Dynamic component models for forecasting trading volumes

Modelli dinamici a componenti per la previsione dei volumi

Antonio Naimoli and Giuseppe Storti

Abstract We propose a new class of models for high-frequency trading volumes. Namely we consider a component model where the long-run dynamics are based on a Heterogeneous MIDAS polynomial structure based on an additive cascade of MIDAS filters moving at different frequencies. The merits of the proposed approach are illustrated by means of an application to three stocks traded on the XETRA market characterised by different degrees of liquidity.

Abstract *Viene proposta una nuova classe di modelli per volumi azionari ad alta frequenza. In particolare viene proposto un modello a componenti dove le dinamiche di lungo periodo sono basate su una struttura polinomiale di tipo MIDAS costituita da una cascata additiva di filtri MIDAS che si muovono a diverse frequenze. I vantaggi dell'approccio proposto vengono illustrati attraverso una applicazione a tre azioni contrattate sul mercato XETRA e caratterizzate da diversi livelli di liquidità.*

Key words: Intra-daily volume, component models, forecasting.

1 Introduction

Aim of this paper is to propose a novel dynamic component model for high-frequency trading volumes and assess its effectiveness for trading by means of an out-of-sample forecasting exercise. Volumes are indeed a crucial ingredient for the implementation of volume-weighted average price (VWAP) trading strategies.

Antonio Naimoli
Università di Salerno, Dipartimento di Scienze Economiche e Statistiche (DISES), Via Giovannni Paolo II, 132, 84084 Fisciano (SA), Italy. e-mail: anaimoli@unisa.it

Giuseppe Storti
Università di Salerno, Dipartimento di Scienze Economiche e Statistiche (DISES), Via Giovannni Paolo II, 132, 84084 Fisciano (SA), Italy. e-mail: storti@unisa.it

VWAP is one of the most common benchmarks used by institutional investors for judging the execution quality of individual stocks. The VWAP of a stock over a particular time horizon (usually one day) is simply given by the total traded value divided by the total traded volume during that period, i.e. the price of each transaction is weighted by the corresponding traded volume. The aim of using a VWAP trading target is to minimize the price impact of a given order by slicing it into smaller transaction sizes, reducing, in this way, the difference between expected price of a trade and its actual traded price. Investors, spreading the timing of transactions throughout the day, seek to achieve an average execution price as close as possible to the VWAP in order to lower market impact costs. Therefore, in this context, the key for a good strategy relies on accurate predictions of intra-daily volumes, since prices are substantially unpredictable.

The proposed specification, called the Heterogeneous MIDAS Component Multiplicative Error Model (H-MIDAS-CMEM), falls within the class of component MEM models as discussed in Brownlees et al. (2011). The most notable differences with respect to the latter are in the specification of the long-run component that is now modelled as an additive cascade of MIDAS filters moving at different frequencies from which the *heterogeneous* quality of the model comes. This specification is motivated by the empirical regularities arising from the analysis of high-frequency time series of trading volumes. After accounting for intra-day seasonality, treated employing a Fourier Flexible Form, these are typically characterised by two prominent and related features: a slowly moving long-run level and a highly persistent autocorrelation structure. In our model, these features are accounted by the heterogeneous MIDAS specification of the long-run component. Residual short term autocorrelation is then explained by an intra-daily non-periodic component that follows a unit mean reverting GARCH-type process. In addition, from an economic point of view, the cascade structure of the long-run component reproduces the natural heterogeneity of financial markets characterised by different categories of agents operating in the market at different frequencies. This results in a variety of sources separately affecting the variation of the average volume at different speeds. On a statistical ground, the cascade structure has the advantage of increasing model's flexibility since it allows to separately parametrize the dynamic contribution of each of these sources.

The estimation of model parameters is performed by the method of maximum likelihood under the assumption that the innovations are distributed according to the Zero-Augmented Generalized F (ZAF) distribution by Hautsch et al. (2014). The reason for this choice is twofold. First, it delivers a flexible probabilistic model for the conditional distribution of volumes. Second, it allows to control for the relevant proportion of zeros present in our data. In order to assess the relative merits of the proposed approach we have performed a forecasting exercise considering high-frequency trading volume for three stocks traded on the Xetra Market in the German Stock Exchange. The stocks have been selected to reflect different liquidity conditions as measured in terms of the number of non trading intra-daily intervals.

Our results show that the H-MIDAS-CMEM model is able to explain the the salient empirical features of the dynamics of high-frequency volumes. Also, we find

that the H-MIDAS-CMEM is able to outperform its main competitors in terms of the usual Mean Squared Error and of the Slicing loss function proposed by Brownlees et al. (2011). Assessing the significance of differences in the predictive performance of models by the Model Confidence Set (MCS) of Hansen et al. (2011), it turns out that the H-MIDAS-CMEM is the only model always included in the set of superior models at different confidence levels.

In the reminder of the paper section 2 describes the proposed H-MIDAS-CMEM model defining its components as intra-daily periodic (subsection 2.1), intra-daily dynamic non-periodic (subsection 2.2) and long-run (subsection 2.3), respectively, while section 3 is dedicated to the out-of-sample forecasting exercise.

2 Model formulation

Let $\{x_{t,i}\}$ be a time series of intra-daily trading volumes. We denote days by $t \in \{1,\dots,T\}$, where each day is divided into I equally spaced intervals indexed by $i \in \{1,\dots,I\}$, then the total number of observations is given by $N = T \times I$.

The empirical regularities of high persistence and clustering of trading activity characterising intra-daily volumes lead us to build a Multiplicative Error Model consisting of multiple components that move at different frequencies. Extending the logic of the Component Multiplicative Error Model (CMEM) by Brownlees et al. (2011) and MIDAS regression models, we propose the H-MIDAS-CMEM which is formulated as

$$x_{t,i} = \tau_t \, g_{t,i} \, \phi_i \, \varepsilon_{t,i}. \tag{1}$$

The multiplicative innovation term $\varepsilon_{t,i}$ is assumed to be conditionally i.i.d., non-negative and to have unit mean and constant variance σ^2, i.e. $\varepsilon_{t,i}|\mathscr{F}_{t,i-1} \sim \mathscr{D}^+(1,\sigma^2)$, where $\mathscr{F}_{t,i-1}$ is the sigma-field generated by the available information until interval $i-1$ of day t. Then, the expectation of $x_{t,i}$, given the information set $\mathscr{F}_{t,i-1}$, is the product of three components characterised by a different dynamic structure. In particular, ϕ_i is an intra-daily periodic component parametrized by a Fourier Flexible Form, which reproduces the approximately U-shaped intra-daily seasonal pattern typically characterising trading activity. The $g_{t,i}$ component represents an intra-daily dynamic non-periodic component, based on a unit mean reverting GARCH-type process, that reproduces autocorrelated and persistent movements around the current long-run level. Finally, τ_t is a lower frequency component given by the sum of MIDAS filters moving at different frequencies. This component is designed to track the dynamics of the long-run level of trading volumes. Furthermore, the use of a time-varying intercept allows to reproduce sudden switches from very low to high trading intensity periods that typically occur in time series of high-frequency trading volumes. The structure of these components is described in more detail in the remainder of this section.

2.1 Intra-daily periodic component

Intra-daily volumes usually exhibit a U-shaped daily seasonal pattern, i.e. the trading activity is higher at the beginning and at the end of the day than around lunch time. To account for the periodic intraday factor we divide volumes $x_{t,i}$ by a seasonal component ϕ_i that is specified via a Fourier Flexible Form as proposed by Gallant (1981)

$$\phi_i = \sum_{q=0}^{Q} a_{0,q}\, \iota^q + \sum_{p=1}^{P} \left[a_{c,p}\, cos(2\pi p\, \iota) + a_{s,p}\, sin(2\pi p\, \iota)\right] \tag{2}$$

where $\iota = i/I \in (0,1]$ is a normalized intraday time trend.

Andersen et al. (2000) suggest that the Fourier terms in (2) do not add any significant information for $Q > 2$ and $P > 6$, so the model precision by using $Q = 2$ and $P = 6$ is enough to capture the behaviour of the intra-day periodicities.[1] Thus, assuming a multiplicative impact of intra-day periodicity effects, diurnally adjusted trading volumes are computed as

$$y_{t,i} = x_{t,i}/\phi_i. \tag{3}$$

2.2 Intra-daily dynamic non-periodic component

The intra-daily non-periodic component, unlike the seasonal component, takes distinctive and non-regular dynamics. In order to make the model identifiable, as in Engle et al. (2013), the intra-daily dynamic component follows a unit mean reverting GARCH-type process, namely $g_{t,i}$ has unconditional expectation equal to 1.

Then, the short-run component, in its simplest form, is formulated as

$$g_{t,i} = \omega^* + \alpha_1 \frac{y_{t,i-1}}{\tau_t} + \alpha_0 I(y_{t,i-1} = 0) + \beta_1 g_{t,i-1}, \tag{4}$$

where $\omega^* = (1 - \alpha_1 - (1-\pi)\alpha_0 - \beta_1)$, π is the probability that $y_{t,i} > 0$ and $I(y_{t,i-1} = 0)$ denotes an indicator function which is equal to 1 if the argument is true and to 0 otherwise.

2.3 The low frequency component

The low frequency component is modelled as a linear combination of MIDAS filters of past volumes aggregated at different frequencies. In this framework, a relevant issue is related to the identification of the frequency of the information to be used

[1] This result is confirmed by computing the Bayesian Information Criterion (BIC) for the estimation of P and Q lags.

by the filters, that notoriously acts a smoothing parameter. Therefore, using trading volumes moving at daily and hourly frequencies, the trend component τ_t is defined as

$$\begin{aligned} log\,\tau_t = m + \theta_d \sum_{k=1}^{K_d} \varphi_k(\omega_{1,d}, \omega_{2,d}) YD_{t-k} \\ + \theta_h \sum_{k=1}^{K_d} \sum_{j=1}^{H} \varphi_{[j+(k-1)H]}(\omega_{1,h}, \omega_{2,h}) YH_{t-k}^{(H-j+1)}, \end{aligned} \tag{5}$$

where $YD_t = \sum_{i=1}^{I} y_{t,i}$, denotes the daily cumulative volume, with the subscript d referring to the daily frequency parameters. The subscript h is related to the parameters corresponding to the *hourly* frequency. If we let $t/H \in \{1,\ldots,H \times T\}$ denote the hourly frequency, with H being the number of intervals in which the day is divided, the variable $YH_t^{(j)}$ corresponds to the (j)-th hourly cumulative volume of the day t, that is $YH_t^{(j)} = \sum_{i=I\frac{(j-1)}{H}+1}^{I\frac{j}{H}} y_{t,i}$, for $j = 1,\ldots,H$. This multiple frequency specification is compatible with the heterogeneous market assumption of Müller et al. (1993), enforcing the idea that market agents can be divided in different groups characterised by different interests and strategies. Also, as pointed out in Corsi (2009), an additive cascade of linear filters moving at different frequencies allows to reproduce very persistent dynamics such as those typically observed for high-frequency trading volumes.

A common choice for determining $\varphi_k(\omega)$ is the Beta weighting scheme

$$\varphi_k(\omega) = \frac{(k/K)^{\omega_1 - 1}(1-k/K)^{\omega_2 - 1}}{\sum_{j=1}^{K}(j/K)^{\omega_1 - 1}(1-j/K)^{\omega_2 - 1}}. \tag{6}$$

As discussed in Ghysels et al. (2007), this Beta-specification is very flexible, being able to accommodate increasing, decreasing or hump-shaped weighting schemes. The Beta lag structure in (6) includes two parameters, but in our empirical applications ω_1 is always set equal to 1 such that the weights are monotonically decreasing over the lags. Furthermore, the number of lags K is properly chosen by information criteria to avoid overfitting problems.

The clustering of the trading activity involves a continuous variation of the average volume level and thus the dynamics of trading volumes are typically characterised by sudden transitions from states of very low trading activity to states of intense trading. In order to account for this switching-state behaviour we further extend the proposed modelling approach introducing a time-varying intercept in the formulation of the long-run component. This is specified as a convex combination of two different unknown parameters m_1 and m_2, that is $m_t = \lambda_t m_1 + (1-\lambda_t) m_2$. The combination weights are time-varying, since they change as a function of observable state-variables. The weight function λ_t follows a logistic specification of the type

$$\lambda_t = \frac{1}{1 + exp(\gamma(\delta - s_{t-1}))}, \qquad (\gamma, \delta) > 0 \tag{7}$$

where γ and δ are unknown coefficients and s_{t-1} is an appropriately chosen state-variable.[2]

3 Out-of-sample forecasting comparison

High-frequency trading volume data used in our analysis refer to the stocks Deutsche Telekom (DTE), GEA Group (G1A) and Salzgitter (SZG) traded on the Xetra Market in the German Stock Exchange. An important feature of the data is the different number of zeros induced by non-trading intervals, since for DTE proportion of zero observations is 0.03%, for G1A 7.046% and for SZG 15.78%.

The raw tick-by-tick data have been filtered employing the procedure proposed by Brownlees and Gallo (2006), only considering regular trading hours from 9:00 am to 5:30 pm. Tick-by-tick data are aggregated computing intra-daily volumes over 10-minutes intervals, which means 51 observations per day. The data have been seasonally adjusted using the Fourier Flexible Form described in equation (2).

To evaluate the predictive ability of the H-MIDAS-CMEM models and their relative merits with respect to competitors, we perform an out-of-sample forecasting comparison over the period January-December 2007, which includes 251 days. In order to capture the salient features of the data and to safeguard against the presence of structural breaks, the model parameters are recursively estimated every day starting from January 2006 with a 1-year rolling window. Therefore at each step we predict 51 intra-daily volumes before re-estimating the models, for a total of 251 days and 12801 intra-daily observations. The out-of-sample performance of the examined models is evaluated by computing some widely used forecasting loss functions. The significance of differences in forecasting performance is assessed by the Model Confidence Set (MCS) approach (Hansen et al., 2011) which relies on a sequence of statistic tests to construct a set of superior models, in terms of predictive ability, at certain confidence level $(1-\alpha)$.

To compare the out-of-sample predictive performances we use the following loss functions

$$L^{MSE} = \sum_{t=1}^{T}\sum_{i=1}^{I}(x_{t,i} - \hat{x}_{t,i})^2$$

$$L^{Slicing} = -\sum_{t=1}^{T}\sum_{i=1}^{I}(w_{t,i}\, log\, \hat{w}_{t,i})$$

where L^{MSE} is the Mean Squared Error (MSE) of the volumes, while $L^{Slicing}$ is the Slicing loss function developed by Brownlees et al. (2011) to evaluate VWAP trading strategies. The slicing weights $\hat{w}_{t,i}$ are computed under both the static and dynamic VWAP replication strategies. The loss functions for single model shown in the top panel of Table 1 point out that the H-MIDAS-CMEM with fixed and, mainly,

[2] A suitable choice for the state-variable is the daily average of intra-daily volumes $\bar{y}_t$.

Table 1: Out-of-sample loss functions comparison

Loss functions average values									
	DTE			G1A			SGZ		
	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}
MEM	0.481	3.920	2.767	1.997	3.921	2.765	0.869	3.921	2.764
CMEM	0.477	3.918	2.766	1.966	3.916	2.762	0.858	3.918	2.762
HAR-MEM	0.476	3.918	2.766	1.963	3.916	2.762	0.857	3.918	2.762
MIDAS-MEM	0.477	3.918	2.766	1.977	3.917	2.762	0.858	3.918	2.762
H-MIDAS-CMEM	0.465	3.915	2.764	1.958	3.909	2.757	0.850	3.912	2.758
H-MIDAS-CMEM-TVI	**0.455**	**3.914**	**2.763**	**1.850**	**3.907**	**2.756**	**0.799**	**3.911**	**2.757**
MCS p-values									
	DTE			G1A			SGZ		
	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}
MEM	0.000	0.000	0.000	0.002	0.000	0.000	0.004	0.000	0.000
CMEM	0.001	0.002	0.000	0.010	0.000	0.002	0.019	0.000	0.000
HAR-MEM	0.001	0.000	0.000	0.014	0.000	0.004	0.025	0.000	0.000
MIDAS-MEM	0.000	0.000	0.000	0.006	0.000	0.002	0.016	0.000	0.000
H-MIDAS-CMEM	0.027	0.059	0.086	0.263	0.232	0.334	0.450	0.378	0.592
H-MIDAS-CMEM-TVI	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Top panel: loss functions values for Mean Squared Error (L^{MSE}) and Slicing Loss with weights computed under the static (L^{SL}_{stc}) and dynamic (L^{SL}_{dyn}) VWAP replication strategy. In **bold** the best model. Bottom panel: MCS p-values for the examined loss functions. In model $\in$ 95% MCS and in model $\in$ 75% MCS.

with time-varying intercept returns the lowest values for both the Mean Squared Error (L^{MSE}) and the Slicing loss using weights computed under the static (L^{SL}_{stc}) and dynamic (L^{SL}_{dyn}) VWAP replication strategy. A lower value of L^{MSE} provides evidence of a greater ability to capture the continuous variation from calms to storms periods, since intra-daily volume series are highly volatile, whereas minimizing the Slicing loss function increases the chances to achieve the VWAP target for a given trading strategy. In order to evaluate if the differences in terms of the considered loss functions are statistically significant, the MCS approach has been used. The MCS results confirm the strength of the H-MIDAS-CMEM, since the model with time-varying intercept is always included into the 75% MCS referring to the set of loss functions employed to measure the predictive ability of the models. For what concerns the H-MIDAS-CMEM with fixed intercept, it falls in the set of the superior models at the 0.75 confidence level for SZG according to the considered loss functions. This also applies to G1A, with the exception of the static Slicing loss entering at the 0.95 level. Finally, for DTE the H-MIDAS-CMEM is out of the MCS

for the L^{MSE}, while falling into the 95% MCS for both the Slicing. Furthermore, the benchmark models never fall into the MCS according to the loss functions and the confidence levels considered.

References

Andersen, T. G., T. Bollerslev, and J. Cai (2000). Intraday and interday volatility in the japanese stock market. *Journal of International Financial Markets, Institutions and Money 10*(2), 107–130.

Brownlees, C. T., F. Cipollini, and G. M. Gallo (2011). Intra-daily volume modeling and prediction for algorithmic trading. *Journal of Financial Econometrics 9*(3), 489–518.

Brownlees, C. T. and G. M. Gallo (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis 51*(4), 2232–2245.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 174–196.

Engle, R. F., E. Ghysels, and B. Sohn (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics 95*(3), 776–797.

Gallant, A. R. (1981). On the bias in flexible functional forms and an essentially unbiased form: the fourier flexible form. *Journal of Econometrics 15*(2), 211–245.

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

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Hautsch, N., P. Malec, and M. Schienle (2014). Capturing the zero: A new class of zero-augmented distributions and multiplicative error processes. *Journal of Financial Econometrics 12*(1), 89–121.

Müller, U. A., M. M. Dacorogna, R. D. Davé, O. V. Pictet, R. B. Olsen, and J. R. Ward (1993). Fractals and intrinsic time: A challenge to econometricians. *Unpublished manuscript, Olsen & Associates, Zürich*.