Methods for Backcasting, Nowcasting and Forecasting Using Factor-MIDAS: With An Application To Korean GDP*

Hyun Hak Kim[1] and Norman R. Swanson[2]

[1]Kookmin University and [2]Rutgers University

July 2017

Abstract

We utilize mixed frequency factor-MIDAS models for the purpose of carrying out backcasting, nowcasting, and forecasting experiments using real-time data. We also introduce a new real-time Korean GDP dataset, which is the focus of our experiments. The methodology that we utilize involves first estimating common latent factors (i.e., diffusion indices) from 190 monthly macroeconomic and financial series using various estimation strategies. These factors are then included, along with standard variables measured at multiple different frequencies, in various factor-MIDAS prediction models. Our key empirical findings are that: (i) When using real-time data, factor-MIDAS prediction models outperform various linear benchmark models. Interestingly, the 'MSFE-best' MIDAS models contain no AR lag terms when backcasting and nowcasting. AR terms only begin to play a role in 'true' forecasting contexts. (ii) Models that utilize only 1 or 2 factors are 'MSFE-best' at all forecasting horizons, but not at any backcasting and nowcasting horizons. In these latter contexts, much more heavily parameterized models with many factors are preferred. (iii) Real-time data are crucial for forecasting Korean GDP, and the use of 'first available' versus 'most recent' data 'strongly' affects model selection and performance. (iv) Recursively estimated models are almost always 'MSFE-best', and models estimated using autoregressive interpolation dominate those estimated using other interpolation methods. (v) Factors estimated using recursive principal component estimation methods have more predictive content than those estimated using a variety of other (more sophisticated) approaches. This result is particularly prevalent for our 'MSFE-best' factor-MIDAS models, across virtually all forecast horizons, estimation schemes, and data vintages that are analyzed.

Keywords: nowcasting, forecasting, factor model, MIDAS.

JEL Classification: C53, G17.

____

* Hyun Hak Kim (hyunhak.kim@kookmin.ac.kr), Department of Economics, 77 Jeongneung-Ro, Seongbuk-Gu, Seoul, 02707, Korea. Norman R. Swanson (nswanson@econ.rutgers.edu), Department of Economics, 75 Hamilton Street, New Brunswick, NJ, 08901 USA. Much of this paper was written while the first author was a member of staff at the Bank of Korea, and the authors would like to thank the bank for providing a stimulating research environment, as well as for financial assistance. This paper has benefited from useful comments made by the editor, Massimiliano Marcellino, as well as 2 referees. The authors also owe many thanks to Christian Schumacher for providing his MATLAB code.

1 Introduction

In this paper, we utilize a combination of real-time data and mixed frequency modeling methods together with a variety of principal component analyses, in order to provide new evidence on the usefulness of these techniques for forecasting. More specifically, we introduce a new real-time Korean GDP dataset, which is used together with a large monthly dataset including 190 variables[1], to backcast, nowcast, and forecast Korean GDP. Our prediction models combine the mixed data sampling (MIDAS) framework of Ghysels et al. (2004), that allows for the incorporation of variables of differing frequencies, with the diffusion index framework of Stock and Watson (2002).

The difference between backcasting, nowcasting, and forecasting can be explained as follows. Suppose that the objective is to predict GDP for 2016:Q2, using a simple autoregressive model of order one, say. In a conventional setting where real-time data are not available, it is assumed that information up to 2016:Q1 is available at the time the prediction is made, so that $\widehat{\text{GDP}}_{2016:\text{Q2}} = \hat{\alpha} + \hat{\beta}\text{GDP}_{2016:\text{Q1}}$, where $\hat{\alpha}$ and $\hat{\beta}$ are parameters estimated using maximum likelihood based on recursive or rolling data windows. In a real-time context, however, this prediction is not feasible. Namely, if the prediction is to be made in April or even May of 2016, then $\text{GDP}_{2016:\text{Q1}}$ is not yet available, even in preliminary release. This issue leads to the convention of defining three different types of predictions, including backcasts (predicting past observations, which are not yet available in real-time), nowcasts (predicting concurrent observations), and forecasts (see Giannone et al. (2008) for a comprehensive discussion of backcasting and nowcasting). One advantage of carefully analyzing the data structure used in the formulation of prediction models is that we are able to simulate real-time decision making processes. In addition to Giannone et al. (2008), the reader is referred to Girardi et al. (2016) for an overview of this literature, within the context of nowcasting Euro area GDP in pseudo real-time using dimension reduction techniques.

Various key macroeconomic indicators in many countries, including Korea, are published with considerable delay and at low frequency. One such example is Korean Gross Domestic Product (GDP), which is a component of the so-called system of national accounts (SNA), and has been published quarterly by the Bank of Korea since 1955. These GDP data are 'real-time', in the sense that they are regularly updated and revised. For example, the base year of SNA data is updated every 5 years. Additionally, since the first GDP release in the 1950s, there have been 11 definitional changes affecting the entire historical record. Finally, since 2005, 'first vintage' or first release real GDP has been regularly announced about 28

[1] This large monthly Korean macroeconomic dataset, which resembles the well-known U.S. Stock and Watson (2002)' dataset, is introduced in Kim (2017).

days after the end of the corresponding calendar quarter. Second vintage data is generally released about 70 days after the end of the quarter (at which time nominal GDP is also released). In approximate conjunction with this second release, the whole prior year of data is also revised and released. Finally, another revision is made approximately 15 months later.

There are several approaches to forecasting lower frequency variables using higher frequency variables. The first approach involves use of the so-called 'bridge' model, which aggregates higher frequency variables with lower frequency variables, such as GDP. This aggregation is called a 'bridge', and this method is commonly used by central banks, since implementation and interpretation is straightforward (see e.g., Rünstler and Sédillot (2003), Golinelli and Parigi (2005) and Zheng and Rossiter (2006)). Indeed, this approach offers a very convenient solution for filtering, or aggregating, variables characterized by different frequencies. However, aggregation may lead to the loss of useful information. This issue has led to the recent development of alternative mixed frequency modeling approaches. One important approach, which is mentioned above, is called MIDAS. This approach involves the use of a regression framework that directly includes variables sampled at different frequencies. Broadly speaking, MIDAS regression offers a parsimonious means by which lags of explanatory variables of differing frequencies can be utilized; and its use for macroeconomic forecasting is succinctly elucidated by Clements and Galvao (2008). Additional recent papers in this area of forecasting include Kuzin et al. (2011), who predict Euro area GDP, Ferrara and Marsilli (2013) who predict French GDP, and Pettenuzzo et al. (2014), who discuss Bayesian implementation of MIDAS. One interesting feature of MIDAS is that the technique readily allows for the inclusion of diffusion indices. For discussion of the combination of factor and MIDAS approaches, see Marcellino and Schumacher (2010), and Section 5 of this paper. For an interesting application to the prediction of German GDP, see Schumacher (2007). The reader is additionally referred to Giannone et al. (2008), Bańbura et al. (2012), and Bańbura and Modugno (2014) for interesting discussions on the use of mixed frequeny modeling for nowcasting, and to Kuzin et al. (2013), Aastveit et al. (2014), and Mazzi et al. (2014) for a discussion on forecast combination in the current context.

Following Stock and Watson (2002), we extract diffusion indices for use in our mixed frequency forecasting models via implementation of principal component analysis (PCA). However, it is worth noting that our diffusion indices are constructed using the monthly dataset mentioned above, which is not a real-time dataset. For related evidence on the usefulness of factors thus constructed, see Stock and Watson (2012), Boivin and Ng (2005) and Kim and Swanson (2014). A further issue, in the context of using our real-time GDP dataset together with the aforementioned monthly dataset concerns the staggered availability of variables that are published at the same frequency in the latter dataset. For example, some of the predictor variables that we use are not available, even in the middle of the current month, while others are. This type of missing data leads to the so-called ‘ragged-edge’ type problem. In this paper, we tackle this issue following Wallis (1986) and Marcellino and Schumacher (2010), and estimate monthly common factors using PCA coupled with either vertical data realignment, AR data interpolation, EM algorithm based missing value estimation, or a standard state space model. MIDAS prediction models are then implemented, yielding ‘factor-MIDAS’ predictions that are available at a monthly frequency for our quarterly GDP target variable.

Our empirical findings can be summarized as follows. First and foremost, real-time data makes a difference. The utilization of real-time data in a recursive estimation framework, coupled with MIDAS, leads to the ‘MSFE-best’ predictions in our experiments. This finding is due in large part to the fact that many important economic indicators, such as CPI and Industrial Production are sampled at monthly or higher frequencies, and are useful for real-time GDP prediction at both monthly and quarterly frequencies. Indeed, when using real-time data, factor-MIDAS prediction models outperform various linear benchmark models. Interestingly, our ‘MSFE-best’ MIDAS models contain no AR lag terms when backcasting and nowcasting. AR terms only begin to play a role in ‘true’ forecasting contexts.

Second, models that utilize only 1 or 2 factors are ‘MSFE-best’ at all forecasting horizons, but not at any backcasting and nowcasting horizons. In these latter contexts, much more heavily parameterized models with many factors are preferred. In particular, while 1 or 2 factors are selected around 1/2 of the time in the cases, 5 or 6 factors are also selected around 1/2 of the time. Interestingly, there is little evidence that using an intermediate number of factors is useful. One should either specify very parsimonious 1 or 2 factor models, or one should go with our maximum of 5 or 6 factors. In summary, forecast horizon matters, in the sense that when uncertainty is most prevalent (i.e., longer forecast horizons), then parsimony ‘wins’ and ‘MSFE-best’ models utilize only 1 or 2 factors. The reverse holds as the forecast horizon reduces and instead nowcasts and backcasts are constructed. This finding is quite sensible, given the vast literature indicating that more parsimonious models are usually preferred, particularly when forecasting at longer horizons.

Third, the variable being predicted makes a difference. For Korean GDP, the use of ‘first available’ versus ‘most recent’ data ‘strongly’ affects model selection and performance. One reason for this is that ‘first available’ data are never revised, and can thus in many cases be viewed as ‘noisy’ versions of later releases of observations for the same calendar date. This is particularly true if rationality holds (see, e.g. Swanson and van Dijk (2006)).

Interestingly, when predictions are constructed using only ‘first available’ data, and when predictive accuracy is correspondingly carried out with ‘first available’ data, factor-MIDAS models without AR terms as well as other benchmark models do not work well, regardless of the number of factors specified. In these cases, pure autoregressive models dominate, in terms of MSFE. This suggests that for short forecast horizons, the persistence of Korean GDP growth is strong, and well modeled using linear AR components. Indeed, in many of these cases, our simplest linear AR models are ‘MSFE-best’. As the forecast horizon gets longer, simple linear models are no longer ‘MSFE-best’, and models without AR terms in some cases outperform models with AR terms. This suggests that uncertainty in autoregressive parameters does not carry over to other model parameters, as the horizon increases, and the role for MIDAS thus increases in importance. However, when ‘most recent’ real-time data are used exclusively in our prediction experiments, MIDAS models dominate at all forecast horizons, as mentioned above, and autoregressive lags of GDP are only useful at longer forecast horizons (of at least 6 months). Given that ‘most recent’ data are those that are most often used by empirical researchers, we thus have direct empirical evidence of the usefulness of factor-MIDAS coupled with real-time data.

Fourth, recursively estimated models are almost always ‘MSFE-best’, and models estimated using autoregressive interpolation dominate those estimated using other interpolation methods. In particular, models estimated using rolling data windows are only ‘MSFE-best’ at 3 forecast horizons, when using ‘first available’ data, and are never ‘MSFE-best’ when ‘most recent’ data are used. Also, when comparing MSFEs, only approximately 10% of models perform best when using vertical alignment or VA interpolation, with 90% favoring autoregressive or AR interpolation.

Fifth, factors constructed using recursive principal component estimation methods have more predictive content than those estimated using a variety of other (more sophisticated) approaches. This result is particularly prevalent for our ‘MSFE-best’ factor-MIDAS models, across virtually all forecast horizons, estimation schemes, and data vintages that are analyzed.

In summary, this paper introduces a new real-time dataset, offers a first look at the issue of backcasting, nowcasting, and forecasting real-time Korean GDP, and is meant to add to the burgeoning literature on the usefulness of MIDAS, diffusion indices, and real-time data for prediction. Future research questions include the following: Are robust shrinkage methods such as the lasso and elastic net useful in the context of real-time prediction, and can the methods discussed herein be modified to utilize these sorts of machine learning and shrinkage techniques? Can predictions be improved by utilizing even higher frequency data than those used here, including high frequency financial data? In the context of high frequency data, are measures of risk such as so-called realized volatility useful as predictors? Finally, are alternative "sparse" diffusion index methodologies, such as sparse principal components analysis and independent component analysis useful in real-time prediction (see, e.g. Kim and Swanson (2016))?

The rest of the paper is organized as follows. Our real-time Korean GDP dataset is introduced in Section 2. Section 3 briefly describes how to estimate common factors using recursive and non-recursive PCA methods, and discusses approaches to addressing ragged-edge data. The MIDAS framework for backcasting, nowcasting, and forecasting is discussed in Section 4. Finally, Section 5 presents the results of our forecasting experiments, and Section 6 concludes the paper.

2 Real-Time Korean Data

2.1 Notation

When constructing real-time datasets, both the data vintage (which 'release' of data we are referring to, and when it was released) and the calendar date (the actual calendar date to which the data pertains) must be delineated. Figure 1 depicts this relationship for Korean GDP.

Moreover, when constructing growth rates (e.g., log differences), data vintage is clearly relevant. It is thus important to carry forward a consistent and sensible notation, when using real-time data in model specification and estimation. Let Z be the level of a variable and z be the log difference thereof. Define:

$$z_t^{(1)} = \ln Z_t^{(1)} - \ln Z_{t-d}^{(1)}, \tag{1}$$

where $Z_t^{(1)}$ denotes the first release of Z_t, for calendar date t, and d denotes the difference taken (i.e., $d = 1$ for quarterly growth rates, and $d = 4$ for annual growth rates, when data are measured at a quarterly frequency). In practice, $z_t^{(1)}$ is not commonly used in empirical analysis, since, at calendar date t, a more recent release than $1st$ may be available for Z_{t-d}. If Z_{t-d} has already been revised once, then use of updated data may be preferred, leading to the following definition:

$$z_t^{(2)} = \ln Z_t^{(1)} - \ln Z_{t-1}^{(2)}. \tag{2}$$

For annual growth rates based on quarterly data, utilizing the latest available revision equates with constructing $z_t^{(3)} = \ln Z_t^{(1)} - \ln Z_{t-4}^{(3)}$. In summary, when we are at calendar date t, the latest observation available for date t is the first release.

In subsequent prediction experiments involving GDP, we update our forecasts at a monthly frequency, even though raw data are accumulated at only a quarterly frequency. It is thus necessary to specify monthly subscripts denoting data vintage. In particular, define:

$$_{t_m}Y_{t_q} = {}_{t_m}\mathbf{Y}_{t_q} - {}_{t_m}\mathbf{Y}_{t_q-d}, \tag{3}$$

where $\mathbf{Y}$ and Y denote the log level and the growth rates of a variable, say GDP, respectively. Here, when $d = 4$, $_{t_m}Y_{t_q}$ are annual growth rates. Suppose that t_m =2016:05 and t_q is 2016:Q1. In practice, we do not know the value of $\mathbf{Y}$ for 2016:Q2, as t_m = May 2016. In light of this, we redefine (3), taking into account the publication lag, k, as follows:

$$_{t_m}Y_{t_q} = {}_{t_m}\mathbf{Y}_{t_q-k} - {}_{t_m}\mathbf{Y}_{t_q-k-d}. \tag{4}$$

Therefore, the annual growth rate of GDP for 2016:Q1 in May 2016 is:

$$_{2014:05}Y_{2016:Q1} = {}_{2014:05}\mathbf{Y}_{2016:Q1} - {}_{2014:05}\mathbf{Y}_{2015:Q1}. \tag{5}$$

Now, let the data release be denoted by adding a superscript to the above expression, as follows:

$$_{2014:05}Y^{(3)}_{2016:Q1} = {}_{2014:05}\mathbf{Y}^{(1)}_{2016:Q1} - {}_{2014:05}\mathbf{Y}^{(3)}_{2015:Q1}, \tag{6}$$

where the superscript 3 corresponds to a third release or vintage of real-time data. Figure 2 depicts the construction of real-time GDP growth rates. Putting it all together, our real-time nomenclature for is:

$$_{t_m}Y^{(v)}_{t_q-d}, \tag{7}$$

where the sub- and super-scripts are defined above. Finally, and in order to simplify our notation, we redefine the superscript "v" so that it corresponds directly to the vintage of the growth rate, rather than the vintage of the raw data used in the construction of the growth rate. Namely, let $_{t_m}Y^{(1)}_{t_q-d}$ denote the first vintage growth rate of GDP, instead of (7). Thus, $_{t_m}Y^{(1)}_{t_q-d}$ is simply the first available growth rate of GDP for a particular calendar date, given data reporting agency release lags, d. Accordingly, $_{t_m+m}Y^{(i)}_{t_q-d}$ is the $i-$th vintage growth rate for calendar date $t_q - d$, at time $t_m + m$, where m is the feasible month for which $i-$ th vintage data are available. In the sequel, when the superscript for the vintage is omitted, we mean first vintage.

Given the above notation, we can specify forecast models using real-time data. Suppose that the objective is to predict h_q steps ahead at time t_m, using an AR(1) model. Then, the prediction model is:

$$_{t_m}Y_{t_q-d+h_q} = \alpha + \beta_{h_q} \cdot {}_{t_m}Y_{t_q-d} + \epsilon_{t_q}, \tag{8}$$

where ϵ_{t_q} is a stochastic noise term, α and β_{h_q} are coefficients estimated using maximum likelihood, and Y is defined as above. Here, vintage notation is omitted for brevity. Note that we forecast h_q periods ahead at time t_m (or t_q), but we do not have real-time information up to t_q. Therefore, the explanatory variable is lagged d quarters. Equation (8) is one of our benchmark forecasting models. Assume that we are at time t_m in the first month of the quarter, t_q. If there is a publication lag equal to 1 (i.e., $d = 1$), we 'backcast' a value of Y, for time period $t_q - d$, 'nowcast' a value of Y, for time period t_q, and 'forecast' a value of Y, for time period $t_q + h_q$.

2.2 The dataset

We have collected real-time Korean GDP beginning with the vintage available in January 2000. The calendar start date of our dataset is 1970:Q1, and data are collected through June 2014. As discussed in the introduction, first release GDP is announced 28 days after the end of the quarter, second GDP release is announced 70 days subsequent to the end of the quarter, and the third release is made available 50 days after a calendar year has passed. Finally, a fourth release is made available a full year later. These release dates have been fixed since 2005. Before then, release dates were relatively irregular, although the first release was usually around 60 days after the end of the quarter, and the second release was around 90 days after the end of the quarter. Even though GDP is finalized after approximately 2 years, there are several definitional changes, as well as regular base-year changes that subsequently affected our dataset. The revision history for Korean GDP is depicted in Figure 3. Panel (a) of the figure shows the growth rate of GDP by vintage. The plot denoted as '1st' is first release GDP, and so on. In Panel (b), revision errors are depicted. Plots denoted as '2nd', '12th' and '24th' all refer to differences relative to the first release. Prior to the 1990's, the differences were relatively large; with notable narrowing of these 'revision errors' more recently. It seems that along with the imposition of stricter release and announcement protocol, early releases have become more accurate. Panel (c) of Figure 3 depicts how GDP for certain calendar dates (i.e., 2001:Q1, 2003:Q1 and 2005:Q1) has evolved across releases. The GDP release dynamics observable in Panels (a), (b) and (c) is indicative of the fact that policy decision-making should carefully account for the real-time nature of GDP data. Panel (d) contains a histogram of first revision errors, which are the difference between first and second releases, over time. Interestingly, the first vintage is biased, as indicated by the asymmetric nature of the histogram. This suggests that the revision error history may be useful for prediction.

3 Estimating Diffusion Indexes

We estimate common latent factors (i.e., diffusion indices) using 190 monthly macroeconomic and financial variables.[2] Thereafter, we utilize our estimated factors, along with various additional variables measured at multiple different frequencies, in MIDAS prediction regressions (see Section 5 for complete details). One conventional way to estimate common factors is via the use of PCA. In order to avoid potential computational burdens associated with matrix inversions, and in order to simulate a 'real-time' environment, we use a variant thereof, called recursive PCA, following Peddaneni et al. (2004). In this section, we discuss recursive PCA and other key details associated with factor estimation, in our context.

Before continuing our discussion, it is worth noting that many other data dimension reduction, variable selection, machine learning and shrinkage methods, including partial least squares, the elastic net, and bagging, to name only a few, can be applied to the problem of constructing factor-MIDAS models. Theoretical and empirical research in this area, however, is left to future research. For related discussion of some of these methods, see Cubadda and Guardabascio (2012).

3.1 Constructing factors using ragged-edge data

Since we model real-time GDP, it is critical to match monthly data availability with GDP release vintages. In particular, some of our monthly variables are not available at certain calendar dates even though new vintages of GDP have been released by said calendar dates. For example, the consumer price index for the previous month is released early in the current

[2] Our largescale monthly predictor dataset is not measured in real-time, as is the case with our real-time Korean GDP, due to data availability. As discussed above, the predictor dataset is discussed in Kim (2017). These data have been categorized into 12 groups: interest rates, imports/exports, prices, money, exchange rates, orders received, inventories, housing, retail and manufacturing, employment, industrial production, and stocks. We extend this monthly dataset through June 2014 in the current paper. All variables are transformed to stationarity, and the final dataset closely resembles the well-known Stock and Watson dataset, which has been extensively used to estimate common factors for the U.S. economy. Note additionally that select monthly indicators in our dataset are seasonally adjusted using the Bank of Korea's X-13 ARIMA filter, and further treatment of potential seasonality in our data is left to future research. For complete details, see Kim (2017) Note also that cointegration is not accounted for in any of our models, and further exploration of the usefulness of applying cointegration restrictions in our experiments is left to future research. For discussion of cointegration and its usefulness in MIDAS models, see Götz et al. (2016), and the references cited therein.

month, whereas the producer price index is released in the middle of the month. In between these releases, new vintages of GDP are often released. This is called a ragged-edge data problem. Denote our N-dimensional monthly dataset as X_{t_m}, where time index t_m denotes the monthly frequency. Assume that the monthly observations have the following factor structure:

$$X_{t_m} = \Lambda F_{t_m} + \xi_{t_m}, \tag{9}$$

where the r-dimensional factor vector is denoted by $F_{t_m} = \left(f'_{1,t_m}, \dots, f'_{r,t_m}\right)$, Λ is an $(N \times r)$ factor loading matrix, and $r << N$. Note that we do not have monthly indicators in real-time, so that there is no prefix subscript in 9. In this formulation, the common components of X_{t_m} consist of the diffusion indices, F_{t_m}. The idiosyncratic components, ξ_{t_m}, are that part of X_{t_m} not explained by the factors. Let data matrix X be a balanced one with dimension $T_m \times N$. The most widely used methods for estimating F_{t_m} are based on static PCA, as in Stock and Watson (2002); and dynamic PCA, as in Forni et al. (2005). However, PCA is based on an eigenvalue/eigenvector decomposition of the covariance matrix of X_{t_m}, which requires inversion of this matrix. This means that the dataset must not be ragged. Therefore, we need to resolve the ragged-edge problem prior to obtaining diffusion index estimates. In this paper, we use *vertical alignment* and *AR interpolation* for missing values. Another convenient way to solve the ragged-edge problem is proposed by Stock and Watson (2002), who use the EM algorithm together with standard PCA. Additionally, one can write the factor model in state-space form in order to handle missing values at the end of each variables' sample, following Doz et al. (2012).[3] Our approaches to the ragged-edge problem are the following:

Vertical alignment (VA) interpolation of missing data:

The simplest way to solve the ragged-edge problem is to directly balance any unbalanced datasets. In particular, assume that variable i is released with a k_i month publication lag. Thus, given a dataset in period T_m, the final observation available for variable i is for period $T_m - k_i$. The realignment proposed by Altissimo et al. (2010) is:

$$\tilde{X}_{i,t_m} = X_{i,t_m - k_i}, \quad \text{for} \quad t_m = k_i + 1, ..., T_m. \tag{10}$$

Applying this procedure for each series, and harmonizing at the beginning of the sample, yields a balanced dataset, $\tilde{X}_{t_m}$, for $t_m = \max\left(\{k_i\}_i = 1^N\right) + 1, ..., T_m$. Given this new dataset, PCA can be immediately implemented. Although easy to use, a disadvantage of this method is that the availability of data determines dynamic cross-correlations between

[3] Doz et al. (2012) use the Kalman filter and smoother for estimation.

variables. Furthermore, statistical release dates for each variable are not the same over time, for example, due to major revisions.

Autoregressive (AR) interpolation of missing data:

As an alternative to vertical alignment, we use univariate autoregressive models for individual monthly indicators, X_i. Namely, specify and estimate the following models:

$$X_{i,t} = \sum_{s=1}^{p_i} \rho_s X_{i,t-s} + u_{i,t}, \qquad i = 1, \ldots, k, \tag{11}$$

where p_i is the lag length, and is selected using the Schwarz Information Criterion (SIC), coefficients ρ are estimated using maximum likelihood, and $u_{i,t}$ is a white noise error term. This AR method depends only on the univariate characteristics of the variable in question, and not on the broader macroeconomic environment from within which the data are generated. However, it is very easy to implement and is an intuitive approach.

EM algorithm for estimating missing data:

The ragged-edge problem essentially concerns estimating missing values. Stock and Watson (2002) propose using the EM algorithm to replace missing values and subsequently carry out PCA. The EM algorithm is initialized with an estimate of the missing data, which is usually set equal to the unconditional mean (this is also the approach that we use). Then, the completed dataset is used to estimate factors using PCA. This algorithm is repeated in two steps, the E-step and the M-step. We briefly explain these steps, and the reader is refereed Schumacher and Breitung (2008) for details. Consider a dataset, X_{t_m}, and pick variable i, say $X_i = (x_{i,1}, ..., x_{i,t_m})'$. Suppose that variable i has missing values due to publication lags. Set $X_i^{obs} = P_i X_i$, where P_i represents the relationship between the full vectors and the ones with missing values. If no missing values are found, then P_i is the identity matrix. As we only observe a subset of X, initialize the EM algorithm by replacing missing values with the unconditional mean of X_i^{obs}, yielding initial estimates of factors and loadings (using PCA), say F^0 and Λ^0. Now iterate this procedure. In the $(j-1)$-th iteration, the E-step updates the estimates of the missing observations using the expectation of the variable X_i conditional on X_i^{obs}, with factors and loadings from the $j-th$ iteration, F^{j-1} and Λ^{j-1}, as follows:

$$X_i^j = F^{j-1}\Lambda^{j-1} + P_i' \left(P_i' P_i\right)^{-1} \left(X_i^{obs} - P_i F^{j-1} \Lambda_i^{j-1}\right), \tag{12}$$

Run the E-step for all i, in each iteration. The M-step involves re-estimating the factors and loadings using ordinary PCA. Continue until convergence is achieved.

State-space model (Kalman filtering) for estimating missing data:

Another popular approach for estimating factors from large datasets is the state-space approach based on Doz et al. (2012) and Giannone et al. (2008). The factor model represented in state-space form is based on the (9), with factors represented using an autoregressive structure, as follows:

$$\Psi(L_m)F_{t_m} = \mathbf{A}\eta_{t_m}, \tag{13}$$

where $\Psi(L_m)$ is a lag polynomial, given by $\sum_{i=1}^{p} \Psi_i L_m^i$, and η_{t_m} is an orthogonal dynamic shock. The state and transition equation can easily be estimated via maximum likelihood (ML). Doz et al. (2012) propose using quasi-ML for large datasets, when conventional ML is not feasible. In particular, as ML estimation involves initialization of factors based on the use of ordinary PCA, one needs a completed data matrix. Marcellino and Schumacher (2010) remove missing values from the end of sample to make it balanced, and estimate initial factors using ordinary PCA. In our forecasting experiments, initial factors are extracted from the completed matrix that is completed using VA and AR interpolation. Then, likelihoods are calculated and evaluated using the Kalman filter. More specifically, given an initial set of factors, estimate loadings by regressing X_{t_m} on the factors. Then, obtain the covariance matrix of the idiosyncratic part from (9), $\sum_\xi$, where $\xi_{t_m} = X_{t_m} - \Lambda F_{t_m}$. Now, estimate a vector AR(p) on the factors, F_{t_m}, yielding coefficient matrix, $\Psi(L)$, and residual covariance matrix, $\sum_\varsigma$ where $\varsigma_{t_m} = \Psi(L_m)F_{t_m}$. Let V be the eigenvectors corresponding to E, where E is a diagonal matrix whose diagonal elements are the eigenvalues in descending order, and zero otherwise. Then, set $P = VE^{-1/2}$. As a final step, the Kalman smoother is used to yield new estimates of the factors.

3.2 Recursive and standard principal component analysis (RPCA and OPCA)

PCA is widely used to estimate factors or diffusion indices in large data environments (see Kim and Swanson (2016) and the references cited therein). In this paper, we utilize PCA, called OPCA in our later discussion of our empirical findings to differentiate it from recursive PCA (discussed below). Our implementation follows the approach of Stock and Watson (2002), and is standard to the literature. For a discussion of the generated regressor problem associated with our empirical implementation of factor augmented prediction models, see Bai and Ng (2008). We conjecture that their results also apply in the context of the MIDAS models considered in this paper, although proof thereof is left to future research.

In general, PCA is quite convenient as it uses standard eigenvalue decompositions of data covariance matrices. However, these matrix operations may be time consuming in certain real-time environments. In light of this, RPCA has been proposed by Peddaneni et al. (2004), and is a natural approach to use in our context, as new data arrive in real-time and need to be incorporated into our prediction models. Also, suppose that F_{t_m} is estimated using PCA. Principal components (factors) in this context are linear combinations of variables that maximize the variance of the data, and there is no guarantee that factor loadings are stationary at each point in time, particularly with large datasets. For example, the factor loadings at times t and $t+1$ may have different signs. Recursive PCA attempts to address these issues, in part by not requiring the calculation of the whole covariance matrix of data with the arrival of each new datum. Without loss of generality, consider a standardized random vector at time t, say x_t, with dimension n. Our aim is to find the principal components of x at time t. To begin, define the covariance (or correlation) matrix of x as:

$$\boldsymbol{R}_t = \frac{1}{t}\sum_{i=1}^{t} x_i x_i' = \frac{t-1}{t}\boldsymbol{R}_{t-1} + \frac{1}{t} x_t x_t'. \tag{14}$$

If $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}$ are the orthonormal eigenvector and diagonal eigenvalue matrices of $\boldsymbol{R}$, respectively, then: $\boldsymbol{R}_t = \boldsymbol{Q}_t \boldsymbol{\Lambda}_{\boldsymbol{t}} \boldsymbol{Q}_t'$ and $\boldsymbol{R}_{t-1} = \boldsymbol{Q}_{t-1}\boldsymbol{\Lambda}_{t-1}\boldsymbol{Q}_{t-1}'$. We can rewrite (14) as:

$$\boldsymbol{Q}_t \left(t\boldsymbol{\Lambda}_t\right) \boldsymbol{Q}_t' = x_t x_t' + (t-1)\, \boldsymbol{Q}_{t-1}\boldsymbol{\Lambda}_{t-1}\boldsymbol{Q}_{t-1}'. \tag{15}$$

If we let $\alpha_t = \boldsymbol{Q}_{t-1}' x_t$, (15) can be written as: $\boldsymbol{Q}_t \left(t\boldsymbol{\Lambda}_t\right) \boldsymbol{Q}_t' = \boldsymbol{Q}_{t-1}\left[(t-1)\,\boldsymbol{\Lambda}_{t-1} + \alpha_t\alpha_t'\right]\boldsymbol{Q}_{t-1}'$. If $\boldsymbol{V}_t$ and $\boldsymbol{D}_t$ are the orthonormal eigenvector and diagonal eigenvalue matrices of $(t-1)\,\boldsymbol{\Lambda}_{t-1} + \alpha_t\alpha_t'$, then:

$$(t-1)\,\boldsymbol{\Lambda}_{t-1} + \alpha_t\alpha_t' = \boldsymbol{V}_t\boldsymbol{D}_t\boldsymbol{V}_t'. \tag{16}$$

Therefore,

$$\boldsymbol{Q}_t \left(t\boldsymbol{\Lambda}_t\right) \boldsymbol{Q}_t' = \boldsymbol{Q}_{t-1}\boldsymbol{V}_t\boldsymbol{D}_t\boldsymbol{V}_t\boldsymbol{Q}_{t-1}'. \tag{17}$$

By comparing both sides of (17), the recursive eigenvector and eigenvalue update rules turn out to be $\boldsymbol{Q}_t = \boldsymbol{Q}_{t-1}\boldsymbol{V}_t$ and $\boldsymbol{\Lambda}_t = \boldsymbol{D}_t/t$. Now,it remains to estimate the eigenvectors and eigenvalues of $(t-1)\,\boldsymbol{\Lambda}_{t-1} + \alpha_t\alpha_t'$, which is equivalent to estimating $\boldsymbol{V}_t$ and $\boldsymbol{D}_t$. It is very difficult to analytically solve for $\boldsymbol{V}_t$ and $\boldsymbol{D}_t$, and so Peddaneni et al. (2004) instead use first order perturbation analysis. Consider the following sample perturbation to the eigenvalue matrix, $(t-1)\,\boldsymbol{\Lambda}_{t-1} + \alpha_t\alpha_t'$. When t is large, this matrix is essentially a diagonal matrix, which means that $\boldsymbol{D}_t$ will be close to $(t-1)\,\boldsymbol{\Lambda}_{t-1}$, and $\boldsymbol{V}_t$ will be close to the identity matrix, $\boldsymbol{I}$. The matrix $\alpha_t\alpha_t'$ is said to perturb the diagonal matrix $(t-1)\,\boldsymbol{\Lambda}_{t-1}$, and as a result, $\boldsymbol{D}_t = (t-1)\,\boldsymbol{\Lambda}_{t-1} + \boldsymbol{P}_\Lambda$ and $\boldsymbol{V}_t = \boldsymbol{I} + \boldsymbol{P}_V$, where $\boldsymbol{P}_\Lambda$ and $\boldsymbol{P}_V$ are small perturbation matrices. Once we find these perturbation matrices, we can solve the problem. Let $\mathbf{\Lambda} = (t-1)\,\mathbf{\Lambda}_{t-1}$. Then:

$$\begin{aligned}\boldsymbol{V}_t\boldsymbol{D}_t\boldsymbol{V}_t' &= (\boldsymbol{I}+\boldsymbol{P}_V)(\mathbf{\Lambda}+\boldsymbol{P}_\Lambda)(\boldsymbol{I}+\boldsymbol{P}_V)' \\ &= \mathbf{\Lambda}+\mathbf{\Lambda}\boldsymbol{P}_V'+\boldsymbol{P}_\Lambda+\boldsymbol{P}_\Lambda\boldsymbol{P}_V'+\boldsymbol{P}_V\mathbf{\Lambda}+\boldsymbol{P}_V\mathbf{\Lambda}\boldsymbol{P}_V'+\boldsymbol{P}_V\boldsymbol{P}_\Lambda+\boldsymbol{P}_V\boldsymbol{P}_\Lambda\boldsymbol{P}_V' \qquad (18)\\ &= \mathbf{\Lambda}+\boldsymbol{P}_\Lambda+\boldsymbol{D}\boldsymbol{P}_V'+\boldsymbol{P}_V\boldsymbol{D}+\boldsymbol{P}_V\mathbf{\Lambda}\boldsymbol{P}_V'+\boldsymbol{P}_V\boldsymbol{P}_\Lambda\boldsymbol{P}_V'\end{aligned}$$

Substituting this equation into (16), and assuming that $\boldsymbol{P}_V\mathbf{\Lambda}\boldsymbol{P}_V'$ and $\boldsymbol{P}_V\boldsymbol{P}_\Lambda\boldsymbol{P}_V'$ are negligible, we have that: $\alpha_t\alpha_t' = \boldsymbol{P}_\Lambda + \boldsymbol{D}\boldsymbol{P}_V' + \boldsymbol{P}_V\boldsymbol{D}$**.** The fact that $\boldsymbol{V}$ is orthonormal yields an additional characterization of $\boldsymbol{P}_V$. Substituting $\boldsymbol{V} = \boldsymbol{I}+\boldsymbol{P}_V$ into $\boldsymbol{V}\boldsymbol{V}' = \boldsymbol{I}$, and assuming that $\boldsymbol{P}_V\boldsymbol{P}_V' \approx 0$, we have that $\boldsymbol{P}_V = -\boldsymbol{P}_V'$. Thus, combining the fact that the $\boldsymbol{P}_V$ is antisymmetric with the fact that $\boldsymbol{P}_\Lambda$, and $\boldsymbol{D}_t$ are diagonal, yields the following solution to our problem:

$$\begin{aligned}\alpha_i^2 &= (i,i)^{th} \text{ element of } \boldsymbol{P}_\Lambda \qquad (19)\\ \frac{\alpha_i\alpha_j}{\lambda_j+\alpha_j^2-\lambda_i-\alpha_i^2} &= (i,j)^{th} \text{ element of } \boldsymbol{P}_V,\ i\neq j, \text{ and } 0 = (i,i)^{th} \text{ element of } \boldsymbol{P}_V.\end{aligned}$$

At time t, use the covariance matrix, $\mathbf{R}_{k-1}$, which is available for period $t-1$, and collect eigenvalues and eigenvectors into Λ_{t-1} and $\mathbf{Q}_{k-1}$, respectively. More specifically, with each a new datum, x_t, calculate $\alpha_t = \mathbf{Q}_{t-1}'x_t$. Next, use (19), to find the perturbation matrices, $\mathbf{P}_V$ and $\mathbf{P}_\Lambda$. Then, estimate the eigenvector matrix, $\tilde{\mathbf{Q}}_t = \mathbf{Q}_{t-1}(I+\mathbf{P}_\Lambda)$. Then, standardize $\tilde{\mathbf{Q}}_t$, using $\hat{\mathbf{Q}}_t = \tilde{\mathbf{Q}}_t\tilde{\mathbf{S}}_t$, where $\tilde{\mathbf{S}}_t$ is a diagonal matrix containing the inverse of the norms of each column of $\tilde{\mathbf{Q}}_t$. Finally, estimate the eigenvalue, $\hat{\Lambda}_t = \hat{\mathbf{Q}}_t'\mathbf{R}_t\hat{\mathbf{Q}}_t$.

4 Backcasting, Nowcasting, and Forecasting Using MIDAS

4.1 Factor-MIDAS

The MIDAS approach for forecasting with real-time data was developed by Clements and Galvao (2008, 2009). Building on their work, the factor-MIDAS approach utilized in the sequel was developed by Marcellino and Schumacher (2010). Note that factor-MIDAS is essentially conventional MIDAS augmented to include explanatory variables that are common factors extracted from higher frequency variables and datasets. More specifically, suppose that Y_{t_q} is sampled at a quarterly frequency. Let X_{t_m} be sampled at a higher frequency - for example, if it is sampled at a monthly frequency, then $m = 3$. The factor-MIDAS model for forecasting h_q quarters ahead is:

$$Y_{t_q+h_q} = \beta_0 + \beta_1 B\left(L^{1/m}, \theta\right) \hat{F}^{(3)}_{t_m} + \varepsilon_{t_q}, \tag{20}$$

where $B\left(L^{1/m}, \theta\right) = \sum\limits_{j=0}^{j^{\max}} b(j,\theta) L^{j/m}$ is the exponential Almon lag with

$$b(j,\theta) = \frac{\exp\left(\theta_1 j + \theta_2 j^2\right)}{\sum_{j=0}^{j^{max}} \exp\left(\theta_1 j + \theta_2 j^2\right)}, \tag{21}$$

and with $\theta = (\theta_1, \theta_2)$. Here, $\hat{F}_{t_m}$ is a set of monthly factors estimated using one of the various approaches discussed in the previous section, $L^{j/m} X^{(m)}_t = X^{(m)}_{t-j/m}$, and $\hat{F}^{(3)}_{t_m}$ is skip sampled from the monthly factor vector, $\hat{F}_{t_m}$. That is, every third observation starting from the final one is included in the predictor, $\hat{F}^{(3)}_{t_m}$. In this formulation, all monthly factors are in the set of predictors, and are appropriately lagged. If we apply our real-time dataset structure in this framework, the model in (20) is:

$$_{t_m}Y_{t_q+h_q} = \beta_0 + \beta_1 B\left(L^{1/m}, \theta\right) F^{(3)}_{t_m} + \varepsilon_{t_q}, \tag{22}$$

and assuming that there are r factors, $F_{t_m,1}, F_{t_m,2}, ..., F_{t_m,r}$, we have that:

$$_{t_m}Y_{t_q+h_q} = \beta_0 + \sum_{i=1}^{r} \beta_{1,i} B_i\left(L^{1/m}, \theta_i\right) F^{(3)}_{t_m,i} + \varepsilon_{t_q+h_q}. \tag{23}$$

Since we do not have monthly real-time data and we interpolate missing values at the end of each monthly indicator, F_{t_m} always exists at time t_m. If we are in the first month of the quarter and the dependent variable from previous quarter is not available, we ‘backcast’ the previous quarter’s value, ‘nowcast’ the current quarter, and ‘forecast’ future quarters, as discussed above. For example, the backcast of Y_{t_q-1} at time t_m, where t_m is the first month of the quarter is:

$$_{t_m}Y_{t_q-1} = \beta_0 + \beta_1 B\left(L^{1/m}, \theta\right) F^{(3)}_{t_m-1} + {}_{t_m}\varepsilon_{t_q-1}. \tag{24}$$

Note that $t_q - 1$ denotes the previous quarter and $t_m - 1$ denotes the previous month. The nowcast of Y_{t_q} at time t_m, where t_m is the first month of the quarter is:

$$_{t_m}Y_{t_q} = \beta_0 + \beta_1 B\left(L^{1/m}, \theta\right) {}_{t_m}F^{(3)}_{t_m} + {}_{t_m}\varepsilon_{t_q}, \tag{25}$$

and for the second month of the quarter, the nowcast is:

$$ {}_{t_m+1}Y_{t_q} = \beta_0 + \beta_1 B\left(L^{1/m}, \theta\right) {}_{t_m+1}F^{(3)}_{t_m+1} + {}_{t_m+1}\varepsilon_{t_q}. \tag{26} $$

Now, define the h_q-ahead forecast at time t_m as follows:

$$ {}_{t_m}Y_{t_q+h_q} = \beta_0 + \beta_1 B\left(L^{1/m}, \theta\right) {}_{t_m}F^{(3)}_{t_m} + {}_{t_m}\varepsilon_{t_q+h_q}. \tag{27} $$

Finally, Clements and Galvao (2008) extend MIDAS by adding autoregressive (AR) terms, yielding models of the following variety:

$$ {}_{t_m}Y_{t_q+h_q} = \beta_0 + \eth Y_{t_q} + \sum_{i=1}^{r} \beta_{1,i} B_i\left(L^{1/m}, \theta_i\right) F^{(3)}_{t_m,i} + \varepsilon_{t_q+h_q}. \tag{28} $$

All of the above models are analyzed in our forecasting experiments.

In closing this section, it should be noted that, according to Ghysels et al. (2004) and Andreou et al. (2010), given θ_1 and θ_2, the exponential lag function, $B(L^{1/m}, \theta)$, provides a parsimonious estimate that can proxy for monthly lags of the factors, as long as j is sufficiently large. It remains how to estimate θ and β. Marcellino and Schumacher (2010) suggest using nonlinear least squares (NLS), yielding coefficients, $\hat{\theta}$ and $\hat{\beta}$. In our experiments, all coefficients are estimated using NLS, except in cases where least squares can directly be applied.

4.2 Other MIDAS specifications

Marcellino and Schumacher (2010) utilize two different MIDAS specifications, including smoothed MIDAS, which is a restricted form of the above MIDAS model with different weights on monthly indicators, and unrestricted MIDAS, which relaxes restrictions on the lag polynomial used. These MIDAS models are explained in the context of the models we implement, as given in equations (25).

Unrestricted MIDAS

Another alternative version of MIDAS involves using an unrestricted lag polynomial when weighting the explanatory variables (i.e. the factors). Namely, let:

$$ {}_{t_m}Y_{t_q+h_q} = \beta_0 + \mathbf{C}\left(L_m\right) \hat{F}^{(3)}_{t_m} + \varepsilon_{t_q+h_q}, \tag{29} $$

where $\mathbf{C}(L_m) = \sum_{j=0}^{j^{max}} \mathbf{C}_j L_m^j$ is an unrestricted lag polynomial of order j. Koenig et al. (2003) propose a similar model in the context of forecasting with real-time data, but not with factors. Marcellino and Schumacher (2010) and Foroni et al. (2015) provide a theoretical justification for this model and derive MIDAS as an approximation to a forecast equation from a high-frequency factor model in the presence of mixed sampling frequencies. Here, $\mathbf{C}(L_m)$ and β_0 are estimated using least squares. Lag order specification in our forecasting experiments is done in two different ways. When using a fixed scheme where $j = 0$, automatic lag length selection is carried out using the SIC and our model only uses t_m dated factors in forecasting.
Smoothed MIDAS

Altissimo et al. (2010) propose a new Eurocoin Index, an indicator of economic activity in real-time. The index is based on a method to obtain a smoothed stationary time series from a large data set. Their index and methodology builds on that discussed in Marcellino and Schumacher (2010), and is used to nowcast and forecast German GDP. In particular, their model can be written as:

$$\begin{aligned} {}_{t_m}Y_{t_q+h_q} &= \hat{\mu}_Y + \mathbf{G}\hat{F}_{t_m}, \quad \text{and} && (30) \\ \mathbf{G} &= \tilde{\Sigma}_{Y,F}(h_m) \times \hat{\Sigma}_F^{-1}, && (31) \end{aligned}$$

where $\hat{\mu}_Y$ is the sample mean of GDP, assuming that the factors are standardized, and $\mathbf{G}$ is a projection coefficient matrix. Here, $\hat{\Sigma}_F$ is the estimated sample covariance of the factors, and $\tilde{\Sigma}_{Y,F}(j)$ is a particular cross-covariance with j monthly lags between GDP and the factors, defined as follows:

$$\tilde{\sum}_{Y,F}(j) = \frac{1}{t^* - 1} \sum_{m=M+1}^{t_m} {}_mY_{t_q} \hat{F}_{m-j}^{(3)\prime}, \qquad (32)$$

where $t^* = \text{floor}\left[(t_m - (M+1)/3)\right]$ is the number of observations available to compute the cross covariance, for $j = -M, ..., M$; and $M \geq 3h_q = h_m$, under the assumption that both GDP and the factors are demeaned. Note that $h_m = 3 \cdot h_q$. Complete computational details are given in Altissimo et al. (2010) and Marcellino and Schumacher (2010). This so-called ‘smoothed MIDAS’ is a restricted form of the MIDAS model given in (20), with a different lag structure.

Note that the models estimated in this paper are univariate models, and the reader is referred to Ghysels (2016) and Schorfheide and Song (2015) for a discussion of the use of multivariate models in the current context. The former authors use a direct generalization of the methods used in this paper, while the latter authors implement their analysis in a

Bayesian framework.

5 Empirical Results

5.1 Benchmark models and experimental setup

In addition to the MIDAS models discussed above, we specify and estimate a number of benchmark models, when forecasting real-time Korean GDP. These include:

- *Autoregressive Model:* We backcast, nowcast and forecast GDP growth rates, ${}_{t_m}\hat{Y}_{t_q+h_q}$, h_q-steps ahead, using autoregressions with p lags, where p is selected using the SIC. Note that our AR model does not use monthly indicators; but since lagged GDP, as well as revised GDP, are available at various dates throughout the quarter, we still update our predictions on a monthly basis. The model is:

  $$ {}_{t_m}\widehat{Y}_{t_q+h_q} = \hat{\beta}_0 + \hat{\beta}_1 \cdot {}_{t_m}Y_{t_q-1} + \ldots + \hat{\beta}_p \cdot {}_{t_m}Y_{t_q-p} \tag{33} $$

  Note that although AR forecast are not updated within a given month, the real-time data are revised over time. Hence, we still utilize our earlier monthly "vintage" notation in (33)

- *Random Walk Model:* We implement a standard random walk model, in which the growth rate is assumed to be constant, although this constant value is re-estimated recursively, at each point in time.

- *Combined Bivariate Autoregressive Distributed Lag (CBADL) Model:* We use the so-called bridge equation, since it is widely used to forecast quarterly GDP using monthly data (see, e.g. Baffigi et al. (2004) and Barhoumi et al. (2008)), particularly at central banks. The CBADL model, which is a standard bridge equation, uses monthly indicators as regressors to predict GDP. Forecasts are constructed using a three step procedure, as follows:

  Step 1 - Construct forecasts of all N monthly explanatory variables, where m is selected using the SIC. Namely, specify and estimate: $X_{i,t_m} = \rho_1 X_{i,t_m-1} + \cdots \rho_m X_{i,t_m-m} + \zeta_{i,s}$, for all $i = 1, ..., N$.

  Step 2 - Use lagged values of GDP as well as predictions of each individual monthly explanatory variable, order to obtain N alternative quarterly forecasts of GDP. Namely, specify and estimate:

$${}_{t_m}Y_{i,t_q+h_q} = \mu_Y + \gamma_1 Y_{t_q-1} + \cdots + \gamma_{q_y} Y_{t_q-q} + \beta_{i,0}\widehat{X}_{i,t_m} + \cdots + \beta_{i,m}\widehat{X}_{i,t_m-m} + \upsilon_{i,t_q+h_q}.$$

Note that q is also selected by BIC

Step 3 - Construct a weighted average of the above predictions. Namely:

$${}_{t_m}\hat{Y}^{CBADL}_{t_q+h_q} = \frac{1}{N}\sum_{i=1}^{N} {}_{t_m}\hat{Y}_{i,t_q+h_q}.$$[4]

- *Bridge Equation with Exogenous Variables (BEX) :* This method is identical to the above CBADL model except that the model in Step 2 is replaced with:

$${}_{t_m}Y_{i,t_q+h_q} = \mu_Y + \beta_{i,0}\widehat{X}_{i,t_m} + \cdots + \beta_{i,m}\widehat{X}_{i,t_m-m} + \upsilon_{i,t_q+h_q}. \tag{34}$$

- *Forecast Combination (Mean):* It is well known that forecast combination can be useful for forecasting macroeconomic variables. Timmermann (2006), Kim and Swanson (2014), and many others discuss this phenomenon. We consider various forecast combinations made by forming equally weighted averages of the predictions constructed using subsets of our prediction models. See below for further discussion.

Note that the real-time nature of our experiments is carefully maintained when specifying and estimating these models. Additionally, in all experiments, prediction model estimation is carried out using both recursive and rolling data windows, with the rolling window length set equal to 8 years (i.e., 32 periods of quarterly GDP and 96 monthly observations). All recursive estimations begin with 8 years of data, with windows increasing in length prior to the construction of each new real-time forecast. Out-of-sample forecast performance is evaluated using predictions beginning in 2000:Q1 and ending in 2013:Q4, and for each quarter, three monthly predictions are made. Figure 4 depicts the monthly/quarterly structure of our prediction experiments.

Table 1 summarizes the forecast models and estimation methods used. In this table, AR, CBADL, and BEX denote the benchmark models, which do not use any factors, and are our alternatives to MIDAS. The two interpolation methods discussed above (i.e., AR and VA interpolation) for addressing the ragged-edge problem are used when estimating factors via implementation of OPCA and RPCA. In addition, the EM algorithm and Kalman Filtering (KF) are used to estimate factors, without interpolation. Once factors are estimated, they are plugged into five different varieties of MIDAS regression model, including: Basic MIDAS w/o AR terms, Basic MIDAS w/ AR terms, Smooth MIDAS, Unrestricted MIDAS w/o AR terms, and Unrestricted MIDAS w/ AR terms. This setup is summarized in Table 1.

[4] Stock and Watson (2012) and Kim and Swanson (2016) implement a version of this model.

In order to assess predictive performance, we construct mean square forecast errors (MSFEs). In conventional datasets that do not contain real-time data, MSFE statistics can be constructed by simply comparing forecasts with actual values of GDP. In the current context, we have two issues. First, we can estimate our forecasting models, in real-time, using only first available data. This is one case considered, and is referred to as our 'first available' case. In this case, when constructing MSFEs, we compare predictions with first available GDP. Second, we can estimate our forecasting models using currently available data, at each point in time. When using currently available data, the most recent observations in any given dataset have undergone the least revision, while the most distant observations have potentially been revised many times. This is the second case considered, and is referred to as our 'most recent' case. In the second case, when constructing MSFEs, we compare predictions with the most recently available (and fully revised or 'recent') GDP observations. The second case is closest to that implemented by practitioners that wish to use as much information as possible when constructing forecasts, and in this case, given that Korean GDP is fully revised after 2 years, which corresponds to the 5^{th} vintage, we compare forecasts with actual data defined as ${}_{t_m}Y_{t_q}^{(5)}$. In general, the MSFE of the $i-$th model for h_q-step ahead forecasts is defined as follows:

$$MSFE_{i,h_q}^{(j)} = \sum_{t=R-h_q+2}^{T_q-h_q+1} \left({}_{t_m+3h_q+s}Y_{t_q+h_q}^{(j)} - {}_{t_m}\hat{Y}_{i,t_q+h_q} \right)^2, \quad j = 1, ... \tag{35}$$

where $R-h_q+2$ is the in-sample period, T_q-h_q+1 denotes the total number of observations, ${}_{t_m+3h_q+s}Y_{t_q+h_q}^{(j)}$ is the observed value of the GDP growth rate, for calendar date t_q+h_q when it is available, so that s denotes the smallest integer value needed in order to ensure availability of actual GDP growth rate data, $Y_{t_q+h_q}^{(j)}$ in real-time, and ${}_{t_m}\hat{Y}_{i,t_q+h_q}$ is the predicted value at t_q+h_q, for the $i-$th model. For example, we forecast the GDP growth rate in 2016:Q1 at 2015:04, called ${}_{2015:04}\hat{Y}_{2016:Q1}$, and the first calendar date at which time we can observe data for 2016:Q1 is May 2016, i.e. ${}_{2016:05}Y_{2016:Q1}^{(1)}$. As discussed above, we evaluate model performance using 'first available', and 'most recent' data. In practice, we construct $MSFE_{i,h_q}^{(first)}$ and $MSFE_{i,h_q}^{(recent)}$, respectively.

Our strawman model for carrying out statistical inference using MSFEs is the autoregressive model, and said inference is conducted using the Diebold and Mariano (1995) test (hereafter, the DM test). The null hypothesis of the DM test is that two models perform equally, when comparing squared prediction loss. Namely, we test:

$$H_0 : E\left[l\left(\varepsilon_{t+h|t}^{AR}\right)\right] - E\left[l\left(\varepsilon_{t+h|t}^{i}\right)\right] = 0, \tag{36}$$

where $\varepsilon^{AR}_{t+h|t}$ is the prediction error associated with the strawman autoregressive model, $\varepsilon^{i}_{t+h|t}$ is the prediction error of the $i-$th alternative model, and $l(\cdot)$ is the quadratic loss function. If a DM statistic under the null hypothesis is positive and significantly different from zero, then we have evidence that model i outperforms the strawman model. The DM statistic is $DM = \frac{1}{P}\sum\limits_{i=1}^{P}\frac{d_t}{\hat{\sigma}_{\bar{d}}}$, where $d_t = \left(\widehat{\varepsilon^{AR}_{t+h|t}}\right)^2 - \left(\widehat{\varepsilon^{i}_{t+h|t}}\right)^2$, $\bar{d}$ is the mean of d_t, $\hat{\sigma}_{\bar{d}}$ is a heteroskedasticity and autocorrelation robust estimator of the standard deviation of $\bar{d}$, and $\widehat{\varepsilon^{AR}_{t+h|t}}$ and $\widehat{\varepsilon^{i}_{t+h|t}}$ are the estimated prediction errors corresponding to $\varepsilon^{AR}_{t+h|t}$ and $\varepsilon^{i}_{t+h|t}$, respectively.

As pointed out by a referee, one important area of ongoing research in the current context involves the construction of density forecasts. A key paper on this topic is Aastveit et al. (2016). Although this topic is left to future research, it is useful to note that many types of density forecasts are available to the practitioner. For example, in the online appendix to this paper we provide selected kernel density plots, for various values of $h = 1$, which are simply based on the distribution of all models' predictions from our experiments. Interestingly, at least based on these naive figures, it appears that downside risk is greater than upside risk, prior to the Great Recession around 2008.

5.2 Experimental findings

There are a number of methodological as well as empirical conclusions that emerge upon examination of the results from our forecasting experiments. Prior to listing these findings, however, it is useful to recall the structure of our experiments. In particular, recall that we construct backcasts, nowcasts, and forecasts. Each of these differ only in the timing of the predictions, relative to currently available data. To be specific, recall that in following our above notational setup, we construct three types of MSFEs. Consider construction of MSFEs using 'first available' data as the 'actual data' against which predictions are compared.[5]

Suppose that we backcast the growth rate for 2010:Q4 in January 2011, which is ${}_{2011:01}\hat{Y}_{2010:Q4}$. Accuracy is then evaluated with 'first available' or 'most recent' data. The prediction error is then

$$\varepsilon^{(first)} = {}_{2011:02}Y^{(first)}_{2010:Q4} - {}_{2011:01}\hat{Y}_{2010:Q4} \qquad (37)$$

Note that in January 2011, $Y_{2010:Q4}$ is not available, and so evaluation is made using actual data from February. Now, assume that we are interested instead with evaluating the accuracy of our prediction in January 2016, with data available at that later date, called *recent* data,

[5] We also use 'most recent' data as our actual data, when constructing MSFEs. This approach is probably the most consistent with actual practice at central banks, for example.

say. This is typically done in non real-time data contexts. The prediction error is now

$$\varepsilon^{(recent)} = {}_{2016:01}Y_{2010:Q4}^{(recent)} - {}_{2011:01}\hat{Y}_{2010:Q4} \tag{38}$$

With this distinction, we can construct $MSFE_{-1}^{first}$ and $MSFE_{-1}^{recent}$, where -1 denotes 'backcast'. In same way, we can construct $MSFE_{h}^{vintage}$, where h is forecast horizon and $vintage = 1, ..., recent$. Here, $h = 1, 2, 3$ denotes current quarter nowcast predictions in the first, second and third months of the quarter. Naturally, once we reach $h = 4, ...$, forecasts are for the next quarter, etc. In our out-of-sample experiments, we consider many vintages of data, but report only on accuracy using 'first' and most 'recent' vintages. See Croushore and Stark (2001) and Stark and Croushore (2002) for a detailed discussion of the different methods available for assessing predictive accuracy in the types of experiments carried out in this paper.

Before turning to a discussion of our main prediction experiment results, we summarize three methodological findings that are potentially useful for applied practitioners. First, recall that the ragged-edge data problem can be addressed in a number of ways. One involves use of either AR or VA interpolation of missing data. Another involves directly accounting for this data problem via the use of the EM algorithm or Kalman filtering. Table 2 summarizes the results of a small experiment designed to compare AR and VA interpolation (EM and Kalman filtering methods are discussed later). In this experiment, both AR and VA interpolation are used to construct missing data, and all forecasting models are implemented in order to construct predictions, including MIDAS models, as well as benchmark models. Indeed, the only models not included in this experiment are MIDAS variants based on use of the EM algorithm and Kalman filtering. Entries in the table denote the proportion of forecasting models for which VA interpolation yields lower MSFEs than AR interpolation. Interestingly, proportions are always less than 0.5, regardless of whether backcasts, nowcasts, or forecasts are compared, and whether 'first available' or 'most recent' data are used. Indeed, in most cases, only approximately 10% of models or less 'prefer' VA interpolation. This is taken as strong evidence in favor of using AR interpolation, and, thus, the remainder of results presented only interpolate data using the AR method. Complete results using both varieties of interpolation are available upon request from the authors.

Second, we compare forecasting performance by estimation type in an experiment for which results are summarized in Table 3. In particular, we are cognizant of the fact that issues relating to structural breaks, model stability, and generic misspecification play an important role on the choice of using either rolling or recursive data windows when constructing real-time forecasting models. In lieu of this fact, we estimated all of our models using both recursive and rolling data windows, and entries in the table report the proportion of models for which the recursive estimation strategy is 'MSFE-best'. In the Korean case it turns out that recursive estimation yields more precise predictions when $forecasing$, while rolling window estimation yields more precise predictions when *backcasting* and *nowcasting*, regardless of whether 'first available' or 'most recent' data are used. This result becomes very pronounced as h increases. This finding suggests that further investigation via the use of so-called data rationality tests may be useful.

Third, a crucial aspect of forecasting models that utilize diffusion indices is exactly how many factors to specify. Bai and Ng (2002) and many others provide statistics that can be used for selecting the number of factors. However, there is no guarantee that the use of any of the exact tests will yield the 'MSFE-best' forecasting model. In one recent experiment, Kim (2017) uses Bai and Ng (2002), and finds that five to six factors are selected for a large scale Korean dataset. In this paper (see Table 4), we directly examine how many factors are used in 'MSFE-best' forecasting models. In particular, entries in Table 4 denote the proportion of times that models with a given fixed number of factors are MSFE-best among all of our factor-MIDAS models, including those estimated using the EM algorithm, the Kalman filter, AR interpolation (with each of OPCA and RPCA), and those estimated both with and without autoregressive lags. It is very clear from inspection of the results that either 1 or 2 factors, at most, are needed when the prediction horizon in more than 1 quarter ahead. On the other hand, for horizons -1 to 3 (i.e. all backcasts and nowcasts), the evidence is more mixed. While 1 or 2 factors are selected around 1/2 of the time, 5 or 6 factors are also selected around 1/2 of the time. Interestingly, there is little evidence that using an intermediate number of factors is useful. One should either specify a very parsimonious 1 or 2 factor models, or one should go with our maximum of 5 or 6. It is clear that forecast horizon matters; and this is consistent with the mixed evidence on this issue. Namely, some authors find that very few factors are useful, while others suggest using 5 or more. Both of these results are confirmed in our experiment, with forecast horizon being the critical determining characteristic. The overall conclusion, thus, appears to be that when uncertainty is more prevalent (i.e., longer forecast horizons), then parsimony is the key ingredient to factor selection. This conclusion is not at all surprising, and is in accord with stylized facts concerning model specification when specifying linear models.

We now turn to our forecasting model evaluation. Entries in Tables 5, Panel (a) are MSFEs for all models, relative to the strawman AR(SIC) model. Thus, entries greater than 1 imply that the corresponding model performs worse than the AR(SIC) model. The column headers in the table denote the forecast horizon, ranging from '-1' for backcasts to 9 for two quarter ahead predictions. In this framework, horizons 1, 2, and 3 are monthly nowcasts for the current quarter, and subsequent horizons pertain to monthly forecasts made during the subsequent two quarters. Notice that the first three rows in the table correspond to our other benchmark models (i.e., the RW, CBADL and BEX models). The rest of the rows in the table report findings for our MIDAS models, constructed with one factor. Results with 2 through 6 factors are provided in an on-line appendix, although it should be noted that precision of predictions generally decreases as the number of factors is increased. Recall that there are 5 different MIDAS specifications: 'Basic MIDAS with and without AR terms', 'Unrestricted MIDAS with and without AR terms', and 'Smoothed MIDAS'. Estimation is done recursively, the ragged-edge problem is solved by AR interpolation, four different factor estimation methods are reported on, including OPCA, RPCA, EM and KF, and data utilized in these experiments are assumed to be 'first available' data, for the purpose of both estimation and forecast evaluation.

Various forecast combinations are also reported on, as discussed above. In particular, 'Mean of Benchmarks' is an average based on our AR, RW, CBADL and BEX models. 'Mean of MIDAS' is the average forecast of all five MIDAS specifications which are constructed using all of our different factor estimation methods (i.e., 20 models). 'Mean of All MIDAS' is the average of the above 20 models, across all factor permutations (i.e., 1 through 6 factors) for a total of 120 models. Finally, 'Mean of All' includes all MIDAS models and benchmark models (i.e. 124 models). Table 5, Panel (b) is the same as Panel (a), except that 'most recent' instead of 'first available' data are used in all experiments reported on. Complete results pertaining to other permutations such as the use of alternative interpolation methods, estimation strategies, and numbers of factors are collected in the aforementioned online appendix.

Digging a bit further into the layout of this table, note that bold entries denote models that are 'MSFE-better' than the AR(SIC) model, entries with superscript 'FB' are 'MSFE-best' for a given forecast horizon and number of factors, and entries with the superscript 'GB' denote models that are 'MSFE-best' across all model permutations, including those reported on in the online appendix.

When forecast experiments are carried out using 'first available' data (see Panel (a) of Table 5), it turns out that for backcasting and nowcasting, factor-MIDAS models without AR terms as well as other benchmark models do not work well, regardless of the number of factors specified. This suggests that for short forecast horizons, the persistence of GDP growth is strong, and well modeled using linear AR components. As the forecast horizon gets longer, models without AR terms benefit from substantial performance improvement of the other components of the models, such as the MIDAS component. Indeed, in some cases, models without AR terms outperform models with AR terms. This is interesting, as it suggests that uncertainty in autoregressive parameters does not carry over as much to other model parameters, as the horizon increases, and the role for MIDAS thus increases in importance.

Evidently, upon inspection of MSFEs in Panel (a) of Table 5, there is little to choose between OPCA and RPCA estimation methods. Thus, given computing considerations[6], RPCA is preferred when analyzing large datasets. Among the other factor estimation methods, the KF and EM algorithms perform well for longer forecast horizons, but KF outperforms EM for shorter horizons. For our Korean GDP data, forecast combination does not result in improved predictive accuracy, as can be seen upon inspection of the findings of the table. However, it is certainly possible that more sophisticated combination methods may yield better results, although investigation of this is left to future research.

Panel (b) of Table 5 contains MSFEs that are based on the use of 'most recent' data. In most practical settings, forecasters assess predictive accuracy using this variety of data. Interestingly, in this set of results, we immediately observe that the 'MSFE-best' model is almost never the AR(SIC) model, although AR terms do enter into preferred (more complicated) models in most cases. Additionally, overall overall performance of our different models is similar to that reported in Panel (a) of the table.

In Table 6 the 'GB' models that are 'MSFE-best' across all permutations (including those reported in the online appendix), for a particular forecast horizon, are given in the rows labeled 'All'. The remainder of the table summarizes associated 'MSFE-best' models (and corresponding factor estimation schemes) for a given number of factors, for the cases where both 'first available' and 'most recent' data are used, and for a variety of forecast horizons. Finally, bold entries denote the best performer among all models, for a given forecast horizon. The results summarized in our discussion of Table 5 are made even more clear in this summary table. Namely, factor-MIDAS models are almost everywhere 'MSFE-best', with the exception of backcasts and nowcasts. Finally, PCA factor estimation methods are almost always preferred, and smoothed MIDAS type models are only useful if including many factors when predicting at the longest horizons. Of course, we do not recommend this, as using many factors for long horizon forecasting has been shown to yield more imprecise predictions than when fewer factors are used.

Figure 5 plots MSFE values that are not relative to the strawman AR(SIC) model, for var-

[6] Computation when using RPCA is around 10% faster that when using OPCA, based on a run using an Intel i7-3700 processor with 16GB of RAM.

ious prediction models. In the figure, 'Basic' and 'Unrestricted' denote factor-MIDAS models with two factors (refer to above discussion, and to Table 5, for further discussion of this terminology), and AR interpolation with OPCA estimation is used throughout. Panels (a) and (b) correspond to recursively estimated models using 'first available' and 'most recent' data, respectively. Panels (c) and (d) are same, but use rolling estimation. In this figure, $h = -1$ corresponds to backcasts, $h = 1, 2, 3$ correspond to nowcasts, and $h = 4, ..., 9$ correspond to forecasts. As discussed above, in conventional forecasting experiments, most forecasters use fully revised data for forecasting evaluation. With these data, factor-MIDAS dominates all other benchmark models, at all horizons, as seen in Panel (b); and RW and CBADL perform poorly at all horizons. Also, among the factor-MIDAS models, 'Basic' factor-MIDAS dominates. If we instead use 'first available' data, factor-MIDAS models as well as BEX models dominate the AR(SIC) model, particularly at long forecast horizons (see Panel (a)). However, as the forecast horizon gets shorter (i.e., we move from forecast→ nowcast→backcast), AR(SIC) and RW models perform better than other models, as confirmed in our discussion of the results presented in Table 5.

For the rolling estimation scheme, the forecast performance of factor-MIDAS models and AR(SIC) models are similar for all horizons. Moreover, we see that that rolling estimation performs slightly 'better' when 'most recent' data are used.

In Figure 6, MSFE values are plotted for the same set of models as in Figure 5. However, in this figure, Panels (a)-(d) contain plots based on the use of different factor estimation methods when specifying the models (i.e., OPCA, RPCA, EM and KF), only first available data are used for MSFE construction, all models are specified with one factor, and AR interpolation is implemented. In light of this, Panel (a) in Figures 5 and 6 is the same. A number of conclusions emerge upon inspection of this figure. First, the pattern of increasing MSFE as forecast horizon increases is observed for all factor estimation methods (compare all 4 panels in the figure), as expected. Also, all estimation methods appear to be rather similar, when faced with 'first available' data. However, even though MSFEs are similar across factor estimation methods, the MSFE magnitudes are slightly higher when using EM and KF, than when using OPCA and RPCA are used for estimation. Interestingly, only our top two MIDAS models (that include AR terms) outperform the benchmark AR(SIC) model at all forecast horizons, as can also be seen by inspection of the results in Table 5.

Finally, Figures 7 plots MSFEs of selected MIDAS models, with one factor. In these figures, MIDAS results are presented with factors estimated using OPCA, RPCA, EM, and KF. Additionally, various benchmark models are included (i.e., AR(SIC), RW, CBADL, and BEX). Using these figures, we can compare the performance of factor estimation methods for a given MIDAS model and value of r. Analogous figures reporting results for models with more than one factor are provided in the online appendix. Based on all of these figures, we find that for $r = 1$, RPCA or OPCA are clearly preferred. However, when $r = 2$, Kalman filtering also works well at many forecast horizons. Finally, as previously observed, when the number of factors is increased, forecast performance worsens substantially for 'Basic MIDAS' and 'Unrestricted MIDAS', as seen in Figure 6. Interestingly, 'Smoothed MIDAS' continues to perform well, even when $r = 6$. This again points to the importance of smoothing when the number of factors is large.

6 Concluding Remarks

We introduce a real-time dataset for Korean GDP, and analyze the usefulness of the dataset for forecasting, using a large variety of factor-MIDAS models, as well as linear benchmark models. Various factor estimation schemes, data interpolation approaches, and data windowing methods are analyzed, and methodological recommendations made. For example, we find that only approximately 10% of the forecasting models examined are 'MSFE-best' when using VA interpolation instead of AR interpolation. Additionally, models estimated using rolling data windows are only 'MSFE-best' at 3 forecast horizons, when comparing real-time predictions to 'first available' data, and are never 'MSFE-best' when comparing predictions to 'most recent' data. Given the usual preference amongst empirical researchers to use 'most recent' data in predictive accuracy analyses, it is clear that, at least in the case of Korean GDP, recursive estimation is preferred. With regard to the number of factors to specify in prediction models, either 1 or 2 factors, at most, are needed when the prediction horizon is more than 3 months ahead. On the other hand, for horizons -1 to 3 (i.e. all backcasts and nowcasts), the evidence is more mixed, and 5 or 6 factors are also selected around 1/2 of the time. Interestingly, there is little evidence that an intermediate number of factors is useful. One should either specify a very parsimonious 1 or 2 factor models, or one should go with our maximum of 5 or 6. In summary, forecast horizon matters, in the sense that when uncertainty is more prevalent (i.e., longer forecast horizons), then parsimony is the key ingredient to factor selection, and more than 1 or 2 factors leads to worsening predictive performance. This is consistent with the stylized notion that prediction multiple periods ahead becomes very uncertain when forecasting macroeconomic aggregates. Most importantly, we find that MIDAS models dominate at all forecast horizons.

References

Aastveit, K. A., Foroni, C., and Ravazzolo, F. (2016). Density forecasts with midas models. *Journal of Applied Econometrics*, (2014/10).

Aastveit, K. A., Gerdrup, K. R., Jore, A. S., and Thorsrud, L. A. (2014). Nowcasting gdp in real time: A density combination approach. *Journal of Business & Economic Statistics*, 32(1):48–68.

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., and Veronese, G. (2010). New eurocoin: Tracking economic growth in real time. *The Review of Economic and Statistics*, 92(4):1024–1034.

Andreou, E., Ghysels, E., and Kourtellos, A. (2010). Should macroeconomic forecasters use daily financial data and how? Technical report, Manuscript, University of Cyprus.

Baffigi, A., Golinelli, R., and Parigi, G. (2004). Bridge models to forecast the euro area gdp. *International Journal of Forecasting*, 20(3):371–401.

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

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

Bańbura, M., Giannone, D., and Reichlin, L. (2012). *Nowcasting*. Oxford University Press.

Bańbura, M. and Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, 29(1):133–160.

Barhoumi, K., Benk, S., Cristadoro, R., Reijer, A. D., Jakaitiene, A., Jelonek, P., Rua, A., Rüstler, G., Ruth, K., and Nieuwenhuyze, C. V. (2008). Short-term forecasting of gdp using large monthly datasets: A pseudo real-time forecast evaluation exercise. ECB Occasional Paper 84, European Central Bank.

Boivin, J. and Ng, S. (2005). Understanding and comparing factor-based forecasts. *International Journal of Central Banking*, 1(3):117–152.

Clements, M. P. and Galvao, A. B. (2008). Macroeconomic forecasting with mixed frequency data. *Journal of Business & Economic Statistics*, 26:546–554.

Clements, M. P. and Galvao, A. B. (2009). Forecasting us output growth using leading indicators: An appraisal using midas mode. *Journal of Applied Econometrics Journal*, 24(7):1187–1206.

Croushore, D. and Stark, T. (2001). A real-time data set for macroeconomists. *Journal of*

Econometrics, 105:111–130.

Cubadda, G. and Guardabascio, B. (2012). A medium-n approach to macroeconomics forecasting. *Economic Modelling*, 29(4):1099–1105.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63.

Doz, C., Giannone, D., and Reichlin, L. (2012). A quasi-maximum likelihood approach for large, approximate dynamic factor models. *The Review of Economics and Statistics*, 94(4):1014–1024.

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.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, 100:830–840.

Foroni, C., Marcellino, M., and Schumacher, C. (2015). Unrestricted mixed data sampling (midas): Midas regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society Series A*, 178(1):57–82.

Ghysels, E. (2016). Macroeconomics and the reality of mixed frequency data. *Journal of Econometrics*, 193(2):294–314.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The midas touch: Mixed data sampling regression models. CIRANO Working Papers 2004s-20, CIRANO.

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

Girardi, A., Golinelli, R., and Pappalardo, C. (2016). The role of iindicator selection in nowcasting euro area gdp in pseudo real time. *Empirical Economics*, forthcoming.

Golinelli, R. and Parigi, G. (2005). Short-run italian gdp forecasting and real-time data. CEPR Discussion Papers 5302, C.E.P.R. Discussion Papers.

Götz, T., Hecq, A., and Urbain, J.-P. (2016). Combining forecasts from successive data vintages: An application to u.s. growth. *International Journal of Forecasting*, 32(1):61–74.

Kim, H. H. (2017). Looking into the black box of the korean economy: The sparse factor model approach. *Journal of Asia-Pacific Economy*, forthcoming.

Kim, H. H. and Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, 178(2):352–367.

Kim, H. H. and Swanson, N. R. (2016). Mining big data using parsimonious factor and shrinkage methods. *International Journal of Forecasting*, forthcoming.

Koenig, E. F., Dolmas, S., and Piger, J. (2003). The use and abuse of real-time data in economic forecasting. *The Review of Economics and Statistics*, 85(3):618–628.

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

Kuzin, V., Marcellino, M., and Schumacher, C. (2013). Pooling versus model selection for nowcasting gdp with many predictors: empirical evidence for six industrialized countries. *Journal of Applied Econometrics*, 28(3):392–411.

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.

Mazzi, G. L., Mitchell, J., and Montana, G. (2014). Density nowcasts and model combination: Nowcasting euro-area gdp growth over the 2008-09 recession. *Oxford Bulletin of Economics and Statistics*, 76(2):233–256.

Peddaneni, H., Erdogmus, D., Rao, Y. N., Hegde, A., and Principe, J. (2004). Recursive principal components analysis using eigenvector matrix perturbation. In *Machine Learning for Signal Processing*, pages 83–92. IEEE.

Pettenuzzo, D., Timmermann, A. G., and Valkanov, R. (2014). A Bayesian MIDAS Approach to Modeling First and Second Moment Dynamics. CEPR Discussion Papers 10160, C.E.P.R. Discussion Papers.

Rünstler, G. and Sédillot, F. (2003). Short-term estimates of euro area real gdp by means of monthly data. Working Paper Series 0276, European Central Bank.

Schorfheide, F. and Song, D. (2015). Real-time forecasting with a mixed-frequency var. *Journal of Business & Economic Statistics*, 33(3):366–380.

Schumacher, C. (2007). Forecasting german gdp using alternative factor models based on large datasets. *Journal of Forecasting*, 26(4):271–302.

Schumacher, C. and Breitung, J. (2008). Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data. *International Journal of Forecasting*, 24(3):386–398.

Stark, T. and Croushore, D. (2002). Forecasting with a real-time data set for macroeconomists. *Journal of Macroeconomics*, 24(4):507–531.

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. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business and Economic Statistics*, 30(4):481–493.

Swanson, N. and van Dijk, D. (2006). Are statistical reporting agencies getting it right? data rationality and business cycle asymmetry. *Journal of Business and Economic Statistics*, 24:24–42.

Timmermann, A. G. (2006). Forecast combinations. In Elliott, G., C., G., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, chapter 4, pages 135–196. Elsevier.

Wallis, K. (1986). Forecasting with an econometric model: the "ragged edge" problem. *Journal of Forecasting*, 5:1–13.

Zheng, I. Y. and Rossiter, J. (2006). Using monthly indicators to predict quarterly gdp. Working Papers 06-26, Bank of Canada.

Table 1: Summary of Models and Estimation Methods*

Estimation Scheme	MIDAS	Factor Estimation	Interpolation
	Basic w/o AR term	OPCA	AR
	Basic w/ AR term		
		RPCA	VA
Recursive	Unrestricted w/o AR term	EM algorithm	
	Unrestricted w/ AR term		
Rolling	Smoothed	Kalman Filtering	
	AR		
	CBADL		
	BEX		

Model	Description
AR(SIC)	Autoregressive model with length of lags determined by SIC
RW	Random Walk
CBADL	Combined Bivariate Autoregressive Distributed Lag model
BEX	Bridge Equation with Exogenous Variable
Basic w/o AR	Basic MIDAS model without AR terms
Basic w/ AR	Basic MIDAS model with AR terms
Unrestricted w/o AR	Unrestricted MIDAS model without AR terms
Unrestricted w/ AR	Unrestricted MIDAS model with AR terms
Smoothed	Smoothed MIDAS model

* Notes: Non-factor-MIDAS type models include AR(SIC), RW, CBADL and BEX. Three types of factor-MIDAS models are specified ('Basic', 'Unrestricted', and 'Smoothed'), and each of these are estimated using each factor estimation method (OPCA and RPCA), interpolation method (AR and VA), and factor-MIDAS estimation method (EM algorithm and Kalman filter). Finally, all of these permutations are implemented using each of recursive and rolling data windowing strategies. For complete details see Section 5.

Table 2: Comparison of Forecasting Performance with AR and VA Interpolation*

	Backcast	Nowcast			Forecast					
	prev. qtr.	current quarter			1 quarter ahead			2 quarter ahead		
Horizon	-1	1	2	3	4	5	6	7	8	9
First available	0.285	0.337	0.110	0.145	0.151	0.134	0.105	0.134	0.093	0.186
Most Recent	0.081	0.163	0.064	0.122	0.134	0.105	0.105	0.128	0.081	0.186

* Notes: See notes to Table 1. Forecasting performance is evaluated by comparing MSFEs across all models which use interpolated missing values, including CBADL, BEX and factor-MIDAS models with OPCA and RPCA. Entries in the table report the proportion of times that the MSFE of models with AR interpolation is greater than 'like' models with VA interpolation. Thus, entries less than 0.5 indicate that AR interpolation performs better than VA, on average, across all model permutations. Prediction models are estimated in real-time using either 'first available' or 'most recent' historical data, and MSFEs are constructed by comparing these predictions with actual 'first available' or 'most recent' data, corresponding to the type of data used in estimation.

Table 3: Comparison of Forecasting Performance by Estimation Type*

	Backcast	Nowcast			Forecast					
	prev.	current quarter			1 quarter ahead			2 quarter ahead		
	qtr.									
Horizon	-1	1	2	3	4	5	6	7	8	9
First available	0.747	0.782	0.653	0.465	0.400	0.147	0.059	0.018	0.029	0.082
Most Recent	0.771	0.759	0.618	0.518	0.424	0.247	0.059	0.018	0.029	0.059

* Notes: See notes to Table 2. Forecasting performance is evaluated by comparing MSFEs across all models, with MSFEs calculated by estimating prediction models using either 'first available' or 'most recent' actual data, as discussed in the footnote to Table 2. In this table, entries report the proportion of times that the MSFEs of models estimated recursively are greater than when 'like' models are estimated using rolling data windows. Thus, entries less than 0.5 indicate that recursive estimation yields lower MSFEs.

Table 4: Comparison of Forecasting Performance Using Differing Numbers of Factors *

		Backcast	Nowcast			Forecast					
		prev.	current quarter			1 quarter ahead			2 quarter ahead		
		qtr.									
	Factor #	-1	1	2	3	4	5	6	7	8	9
First available	1	0.20	0.25	0.20	-	-	0.20	-	0.20	-	-
	2	0.20	-	0.20	-	0.60	0.60	1.00	0.60	1.00	1.00
	3	-	0.15	-	0.20	0.20	0.20	-	0.20	-	-
	4	-	-	0.20	0.40	-	-	-	-	-	-
	5	0.20	0.20	-	-	0.20	-	-	-	-	-
	6	0.40	0.40	0.40	0.40	-	-	-	-	-	-
Most Recent	1	-	0.05	0.05	-	-	0.20	-	0.20	0.20	-
	2	0.20	0.20	0.15	0.20	1.00	0.60	1.00	0.80	0.80	1.00
	3	0.20	0.15	-	-	-	0.20	-	-	-	-
	4	-	-	0.40	0.40	-	-	-	-	-	-
	5	0.35	0.40	-	0.20	-	-	-	-	-	-
	6	0.25	0.20	0.40	0.20	-	-	-	-	-	-

* Notes: See notes to Table 3. The proportion of factor-MIDAS 'MSFE-best' models, when comparing 'like' models with the number of factors varying from 1 to 6, is reported in this table. This using either 'first available' or 'most recent' data (as discussed in the footnote to Table 2), as well as for a number of backcast, nowcast, and forecast horizons. See Section 6 for a detailed discussion of the different horizons reported on. All results are based on OPCA and RPCA using AR interpolation, under a recursive estimation scheme.

Table 5: Relative MSFEs When Backcasting, Nowcasting, and Forecasting Korean GDP*

Panel (a): First Available

		Backcast	Nowcast			Forecast					
		prev. qtr.	current quarter			1 quarter ahead			2 quarter ahead		
Recursive		-1	1	2	3	4	5	6	7	8	9
RW		1.45	1.35	1.12	**0.94**	**0.94**	1.01	1.14	1.13	1.20	1.68
CBADL		5.49*	4.67*	3.28*	1.73	1.63	1.63	1.36	1.32	1.37	1.46
BEX		3.48*	3.25*	2.16	**0.89**	**0.87**	**0.85**	**0.64***	**0.62****	**0.64***	**0.71****
Mean of Benchmarks		4.38	3.85	2.73	1.55	1.47	1.42	**0.92**	**0.89**	**0.93**	1.10
Basic w/o AR	OPCA	3.01	2.46	1.70	**0.80***	**0.77***	**0.84***	**0.72***	**0.69***	**0.73***	**0.87**
	RPCA	3.01	2.46	1.70	**0.80***	**0.77***	**0.84***	**0.72***	**0.69***	**0.73***	**0.87**
	EM	3.25	2.96**	2.00**	1.07	1.08	1.09	**0.88**	**0.82**	**0.81**	**0.86***
	KF	2.72	2.32**	1.73**	**0.91**	**0.90**	**0.98**	**0.82**	**0.77**	**0.80**	**0.90**
Basic w/ AR	OPCA	**0.70**$_{\mathbf{GB}}$	**0.83**$_{\mathbf{FB}}$	**0.70***	**0.49****	**0.57****	**0.66***	**0.75***	**0.75***	**0.80***	**0.97**
	RPCA	**0.70**	**0.83**	**0.70***	**0.49****	**0.57****	**0.66***	**0.75***	**0.75***	**0.80***	**0.97**
	EM	1.18	1.45*	1.12	1.00	1.09	1.07	1.05	**0.95**	**0.97**	1.12
	KF	**0.90**	**0.99**	**0.93**	**0.71**	**0.78**	1.06	1.02	**0.98**	**0.98**	1.21
Unrestricted w/o AR	OPCA	3.10	2.63	1.82	**0.83***	**0.75****	**0.75***	**0.61****	**0.58****	**0.62***	**0.74***
	RPCA	3.10	2.63	1.82	**0.83***	**0.75****	**0.75***	**0.61****	**0.58****	**0.62***	**0.74***
	EM	3.33	3.10**	2.14**	1.13	**0.98**	1.05	**0.78**	**0.69**	**0.78**	**0.80***
	KF	2.81	2.45**	1.80**	**0.90***	**0.75****	**0.80****	**0.72***	**0.57***	**0.62***	**0.79**
Unrestricted w/o AR	OPCA	**0.76**	**0.88**	**0.68**$^{*}_{\mathbf{FB}}$	**0.47**$^{**}_{\mathbf{FB}}$	**0.49****	**0.59**$^{*}_{\mathbf{FB}}$	**0.49**$^{**}_{\mathbf{FB}}$	**0.53****	**0.70***	**0.73**$^{*}_{\mathbf{FB}}$
	RPCA	**0.76**	**0.88**	**0.68***	**0.47****	**0.49**$^{**}_{\mathbf{FB}}$	**0.59***	**0.49****	**0.53****	**0.70***	**0.73**$^{*}_{\mathbf{FB}}$
	EM	1.33	1.39	1.03	**0.80**	**0.75**	**0.85**	**0.75**	**0.82**	**0.73**	1.09
	KF	**0.91**	**0.99**	**0.76**	**0.57****	**0.50****	**0.66**	**0.58****	**0.48**$^{**}_{\mathbf{FB}}$	**0.55**$^{*}_{\mathbf{FB}}$	1.07
Smoothed	OPCA	2.47	2.16	1.51	**0.70***	**0.71****	**0.77****	**0.64****	**0.65***	**0.69***	**0.81***
	RPCA	2.47	2.16	1.51	**0.70***	**0.71****	**0.77****	**0.64****	**0.65***	**0.69***	**0.81***
	EM	2.44	2.43**	1.75*	**0.84***	**0.92**	**0.95**	**0.75**	**0.75**	**0.74**	**0.83***
	KF	2.32	2.11**	1.58*	**0.78****	**0.81**	**0.89**	**0.72***	**0.72**	**0.74**	**0.84***
Mean with 1 factor		1.43	1.42	1.08	**0.62****	**0.62****	**0.72****	**0.67***	**0.65***	**0.70***	**0.84**
Mean of All MIDAS		1.16	1.25	**0.86**	**0.45****	**0.50****	**0.52****	**0.47****	**0.61***	**0.77**	1.12
Mean of All		1.19	1.26	**0.93**	**0.53****	**0.58****	**0.61****	**0.53****	**0.60****	**0.70***	**0.86**

Panel (b): Most Recent

		Backcast		Nowcast				Forecast			
		prev. qtr.		current quarter			1 quarter ahead			2 quarter ahead	
Recursive		-1	1	2	3	4	5	6	7	8	9
RW		1.07	1.11	1.15	1.25**	1.29**	1.27**	1.41**	1.49**	1.49**	1.60**
CBADL		**0.97**	**0.97**	1.10	1.16	1.12	1.21*	1.33**	1.29**	1.38**	1.51**
BEX		**0.65****	**0.67****	**0.74****	**0.80**	**0.82**	**0.85**	**0.96**	**0.98**	1.02	1.11
Mean of Benchmarks		3.22*	2.69*	2.01*	1.62	1.54	1.44	**0.96**	**0.89**	**0.94**	1.25
	OPCA	2.26	1.78	1.17	**0.76***	**0.73***	**0.76****	**0.67***	**0.63***	**0.69***	**0.83***
Basic	RPCA	2.26	1.78	1.17	**0.76***	**0.73***	**0.76****	**0.67***	**0.63***	**0.69***	**0.83***
w/o AR	EM	2.71**	2.38**	1.40	1.06	1.07	**1.00**	**0.83**	**0.76**	**0.77**	**0.82****
	KF	2.04	1.69	1.16	**0.86***	**0.85**	**0.89**	**0.77**	**0.71**	**0.76**	**0.86***
	OPCA	**0.75**$_{FB}$	**0.81**$^{*}_{GB}$	**0.61****	**0.51****	**0.59****	**0.67***	**0.70***	**0.70***	**0.76***	**0.94**
Basic	RPCA	**0.75**	**0.86***	**0.61****	**0.51****	**0.59****	**0.67***	**0.70***	**0.70***	**0.76***	**0.94**
w/ AR	EM	1.17	1.31	**0.83**	**0.99**	1.21	1.00	1.01	**0.90**	**0.94**	1.11
	KF	**0.84**	**0.84**	**0.72****	**0.71**	**0.84**	1.06	**0.99**	**0.95**	**0.97**	1.22
	OPCA	2.32	1.88	1.25	**0.79***	**0.70****	**0.66****	**0.56****	**0.51****	**0.56***	**0.68****
Unrestricted	RPCA	2.32	1.88	1.25	**0.79***	**0.70****	**0.66****	**0.56****	**0.51****	**0.56***	**0.68****
w/o AR	EM	2.74*	2.48**	1.50	1.10	**0.94**	**0.93**	**0.74**	**0.60**	**0.72**	**0.76****
	KF	2.10	1.79	1.19	**0.82****	**0.68****	**0.67****	**0.64***	**0.48****	**0.54***	**0.71****
	OPCA	**0.84**	**0.89***	**0.57****	**0.46**$^{**}_{FB}$	**0.46****	**0.54**$^{**}_{FB}$	**0.44**$^{**}_{FB}$	**0.46****	**0.67**	**0.68**$^{**}_{FB}$
Unrestricted	RPCA	**0.84**	**0.89***	**0.57****	**0.46****	**0.46**$^{**}_{FB}$	**0.54****	**0.44****	**0.46****	**0.67**	**0.68**$^{**}_{FB}$
w/o AR	EM	1.23	1.18	**0.76**	**0.82**	**0.77**	**0.85**	**0.74**	**0.79**	**0.69**	1.08
	KF	**0.85**	**0.86**	**0.55**$^{**}_{FB}$	**0.54****	**0.46****	**0.59***	**0.52****	**0.40**$^{**}_{FB}$	**0.49**$^{**}_{FB}$	1.06
	OPCA	1.65	1.45	**0.99**	**0.63****	**0.64****	**0.68****	**0.58****	**0.59****	**0.64***	**0.76****
	RPCA	1.65	1.45	**0.99**	**0.63****	**0.64****	**0.68****	**0.58****	**0.59****	**0.64***	**0.76****
Basic	EM	1.74	1.78	1.15	**0.75****	**0.86**	**0.84**	**0.69**	**0.68**	**0.70**	**0.79****
w/ AR	KF	1.53	1.43	1.00	**0.69****	**0.74**	**0.78**	**0.66**	**0.66**	**0.70**	**0.80****
Mean of MIDAS		1.03	1.02	**0.69****	**0.54****	**0.56****	**0.62****	**0.61***	**0.57***	**0.64***	**0.78****
Mean of All MIDAS		**0.84**	**0.89**	**0.52****	**0.40****	**0.49****	**0.47****	**0.45****	**0.63**	**0.86**	1.27
Mean of All		**0.76**	**0.82**	**0.58****	**0.46****	**0.54****	**0.56****	**0.50****	**0.58***	**0.72**	**0.91**

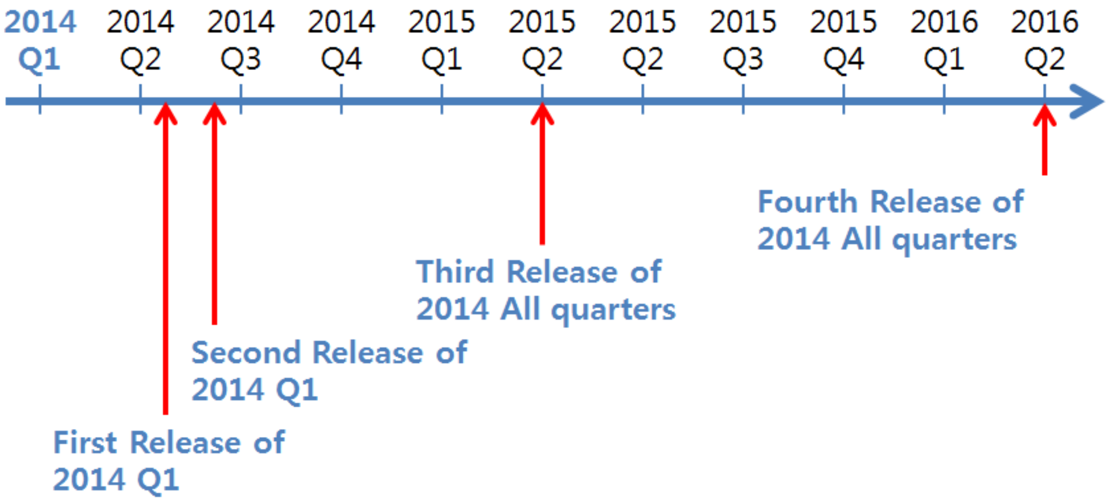
* Notes: See notes to Tables 1-4. Entries in this table are ratios of point MSFEs of our benchmark or 'strawman' AR(SIC) model to each other model, for various estimation methods and horizons. Panel (a) reports MSFEs based on experiments using 'first available' real-time quarterly historical data, and Panel (b) reports results based on the use of 'most recent' real-time quarterly historical data. All results are based on recursively estimated models. The column denoted by 'Backcast' contains MSFEs for quarterly forecasts of GDP made 1-month prior to the calendar date of the quarterly GDP datum being predicted, and the columns denoted by 'Nowcast' contain MSFEs for forecasts of the first, second and third months of each quarterly calendar dated GDP observation. Finally, the columns denoted by 'Forecast' contain MSFEs based on 1-quarter ahead predictions made from 1 month after the end of the quarter (called month 4) to 3 month ahead (called month 6). Months 7-9 correspondingly refer to 2-quarter ahead predictions. Bold entires denote cases for which the point MSFE of a given model is lower than the point MSFE of the AR(SIC) model. Entries superscripted by a ** (5% level) and a * (10% level) are significantly better than the AR(SIC) model, based on application of the DM predictive accuracy test. Finally, entries subscripted with 'FB' denote the MSFE-best models for a given number of estimated factors and for each horizon, while entries subscripted with 'GB' denote MSFE-best models across all specification permutations, for a given horizon. See Section 5 for complete details.

Table 6: Summary of MSFE-Best Models Across All Modelling Permutations*

		Backcast		Nowcast				Forecast		
		prev. qtr.		current quarter			1 quarter ahead			2 quarter ahead
	Fac. No.	-1	1	2	3	4	5	6	7	8	9
First Available	1	**Basic** **w/ AR** **OPCA**	Basic w/ AR OPCA	Unrestricted w/ AR OPCA	Unrestricted w/ AR OPCA	Unrestricted w/ AR RPCA	Unrestricted w/ AR OPCA	Unrestricted w/ AR OPCA	Unrestricted w/ AR KF	Unrestricted w/ AR KF	Unrestricted w/ AR Both PCAs
	2	Basic w/ AR OPCA	Basic w/ AR RPCA	Basic w/ AR KF	Basic w/ AR OPCA	**Basic** **w/ AR** **OPCA**	**Basic** **w/ AR** **KF**	Basic w/o AR RPCA	**Basic** **w/ AR** **OPCA**	**Basic** **w/ AR** **KF**	**Basic** **w/o AR** **OPCA**
	3	Basic w/ AR RPCA	**Basic** **w/ AR** **RPCA**	**Basic** **w/ AR** **KF**	**Basic** **w/ AR** **KF**	Basic w/ AR KF	Basic w/ AR KF	**Smoothed** **KF**	Smoothed Both PCAs	Smoothed KF	Smoothed Both PCAs
	4	Basic w/ AR OPCA	AR -	Unrestricted w/ AR RPCA	Unrestricted w/ AR RPCA	Basic w/ AR KF	Mean -	Smoothed KF	Smoothed KF	Smoothed KF	Smoothed RPCA
	5	Basic w/ AR OPCA	AR -	Basic w/ AR EM	Basic w/o AR KF	Basic w/ AR OPCA	Smoothed KF	Smoothed KF	Smoothed KF	Smoothed KF	Smoothed Both PCAs
	6	Basic w/ AR OPCA	AR -	Basic w/ AR KF	Mean -	Smoothed Both PCAs	Smoothed Both PCAs	Smoothed Both PCAs	Smoothed Both PCAs	Smoothed Both PCAs	Smoothed Both PCAs
Most Recent	1	Basic w/ AR OPCA	**Basic** **w/ AR** **OPCA**	Unrestricted w/ AR KF	Unrestricted w/ AR OPCA	Unrestricted w/ AR RPCA	Unrestricted w/ AR OPCA	Unrestricted w/ AR OPCA	Unrestricted w/ AR KF	Unrestricted w/ AR KF	Unrestricted w/ AR Both PCAs
	2	**Unrestricted** **w/ AR** **OPCA**	Basic w/ AR KF	Basic w/ AR KF	**Basic** **w/ AR** **KF**	**Basic** **w/ AR** **OPCA**	Basic w/ AR KF	Unrestricted w/ AR OPCA	**Basic** **w/ AR** **OPCA**	**Basic** **w/ AR** **KF**	**Basic** **w/ AR** **RPCA**
	3	Basic w/ AR RPCA	Basic w/ AR KF	Basic w/ AR KF	Basic w/ AR KF	Mean -	Smoothed KF	Smoothed KF	Smoothed KF	Smoothed KF	Smoothed Both PCAs
	4	Basic w/ AR OPCA	Basic w/ AR KF	Unrestricted w/ AR RPCA	Smoothed KF	Basic w/ AR KF	**Mean** **-**	**Smoothed** **KF**	Smoothed KF	Smoothed KF	Smoothed KF
	5	Basic w/ AR OPCA	Basic w/ AR KF	**Basic** **w/o AR** **KF**	Smoothed KF	Smoothed KF	Smoothed KF	Smoothed KF	Smoothed KF	Smoothed KF	Smoothed OPCA
	6	Basic w/ AR OPCA	Mean -	Basic w/o AR KF	Smoothed Both PCAs	Smoothed RPCA	Smoothed RPCA	Smoothed Both PCAs	Smoothed KF	Smoothed Both PCAs	Smoothed RPCA

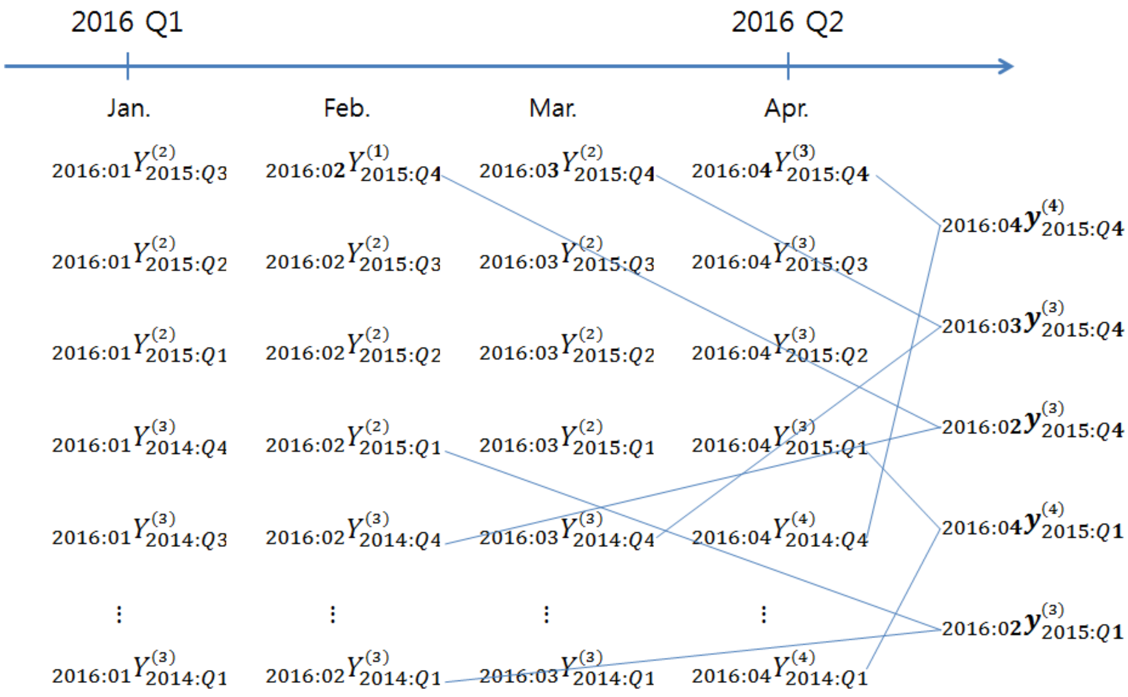
* Notes: See notes to Table 5. Entries indicate the model and estimation methods for all 'MSFE-best' specifications, by historical data type, number of factors used, and horizon. Entries in the row labeled 'All' are the 'MSFE-best' models across all factor specifications, for a given historical data type. All model estimation is done recursively and AR interpolation is used for missing value construction. For example, for the 'Backcast' horizon, the 'Basic factor-MIDAS' model with AR terms and with factors estimated using OPCA is the 'globally best' performer when experiments are conducted using 'first available' real-time historical data. When MSFEs based on the use of OPCA and RPCA are the same up to three decimal places, the PCA method is denoted by 'both PCA'. Bold entries denote 'MSFE-best' models across all different models, including those with 1-6 factors (for complete tabulated results, see the online appendix).

Figure 1: Release Dates for Real-Time Korean GDP*



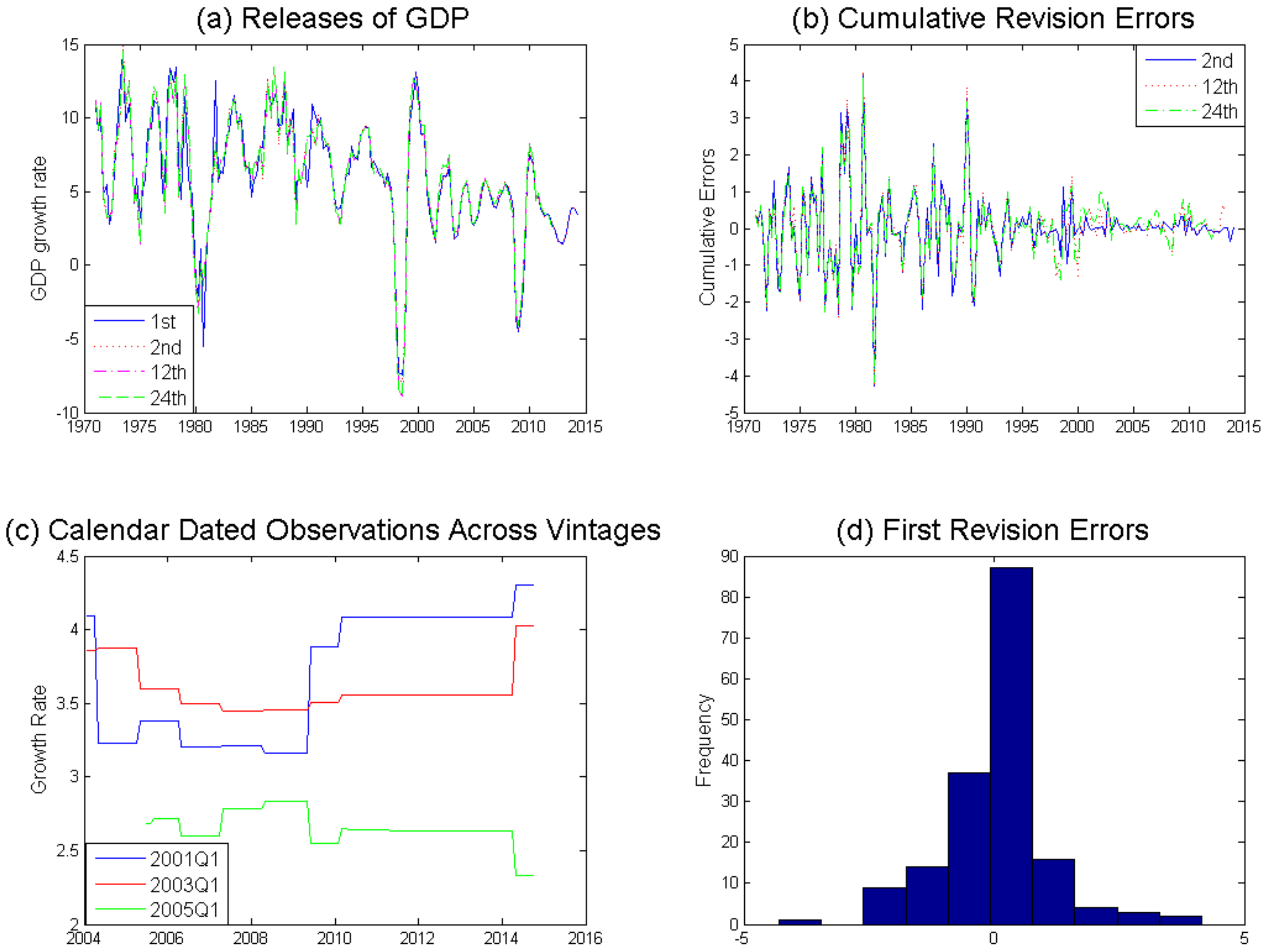
* Notes: See Sections 2 and 3 for complete details.

Figure 2: Depiction of Annualized GDP Growth Rates Based on Real-Time Data*


2016 Q1
2016 Q2
Jan.
Feb.
Mar.
Apr.
2016:01$Y^{(2)}_{2015:Q3}$
2016:02$Y^{(1)}_{2015:Q4}$
2016:03$Y^{(2)}_{2015:Q4}$
2016:04$Y^{(3)}_{2015:Q4}$
2016:01$Y^{(2)}_{2015:Q2}$
2016:02$Y^{(2)}_{2015:Q3}$
2016:03$Y^{(2)}_{2015:Q3}$
2016:04$Y^{(3)}_{2015:Q3}$
2016:01$Y^{(2)}_{2015:Q1}$
2016:02$Y^{(2)}_{2015:Q2}$
2016:03$Y^{(2)}_{2015:Q2}$
2016:04$Y^{(3)}_{2015:Q2}$
2016:01$Y^{(3)}_{2014:Q4}$
2016:02$Y^{(2)}_{2015:Q1}$
2016:03$Y^{(2)}_{2015:Q1}$
2016:04$Y^{(3)}_{2015:Q1}$
2016:01$Y^{(3)}_{2014:Q3}$
2016:02$Y^{(3)}_{2014:Q4}$
2016:03$Y^{(3)}_{2014:Q4}$
2016:04$Y^{(4)}_{2014:Q4}$
⋮
⋮
⋮
⋮
2016:01$Y^{(3)}_{2014:Q1}$
2016:02$Y^{(3)}_{2014:Q1}$
2016:03$Y^{(3)}_{2014:Q1}$
2016:04$Y^{(4)}_{2014:Q1}$
2016:04$y^{(4)}_{2015:Q4}$
2016:03$y^{(3)}_{2015:Q4}$
2016:02$y^{(3)}_{2015:Q4}$
2016:04$y^{(4)}_{2015:Q1}$
2016:02$y^{(3)}_{2015:Q1}$

* Notes: See Section 2 for a detailed discussion of the dating conventions used in this diagram.

Figure 3: Historical Real-Time Data Releases for Korean GDP*


(a) Releases of GDP

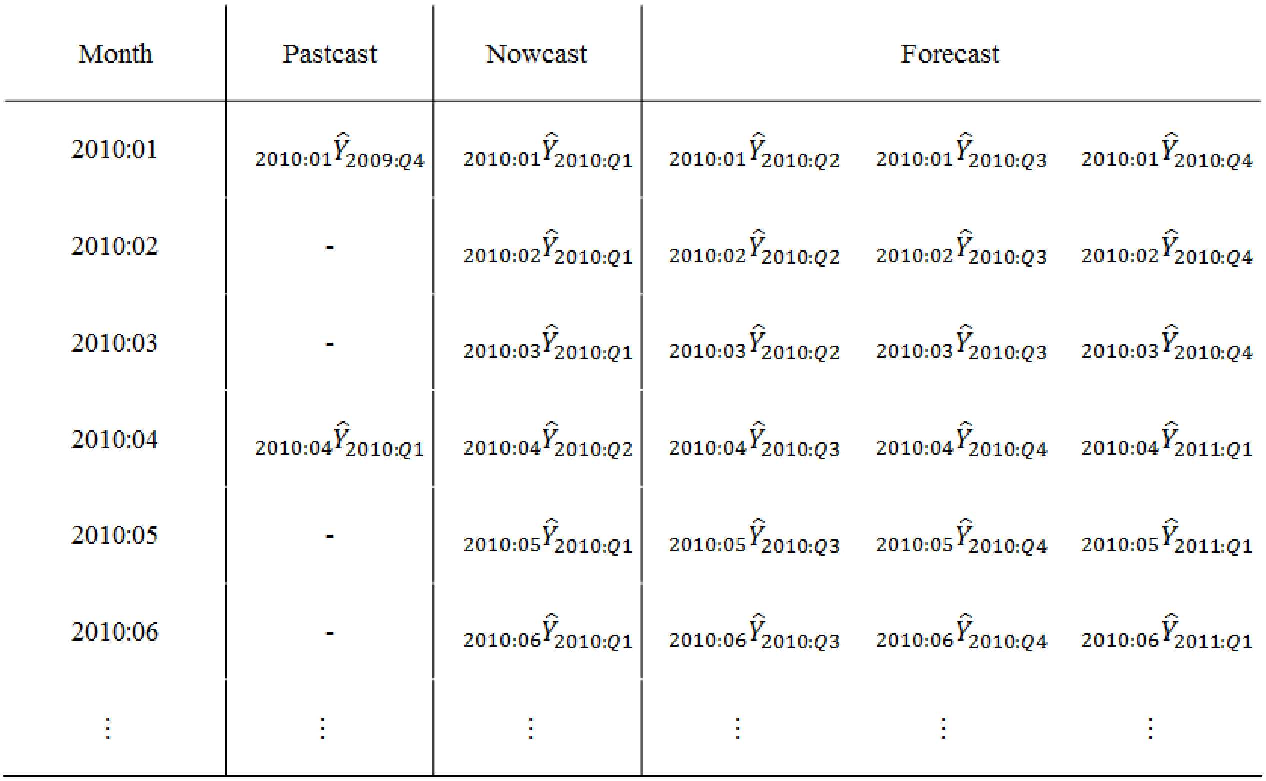
(b) Cumulative Revision Errors

(c) Calendar Dated Observations Across Vintages

(d) First Revision Errors

* Notes: In Figure (a), the solid line depicts 1st release GDP, the dotted line depicts 2nd release data, and the dot-dash line depicts final release data. Figure (b) depicts cumulative revision errors between the 1st and either the 2nd, 12th, or 24th releases. Figure (c) shows how the growth rate of 2001:1, 2003:1 and 2005:1 calendar dated GDP evolves as the series is revised. Figure (d) plots the distribution of first revision errors (i.e., the differences between 1st and 2nd data releases).

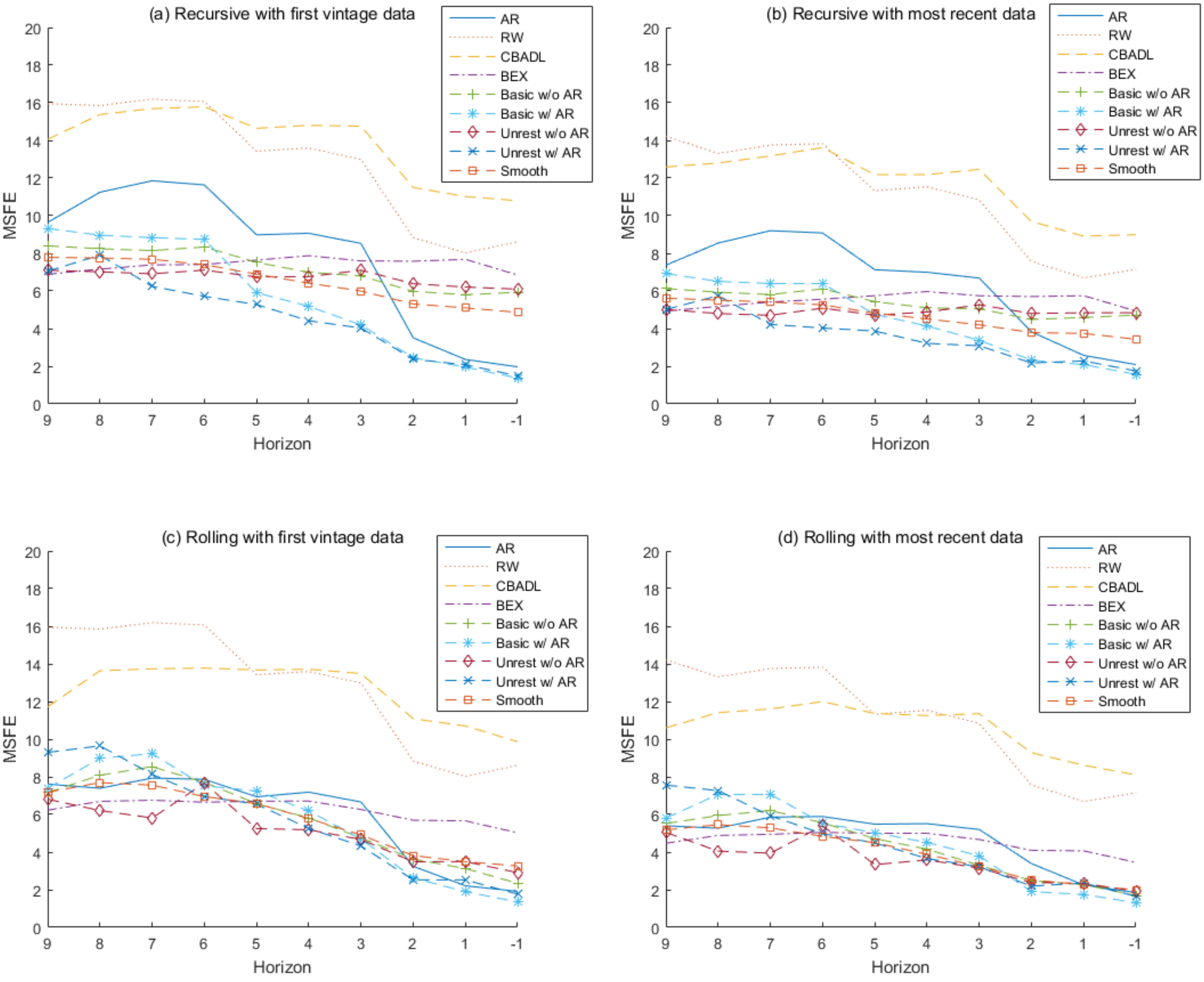
Figure 4: Structure of Monthly/Quarterly Prediction Experiments*



Month	Pastcast	Nowcast	Forecast		
2010:01	$_{2010:01}\hat{Y}_{2009:Q4}$	$_{2010:01}\hat{Y}_{2010:Q1}$	$_{2010:01}\hat{Y}_{2010:Q2}$	$_{2010:01}\hat{Y}_{2010:Q3}$	$_{2010:01}\hat{Y}_{2010:Q4}$
2010:02	-	$_{2010:02}\hat{Y}_{2010:Q1}$	$_{2010:02}\hat{Y}_{2010:Q2}$	$_{2010:02}\hat{Y}_{2010:Q3}$	$_{2010:02}\hat{Y}_{2010:Q4}$
2010:03	-	$_{2010:03}\hat{Y}_{2010:Q1}$	$_{2010:03}\hat{Y}_{2010:Q2}$	$_{2010:03}\hat{Y}_{2010:Q3}$	$_{2010:03}\hat{Y}_{2010:Q4}$
2010:04	$_{2010:04}\hat{Y}_{2010:Q1}$	$_{2010:04}\hat{Y}_{2010:Q2}$	$_{2010:04}\hat{Y}_{2010:Q3}$	$_{2010:04}\hat{Y}_{2010:Q4}$	$_{2010:04}\hat{Y}_{2011:Q1}$
2010:05	-	$_{2010:05}\hat{Y}_{2010:Q1}$	$_{2010:05}\hat{Y}_{2010:Q3}$	$_{2010:05}\hat{Y}_{2010:Q4}$	$_{2010:05}\hat{Y}_{2011:Q1}$
2010:06	-	$_{2010:06}\hat{Y}_{2010:Q1}$	$_{2010:06}\hat{Y}_{2010:Q3}$	$_{2010:06}\hat{Y}_{2010:Q4}$	$_{2010:06}\hat{Y}_{2011:Q1}$
⋮	⋮	⋮	⋮	⋮	⋮

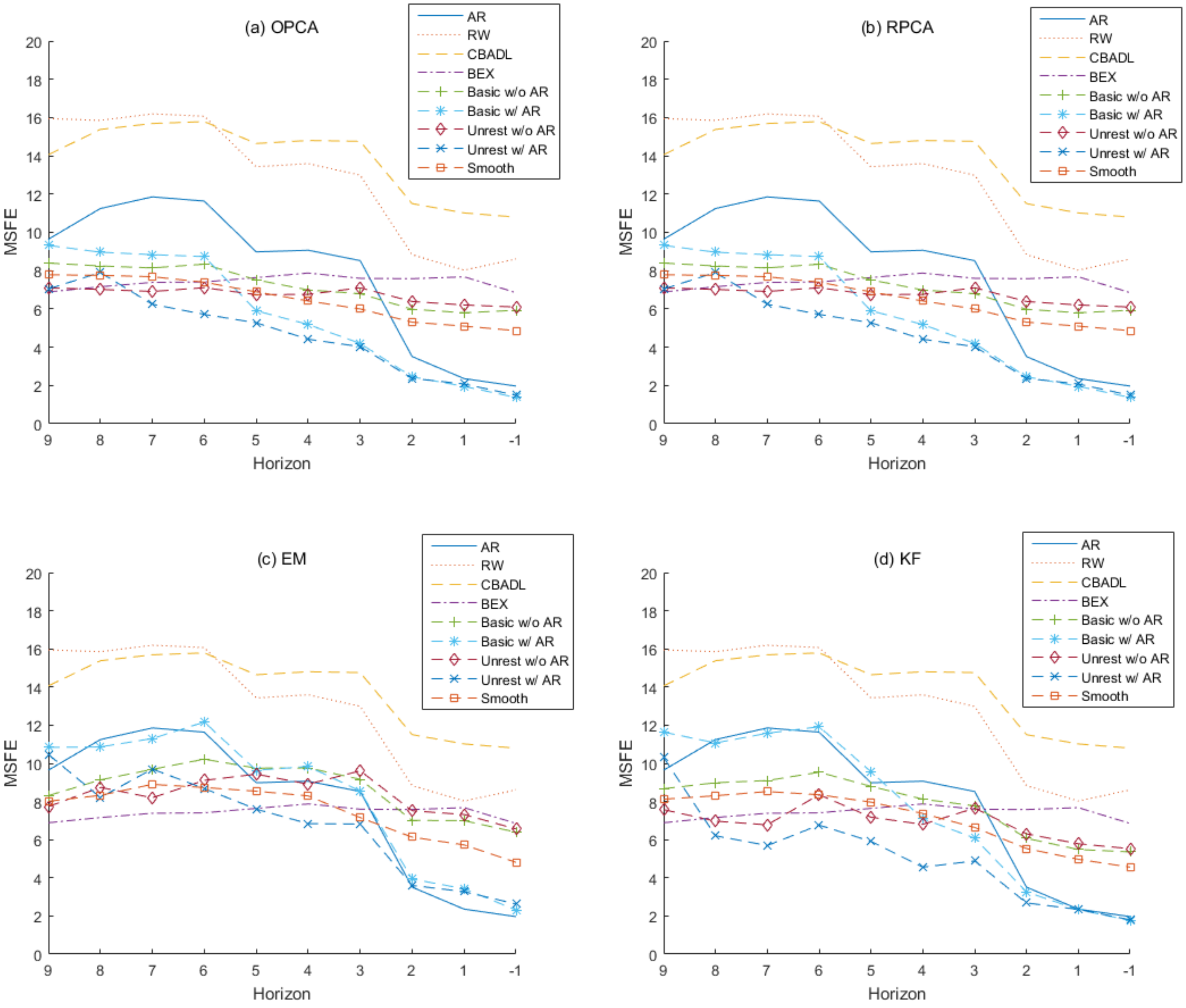
* Notes: This table describes the timing of monthly backcasts, nowcasts, and forecasts of quarterly GDP. For example, in 2010:01, we backcast the GDP growth of 2009:Q4, since its value is not available yet in 2009:Q4; and we nowcast all three months in 2010:Q1. Finally, at the same point in time, we also create monthly forecasts of GDP at 2010:Q2 and 2010:Q3. In the next month, 2010:02, we do not backcast 2009:Q4, since its value is now available. For complete details, refer to Section 5.

Figure 5: Forecasting Using ‘First Available vs. ‘Most Recent’ Real-Time Data*



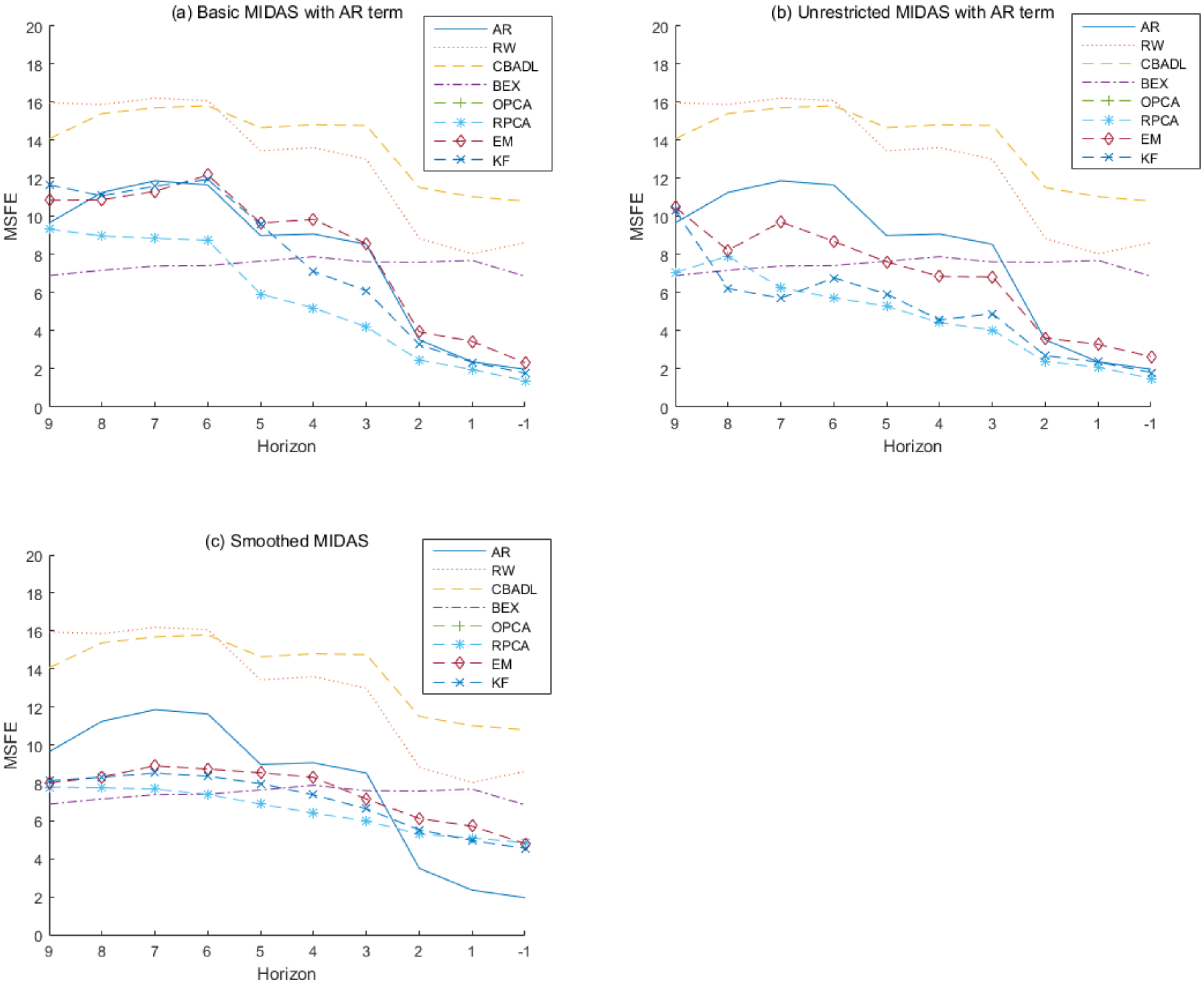
* Notes: This figure plots MSFEs for various models estimated using recursive and rolling data windows, based on either ‘first available’ or ‘most recent’ real-time historical data. Factor-MIDAS models are estimated using OPCA and AR interpolation, and horizons (depicted on the horizontal axes of the graphs) range from nine months ahead (forecasts) to a negative month ahead (backcasts). See Section 5 for complete details.

Figure 6: MSFEs of Forecasting Models Constructed Using One Factor ($r = 1$)*



* Notes: Notes: See notes to Figure 5. Each panel plots the MSFEs of various models for a different estimation method. Benchmark models, including AR, CBADL and BEX, are redundantly included in all panels of this figure, for comparability. ‘Basic w/o AR’ and ‘Basic w/ AR’ are the basic factor-MIDAS models with and without AR terms. ‘Unrest’ and ‘Smooth’ denote alternative factor-MIDAS specifications (see Section 4). OPCA and RPCA are implemented with AR interpolation, and all forecasts are based on recursively estimated models. See Section 5 for complete details.

Figure 7: MSFEs of Factor-MIDAS Models with with One Factor ($r = 1$)*



* Notes: See the Notes to Figure 6.