing an objective variable selection with broad applicability.

Keywords Forecasting, mixed frequency data, MIDAS, variable selection, Lasso, US growth, US database.

JEL Codes C53, E37.

*Banque de France, International Macroeconomics Division, clement.marsilli@banque-france.fr. I would like to thank Juan-Pablo Ortega, Laurent Ferrara, Gong Cheng, Eric Ghysels, Domenico Giannone, Claude Lopez, Luc Bauwens, Frederic Bec, Massimiliano Marcellino, Christian Schumacher, Marie Bessec, Lyudmila Grigoryeva and Matteo Mogliani for their very helpful comments. The views expressed herein are those of the authors and do not necessarily reflect those of the Banque de France and the Eurosystem.

1 Introduction

Short-term analysis aims at providing forecasts based on all the available information, it is usually required to use data sampled at different frequencies. Recent literature exploits either large panels of time series, by using factor models for instance, or mixed-frequency setup, such as the MIDAS. A growing literature suggests that adding more variables do not necessarily lead to better forecasting results. In macroeconomic forecasting, empirical models are generally based on dimension reduction methods. The main goal of this paper is introducing one variable selection method, the Lasso, within the mixed-frequency framework for macroeconomic forecasting.

We focus on the Mixed Data Sampling (MIDAS) method that allows the use of high frequency series to explain low frequency variables. Introduced by Ghysels et al. (2002) and more formally conceptualized by Ghysels et al. (2007) and by Andreou et al. (2010), the MIDAS approach constitutes a parsimonious weighting framework for handling distributed lags. This method allows us to explain a low frequency variable by using exogenous variables sampled at higher frequencies without resorting to any aggregation procedure. It has proved to be particularly suitable in macroeconomic forecasting and in capturing early signals of turning points using multifrequency explanatory variables (see for example Ferrara and Marsilli, 2013). Furthermore, the MIDAS regression has been used to predict quarterly GDP fluctuations using both monthly real economic data and daily financial series; Andreou et al. (2013) and Ferrara et al. (2014) showed that this combination of information significantly improves the prediction results. Many works prove that an appropriate selection of explanatory variables, regardless of their sampling frequencies, has a major impact on the performance of this forecasting method. In this respect, Banbura et al. (2012) have recently reviewed the existing mixed-frequency models designed for handling immediate past data (usually referred to as *ragged edge* data) and for nowcasting.

Based on both distributed lags and temporal aggregation literature, Ghysels and his coauthors have developed the MIDAS regression model (see Ghysels et al., 2002; Andreou et al., 2010). The MIDAS aims at accommodating with parsimony data sampled at different frequencies without any aggregation procedure. The standard MIDAS regression for explaining the stationary low-frequency variable y_t using a set of n high-frequency covariates is given by:

$$y_t = \alpha + \sum_{i=1}^{n} \beta_i \; m_{K_i}(\theta, L)\, x_{t,i}^{(\kappa_i)} + \varepsilon_t, \tag{1}$$

where $x_{i,t}^{(\kappa_i)}$ is one of the n exogenous stationary covariates sampled at the high frequency κ_i, α and $(\beta_1, \ldots, \beta_n)$ are the regression coefficients, and ε_t is the idiosyncratic term. The MIDAS function $m_K(\theta, L)$ controls the polynomial weights that allows the frequency mixing. It smoothes the K_i past values of $x_{t,i}^{(\kappa_i)}$ on which the regression is based. Different parameterizations of this weight function has been developed in the literature (see Ghysels et al., 2007, for a review of the most current ones). In this study we implement the two-parameters exponential Almon lag polynomial which can be defined as:

$$m_K(\theta, L) \equiv m_K(\theta_1, \theta_2, L) = \sum_{k=1}^{K} \frac{\exp(\theta_1 k + \theta_2 k^2)}{\sum_{l=1}^{K} \exp(\theta_1 l + \theta_2 l^2)} L^{k-1} \tag{2}$$

where L is the lag operator. Considering that we observe κ times the explanatory variable $x_t^{(\kappa)}$ during the period $[t-1,t]$, the lag operator L can be defined such that: $L^s x_t^{(\kappa)} = x_{t-s/\kappa}^{(\kappa)}$.

In macroeconomic forecasting, empirical models are generally based on dimension reduction methods coming from either variable or model selection. The difference between these two close schemes is mainly methodological: while variable selection aims at an a priori determination of the relevant predictors, model selection provides an algorithmic approach to combine models which are typically univariate. In this paper, we will focus on a dimension reduction technique that allows the use of high frequency variables through the MIDAS regression structure. A large family of widely used techniques in the literature for economic forecasting is based on principal component analysis and factor models. We refer, among others, to Forni et al. (2000) or Stock and Watson (2002). In the context of mixed-frequency models, Marcellino and Schumacher (2010) have put forward a Factor -augmented MIDAS model (FaMIDAS) as a way to tackle the lack of parsimony associated to the profusion of covariates. FaMIDAS is a method to incorporate in a MIDAS framework standard tools of factor analysis that usually produce very good results for short-term forecasting (see Giannone et al., 2008 or Barhoumi et al., 2010).

As an alternative to principal components analysis, a large body of literature suggests that the selection of relevant variables from a large set improves forecast efficiency. Bai and Ng (2008) used *targeted predictors* within a factor forecasting model. Their *pre-selection* is based on hard and soft thresholding rules carried out using the Lasso. The Lasso (*Least Absolute Shrinkage and Selection Operator*) introduced by Tibshirani, 1996 is a dimension reduction technique. Its asymptotically comes close to being an ideal subset selector in terms of its function as an oracle. De Mol et al. (2008) also suggested the use of penalized regressions, such as the Lasso, to perform variable selection and enhance the prediction accuracy and interpretability.

The paper makes use of the method of penalized variable selection by developing a combined strategy: the Lasso augmented MIDAS model. The estimation accuracy and empirical performances of this Lasso-MIDAS model are assessed in a predictive simulation exercise on various data sample using a data generating process (DGP). Then the operational efficiency of Lasso-type penalty in the context of MIDAS regression model is evaluated for macroeconomic forecasting purposes. We especially develop a predictive cross-validation procedure to forecast the US GDP growth. Our setup relies on the recent best forecasting performances and hence is designed for real-time analysis and nowcasting. To evaluate the empirical results of the setup, we perform a backtesting from 2004 to 2015 using an augmented version of the Stock and Watson data set, involving mixed frequencies. This new database merges the quarterly data set of Stock and Watson (2002) with the monthly FRED-MD data from Saint-Louis Fed database (see McCracken and Ng, 2015) and allows each time series variable to be sampled at its original sampling frequency (e.g. daily frequency for stock market indices, weekly frequency for jobless claims as released by the US Department of Labor, monthly frequency for personal consumer expenditure from BEA). Comparing empirical results of Lasso MIDAS model with several benchmarks (especially factor augmented models) allow us to draw several important conclusions: first, we show that mixing frequency is relevant to early capture GDP trends. Second, our large data set is a useful source of information which significantly improves forecasting performances, for all phases of the business cycle observed, whatever the technique for

reducing dimension: both factor and selection. Third, we observe that the Lasso succeeded in identifying informative content with respect to forecasting horizon. Fourth, the set of chosen predictors determined by the proposed variable/model selection procedure reflects the varying nature of the economic outlook.

The paper is structured as follows: Section 2 and Section 3 exhibits behavior of the Lasso and its combination with the MIDAS model using a simulation exercise. Then, in Section 4, we empirically show how the proposed selection method of explanatory variables out of a universe of well-known economic variables can significantly improve short-term forecasts of US GDP.

2 The Lasso augmented MIDAS model

The Lasso (Least Absolute Shrinkage and Selection Operator) has been introduced by Tibshirani (1996) as a covariates selection method in a linear regression setup. Lasso operates by penalizing the optimization problem associated to the regression with a term that involves the ℓ_1-norm of the coefficients. It belongs to the family of penalized regression model which amounts to performing least squares with some additional constraints on the coefficients, the ℓ_1-norm in the case of Lasso. Ng (2012) have shown that Lasso tends to have a lower misspecification risk in forecasting models when compared with usual information criteria. In the econometrics setup Bai and Ng (2008) and Schumacher (2010) have proposed to forecast economic series by using a combination of factor analysis with a LARS (see Efron et al., 2004) implementation of Lasso.

To be more specific, the Lasso takes advantage of the sparsifying properties of the ℓ_1-norm when solving the penalized optimization problem,

$$\begin{aligned}\widehat{b} &= \arg\min_b \sum_t \left(y_t - a - \sum_i b_i \, x_{t,i} \right)^2 + \lambda_{\text{Lasso}} \sum_i |b_i| \\ &= \arg\min_b ||Y - Xb||_2^2 + \lambda_{\text{Lasso}} \, ||b||_1,\end{aligned} \tag{3}$$

where y_t is the dependent variable, x_t is the vector of covariates, b is the vector containing the regression parameters, and λ_{Lasso} is the exogenous parameter which controls the strength of the Lasso penalization. The Lasso method does indeed reduce the dimension of the explanatory matrix X by driving non informative β_i elements to zero. Increasing $\lambda_{\text{Lasso}} \in \mathbb{R}^+$ brings gradually elements of the β vector to zero, hence selecting relevant explanatory variables. The choice of the exogenous parameter λ_{Lasso} that determines the number of covariates that are eliminated is essential and therefore a key issue that we will address later on via cross-validation.

Ridge regression is another popular penalized optimization scheme which, as opposed to the ℓ_1 penalty of Lasso, is based on a ℓ_2-norm penalty. Figure 1 illustrates the underlying principle of both techniques in the case of a multivariate regression model with two variables: b_1 and b_2. The Lasso is on the left, and the ridge regression on the right.

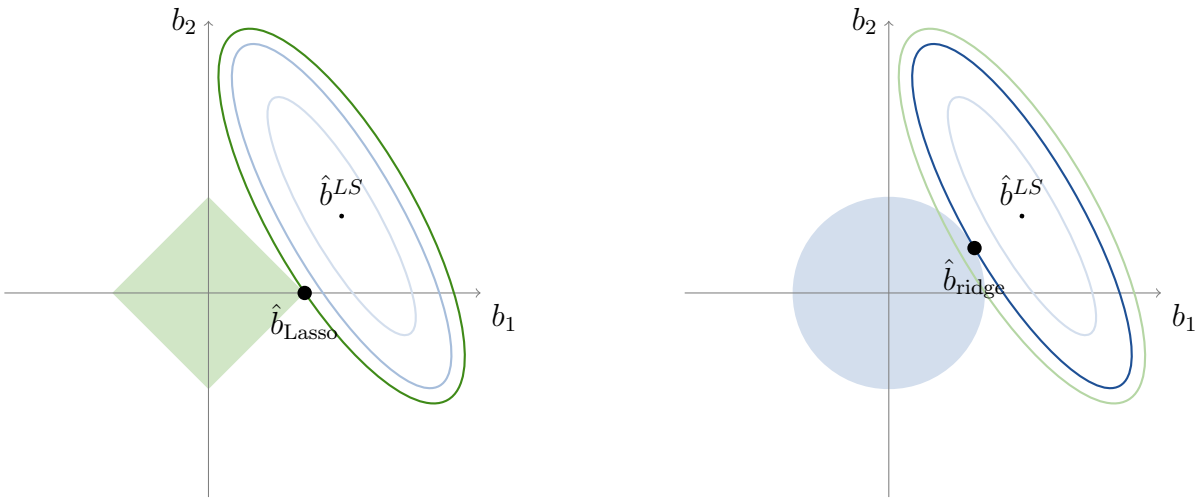

FIGURE 1: Penalized least squares estimate for the ℓ_1-norm (green) and the ℓ_2-norm (blue)

The ellipses around the least square estimator, $\hat{b}^{LS}$ represent the level sets of the squared error function $\|Y - Xb\|_2^2$ and the light colored areas correspond to balls of the ℓ_1 and ℓ_2 norms. In view of expression (3), the solution of the optimization problem that we are interested in takes place at the points in which both surfaces are tangent. The geometry of the problems makes that in the $\hat{b}^{\ell_1}$ case, the solution is generically located at the vertices of the ℓ_1-balls and hence the Lasso penalized solutions have entries equal to zero. The ridge based solutions are generally not located at that kind of specific points and are hence not necessarily sparse.

We put forward an extension of the Lasso model to the nonlinear MIDAS regression context by proposing the following optimization problem:

$$\begin{aligned}[\hat{\beta}, \hat{\theta}] &= \arg\min_{\beta,\theta} \sum_t \left(y_t - \alpha - \sum_{i=1}^{n} \beta_i \, m_{K_i}(\theta_i) \, x_{t,i}^{\kappa_i} \right)^2 + \lambda \sum_i |\beta_i| \\ &= \arg\min_{\beta,\theta} \|Y - \mathsf{X}(\theta)\,\beta\|_2^2 + \lambda \, \|\beta\|_1,\end{aligned} \tag{4}$$

where the matrix $\mathsf{X}(\theta)$ contains the MIDAS specifications that we previously described in (1),

$$\mathsf{X}(\theta) = \begin{pmatrix} 1 & m_{K_1}(\theta_1, L)\, x_{1,1}^{\kappa_1} & \cdots & m_{K_n}(\theta_n, L)\, x_{1,1}^{\kappa_n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & m_{K_1}(\theta_1, L)\, x_{T,n}^{\kappa_1} & \cdots & m_{K_n}(\theta_n, L)\, x_{T,n}^{\kappa_n} \end{pmatrix}. \tag{5}$$

As we will see later on and like in the linear case, the ℓ_1 penalization on the β parameters implies a selection of the most relevant predictors. The number of covariates eliminated can be chosen by tuning the value of the exogenous parameter λ. A technical complication in solving (4) via any gradient descent method arises due to the non-smooth nature of the ℓ_1 norm. We overcome this difficulty using a local regularization technique due to Nesterov (2005). Further details are provided in the Appendix A.

3 Simulations

Design of the simulation exercise

In order to assess the efficiency of Lasso-type penalty in the context of MIDAS regression model, we run some simulations using a data generating process (DGP). Our simulation exercise relies on the following design:

- We sample each exogenous variable to be an AR(1) process such that $X_t = a + bX_{t-1} + e_t$ for which we choose $a \sim \mathcal{U}[0; 0.2]$, $b \sim \mathcal{U}[-1; 1]$ and $e_t \sim \mathcal{N}(0, 0.15)$. We assume a mixed-frequency framework where variables are either daily sampled as $X_{\tau_D,i}$ with $i = 1, \dots, n_D$ and $\tau_D = \kappa_D \times t$, or monthly sampled, as $X_{\tau_M,j}$ where $j = 1, \dots, n_M$ and $\tau_M = \kappa_M \times t$. The MIDAS model mimics a realistic framework in which the dependent variable is quarterly sampled and the covariates includes both daily and monthly variables. We specify $K_D = 30$ (corresponding to the last 30 daily data with respect to outcome of the quarterly variable) and $K_M = 12$ for the monthly variables.
- Then, we sample the dependent variable y_t using the following DGP based a Gaussian MIDAS regression model:

$$\tilde{y}_t \sim \mathcal{N}(y_t, 0.15),$$
$$\text{with } y_t = \alpha + \sum_{i=1}^{n_D} \beta_i^{(D)} \; m_{K_D}(\theta_i, L) X_{t,i}^{(\kappa_D)} + \sum_{i=1}^{n_M} \beta_i^{(M)} \; m_{K_M}(\theta_{n_D+j}, L) X_{t,j}^{(\kappa_M)}, \quad (6)$$

  and where we assume that the vector $\beta = (\beta^{(D)}, \beta^{(M)})$ of regression parameters is sparse: only q percent of the $n = n_D + n_M$ explanatory variables (X_{τ_D}, X_{τ_M}) are relevant indicators. Thus the vector of n coefficients β is defined as, for all $i = 1, \dots, n$, $\beta_i \sim g\mathcal{N}(0, 1)$ with $\pi(g = 1) = q$ and $\pi(g = 0) = 1 - q$.
  In practise, we choose $q = 30\%$ and $n_D = n_M$ (half of the sample is daily, half monthly).
- Given a fixed number of time observations $T = 200$, we simulate 50 models for various numbers n of candidate variables (four cases: $n = 20, 100, 200$ and 300). The selection is carried by using the nonlinear least square estimation of the minimization equation (4) for a range of parameters λ in $\{0, ..., 30\}$.[1]
- We assess our simulations on 50 out-of-sample forecasts for which we compute the deviation of the model prediction $\hat{y}_t$ to the real value y_t using the root mean square error of the forecasts (RMSFE). Then we compare the performances of the forecasts with respect to the parameter λ.

It can be noticed that this setup would require estimating $30\,n_D + 12\,n_M = 42\,n$ parameters using standard linear models (or the U-MIDAS *a la* Foroni et al., 2015) while MIDAS only needs $3\,n$ estimates.

[1] The optimization algorithm is carried out through matlab nonlinear programming solver where initial point is a vector of zeros. The matlab code can be downloaded online `www.seltenhut.com/clement.marsilli/variablesection`.

Results

The Figure 2 shows the simulated results we obtain for the cases $n = 20, 100, 200$ and 300.

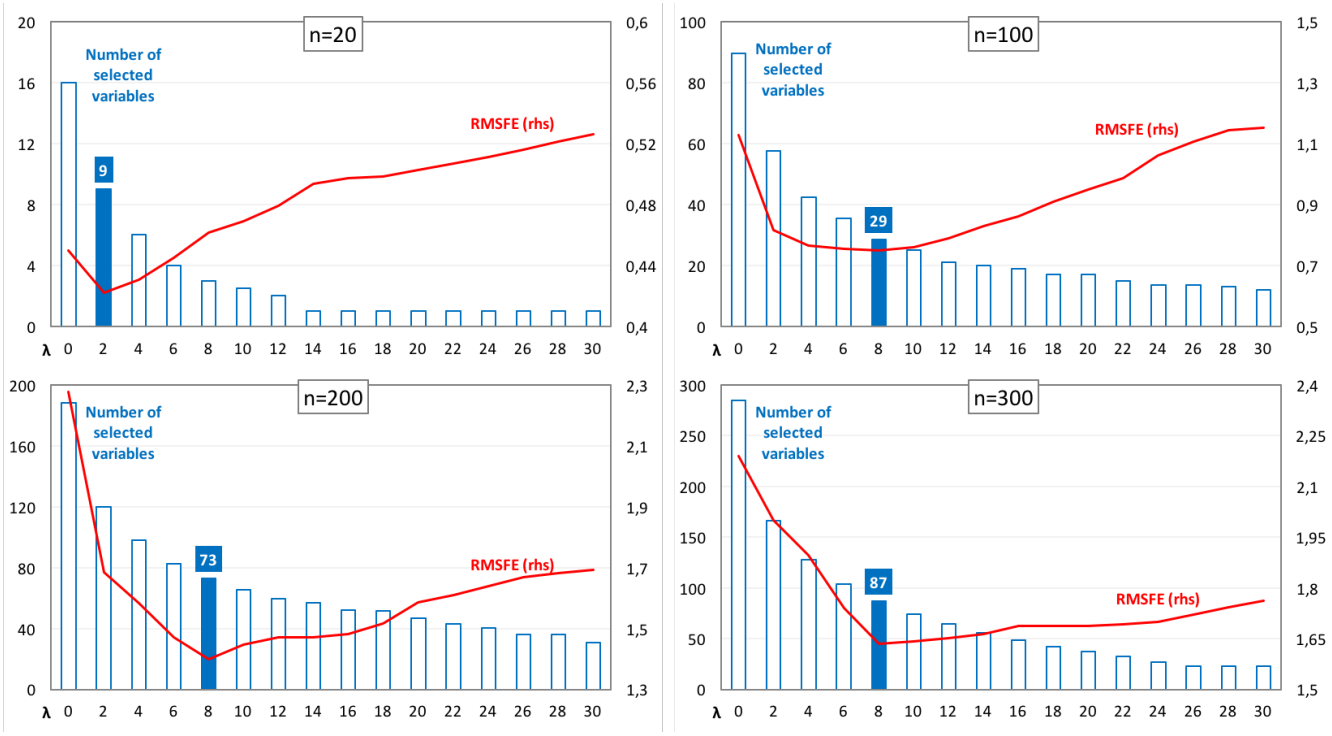

FIGURE 2: Lasso-MIDAS simulated results.
With respect to the Lasso-penalty λ (in the x-axis), the bars features the sparsity of the simulated models (number of the selected variables in the left axis), and the red line is the RMSFE. Both red lines and bars are the median of the 50 simulated models.

The number of selected variables tends to decrease with the Lasso penalty. The sparsity in the covariate matrix is therefore associated the penalty parameter λ, as in the linear case. Regarding the forecasting performances, the best models, which minimize RMSFEs, involve a variable selection for all sample sizes n. The Lasso selection within MIDAS model (ie Lasso-MIDAS with penalty $\lambda > 0$) turns out to significantly improve forecasting accuracy with respect to the "full" MIDAS model (ie Lasso-MIDAS with penalty $\lambda = 0$). And that difference increase with the size n of the covariate matrix.

Moreover, we can see that the improvement in forecasting accuracy is consistent with simulated sparsity within the covariate matrix. For each sample size, the global minimum (of the median of the 50 simulated models) of forecasting residuals corresponds to a specific selection size, as defined by λ. We can notice that this size is in line with the $q = 30$ percent of relevant variables involved in the DGP. For instance in the case of $n = 300$ (see Figure 2), the RMSFE's minimum is reached at $\lambda = 8$ involving a selection of 87 variables out of 300; that is about 29% and roughly corresponds to the imposed shared of relevant indicators of 30 percent.

In this exercise, values of "optimal" λ are not comparable across samples. Our setup relies on Gaussian draws of the parameters β. These draws are generated for each simulation

depending on the number of covariates n. Nevertheless, simulations make also out that sparse Lasso-MIDAS models with large Lasso-penalty λ have lower RMSFE than MIDAS model with $\lambda = 0$, in for large sample size (n=200 and n=300). Leading to the conclusion that, in large dimension, models with "wrong" variable selection (kind of misspecification) can be better in forecasting than models without any selection.

Though Lasso regression models present the advantage to be a one-step procedure, the determination of the penalization strength λ usually sets the most expensive step. This feature affects directly the numerical effort involved in its implementation and usually requires cross-validation-type strategies to be efficient as opposed to other standard iterative variable selection techniques.

4 Forecasting design

We specify a horizon-specific equation that incorporates the real time data flow within a Lasso-MIDAS framework. This direct approach relies on the following minimization problem:

$$\min_{\alpha,\beta,\theta} \sum_{t} \left(GDP^{(q)}_{t+1|t+1-h} - \alpha(h) + \sum_{i=1}^{N} \beta_i(h)\, m_K(\theta_i(h), L)\, x^{\kappa_i}_{t+1-h,i} \right)^2 + \lambda(h) \sum_{i=1}^{N} |\beta_i(h)|, \tag{7}$$

where the exogenous variables $x^{\kappa_i}_{t,i}$ stem from the quasi real-time data set we gathered. In our nowcasting context, we nowcast the dependant variable GDP_t at the period $t+1$ using the information available at time $t+1-h$ where h is the forecasting horizon and is smaller than one.

In fact, the MIDAS component is an extended mixed-frequency form involving four different sampling frequencies of our macroeconomic database. The MIDAS has the following form:

$$\begin{aligned}
\sum_{i=1}^{N} \beta_i(h)\, m_K(\theta_i(h), L)\, x^{\kappa_i}_{t,i} = & \sum_{i=1}^{N_D} \beta^D_i(h)\, m_{K_D}(\theta^D_i(h), L)\, x^D_{t,i} && \text{(Daily data)} \\
& + \sum_{i=1}^{N_W} \beta^W_i(h)\, m_{K_W}(\theta^W_i(h), L)\, x^W_{t,j} && \text{(Weekly data)} \\
& + \sum_{i=1}^{N_M} \beta^M_i(h)\, m_{K_M}(\theta^M_i(h), L)\, x^M_{t,j} && \text{(Monthly data)} \\
& + \sum_{i=1}^{N_Q} \beta^Q_i(h)\, m_{K_Q}(\theta^Q_i(h), L)\, x^Q_{t,j} && \text{(Quarterly data)}
\end{aligned}$$

The Lasso selection does consider similarly all variables regardless of their frequency.

4.1 The factor-augmented MIDAS as a benchmark

The factor-Augmented MIDAS (FaMIDAS) model which uses some factors extracted from a large mixed-frequency data set, constitutes a fair benchmark for our Lasso-MIDAS analysis. The FaMIDAS has been put forward by Marcellino and Schumacher (2010) and recently used by Ferrara and Marsilli (2014) to nowcast economic growth. This approach differs from the Lasso-MIDAS since it requires a two-step proceeding: first, we summarize the information into a certain number of factors for each block of covariates sharing the same frequency. Then, we use these factors as explanatory variables in a MIDAS model to predict the dependant variable.

More specifically, the FaMIDAS model is based on a assumption of a underlying factor structure in the explanatory variables matrix $(X_\tau)_\tau$, such that:

$$X_\tau = \Lambda F_\tau + \eta_\tau,$$

where τ corresponds to a specific frequency. As an example, the case $t = \tau$ means that x_τ is quarterly sampled. and F_τ is the vector of factors $F_\tau = (f_{1,\tau}, \ldots, f_{r,\tau})$.

Then the model incorporates the vector of factors as explanatory variables in the forecasting equation. In our mixed-frequency context, each factor vector only summarises the information of one group of variables sharing the same frequency. The MIDAS nowcasting model is given by the following equation:

$$\widehat{GDP}^{(q)}_{t+1|t+1-h} = \hat{\beta}_0(h) + \sum_{i=1}^{r(h)} \hat{\beta}_i(h) m_{K_i}(\hat{\theta}_i(h), L)\, \hat{f}^{\kappa_i}_{t+1-h,i}, \tag{8}$$

where $r(h)$ is the total number of factors including the daily, the weekly, the monthly and the quarterly factors, for a specific horizon h.

4.2 Predictive cross-validation

As initially suggested by Tibshirani (1996), the Lasso-type shrinkage is based on a targeted number of explanatory variables. Empirical literature usually argued that cross validation is a suitable strategy for variable selection. In our setup, the Lasso aims at providing variable selection which improve forecasting accuracy. To that end, Andreou et al. (2013) suggested a sophisticated method, in the same vein as cross validation, which accounts for the historical performance of each individual in order to define optimal weight in model combination. Using a non-linear strategy, they computed a discount factor attaching greater weight to the recent forecast accuracy. Nevertheless, the choice of some exogenous coefficient to parametrize the discount factor function, can affect severely the results: being either very smooth or very reactive to the recent performances. We propose a predictive cross-validation method based on recent prediction performances. That defines λ_t to provide model specifications at time t. In fact, the cross-validation is performed on the last S quarters within the rolling window estimation scheme; that provides a time varying Lasso selection adapted from the recent best forecasts. Thus, for all t and a given horizon h, we define the optimal $\lambda^\star_t(h)$ as the one

minimizing forecasting residuals over a subsample of size S, such that:

$$\lambda_t^\star(h) = \underset{\lambda>0}{\arg\min} \left(\frac{1}{S} \sum_{s=0}^{S-1} \left(GDP_{t-s} - \widehat{GDP}_{t-s|t-s-h}(\lambda_t) \right)^2 \right)^{1/2} \tag{9}$$

In this setup, the time stability of the parameter λ_t directly depends on the size S of the predictive cross-validation subsample. Empirically, this technique is an automated updated variable selection procedure providing recent best set of the explanatory variables.

To be consistent, we use similar predictive cross-validation procedure to select the number of factors r_t to be included at time t in the FaMIDAS nowcasting model (8). The optimal choice $r^\star$ minimizes forecasting residuals over the last S quarters.

$$r_t^\star(h) = \underset{r}{\arg\min} \left(\frac{1}{S} \sum_{s=0}^{S-1} \left(GDP_{t-s} - \widehat{GDP}_{t-s|t-s-h}(r_t) \right)^2 \right)^{1/2}$$

Note that factors can represent only one family of variables sharing same frequency. Since we mix daily, weekly, monthly and quarterly predictors, we define $r = (r^d, r^w, r^m, r^q)$.

5 Empirical application: nowcasting US GDP growth

The recent empirical literature has shown that time series models such as MIDAS are particularly suitable for small forecasting horizons analysis. This empirical exercise will focus on nowcasting purposes by performing a quasi-real time tracking of the US GDP growth over the period from 2004 to 2015.

5.1 Data set

To perform a reliable tracking of the US activity, we exploit the well-known "Stock and Watson data set" that we adapt to our framework. This US macroeconomic database was first compiled by Stock and Watson (1996) and included 76 variables. Then it was expanded by Stock and Watson (2002), Bernanke and Boivin (2003) and Marcellino et al. (2006) for various modelling purposes. Then the factor analysis literature considered a well-balanced version of this macroeconomic dataset to focus on forecasting (see Stock and Watson, 2006 and Bai and Ng, 2008) and uncertainty analysis (see Jurado et al., 2015). Recently McCracken and Ng (2015) have constructed a monthly database (MD) which mainly uses Stock and Watson data collection with a monthly sampling frequency and features the great advantage to be publicly accessible and monthly updated in the FRED website. In their paper, they showed that a few numbers of factors extracted from the FRED-MD database have equivalent predictive content than the vintages of Stock and Watson data set.

In order to take advantage of this database within our mixed-frequency strategy, we expand the FRED-MD set of McCracken and Ng (2015) to allow each variable for holding its own frequency. For instance, financial time series, such as interest rate and stock indices, are included in the set with their initial daily frequency, without prior high-to-low frequency

aggregation. We construct a database which gathers series from the FRED-MD, sampled at either daily, weekly or monthly frequency with quarterly variables from the original Stock and Watson database.[2] It includes 191 variables nearly arranged in the same eight following categories than FRED-MD database[3].

1. "Consumption and Orders" incorporates monthly BEA consumption data and housing market variables,
2. "Labor Market" includes BLS employment and earnings data,
3. "Money and Credit" includes money supply variables and both consumer and corporate credit-related variables
4. "Orders and Inventories" mostly includes supply side surveys
5. "Output and Income" includes decomposed variables of industrial production
6. "Prices" mostly includes monthly and quarterly inflation data (prices indices and GDP components deflators)
7. "Stock Market" includes main daily financial stock indices
8. "Interest Rates and Exchanges Rates" that includes daily interest rates, daily exchanges rates, and some spreads (term spread, yield spread, etc.),

The data set we consider in this exercise have been downloaded in August 2016 using both Thomson Reuters Datastream and the FRED-MD in the Saint-Louis Fed website. Though the database does not involve real-time vintages, it mimics the real-time data flow by using a stylized calendar of the order in which data releases are typically published over the quarter. Each series is stationnarized following the "data transformation" suggested by McCracken and Ng (2015) for the FRED-MD database.

Regarding the dependant variable, we consider the first release of the quarterly growth of the US GDP such as given by the *Real-Time Data Set for Macroeconomists* of the Federal Reserve Bank of Philadelphia.

5.2 Results

Using this data set, we estimate both Lasso MIDAS and Factor MIDAS models using a rolling-window framework of 20 years from 1984q1. The forecasting results rely on an out-of-sample assessment of US GDP growth from 2005q1 to 2014q4. These regression models include up to $n = 191$ variables sampled at either a daily, a weekly, a monthly or a quarterly frequency, with respect to their availability over the whole sample. We estimate the models under various specifications and assess predictive performances and empirical behaviour of the Lasso in the mixed-frequency context of this well-known exercise of US GDP forecasting.

Operationally, we use the predictive cross validation to define the optimal $\lambda^*(h)$ in the Lasso MIDAS which set the level of sparsity in the model. We use a loop iterating over the range $\lambda \in [0, 12]$, by a step value of 0.2 on each iteration. Forecasts are based recursively on

[2]The database, as well as the matlab code for loading data and forecasting GDP growth within a Lasso-MIDAS model can be downloaded on the website http://www.seltenhut.com/clement.marsilli/VariableSelection.
[3]The exact composition of the eight categories is given in Appendix.

recent best set of predictors as selected by $\lambda^*(h)$ in the cross validation subsample of size S. The estimation of Factor MIDAS models is based on similar cross validation technique to get the optimal number of factors $r^*(h)$ to include within model specifications. In practice, for computational reasons, we only consider up to 2 factors per frequency family.

From the nowcasts of US growth obtained over the period form 2004q1 to 2014q4, we compute standard RMSFE. Table 1 presents the results.

	h=60			h=40			h=20			h=0		
	S=1	S=4	S=8	S=1	S=4	S=8	S=1	S=4	S=8	S=1	S=4	S=8
Lasso MIDAS	0.30	0.27	0.29	0.28	0.24	0.24	0.24	0.23	0.20	0.17	0.16	0.18
Average number of variables	10.3	8.9	7.3	8.3	7.2	8.1	9.8	10.2	7.4	10.9	10.0	9.3
FaMIDAS	0.29	0.33	0.35	0.28	0.28	0.28	0.25	0.26	0.23	0.21	0.17	0.21
# daily factor	0.8	0.7	0.7	0.9	0.7	0.5	1.0	0.4	0.7	0.9	0.6	0.4
# weekly factor	0.8	0.7	0.7	1.1	0.8	0.8	0.9	1.2	1.1	0.8	0.8	0.7
# monthly factor	1.5	1.2	0.9	1.0	1.0	0.9	1.2	0.8	0.6	0.8	1.0	0.8
# quarterly factor	0.9	1.2	1.1	0.9	0.9	0.6	1.1	1.2	1.1	1.1	1.2	0.9
Ratio LassoMIDAS/FaMIDAS	1.03	0.81*	0.82*	1.00	0.85*	0.85*	0.96	0.88*	0.87*	0.81*	0.94*	0.86*

TABLE 1: Comparisons of root mean squared forecast error for the two models over the period 2004-2015. Note: the lowest RMSFE is indicated in bold type. The stars indicates that the Diebold Mariano test rejects the forecasts equivalence hypothesis between both models.

Both models carries out reliable forecasts which fit well the signal and converge towards the target inversely with the horizon. A close examination of this table reveals several results.

First, we find that Lasso-MIDAS outperforms FaMIDAS of about 11% in our exercise. That is particularly significant for longer nowcasting horizons, from 3 to 1 months ahead (from $h = 60$ to $h = 20$). Although the results do not allow to clearly distinguish a specific behaviour of the cross validation strategy, we can notice that the reactive selection based on the small cross validation sample size ($S = 1$) does not provide the more accurate forecasts, both in Lasso and in Factor augmented cases.

Second, the size of the Lasso selection is remarkably stable. The average number of variables varies between 7 and 11 variables regardless of the forecasting horizon. That represents to a shrinkage ratio of about 4% to 6%, and hence clearly indicates an underlying pattern of sparsity within this data set.

Third, the composition of this selection changes. Figure 3 displays kind of heat maps of the inclusion of each predictor over time in various cases. For all horizons, the results improve with larger size S of the cross validation sample, which tends to stabilize the selection. In fact, smoothing the discount factor for historical performances leads to an average better accuracy than a more reactive selection. The comparison between very recent best performing set ($S = 1$ quarter, as displayed in the right hand side charts of Figure 3) and more stable set ($S = 8$ quarters, on the right hand side) shows significant differences in forecasting errors metrics in support of stability, despite almost similar absolute sparsity.

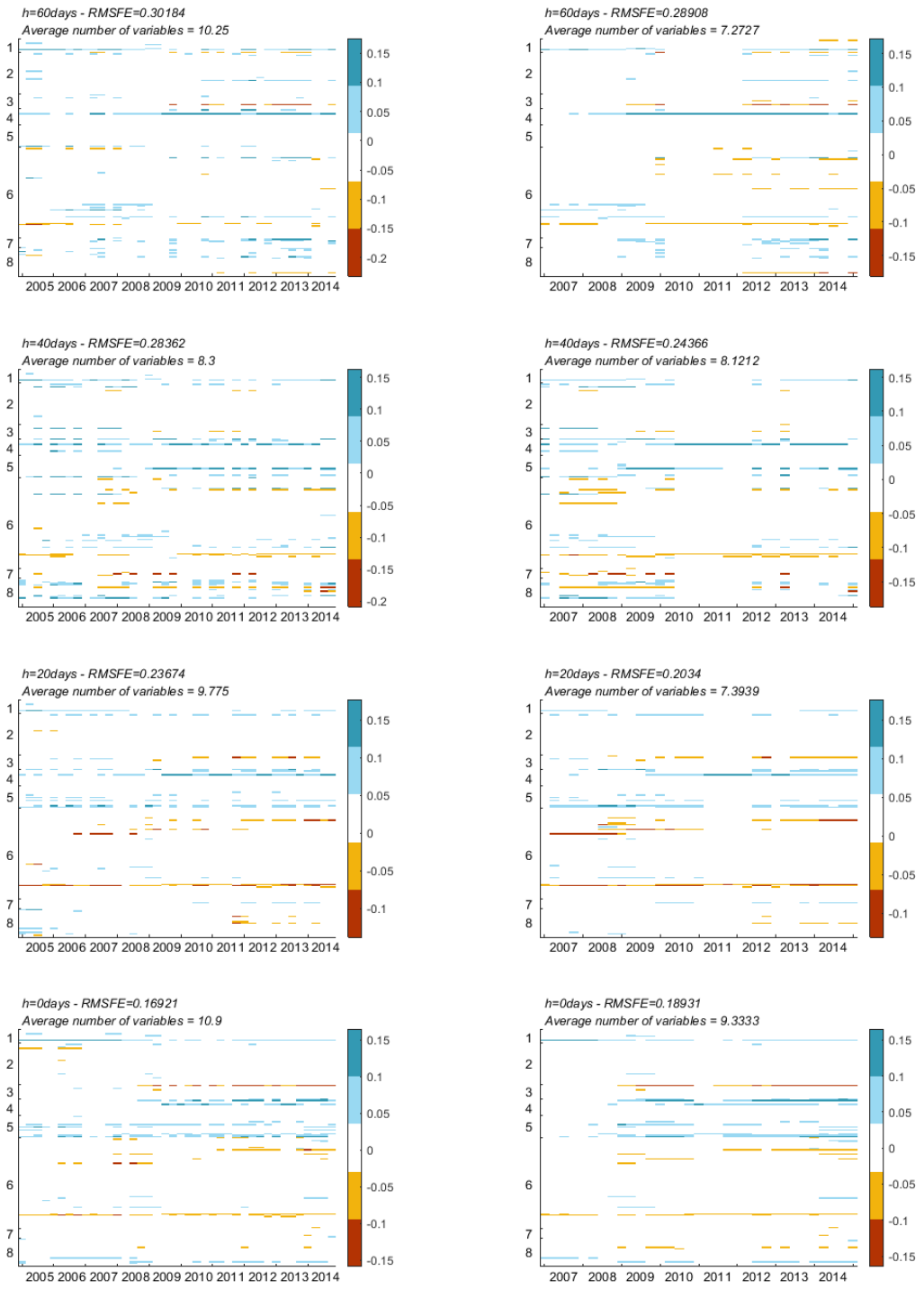

FIGURE 3: Variable selection carried out by Lasso MIDAS models for various predictive cross validation sample sizes and forecasting horizons. Charts on the left correspond to a predictive cross validation sample of size $S = 1$ quarter; and the right, it shows selection of sample of size $S = 8$ quarters. On the y-axis, the groups from 1 to 8 represents the classes of variables of the data set. Every row refers to a variable.

6 Conclusion

The paper makes use of the method of penalized variable selection by developing a combined strategy: the Lasso augmented MIDAS model. The estimation accuracy and empirical performances of this Lasso-MIDAS model are assessed in a predictive simulation exercise on various data sample using a data generating process (DGP), and the well-known application of US GDP growth forecasting.

Our empirical results show a clear pattern of sparsity in the economic nowcasting problem. Our model succeeds in identifying leading indicators and constructing an objective variable selection with broad applicability. However, this patterns of sparsity varies a lot over time, and across forecasting horizon. The usefulness of sparse methods, as the Lasso on which we base our mixed-frequency analysis, is not always obvious. Specifying models with respect to the recent empirical behaviour is essential to provide accurate predictions.

References

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

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

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

Banbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2012). Now-Casting and the Real-Time Data Flow. Technical report, CEPR.

Barhoumi, K., Darné, O., and Ferrara, L. (2010). Are disaggregate data useful for factor analysis in forecasting French GDP? *Journal of Forecasting*, 29(1-2):132–144.

Bernanke, B. S. and Boivin, J. (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics*, 50(3):525–546.

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

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.

Ferrara, L. and Marsilli, C. (2013). Financial variables as leading indicators of GDP growth: Evidence from a MIDAS approach during the Great Recession. *Applied Economics Letters*, 20(3):233–237.

Ferrara, L. and Marsilli, C. (2014). Nowcasting global economic growth: A factor-augmented mixed-frequency approach. *Banque de France Working paper*, 515.

Ferrara, L., Marsilli, C., and Ortega, J.-P. (2014). Forecasting growth during the Great Recession: is financial volatility the missing ingredient? *Economic Modelling*, 36:44–50.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The Generalized Dynamic-Factor Model: Identification And Estimation. *The Review of Economics and Statistics*, 82(4):540–554.

Foroni, C., Marcellino, M., and Schumacher, C. (2015). U-MIDAS: MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society - Series A*, 178(1):57–82.

Ghysels, E., Santa-clara, P., and Valkanov, R. (2002). The MIDAS Touch: Mixed Data Sampling Regression Models.

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

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

Jurado, K., Ludvigson, S. C., and Ng, S. (2015). Measuring Uncertainty. *American Economic Review*, 105(3):1177–1216.

Marcellino, M. and 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.

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

McCracken, M. W. and Ng, S. (2015). FRED-MD: A Monthly Database for Macroeconomic Research. *Federal Reserve Bank of St. Louis Working Papers*.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Math. Program.*, 103(1, Ser. A):127–152.

Ng, S. (2012). Variable Selection in Predictive Regressions.

Schumacher, C. (2010). Factor forecasting using international targeted predictors: The case of German GDP. *Economics Letters*, 107(2):95–98.

Stock, J. H. and Watson, M. W. (1996). Evidence on Structural Instability in Macroeconomic Time Series Relations. *Journal of Business & Economic Statistics*, 14(1):11–30.

Stock, J. H. and Watson, M. W. (2002). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, 97:1167–1179.

Stock, J. H. and Watson, M. W. (2006). Forecasting with Many Predictors. In Elliot, G., Granger, C. W., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1. Elsevier edition.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Appendices

A Nesterov regularization technique

Let us consider the following regression model:

$$\widehat{\beta} = \arg\min_{\beta} \sum_{t} \left(y_t - \beta_0 - \sum_{i} \beta_i \, x_{t,i} \right)^2 + \lambda_{\text{Lasso}} \sum_{i} |\beta_i| \tag{10}$$

where y_t is the dependent variable, x_t is the vector of covariates, β is the vector containing the regression parameters, and λ_{Lasso} is the exogenous parameter which controls the strength of the Lasso sparsifying regularization. To overcome the estimation of the problem 10, we use the following local regularization technique of Nesterov (2005).

We start by noting that the ℓ_1 norm can be expressed using the function g defined as

$$g(\beta) = \|\beta\|_1 = \max_{\|\gamma\|_\infty \leqslant 1} \gamma'\beta.$$

Then, we define the function g_μ such that $g_\mu \longrightarrow g$ with respect to $\mu \to 0$ and $\mu > 0$. We have:

$$g_\mu(\beta) := \max_{\|\gamma\|_\infty \leqslant 1} \gamma'\beta - \frac{\mu}{2}\|\gamma\|_2^2,$$

The Nesterov regularization technique consists of replacing the norm $g(\beta) = \|\beta\|_1$ by $g_\mu(\beta)$ with μ small. The advantage of proceeding in this fashion is that the function g_μ is obviously smooth with a gradient $\nabla g_\mu(\beta)$ whose components are given by

$$\nabla_i g_\mu(\beta) = \begin{cases} \text{sign}(\beta_i) & \text{if } |\beta_i| > \mu, \\ \dfrac{1}{\mu}\beta_i & \text{if } |\beta_i| < \mu. \end{cases}$$

B Simulation results

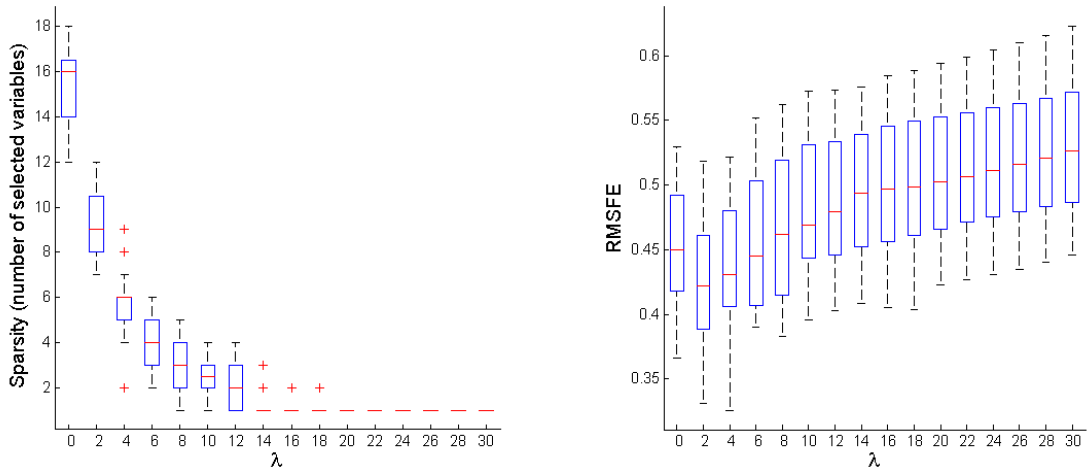

FIGURE 4: Boxplot of RMSFE of 20 simulated models where $n = 20$ variables and $T = 200$

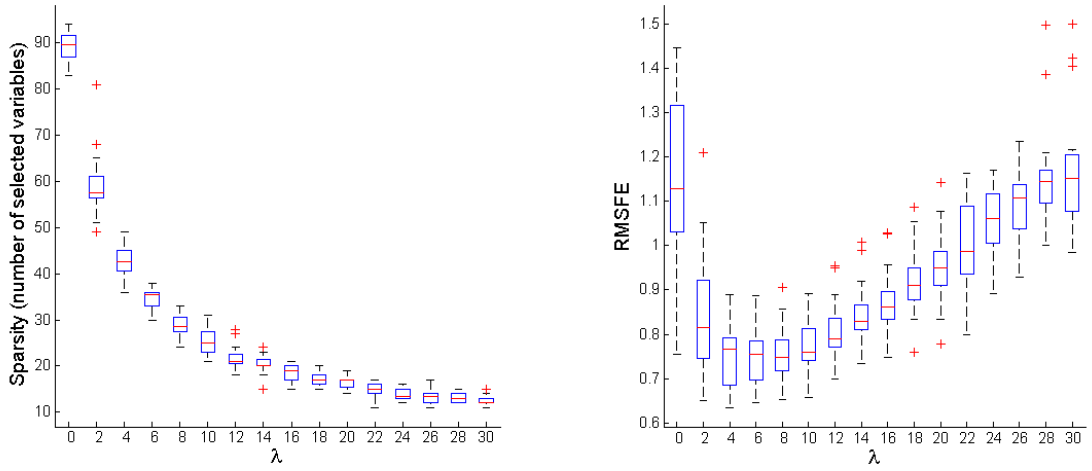

FIGURE 5: Boxplot of RMSFE of 20 simulated models where $n = 100$ variables and $T = 200$

C Data set

Data are transformed following similar ways than proposed in McCracken and Ng (2015) for the FRED-MD database.

1. Consumption and Orders

	Freq.	Variable	Code	Provider
44	M	US PERSONAL CONSUMPTION EXPENDITURES (AR) CONA	USPERCOND	Reuters Datastream
45	M	US PERSONAL CONSUMPTION EXPENDITURES - DURABLES (AR) CONA	USCONDURD	Reuters Datastream
46	M	US PERSONAL CONSUMPTION EXPENDITURES - NONDURABLES (AR) CONA	USCONNDRD	Reuters Datastream
47	M	US PERSONAL CONSUMPTION EXPENDITURES - SERVICES (AR) CONA	USCONSRVD	Reuters Datastream
115	M	Housing Starts Total New Privately Owned	HOUST	FRED MD
116	M	Housing Starts Northeast	HOUSTNE	FRED MD
117	M	Housing Starts Midwest	HOUSTMW	FRED MD
118	M	Housing Starts South	HOUSTS	FRED MD
119	M	Housing Starts West	HOUSTW	FRED MD
120	M	New Private Housing Permits	PERMIT	FRED MD
121	M	New Private Housing Permits Northeast	PERMITNE	FRED MD
122	M	New Private Housing Permits Midwest	PERMITMW	FRED MD
123	M	New Private Housing Permits South	PERMITS	FRED MD
124	M	New Private Housing Permits West	PERMITW	FRED MD

2. Labor Market

	Freq.	Variable	Code	Provider
33	W	US UNEMPLOYMENT (SA) - INITIAL CLAIMS - ECONOMIC SERIES	USUNCLM	Reuters Datastream
87	M	Help-Wanted Index for US	HWI	FRED MD
88	M	Ratio of HW/No. Unemployed	HWIURATIO	FRED MD
89	M	Civilian Labor Forece	CLF16OV	FRED MD
90	M	Civilian Employment	CE16OV	FRED MD
91	M	Civilian Unemployment	UNRATE	FRED MD
92	M	Average Duration of Unemployment (weeks)	UEMPMEAN	FRED MD
93	M	Civilian Unemployed - Less than 5W	UEMPLT5	FRED MD
94	M	Civilian Unemployed for 5-14W	UEMP5TO14	FRED MD
95	M	Civilian Unemployed - 15W and over	UEMP15OV	FRED MD
96	M	Civilian Unemployed for 15-26W	UEMP15T26	FRED MD
97	M	Civilian Unemployed for 27W and over	UEMP27OV	FRED MD
98	M	All Employees Total nonfarm	PAYEMS	FRED MD
99	M	All Employees Goods-Producing Industries	USGOOD	FRED MD
100	M	All Employees Mining and Logging (Mining)	CES1021000001	FRED MD
101	M	All Employees Construction	USCONS	FRED MD
102	M	All Employees Manufacturing	MANEMP	FRED MD
103	M	All Employees Durable Goods	DMANEMP	FRED MD
104	M	All Employees Nondurable Goods	NDMANEMP	FRED MD
105	M	All Employees Services-Providing Industries	SRVPRD	FRED MD
106	M	All Employees Trade, Transportation and Utilities	USTPU	FRED MD
107	M	All Employees Wholesale Traed	USWTRADE	FRED MD
108	M	All Employees Retail Trade	USTRADE	FRED MD
109	M	All Employees Financial Activities	USFIRE	FRED MD
110	M	All Employees Government	USGOVT	FRED MD
111	M	Avg Weekly Hours Goods-Producing	CES0600000007	FRED MD
112	M	Avg Weekly Overtime Hours Manuf	AWOTMAN	FRED MD
113	M	Avg Weekly Hours Manuf	AWHMAN	FRED MD
114	M	ISM Manuf Employment Index	NAPMEI	FRED MD
159	M	Avg Hourly Earnings Goods-Producing	CES0600000008	FRED MD
160	M	Avg Hourly Earnings Construction	CES2000000008	FRED MD
161	M	Avg Hourly Earnings Manufacturing	CES3000000008	FRED MD

3. Money and Credit

	Freq.	Variable	Code	Provider
30	W	US MONEY SUPPLY M1 - ECONOMIC SERIES	USMONEY	Reuters Datastream
31	W	US MONEY SUPPLY M2 SEASONALLY ADJ. - ECONOMIC SERIES	USM2WSA	Reuters Datastream
32	W	US COMMERCIAL BANKS: SECURITIES - ECONOMIC SERIES	USCAS.S	Reuters Datastream
135	M	Real M2 Money Stock	M2REAL	FRED MD
136	M	St Louis Adj Monetary Base	AMBSL	FRED MD
137	M	Total Reserves of Depository Institutions	TOTRESNS	FRED MD
138	M	Reserves of Depository Institutions	NONBORRES	FRED MD
139	M	Commercial and Industrial Loans	BUSLOANS	FRED MD
140	M	Real Estate Loans at All Commercial Banks	REALLN	FRED MD
141	M	Total Revolving Credit	NONREVSL	FRED MD
142	M	Nonrevolving consumer credit to Personal Income	CONSPI	FRED MD
163	M	MZM Money Stock	MZMSL	FRED MD
164	M	Consumer Motor Vehicle Loans Outstanding	DTCOLNVHFNM	FRED MD
165	M	Total Consumer Loans and Leases Outstanding	DTCTHFNM	FRED MD

4. Orders and Inventories

	Freq.	Variable	Code	Provider
69	M	Real Personal Consuption Expenditures	DPCERA3M086SBEA	FRED MD
70	M	Real Manuf and Trade Industries Sales	CMRMTSPLx	FRED MD
71	M	Retail and Food Services Sales	RETAILx	FRED MD
125	M	ISM PMI Composite Index	NAPM	FRED MD
126	M	ISM New Orders Index	NAPMNOI	FRED MD
127	M	ISM Supplier Deliveries Index	NAPMSDI	FRED MD
128	M	ISM Inventories Index	NAPMII	FRED MD
129	M	New Orders for Consumer Goods	ACOGNO	FRED MD
130	M	New Orders for Durable Goods	AMDMNOx	FRED MD
131	M	New Orders for Nondefense Capital Goods	ANDENOx	FRED MD
132	M	Unfilled Orders for Nondefense Capital Goods	AMDMUOx	FRED MD
133	M	Total Business Inventories	BUSINVx	FRED MD
134	M	Total Business Inventories to Sales Ratio	ISRATIOx	FRED MD
162	M	Consumer Sentiment Index	UMCSENTx	FRED MD

5. Output and Income

	Freq.	Variable	Code	Provider
67	M	Real Personal Income	RPI	FRED MD
68	M	Real Personal Income ex transfert receipts	W875RX1	FRED MD
72	M	IP Index	INDPRO	FRED MD
73	M	IP Final Products and Nonindustrial Supplies	IPFPNSS	FRED MD
74	M	IP Final Products (Market Group)	IPFINAL	FRED MD
75	M	IP Consumer Goods	IPCONGD	FRED MD
76	M	IP Nondurable Consumer Goods	IPDCONGD	FRED MD
77	M	IP	IPNCONGD	FRED MD
78	M	Business Equipment	IPBUSEQ	FRED MD
79	M	IP Materials	IPMAT	FRED MD
80	M	IP Durable Materials	IPDMAT	FRED MD
81	M	IP Nondurable Materials	IPNMAT	FRED MD
82	M	IP Manufacturing	IPMANSICS	FRED MD
83	M	IP Residential Utilities	IPB51222S	FRED MD
84	M	IP Fuels	IPFUELS	FRED MD
85	M	ISM Manufacturing: Production Index	NAPMPI	FRED MD
86	M	Capacity Utilization	CUMFNS	FRED MD

8 Interest Rates and Exchanges Rates

	Freq.	Variable	Code	Provider
1	D	US FED FUNDS EFF RATE (D) - MIDDLE RATE	FRFEDFD	Reuters Datastream
2	D	US COMM PAPER FIN 3 MONTH (D) - MIDDLE RATE	FRCPF3M	Reuters Datastream
3	D	US T-BILL SEC MARKET 3 MONTH (D) - MIDDLE RATE	FRTBS3M	Reuters Datastream
4	D	US T-BILL SEC MARKET 6 MONTH (D) - MIDDLE RATE	FRTBS6M	Reuters Datastream
5	D	US TREASURY CONST MAT 1 YEAR (D) - MIDDLE RATE	FRTCM1Y	Reuters Datastream
6	D	US TREASURY CONST MAT 5 YEAR (D) - MIDDLE RATE	FRTCM5Y	Reuters Datastream
7	D	US TREASURY CONST MAT 10 YEAR (D) - MIDDLE RATE	FRTCM10	Reuters Datastream
8	D	USD MAJOR CURRENCY MAR 73=100 (FED) - EXCHANGE INDEX	US$CWMN	Reuters Datastream
9	D	USD TO EURO (WMR&DS) - EXCHANGE RATE	USEURSP	Reuters Datastream
10	D	USD TO UK P (WMR) - EXCHANGE RATE	USDOLLR	Reuters Datastream
11	D	USD TO CANADIAN D (GTIS/TR) - EXCHANGE RATE	CDNDLUS	Reuters Datastream
12	D	USD TO 100 JAPANESE YEN (GTIS/TR) - EXCHANGE RATE	JAPYNUS	Reuters Datastream
13	D	USD TO SWISS FRANC (GTIS/TR) - EXCHANGE RATE	SWISFUS	Reuters Datastream
14	D	USD TO CHINESE YUAN (GTIS/TR) - EXCHANGE RATE	CHINYUS	Reuters Datastream
22	D	3-Month Commercial Paper minus FEDFUNDS	Spr3MCmFFR	Reuters Datastream
23	D	3-Month Treasuary minus FEDFUNDS	Spr3MmFFR	Reuters Datastream
24	D	6-Month Treasuary minus FEDFUNDS	Spr6MmFFR	Reuters Datastream
25	D	1-Year Treasuary minus FEDFUNDS	Spr1YmFFR	Reuters Datastream
26	D	5-Year Treasuary minus FEDFUNDS	Spr5YmFFR	Reuters Datastream
27	D	10-Year Treasuary minus FEDFUNDS	Spr10YmFFR	Reuters Datastream
34	W	US CORP BONDS MOODYS SEASONED AAA (W) - MIDDLE RATE	FRMCAAA	Reuters Datastream
35	W	US CORP BONDS MOODYS SEASONED BAA (W) - MIDDLE RATE	FRMCBAA	Reuters Datastream
36	W	Spread Corporate Bonds Moodys AAA minus FEDFUNDS	SprAAAmFFR	Reuters Datastream
37	W	Spread Corporate Bonds Moodys BAA minus FEDFUNDS	SprBAAmFFR	Reuters Datastream
33	W	US UNEMPLOYMENT (SA) - INITIAL CLAIMS - ECONOMIC SERIES	USUNCLM	Reuters Datastream

6. Prices

	Freq.	Variable	Code	Provider
28	D	Crude Oil-WTI Spot Cushing U$/BBL - DS MID PRICE	CRUDOIL	Reuters Datastream
29	D	Crude Oil-Brent Cur. Month FOB U$/BBL	OILBREN	Reuters Datastream
38	M	US IMPORT PRICE INDEX - ALL COMMODITIES (END USE) NADJ	USIMPPRCF	Reuters Datastream
39	M	US IMPORT PRICE INDEX - EXCLUDING PETROLEUM (END USE) NADJ	USIPEXPTF	Reuters Datastream
40	M	US IMPORT PRICE INDEX - EXCLUDING FUELS (END USE) NADJ	USIPEXFUF	Reuters Datastream
41	M	US EXPORT PRICE INDEX - ALL COMMODITIES (END USE) NADJ	USEXPPRCF	Reuters Datastream
42	M	US EXPORT PRICE INDEX-AUTO VEHICLES,PARTS & ENGINES(END USE)	USEPMOTVF	Reuters Datastream
43	M	US EXPORT PRICE INDEX - CAPITAL GOODS (END USE) NADJ	USEPCAPGF	Reuters Datastream
48	M	US PRICE INDEX PCE,DURABLE GOODS,FURNISHINGS & DUR HHOLD EQP	USU1CO45E	Reuters Datastream
49	M	US PRICE INDEX PCE, DURABLE GOODS, MOTOR VEHICLES AND PARTS SADJ	USUR6KOWE	Reuters Datastream
50	M	US PRICE INDEX PCE,DURABLE GOODS,RECREATIONAL GOODS & VEHICLES	USU5K32SE	Reuters Datastream
51	M	US PRICE INDEX PCE, NONDURABLE GOODS, CLOTHING AND FOOTWEAR SADJ	USU67FUTE	Reuters Datastream
52	M	US PRICE INDEX PCE,NONDURABLE GOODS,GASOLINE & OTHER ENERGY GDS	USUJLNVHE	Reuters Datastream
53	M	US PRICE INDEX PCE,NONDURABLE GOODS, OTHER NONDURABLE GOODS SADJ	USUZXRBGE	Reuters Datastream
54	M	US PRICE INDEX PCE, SERVICES, HHFCE, HOUSING AND UTILITIES SADJ	USU7P5IHE	Reuters Datastream
55	M	US PRICE INDEX PCE,SERVICES, HHFCE, TRANSPORTATION SERVICES SADJ	USUYJFNLE	Reuters Datastream
56	M	US PRICE INDEX PCE, SERVICES, HHFCE, RECREATION SERVICES SADJ	USUMF23TE	Reuters Datastream
57	M	US PRICE INDEX PCE, SERVICES, HHFCE, HEALTH CARE SADJ	USU3NVUFE	Reuters Datastream
58	M	US PRICE INDEX PCE, SERVICES, HHFCE, OTHER SERVICES SADJ	USU1UZXTE	Reuters Datastream
59	M	US PRICE INDEX PCE,SERVICES,HOUSEHOLD CONSMPTN.EXPENDITURES SADJ	USUAVUKLE	Reuters Datastream
60	M	US PRICE INDEX PCE,SERVICES,HHFCE,FINANCIAL SERVICES & INSUR	USUJJDHIE	Reuters Datastream
61	M	US PRICE INDEX PCE,SERVICES,HHFCE, FOOD SERVICES AND ACCOMS SADJ	USUG1WZ7E	Reuters Datastream
62	M	US PRICE INDEX PCE, PERSONAL CONSUMPTION EXPENDITURES SADJ	USUR9RQ8E	Reuters Datastream
63	M	US PRICE INDEX PCE, SERVICES SADJ	USU4H9R7E	Reuters Datastream
64	M	US PRICE INDEX PCE, OTHER DURABLE GOODS SADJ	USUZY6NKE	Reuters Datastream
65	M	US PRICE INDEX PCE, GOODS, NONDURABLE GOODS SADJ	USULFSOQE	Reuters Datastream
66	M	US PRICE INDEX PCE, GOODS, DURABLE GOODS SADJ	USUND7JGE	Reuters Datastream
143	M	PPI Finished Goods	PPIFGS	FRED MD
144	M	PPI Finished Consumer Goods	PPIFCG	FRED MD
145	M	PPI Intermediate Materials	PPIITM	FRED MD
146	M	PPI Crude Materials	PPICRM	FRED MD
147	M	PPI Metals and metal products	PPICMM	FRED MD
148	M	ISM Manuf Price Index	NAPMPRI	FRED MD
149	M	CPI All Items	CPIAUCSL	FRED MD
150	M	CPI Apparel	CPIAPPSL	FRED MD
151	M	CPI Transportation	CPITRNSL	FRED MD
152	M	CPI Medical Care	CPIMEDSL	FRED MD
153	M	CPI Commodities	CUSR0000SAC	FRED MD
154	M	CPI Durables	CUUR0000SAD	FRED MD
155	M	CPI Services	CUSR0000SAS	FRED MD
156	M	CPI All Items less Food	CPIULFSL	FRED MD
157	M	CPI All Items less Shelter	CUUR0000SA0L2	FRED MD
158	M	CPI All Items less Medical Care	CUSR0000SA0L5	FRED MD
166	Q	US CHAIN-TYPE QUANTITY INDEX FOR GROSS PRIVATE DOM.INVESTMENT	USIVP.CQE	Reuters Datastream
167	Q	US CHAIN-TYPE QUANTITY INDEX FOR PRIVATE FIXED INVESTMENT SADJ	USIVF.CQE	Reuters Datastream
168	Q	US CHAIN-TYPE QUANTITY INDEX FOR NONRESIDENTIAL FIXED INVESTMENT	USIVFNCQE	Reuters Datastream
169	Q	US CHAIN-TYPE QUANTITY INDEX-NONRESL FIXED INVESTMENT,STRUCTURES	USISN.CQE	Reuters Datastream
170	Q	US CHAIN-TYPE QUANTITY INDEX-PRIVATE FIXED INVEST,EQUIPMENT SADJ	USIVEQPQE	Reuters Datastream
171	Q	US CHAIN-TYPE QUANTITY INDEX FOR RESIDENTIAL FIXED INVESTMENT	USIVFRCQE	Reuters Datastream
172	Q	US CHAIN-TYPE QUANTITY INDEX FOR STATE & LOCAL GOVT CNSMPT & INV	USPUGZCQE	Reuters Datastream
173	Q	US CHAIN-TYPE QUANTITY INDEX FOR FEDL GOVT CONSMPTN.& INVESTMENT	USPUGCCQE	Reuters Datastream
174	Q	US CHAIN-TYPE QUANTITY INDEX FOR GOVT.CONSMPTN.& INVESTMENT SADJ	USPUG.CQE	Reuters Datastream
175	Q	US QUANTITY INDEX- ADDENDA, FINAL SALES OF DOMESTIC PRODUCT SADJ	USUGZDPRE	Reuters Datastream
176	Q	US QUANTITY INDEX-ADDENDA,FINAL SALES TO DOMESTIC PURCHASES SADJ	USUGZDPUE	Reuters Datastream
177	Q	US QUANTITY INDEX - ADDENDA, GROSS DOMESTIC PURCHASES SADJ	USUGZGDPE	Reuters Datastream
178	Q	US CHAIN-TYPE PRICE INDEX OF GDP SADJ	USGDP..CE	Reuters Datastream
179	Q	US CHAIN-TYPE PRICE INDEX FOR GROSS PRIVATE DOM. INVESTMENT SADJ	USIVP..CE	Reuters Datastream
180	Q	US CHAIN-TYPE PRICE INDEX FOR PRIVATE FIXED INVESTMENT SADJ	USIVF..CE	Reuters Datastream
181	Q	US CHAIN-TYPE PRICE INDEX FOR NONRESIDENTIAL FIXED INVESTMENT	USIVFN.CE	Reuters Datastream
182	Q	US CHAIN-TYPE PRICE INDEX-NONRESL FIXED INVESTMENT,STRUCTURES	USISN..CE	Reuters Datastream
183	Q	US CHAIN-TYPE PRICE INDEX - PRIVATE FIXED INVEST, EQUIPMENT SADJ	USIVEQPCE	Reuters Datastream
184	Q	US CHAIN-TYPE PRICE INDEX FOR RESIDENTIAL FIXED INVESTMENT SADJ	USIVFR.CE	Reuters Datastream
185	Q	US CHAIN-TYPE PRICE INDEX FOR GNP SADJ	USGNP..CE	Reuters Datastream
186	Q	US CHAIN-TYPE PRICE INDEX FOR EXPORTS (NIA) SADJ	USEXN..CE	Reuters Datastream
187	Q	US CHAIN-TYPE PRICE INDEX FOR IMPORTS (NIA) SADJ	USIMN..CE	Reuters Datastream
188	Q	US CHAIN-TYPE PRICE INDEX FOR EXPORTS OF GOODS SADJ	USEXC..CE	Reuters Datastream
189	Q	US CHAIN-TYPE PRICE INDEX FOR EXPORTS OF SERVICES SADJ	USEXS..CE	Reuters Datastream
190	Q	US CHAIN-TYPE PRICE INDEX FOR IMPORTS OF GOODS SADJ	USIMC..CE	Reuters Datastream
191	Q	US CHAIN-TYPE PRICE INDEX FOR IMPORTS OF SERVICES SADJ	USIMS..CE	Reuters Datastream

7. Stock Market

	Freq.	Variable	Code	Provider
15	D	S&P 500 COMPOSITE - PRICE INDEX	S&PCOMP	Reuters Datastream
16	D	S&P INDUSTRIAL - PRICE INDEX	S&PINDS	Reuters Datastream
17	D	DOW JONES INDUSTRIALS - PRICE INDEX	DJINDUS	Reuters Datastream
18	D	DOW JONES COMPOSITE 65 STOCK AVE - PRICE INDEX	DJCMP65	Reuters Datastream
19	D	DOW JONES UTIILITIES - PRICE INDEX	DJUTILS	Reuters Datastream
20	D	NYSE COMPOSITE - PRICE INDEX	NYSEALL	Reuters Datastream
21	D	CBOE SPX VOLATILITY VIX (NEW) - PRICE INDEX	CBOEVIX	Reuters Datastream

D Time-varying specifications of Factor augmented MIDAS model

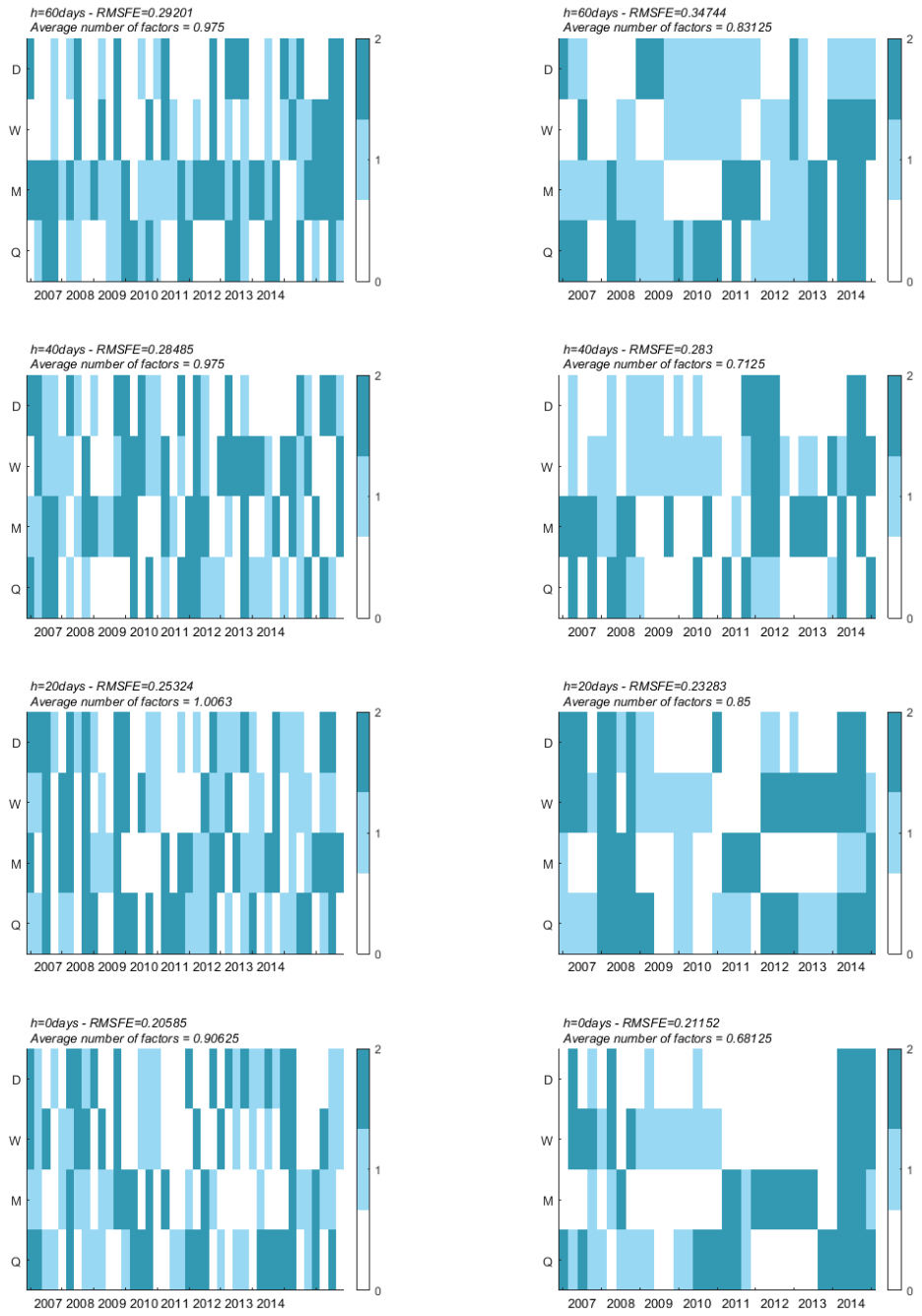

FIGURE 6: FaMIDAS specifications for various predictive cross validation sample sizes and forecasting horizons. Charts on the left correspond to a predictive cross validation sample of size $S = 1$ quarter; and the right, it shows selection of sample of size $S = 8$ quarters. On the y-axis, the groups "D", "W", "M" and "Q" stands respectively for "Daily", "Weekly", "Monthly" and "Quarterly".