

# Matlab Toolbox for Mixed Sampling Frequency Data Analysis using MIDAS Regression Models\*

Arthur Sinko<sup>†</sup>      Michael Sockin<sup>‡</sup>      Eric Ghysels<sup>§</sup>

First Draft: December 2009

This Draft: January 5, 2011

© 2010 All rights reserved

Version 1.0

---

\*The third author benefited from funding by the Federal Reserve Bank of New York through the Resident Scholar Program.

<sup>†</sup>Economics, School of Social Sciences, The University of Manchester, Manchester, UK, M13 9PL. e-mail: *arthur.sinko@manchester.ac.uk*

<sup>‡</sup>Department of Economics, Princeton University, Princeton, NJ 08544

<sup>§</sup>Department of Finance - Kenan-Flagler Business School and Department of Economics, University of North Carolina, McColl Building, Chapel Hill, NC 27599. e-mail: *eghysels@unc.edu*.

# 1 Introduction

Regression models involving data sampled at different frequencies are of general interest. In this document we describe Matlab code for running such regression based on a framework put forward in recent work by Ghysels, Santa-Clara, and Valkanov (2002), Ghysels, Santa-Clara, and Valkanov (2006) and Andreou, Ghysels, and Kourtellos (2010) using so called MIDAS, meaning Mi(xed) Da(ta) S(ampling), regressions.

The original work on MIDAS focused on volatility predictions, see e.g. Alper, Fendoglu, and Saltoglu (2008), Aragó and Salvador (2010), Chen and Ghysels (2010), Engle, Ghysels, and Sohn (2008), Brown and Ferreira (2003), Chen, Ghysels, and Wang (2009b), Chen, Ghysels, and Wang (2009a), Clements, Galvão, and Kim (2008), Corsi (2009), Forsberg and Ghysels (2006), Ghysels, Santa-Clara, and Valkanov (2005), Ghysels, Santa-Clara, and Valkanov (2006), Ghysels and Sinko (2006a), Ghysels and Sinko (2006b), Ghysels, Rubia, and Valkanov (2008), León, Nave, and Rubio (2007), Sinko (2006) among others.

Recent work has used the regressions in the context of improving quarterly macro forecasts with monthly data (see e.g. Armesto, Hernandez-Murillo, Owyang, and Piger (2009), Clements and Galvão (2008a), Clements and Galvão (2008b), Frale and Monteforte (2009), Kuzin, Marcellino, and Schumacher (2009), Galvão (2006), Monteforte and Moretti (2008), Marcellino and Schumacher (2010), Schumacher and Breitung (2008), Tay (2007)), or improving quarterly and monthly macroeconomic predictions with daily financial data (see e.g. Andreou, Ghysels, and Kourtellos (2008), Ghysels and Wright (2009), Hamilton (2008), Tay (2006)).

Other regression applications include Ghysels, Plazzi, and Valkanov (2007) (commercial real estate valuation), Valkanov, Yadav, and Zhang (2009) (option early exercise premia),

In addition, dynamic correlation models featuring mixed data sampling schemes based on MIDAS have been used by Colacito, Engle, and Ghysels (2009) and Baele, Bekaert, and Inghelbrecht (2010).

Econometric analysis of MIDAS regressions appears in Ghysels, Santa-Clara, Sinko, and Valkanov (2003), Andreou, Ghysels, and Kourtellos (2010), Bai, Ghysels, and Wright (2009), Kvedaras and Račkauskas (2010), Rodriguez and Puggioni (2010), Wohlrabe (2009), among others.

MIDAS regression can also be viewed as a reduced form representation of the linear projection that emerges from a state space model approach - by reduced form we mean that the MIDAS regression does not require the specification of a full state space system of equations. Bai, Ghysels, and Wright (2009) show that in some cases the MIDAS regression is an exact representation of the Kalman filter, in other cases it involves approximation errors that are typically small. The Kalman filter, while clearly optimal as far as linear projections goes, has several disadvantages (1) it is more prone to specification errors as a full system of measurement and state equations is required and as a consequence (2) requires a lot more parameters, which in turn results in (3) computational complexities that often limit the scope of applications. In contrast, MIDAS regressions - combined with forecast combination schemes if large data sets are involved (see Andreou, Ghysels, and Kourtellis (2008)) are computationally easy to implement and more prone to specification errors.

## 2 Intro to MIDAS regressions

For illustrative purpose we start with a combination of two sampling frequencies - and select a combination of one quarterly and one daily time series. In particular, suppose we are interested in forecasting a quarterly series, denoted  $Y_{t+1}^Q$ , using daily time series, denoted  $X_{j,t}^D$ , for the  $j^{th}$  day in quarter  $t$  (with  $j = 1$  the first day of the quarter and  $j = N_D$  the last, with  $N_D$  the number of days in a quarter - assumed constant for simplicity). The conventional approach, in its simplest form, consists of aggregating the daily data to a quarterly frequency by computing for example averages to obtain  $X_t^Q = (X_{N_D,t}^D + X_{N_D-1,t}^D + \dots + X_{1,t}^D)/N_D$  and subsequently estimate a regression:

$$Y_{t+1}^Q = \mu + \beta X_t^Q + u_{t+1} \tag{2.1}$$

where  $\mu$  and  $\beta$  are unknown parameters and  $u_{t+1}$  is the error term. In (2.1) uses implicitly an equal weighting scheme of the high frequency data since aggregation is based on an average of the daily data. An alternative approach would consist of estimating a model:

$$Y_{t+1}^Q = \mu + \sum_{j=0}^{N_D-1} \beta_{N_D-j} X_{N_D-j,t}^D + u_{t+1}. \tag{2.2}$$

Such an approach is unappealing because of parameter proliferation: when  $N_D = 66$ , we have to estimate 68 slope coefficients.<sup>1</sup>

The key feature of MIDAS regression models is the use of a parsimonious and data-driven weighting scheme. The parsimonious specification yields a linear projection of high frequency data  $X_{.,t}^D$  onto  $Y_t^Q$  using only a few parameters:

$$Y_{t+1}^Q = \mu + \beta \sum_{j=0}^{N_D-1} w_{N_D-j}(\theta^D) X_{N_D-j,t}^D + u_{t+1}. \quad (2.3)$$

Note that equation (2.3) nests the regression model in equation (2.1) under equal or flat weights. We assume that  $\sum_{j=0}^{N_D-1} w_{N_D-j}(\theta^D) = 1$ , which allows us to identify the slope coefficient  $\beta$  in the MIDAS regression model. The parameters  $(\mu, \beta, \theta^D)$  are estimated by Nonlinear Least Squares (NLS). Later we will discuss various parametric specifications for the weighting schemes

Our understanding of MIDAS regression can be further enhanced by decomposing the conditional mean in equation as the sum of an aggregated term based on flat weights,  $X_t^Q$ , and a weighted sum of (higher order) differences of the high frequency variable. Following Andreou, Ghysels, and Kourtellis (2010), we can easily show that the MIDAS term in equation (2.3) can be written as

$$\begin{aligned} \sum_{j=0}^{N_D-1} w_{N_D-j}(\theta^D) X_{N_D-j,t}^D &= \frac{1}{N_D} (X_{N_D,t}^D + X_{N_D-1,t}^D + \dots + X_{1,t}^D) \\ &+ (w_0 - \frac{1}{N_D}) X_{N_D,t}^D + (w_1 - \frac{1}{N_D}) X_{N_D-1,t}^D + \dots \\ &+ (w_{N_D-2} - \frac{1}{N_D-2}) X_{2,t}^D + (\frac{N_D-1}{N_D} - w_0 - w_1 - \dots - w_{N_D-2}) X_{1,t}^D \end{aligned} \quad (2.4)$$

where the last parenthesis uses the assumption that the weights sum to one. Substituting equation (2.4) into (2.3) we get

$$Y_{t+1}^Q = \mu + \beta X_t^Q + \beta \sum_{j=0}^{N_D-1} (w_{N_D-j}(\theta^D) - \frac{1}{N_D}) \Delta^{N_D-j} X_{N_D-j,t}^D + u_{t+1}. \quad (2.5)$$

---

<sup>1</sup>Typically we have about 66 observations for many daily financial data over a quarter since each month has 22 trading days.

Equation (2.5) shows that the traditional temporal aggregation approach, which imposes flat weights  $w_j = 1/N_D$  and only accounts for  $X_t^Q$ , yields an omitted variable term in the regression model (2.1). This implies a host of econometric estimation issues - pertaining to asymptotic inefficiencies at best or, as typical, asymptotic biases - that are discussed in detail in Andreou, Ghysels, and Kourtellis (2010).

The remainder of this section is structured as follows. In subsection 2.1 we cover DL-MIDAS regressions, followed by ADL-MIDAS in subsection 2.2. The parameterization of the various MIDAS regressions is covered in subsection 2.3. Multiplicative ADL-MIDAS are discussed in subsection 2.4. In the next subsection we introduce MIDAS regressions involving other low frequency regressors followed by a subsection covering so called leads in MIDAS regressions.

## 2.1 DL-MIDAS regressions

MIDAS regressions share some features with distributed lag models, or DL models, and also have unique novel features. A stylized DL model is of the following type:

$$Y_{t+1}^Q = \mu + \sum_{j=0}^{q_X^Q-1} \beta_j(\theta^Q) X_{t-j}^Q + u_{t+1}, \quad (2.6)$$

where  $\sum_j \beta_j(\theta^Q)$  is some finite or infinite lag polynomial operator, usually parameterized by a small set of hyperparameters (see e.g. Dhrymes (1971) for a survey on distributed lag models). The same idea was used in the MIDAS regression we discussed so far, albeit with series sampled at different frequencies.

By analogy with DL models, we can characterize a *DL - MIDAS*( $p_X^D$ ) regression model as:

$$Y_{t+1}^Q = \mu + \beta \sum_{j=0}^{p_X^D-1} \sum_{i=0}^{N_D-1} w_{N_D-i+j*N_D}(\theta^D) X_{N_D-i,t-j}^D + u_{t+1}, \quad (2.7)$$

where the second summation allows for daily lags to extend beyond the last day of quarter  $t$ , but to simplify notation, we will always take lags in blocks of quarterly sets of daily data,  $p_X^D$ . Note that equation (2.7) nests the simple DL model in equation (2.1) under flat-weights. We assume again that  $\sum_{j=0}^{p_X^D-1} \sum_{i=0}^{N_D-1} w_{N_D-i+j*N_D}(\theta^D) = 1$ , which allows for the identification of the slope coefficient  $\beta$  in the DL-MIDAS regression model.

Note that in a general DL-MIDAS regression model the regressor is sampled  $m$  times more frequently than the regressand where the latter has a sample of  $T$  observations. The asymptotics are based on  $T$  for a given  $m$ . Therefore one studies the asymptotic properties of the estimators assuming that the span of the data set  $T$  grows and the high frequency sample size of the regressors would be  $mT$  such that when  $T \rightarrow \infty$  then both the low and high frequency samples become large.

## 2.2 ADL-MIDAS regressions

When  $Y_{t+1}^Q$  is serially correlated, as it is typically the case for time series variables, the simple model in equation (2.1) is extended to a dynamic linear regression or autoregressive distributed lag (ADL) model. Again the conventional approach, in its simplest form, aggregates the high frequency data at the low frequency by computing simple averages and estimates a simple linear regression of  $Y_{t+1}^Q$  on  $X_t^Q$ . Take for instance the ADL(1,1)

$$Y_{t+1}^Q = \mu + \mu Y_t^Q + \beta X_t^Q + u_{t+1}, \quad (2.8)$$

where  $\mu$  and  $\beta$  are unknown parameters and  $u_{t+1}$  is an error term. In a similar manner the *ADL – MIDAS*( $p_Y^Q, p_X^D$ ) is:

$$Y_{t+1}^Q = \mu + \sum_{j=0}^{p_Y^Q-1} \mu_{j+1} Y_{t-j}^Q + \beta \sum_{j=0}^{p_X^D-1} \sum_{i=0}^{N_D-1} w_{N_D-i+j*N_D}(\theta^D) X_{N_D-i,t-j}^D + u_{t+1} \quad (2.9)$$

Note that again the number of daily lags is a multiple of the number of trading days in a quarter,  $N_D$ . As above the slope coefficient  $\beta$  in the MIDAS regression is identified via the scaling of the weights, such that they add up to one. The above model specification generates notation very similar to ARMA models, e.g. ADL-MIDAS(1,1) or ADL-MIDAS(AIC,AIC) (more on model selection later).

## 2.3 Parameterizations the MIDAS polynomial weights

Various parameterizations for the polynomial lag structure appearing in equations such as (2.7), (2.9), (2.23), (2.21), (2.22), etc. have been used and discussed notably in Ghysels, Santa-Clara, Sinko, and Valkanov (2003). We will use  $N$  as the number of lags in the

MIDAS polynomial - without specific reference to say daily lags. The specifications are as follows:

1. Normalized beta probability density function, unrestricted ( $u$ ) and restricted ( $r$ ) cases with non-zero and zero last lag. Please note that for specifications with a small number of MIDAS lags the zero-last-lag assumption may generate significant bias in the weighting scheme.

$$w_i^{u,nz} = w_i(\theta_1, \theta_2, \theta_3) = \frac{x_i^{\theta_1-1}(1-x_i)^{\theta_2-1}}{\sum_{i=1}^N x_i^{\theta_1-1}(1-x_i)^{\theta_2-1}} + \theta_3 \quad (2.10)$$

$$w_i^{r,nz} = w_i(1, \theta_2, \theta_3) \quad (2.11)$$

$$w_i^{u,z} = w_i(\theta_1, \theta_2, 0) \quad (2.12)$$

$$w_i^{r,z} = w_i(1, \theta_2, 0) \quad (2.13)$$

where  $x_i = (i-1)/(N-1)$ .<sup>2</sup>

2. Normalized exponential Almon lag polynomial

$$w_i^u = w_i(\theta_1, \theta_2) = \frac{e^{\theta_1 i + \theta_2 i^2}}{\sum_{i=1}^N e^{\theta_1 i + \theta_2 i^2}} \quad (2.14)$$

$$w_i^r = w_i(\theta_1, 0) \quad (2.15)$$

3. Almon lag polynomial specification of order  $P$  (not normalized, i.e. sum of individual weights is not equal to 1 and  $\beta w_i(\theta)$  from Eq. 2.9 are estimated jointly).

$$\beta w_i(\theta_0, \dots, \theta_P) = \sum_{p=0}^P \theta_p i^p \quad (2.16)$$

---

<sup>2</sup>To eliminate irregular behavior of the polynomial for some values of  $\theta$  at the ends of  $[0,1]$  interval we use instead  $x_i = eps + (i-1)/(N-1)(1-eps)$ , where  $eps$  is a machine 0 for MATLAB.

Note that this can also be written in matrix form:

$$\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2^2 & \cdots & 2^P \\ 1 & 3 & 3^2 & \cdots & 3^P \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & N & N^2 & \cdots & N^P \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_P \end{bmatrix} \quad (2.17)$$

Therefore the use of Almon lags in MIDAS models can be achieved via OLS estimation with properly transformed high frequency data regressors using the matrix representation appearing in the above equation. Once the weights are estimated via OLS, one can always rescale them to obtain a slope coefficient (assuming the weights do not sum up to zero).

#### 4. Polynomial specification with step functions (not normalized)

$$\begin{aligned} \beta w_i(\theta_1, \dots, \theta_P) &= \theta_1 I_{i \in [a_0, a_1]} + \sum_{p=2}^P \theta_p I_{i \in (a_{p-1}, a_p]} \\ a_0 &= 1 < a_1 < \dots < a_P = N \\ I_{i \in [a_{p-1}, a_p]} &= \begin{cases} 1, & a_{p-1} \leq i \leq a_p \\ 0, & \textit{otherwise} \end{cases} \end{aligned} \quad (2.18)$$

where  $a_0 = 1 < a_1 < \dots < a_P = N$ .

We implicitly referred to model selection in the previous subsection when we mentioned ADL-MIDAS(AIC,AIC). Indeed, AIC/BIC criteria will be used to select optimal number of lags for the regressions, as discussed in detail Section 3.

## 2.4 Multiplicative MIDAS regressions

Recall that temporal aggregation imposes a fixed weighting scheme such as equal weights. What would happen if we instead consider a parameter-driven regressor:

$$X_t^Q(\theta_X^D) \equiv \sum_{i=0}^{N_D-1} w_{N_D-i}(\theta_X^D) X_{N_D-i,t}^D$$

where we scale weights again such that they add up to one. This will allow us to estimate a model we call multiplicative MIDAS or *ADL – MIDAS – M*( $p_Y^Q, p_X^Q$ ):

$$Y_{t+1}^Q = \mu + \sum_{k=0}^{p_Y^Q-1} \mu_k Y_{t-k}^Q + \sum_{k=0}^{p_X^Q-1} \beta_k X_{t-k}^Q(\theta_X^D) + u_{t+1} \quad (2.19)$$

The above equation looks like a standard ADL model, except that it involves a parameter-driven regressor that mimics an aggregation scheme. There are pros and cons to the above approach. The pros are that the above multiplicative weighting scheme corresponds to the structure of a steady state Kalman filter with mixed data sampling, see Bai, Ghysels, and Wright (2009) for further details. Furthermore, the specification nests standard aggregation schemes and can also easily handle very complex aggregation schemes, as shown in Chen and Ghysels (2010). The main disadvantage of the multiplicative MIDAS regression scheme is that it is not as parsimonious as the ADL-MIDAS in equation (2.9) which involves a single polynomial and therefore requires the estimation of a few parameters. Bai, Ghysels, and Wright (2009) in fact show that the differences in terms of RMSFE between standard ADL-MIDAS and the multiplicative specification are small, including cases where the multiplicative specification is an exact match of the steady state Kalman filter.

## 2.5 Factors and other regressors in ADL-MIDAS models

Recently, a large body of recent work has developed factor model techniques that are tailored to exploit a large cross-sectional dimension; see for instance, Bai and Ng (2002), Bai (2003), Forni, Hallin, Lippi, and Reichlin (2000), Forni, Hallin, Lippi, and Reichlin (2005), Stock and Watson (1989), Stock and Watson (2003), among many others. These factors are usually estimated at quarterly frequency using a large cross-section of time-series. Following this literature Andreou, Ghysels, and Kourtellis (2008) investigate whether

one can improve factor model forecasts by augmenting such models with high frequency information, especially daily financial data.

We therefore augment the aforementioned MIDAS models with factors,  $F_t$ , obtained by following dynamic factor model

$$\begin{aligned} X_t &= \Lambda_t F_t + u_t \\ F_t &= \Phi F_{t-1} + \eta_t \\ u_{it} &= a_{it}(L)u_{it-1} + \varepsilon_{it}, \quad i = 1, 2, \dots, n \end{aligned} \tag{2.20}$$

where the number of factors is computed using criteria proposed by Bai and Ng (2002). The data used to implement the factor representation will be described in the next section. Suffice it here to say that we use series similar to those used by Stock and Watson (2008a).

Augmenting the MIDAS regression models from the previous subsection with the factors, we obtain a richer family of models that includes monthly frequency lagged dependent variable, quarterly factors, and a daily financial indicator. For instance, equation (2.9) generalizes to the  $FADL - MIDAS(p_F, p_Y^Q, p_X^D)$  model:

$$Y_{t+1}^Q = \mu + \sum_{i=0}^{p_F-1} \beta_i^F F_{t-i}^Q + \sum_{j=0}^{p_Y^Q-1} \mu_j Y_{t-j}^Q + \beta \sum_{j=0}^{p_X^D-1} \sum_{i=1}^{N_D} w_{N_D-i+j*N_D}(\theta^D) X_{N_D-i, t-j}^D + u_{t+1} \tag{2.21}$$

where we use quarterly factors  $F^Q$  to augment the ADL-MIDAS regression. Note that we can also formulate a  $FADL - MIDAS - M(p_F, p_Y^Q, p_X^Q)$  model that involves a multiplicative MIDAS weighting scheme.

Equation (2.21) simplifies to the traditional factor model with additional regressors when the MIDAS features are turned off - i.e. say a flat aggregation scheme is used. When the lagged dependent variable is excluded then we have a projection on daily data, combined with aggregate factors.

It should finally be noted that we can add any low frequency regressor, not just factors. The software is written such that one can add any type of low frequency regressor.

## 2.6 MIDAS with leads

Giannone, Reichlin, and Small (2008), among others, have formalized the process of updating the nowcast and forecasts as new releases of data become available.<sup>3</sup> These studies typically use again state space setup. The process can be mimicked via MIDAS regression models with *leads*. Say we are one or two months into quarter  $t + 1$ . We consider MIDAS models with two months of daily lead data in order to incorporate real-time information available. Consider the Factor ADL model with MIDAS in equation (2.21), which allows for  $J_X^D$  daily leads for the daily predictor. Then we can specify the *FADL – MIDAS*( $p_Y^Q, p_F^Q, p_X^D, J_X^D$ ) model

$$\begin{aligned}
Y_{t+1}^Q &= \mu + \sum_{i=0}^{p_F-1} \beta_i^F F_{t-i}^Q + \sum_{j=0}^{p_Y^Q-1} \mu_j Y_{t-j}^Q \\
&+ \beta \left[ \sum_{i=0}^{J_X^D-1} w_{J_X^D-i}(\theta^D) X_{J_X^D-i, t+1}^D + \sum_{j=0}^{p_X^D-1} \sum_{i=0}^{N_D-1} w_{J_X^D+N_D-i+j*N_D}(\theta^D) X_{i, t-j}^D \right] \\
&+ u_{t+1}
\end{aligned} \tag{2.22}$$

In the above equation, we use a single polynomial that carries the leads as well as the daily lags. In a multiplicative MIDAS scheme, we can write a model with leads as follows:

$$X_{t+1}^{J_X^D}(\theta_X^D) \equiv \sum_{i=0}^{J_X^D-1} w_{J_X^D-i}(\theta_X^D) X_{J_X^D-i, t}^D$$

where we scale weights again such that they add up to one. This will allow us to estimate a model we call multiplicative MIDAS or *ADL – MIDAS – M*( $p_Y^Q, p_X^Q$ ):

$$Y_{t+1}^Q = \mu + \sum_{k=0}^{p_Y^Q-1} \mu_k Y_{t-k}^Q + \beta_L X_{t+1}^{J_X^D}(\theta_X^D) + \sum_{k=0}^{p_X^Q-1} \beta_k X_{t-k}^Q(\theta_X^D) + u_{t+1} \tag{2.23}$$

Hence, the aggregation scheme is used via a partial sum for the lead days into month  $t + 1$ .

To conclude it should be noted that two modes of forecasting can be used in the Matlab MIDAS Toolbox. The first is fixed in-sample estimation and fixed out-of-sample prediction

---

<sup>3</sup>Nowcasting is studied at length by Doz, Giannone, and Reichlin (2008), Doz, Giannone, and Reichlin (2006), Stock (2006), Angelini, Camba-Mendez, Giannone, Rünstler, and Reichlin (2008), Giannone, Reichlin, and Small (2008), Moench, Ng, and Potter (2009), among others.

and the second is a rolling window approach. For details, see Section 3.

## 2.7 Forecast combinations

There is a large literature on forecast combinations, see Timmermann (2006) for an excellent survey. Although there is a consensus that forecast combinations improve forecast accuracy there is no consensus concerning how to form the forecast weights.

Given the findings in Stock and Watson (2004), Stock and Watson (2008b) and Andreou, Ghysels, and Kourtellos (2008) we focus primarily on the Squared Discounted MSFE forecast combinations method, which delivers the highest forecast gains relative to other methods in many applications. The software also includes a BIC-based criterion as an option.

Let  $\hat{y}_{i,t+h|t}$  denote the  $i^{\text{th}}$  individual out-of-sample forecast of  $y_{i,t+h|t}$  computed at date  $t$ . The combination forecasts are weighted averages of the individual forecasts (possibly with time-varying weights):

$$f_{t+h|t} = \sum_{i=1}^n w_{i,t} \hat{y}_{i,t+h|t} \quad (2.24)$$

where  $n$  is the number of individual models. We consider four different weighting schemes:

- Equally weighted weights

$$w_{i,t} = \frac{1}{n} \quad (2.25)$$

- BIC-weighted forecast

$$w_{i,t} = \frac{\exp(-BIC_i)}{\sum_{i=1}^n \exp(-BIC_i)} \quad (2.26)$$

- MSFE-related model averaging:

$$w_{i,t} = \frac{m_{i,t}^{-1}}{\sum_{i=1}^n m_{i,t}^{-1}} \quad (2.27)$$

$$m_{i,t} = \sum_{i=T_0}^t \delta^{t-i} (y_{s+h}^h - \hat{y}_{i,s+h|s}^h)^2$$

where  $T_0$  is the first out-of-sample observation,  $\hat{y}_{i,s+h|s}^h$  – out-of-sample forecast,  $\delta$  – exponential averaging parameter.

1. MSFE averaging:  $\delta = 1$
2. DMSFE averaging:  $\delta = .9$

## 2.8 Nuts and bolts issues

Practical implementation of MIDAS involves issues that are typical for regression analysis, yet there are some not commonly encountered in standard regression problems and they pertain to the mixed sampling nature of the data.

Since the quarterly/daily combination has been used throughout this document, consider the situation of holidays occurring throughout a calendar year. This will create an unequal number of days on a quarter by quarter basis. While one can take different approaches towards this, we treat the holidays as missing values in the MIDAS polynomial. They will be linearly interpolated using various schemes.

The algorithms can be grouped into (1) specifications with the same number of MIDAS lags each period and (2) specifications that cover the same time span each period.

Define a sequence of MIDAS polynomial weights  $w_{\tau_1}, w_{\tau_2}, \dots$ . Then we have the following:

1. Equally-spaced specification.
  - (a) It is characterized by the fact that each observation point  $\{y_t, X factor_t, Xmidas_t\}$  has the same number of MIDAS lags  $Xmidas_t$ . As a result, different periods may have different time span coverage *but the same number of lags*. The sequence of weights  $w_{\tau_i}, w_{\tau_{i+1}}, \dots$  is defined in this case as  $w_i, w_{i+1}, \dots$
2. Real-time specifications. They are characterized by unequal number of MIDAS lags over time that cover *the same time span*.
  - (a) Real time specification. The distance between  $w_{\tau_i}$  and  $w_{\tau_{i+1}}$  is proportionate to  $\tau_i - \tau_{i+1}$ . No artificial observations are inserted in the MIDAS polynomial.
  - (b) Real time specification with zeros at the end. Depending on the number of calendar days within a given time interval all missing days are added as zeros to the end of Xmidas lag structure. MIDAS weights are constructed as in the equally-spaced case.<sup>4</sup>

---

<sup>4</sup>Please note that normalization of the polynomial in this case is different from the equally-spaced

### 3 Software Usage

In this section we describe the implementation of data generation, estimation and forecast averaging algorithms for MIDAS regressions. The model of interest is

$$Y_{t+s}^Q = \mu + \sum_{k=1}^K \sum_{i=0}^{p_F^k-1} \beta_{i,k}^F F_{t-i,k}^Q + \sum_{j=0}^{p_Y^Q-1} \mu_j Y_{t-j}^Q + \beta \sum_{j=0}^{p_X^D-1} \sum_{i=0}^{N_D-1} w_{N_D-i+j*N_D}(\theta^D) X_{N_D-i,t-j}^D + u_{t+s} \quad (3.1)$$

Currently multiplicative MIDAS is not implemented ( $p_x^D = 1$ ). MIDAS with leads specification is implemented via defining proper values of  $s$ . It can take either numeric or string values. Consider the following example. Assume that  $Y$  is observed on a quarterly basis. Numeric values of  $s$  (1, 2, 3, ...) define regression in terms of lags of  $Y$ . If  $s = 1$  and  $Y$  is observed quarterly, the data is constructed using data vector  $X$  observable *prior* to the lagged value of  $Y$  ( $X$  observable at the date when lagged  $Y$  is available). String values of  $s$  define the time distance between  $Y$  and explanatory variables in terms of calendar days, weeks, months, quarters. For example,  $s = '3m'$  or  $s = '1q'$  will generate results that are close to the result for  $s = 1$  for quarterly  $Y$ .<sup>5</sup> MIDAS with leads is estimated when time lag defined by  $s$  is smaller than 3 months. Say, for  $s = '1m'$  in this case will use For  $s < '3m'$  it will perform “nowcasting”. In particular,

An example file `midas_example.m` is located in the main directory. In this file we generate an artificial data set with 2 factors, 30 MIDAS daily lags, and 3 dependent variable lags and subsequently estimate the parameters of the regression.

For the purposes of estimation the input data is stored as a structure with fixed fields. We use the artificial data from `midas_example.m` as an example of such a structure:

```
%Vector of high-frequency data that generates  $X_{i,t-j}^D$ 
data.x=randn(10000,1)*5;
%Corresponding dates in the datenum() form
data.xdate=datenum(1990,1,1:10000)';
%Vector of data to form  $Y_{t+s}^Q$ 
data.y=(1:yperiod:10000)';
```

specification.

<sup>5</sup>Dates of announcement of  $Y$  do not necessarily occur at the same day of the month. Thus, “one quarter ago” and “one lag ago” might refer to different dates.

```

%Corresponding dates in the datenum() form
data.ydate=datenum(1990,1,1:yperiod:10000)';
%K factors to form  $F_{t-i,k}^Q$ . They are defined as a cell array. As
%a result each factor can have different number of observations. The resulting
% dataset will be the intersection of all timespans
data.factors={randn(size(data.y)) randn(size(data.y))};
%corresponding cell array with K vectors of dates
data.factorsd={data.ydate data.ydate};

```

The program allows to choose forecasting/nowcasting horizon  $s$  of  $Y_{t+s}$  (Eq. 3.1) in terms of lags of  $Y$ , number of years, quarters, months, weeks, days. The next step is to define parameters for proper data transformation, i.e.  $s, K, p_F^k$  etc. All options are stored in a special structure. The options are divided into three groups: data generation, polynomial specification, estimation and estimation options.

```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Data generation options %%%%%%%%%%
%Number of HF lags/time span in MIDAS polynomial opt.aggr.xlag
%( $p_x^D - 1$  in Eq.2.21) It can be defined either in numeric
%format or in string format numeric|'Nd'|'Nwk'|'Nmn'|'Nq'|'Ny',
%where N - is a number of days/weeks/months/quarters/years
%For the equally spaced specification p.12 1a it should be numeric.
%For all other specifications (p.12 2a, 2b, 2.8) it should be char.
opt.aggr.xlag=30;

%linear interpolation for missing dates in MIDAS. true|false (See p.12)
opt.aggr.xmidaslip=false;

%equally spaced specification, xlag should be numeric
opt.est.polyconstr='es';
%real time specification with zeros at the end, xlag should be char
opt.est.polyconstr='zae';
%real time specification, xlag should be char
opt.est.polyconstr='rt';
%Number of Y lags ( $p_y^Q - 1$ ) in Eq.2.21)
opt.aggr.ylag=3;
%Forecasting horizon opt.est.horizon s in Eq.3.1)
%Can be defined either in numeric format (a simple forecasting case)
%or in string format (forecasting/nowcasting cases).

```

```

%numeric|'Nd'|'Nwk'|'Nmn'|'Nq'|'Ny',
opt.est.horizon=1;

%number of lags  $p_F^k - 1$  for each factor  $k = 1, \dots, K$  in Eq.3.1)
opt.aggr.factors=[3 2];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Data estimation options %%%%%%%%%%

% Optimal Y lags selection (in terms of AIC or BIC) See Section ??
% If enabled, it invalidates opt.aggr.ylag value.
opt.autoIC.auto=true; %enables automatic IC selection (true|false)
opt.autoIC.ic='bic'; %information criterion to optimize ('aic'|'bic')
opt.autoIC.minlag=0; %minimum number of lags considered
opt.autoIC.maxlag=12; %maximum number of lags considered
%Reports all lags if true (true|false). Otherwise the optimization
%procedure reports only optimal ones.
opt.autoIC.reportall=false;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Polynomial specification (See Subsection 2.3) %%%%%%%%%%

%Generates proper default values for the substructure opt.poly which
%suppresses estimation:
opt.poly=optpolyinit();
%Estimation occurs only if the default values are changed
opt.poly.betann =2;%Non-zero Beta poly degree. Eq.2.11, if 1; Eq.2.10, if 2
opt.poly.beta=2;%Zero Beta poly degree. Eq.2.13, if 1; Eq.2.12, if 2
opt.poly.exp=2; %Exp. Almon poly degree Eq.2.15, if 1; Eq.2.14, if 2.
opt.poly.almon = 3; %Almon polynomial degree  $P$  (Eq.2.16).
opt.poly.stepfun=[5 15 30]; %Defines  $a_1, \dots, a_{P-1}$  (Eq.2.18).

```

Forecasting is defined using the options below. If they are not defined, the whole sample is used for estimation.

```

%start date for the initial window, either datestr() or observation number
opt.est.start = '06/02/1990';
%end date for initial window, either datestr or observation number
opt.est.end = '12/01/1998';
%designates rolling window (1 for recursive estimation),
%0 for the fixed in-sample/out-of-sample
opt.est.estmethod = 0;

```

Given that the opt and data structures are defined as above, the estimation procedure runs as follows:

```
output=midas_interface_adl_new(opt,data);
```

It will generate the output cell structure with the following substructures (if they were defined in the polynomial specification options):

```
output{1}.beta: %Beta polynomial with zero last lag output
output{1}.betann: %Beta polynomial with non-zero last lag output
output{1}.exp: %Exp Almon polynomial output
output{1}.stepfun: %Stepfunction MIDAS output
output{1}.almon: %Almon lag output
output{1}.ylag %Optimal ylag
```

Each of the substructures has the following entries:

```
.aic %AIC
.bic %BIC
.aopt %All optimal parameters as one vector
.aoptxols %All optimal parameters but MIDAS
.aoptmidas %All MIDAS optimal parameters
.error %Errors of estimation
.estvar %Estimated Covariance matrix of parameters using  $(X'X)^{-1}\hat{\sigma}^2$ 
.stderror %St errors of parameters using .estvar
.Tvalues %t-values of parameters using .stderror
.R2 %R2
.robustV %Robust variance using NW kernel and sandwich estimator
.robustSE %Robust se using .robustV
.robustT %Robust t-stat using .robustSE
%Out-of-sample Forecast
.yfor
```

Another set of substructures that partially duplicates the results above in more structured way

```
.COEFFS%structure of coefficients
.STDERR%structure of sterrors
.TVALUES%structure of tvalues
.ROBUSTSE%structure of robust sterrors
.ROBUSTT%structure of robust t-stats
```

Each of these substructures have the following fields:

```
%The most recent parameter is saved in the first element of the vector
.factorN %parameters/t-stats/etc for the nth factor
.factorN-1 %parameters/t-stats/etc for the n-1th factor
...
.factor1%parameters/t-stats/etc for the 1st factor
.ylags%parameters/t-stats/etc for y lags
%Intercept is in .midas(1),  $\beta$  (Exp, Beta, Betann) is in .midas(2), etc
%For stepfunc and Almon specifications they are stored in their natural order
.midas%parameters/t-stats/etc for MIDAS polynomials
```

Model averaging is invoked via the following function:

```
averaging=midas_model_averaging(opt,data);
averaging.dmsfe; %Discounted MSFE average Eq.2.27,  $\delta = .9$ 
averaging.msfe %MSFE average Eq.2.27,  $\delta = 1$ 
averaging.mean %equal weighting Eq.2.25
averaging.BIC %BIC average weighting Eq.2.26,  $\delta = .9$ 
```

. There are three ways to define the models:

- Different model specifications for the same data set (vector of opt structures for the same data structure).
- Different data sets for the same model(s) (Vector of data structures for the same opt structure).
- Different model specifications for different data set.

Note: It is sufficient to define in the consecutive opt elements only fields that are different from the opt(1). For example, to estimate an additional model using the same data with

two factors with maximum lags [1 1] instead of [3 2], it is sufficient to define

```
opt(2).aggr.factors=[1 1];
```

## References

- Alper, C.E., S. Fendoglu, and B. Saltoglu, 2008, Forecasting Stock Market Volatilities Using MIDAS Regressions: An Application to the Emerging Markets, Discussion paper, MPRA Paper No. 7460.
- Andreou, E., E. Ghysels, and A. Kourtellos, 2008, Should macroeconomic forecasters look at daily financial data?, Discussion paper, Discussion Paper UNC and University of Cyprus.
- Andreou, Elena, Eric Ghysels, and Andros Kourtellos, 2010, Regression Models With Mixed Sampling Frequencies, *Journal of Econometrics* 158, 246–261.
- Angelini, Elena, Gonzalo Camba-Mendez, Domenico Giannone, Gerhard Rünstler, and Lucrezia Reichlin, 2008, A Two-Step Estimator for Large Approximate Dynamic Factor Models Based on Kalman Filtering, ECB Working Papers.
- Aragó, V., and E. Salvador, 2010, Re-Examining the Risk-Return Relationship: The Influence of Financial Crisis (2007-2009), Discussion paper, Discussion Paper, Universitat Jaume I, Spain.
- Armesto, Michelle T., Rubén Hernandez-Murillo, Michael Owyang, and Jeremy Piger, 2009, Measuring the Information Content of the Beige Book: A Mixed Data Sampling Approach, *Journal of Money, Credit and Banking* 41, 35–55.
- Baele, L., G. Bekaert, and K. Inghelbrecht, 2010, The determinants of stock and bond return comovements, *Review of Financial Studies* 23, 2374–2428.
- Bai, Jushan, 2003, Inferential theory for factor models of large dimensions, *Econometrica* pp. 135–171.
- Bai, J., E. Ghysels, and J.H. Wright, 2009, State Space Models and MIDAS Regressions, Discussion Paper, NY Fed, UNC and Johns Hopkins.
- Bai, Jushan, and Serena Ng, 2002, Determining the number of factors in approximate factor models, *Econometrica* 70, 191–221.
- Brown, David P., and Miguel A. Ferreira Ferreira, 2003, The Information in the Idiosyncratic Volatility of Small Firms, Working paper, Univesrity of Wisconsin and ISCTE.

- Chen, X., and E. Ghysels, 2010, News-Good or Bad-and its Impact on Volatility Forecasts over Multiple Horizons, .
- , and F. Wang, 2009a, On the role of Intra-Daily Seasonality in HYBRID GARCH Models, Discussion paper, *Journal of Time Series Econometrics*, forthcoming.
- , 2009b, The HYBRID GARCH class of models, Discussion paper, Working Paper, UNC.
- Clements, M.P., and A.B. Galvão, 2008a, Forecasting US output growth using Leading Indicators: An appraisal using MIDAS models, *Journal of Applied Econometrics* (forthcoming).
- , and J.H. Kim, 2008, Quantile forecasts of daily exchange rate returns from forecasts of realized volatility, *Journal of Empirical Finance* 15, 729–750.
- Clements, Michael P., and Ana B. Galvão, 2008b, Macroeconomic Forecasting with Mixed Frequency Data: Forecasting US output growth, *Journal of Business and Economic Statistics* 26, 546–554.
- Colacito, R., R.F. Engle, and E. Ghysels, 2009, A component model for dynamic correlations, .
- Corsi, F., 2009, A simple approximate long-memory model of realized volatility, *Journal of Financial Econometrics* 7, 174–196.
- Dhrymes, Phoebus, 1971, *Distributed Lags: Problems of Formulation and Estimation* (Holden-Day: San Francisco).
- Doz, Catherine, Domenico Giannone, and Lucrezia Reichlin, 2006, A Two-Step Estimator for Large Approximate Dynamic Factor Models Based on Kalman Filtering, CEPR Discussion Papers 6043.
- , 2008, A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor Models, CEPR Discussion Papers 5724.
- Engle, Robert F., Eric Ghysels, and Bumjean Sohn, 2008, On the Economic Sources of Stock Market Volatility, Discussion Paper NYU and UNC.

- Forni, Mario, Marc Hallin, Marco Lippi, and Lucrezia Reichlin, 2000, The generalized dynamic-factor model: Identification and estimation, *Review of Economics and Statistics* 82, 540–554.
- , 2005, The generalized dynamic factor model, *Journal of the American Statistical Association* 100, 830–840.
- Forsberg, Lars, and Eric Ghysels, 2006, Why do absolute returns predict volatility so well?, *Journal of Financial Econometrics* 6, 31–67.
- Frale, C., and L. Monteforte, 2009, FaMIDAS: A Mixed Frequency Factor Model with MIDAS structure, Discussion Paper, Bank of Italy.
- Galvão, Ana B., 2006, Changes in Predictive Ability with Mixed Frequency Data, Discussion Paper Queen Mary.
- Ghysels, E., A. Plazzi, and R. Valkanov, 2007, Valuation in US commercial real estate, *European Financial Management* 13, 472–497.
- Ghysels, E., A. Rubia, and R. Valkanov, 2008, Multi-Period Forecasts of Variance: Direct, Iterated, and Mixed-Data Approaches, Working paper, Alicante, UCSD and UNC.
- Ghysels, Eric, Pedro Santa-Clara, Arthur Sinko, and Rossen Valkanov, 2003, MIDAS Regressions: Further Results and New Directions, Working paper, UNC and UCLA.
- Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov, 2002, The MIDAS touch: Mixed data sampling regression models, Working paper, UNC and UCLA.
- Ghysels, Eric, Pedro Santa-Clara, and Ross Valkanov, 2005, There is a risk-return tradeoff after all, *Journal of Financial Economics* 76, 509–548.
- Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov, 2006, Predicting volatility: getting the most out of return data sampled at different frequencies, *Journal of Econometrics* 131, 59–95.
- Ghysels, E., and A. Sinko, 2006a, Comment on realized variance and market microstructure noise by p. r. hansen and asger lunde, *Journal of Business and Economic Statistics* 24, 192–194.

- Ghysels, Eric, and Arthur Sinko, 2006b, Volatility prediction and microstructure noise, Work in progress.
- Ghysels, Eric, and Jonathan Wright, 2009, Forecasting professional forecasters, *Journal of Business and Economic Statistics* 27, 504–516.
- Giannone, Domenico, Lucrezia Reichlin, and David Small, 2008, Nowcasting: The real-time informational content of macroeconomic data, *Journal of Monetary Economics* 55, 665–676.
- Hamilton, J.D., 2008, Daily monetary policy shocks and new home sales, *Journal of Monetary Economics* 55, 1171–1190.
- Kuzin, V., M. Marcellino, and C. Schumacher, 2009, MIDAS versus mixed-frequency VAR: nowcasting GDP in the euro area, .
- Kvedaras, V., and A. Račkauskas, 2010, Regression Models with Variables of Different Frequencies: The Case of a Fixed Frequency Ratio\*, Discussion paper, *Oxford Bulletin of Economics and Statistics*, forthcoming.
- León, Ángel., Juan M. Nave, and Gonzalo Rubio, 2007, The relationship between risk and expected return in Europe, *Journal of Banking and Finance* 31, 495–512.
- Marcellino, M., and C. Schumacher, 2010, Factor MIDAS for Nowcasting and Forecasting with Ragged-Edge Data: A Model Comparison for German GDP, *Oxford Bulletin of Economics and Statistics* 72, 518–550.
- Moench, Emanuel, Serena Ng, and Simon Potter, 2009, Dynamic hierarchical factor models, Staff Reports 412, Federal Reserve Bank of New York.
- Monteforte, L., and G. Moretti, 2008, Real time forecasts of inflation: the role of financial variables, Discussion Paper, Bank of Italy.
- Rodríguez, A., and G. Puggioni, 2010, Mixed frequency models: Bayesian approaches to estimation and prediction, *International Journal of Forecasting* 26, 293–311.
- Schumacher, Christian, and Jorg Breitung, 2008, Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data, *International Journal of Forecasting* 24, 386–398.

- Sinko, Arthur, 2006, On predictability of market microstructure noise volatility, Work in progress.
- Stock, James H., 2006, Forecasting and Now-casting with Disparate Predictors: Dynamic Factor Models and Beyond, Manuscript, Department of Economics, Harvard University.
- , and Mark W. Watson, 1989, New indexes of coincident and leading economic indicators, *NBER Macroeconomics Annual* pp. 351–394.
- , 2003, Forecasting output and inflation: the role of asset prices, *Journal of Economic Literature* pp. 788–829.
- , 2004, Combination Forecasts Of Output Growth In A Seven-Country Data Set, *Journal of Forecasting* 23, 405–430.
- , 2008a, Forecasting in Dynamic Factor Models Subject to Structural Instability, in Jennifer Castle, and Neil Shephard, ed.: *The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry*. Oxford University Press.
- , 2008b, Phillips Curve Inflation Forecasts, NBER Working paper.
- Tay, Anthony S., 2006, Financial Variables as Predictors of Real Output Growth, Discussion Paper SMU.
- , 2007, Mixing Frequencies: Stock Returns as a Predictor of Real Output Growth, Discussion Paper SMU.
- Timmermann, Allan, 2006, Forecast combinations, in Allan Timmerman Graham Elliott, Clive Granger, ed.: *Handbook of Economic Forecasting* vol. 1 . pp. 136–196 (North Holland: Amsterdam).
- Valkanov, R., P. Yadav, and Y. Zhang, 2009, Does the Early Exercise Premium Contain Information about Future Underlying Returns?, Discussion paper, Discussion Paper UCSD.
- Wohlrabe, K., 2009, Forecasting with mixed-frequency time series models, Discussion paper, Ph. D. Dissertation Ludwig-Maximilians-Universitat Munchen.