Econometric techniques for forecasting financial time series in discrete time.

PhD Thesis in Financial Mathematics, XXX Cycle

Luca Cattivelli

Supervisors: Giampiero M. Gallo, Davide Pirino

April 27, 2019

Contents

Acknowledgments

First I want to thank my advisors Giampiero M. Gallo e Davide Pirino for their ideas, suggestions and useful observations. Their comments have greatly improved the content of the thesis. I wish to express my deepest gratitude for their guidance. Moreover I would like to thank my professors and colleagues of Scuola Normale Superiore, for helpful advice and comments.

Last but not least, I am very grateful to Francesca, Valentino, my family and friends for all their support, patience and love.

Introduction

This thesis is a collection of three essays on financial econometrics with a common background in ultra-high frequency modeling of market activity.

In the first essay, we propose an accurate and fast-to-estimate forecasting model for discrete valued time series with long memory and seasonality.[1] The modelling is achieved with an autoregressive conditional Poisson process that features seasonality and heterogeneous autoregressive components (whence the acronym SHARP: Seasonal Heterogeneous AutoRegressive Poisson). Motivated by the prominent role of the bid-ask spread as a transaction cost for trading, we apply the SHARP model to forecast the bid-ask spreads of a large sample of NYSE equity stocks. Indeed, the possibility of having a good forecasting model is of great importance for many applications, in particular for algorithms of optimal execution of orders. We also present a mixed-data sampling extension of the model, called MIDAS-SHARP, which adopts more efficiently the historical information flow and provides empirically the best (among all the models considered) forecasting performances for the bid-ask spreads of NYSE equity stocks. We conclude with an application by showing how bid-ask spread forecasts can be exploited to reduce the total cost incurred by a trader that is willing to buy (or sell) a given amount of an equity stock.

In the second essay, we propose the adoption of Adaptive Lasso techniques for variable selection in vector Multiplicative Error Models (vMEM).[2] These models are designed to analyse the dynamic interactions of several non-negative-valued financial time

[1] This is a joint work with Davide Pirino.
[2] This is a joint work with Giampiero M. Gallo.

series (e.g. realized volatilities, daily ranges, bid-ask spreads, order-book depths, transaction volumes, durations, etc.). The interdependences are captured with a multiplicative structure, modelling the N dimensional vector of interest as a product of its conditional expectation (which depends on the past values of all the N series) times an i.i.d. random vector with unit mean. When the dimension N of the vector of interest is large, the number of parameters of the models grows rapidly[3] often including zero parameters in the data generating process which leads to inefficient parameter estimates and poor forecasting performances if they are not correctly excluded from the model with a variable selection procedure.

In this context, we suggest to select the model and estimate the parameters with a shrinkage method based on an Adaptive Lasso approach. Following the results of Zou [2006], we prove that the Adaptive Lasso for vMEM provides the oracle property, that is, asymptotic consistency in variable selection and the same efficiency as if the set of true predictors were known in advance. With a Monte Carlo simulation exercise, we analyse the finite sample properties of the Adaptive Lasso estimator demonstrating the good performances of this estimator and proving that it reduces the MSE of the estimates compared to other benchmark approaches.

Finally, in the third essay we show empirically the effectiveness of the Adaptive Lasso approach for vMEMs in the study of the network of volatility spillovers among European financial indices, showing the presence of a time-varying interdependence across European financial markets with notable differences during and after the sovereign debt crisis. We are able to demonstrate the superior volatility forecast ability of Adaptive Lasso techniques in a context where a common trend is removed prior to multivariate volatility spillover analysis.

[3]The number of parameters increases quadratically in N.

Chapter 1

A SHARP model of bid-ask spread forecasts[1]

1.1 Introduction

It is widely accepted that, at moderate sampling frequencies, the dynamics of financial prices is well described by the efficient Ito-semimartingale model. Nevertheless, at high frequencies (e.g., at one minute or more), price paths move away from this assumption: the discrete nature of prices clearly arises and the modelling of the price variations with continuous-valued processes (as implied by the Ito-semimartingale assumption) can lead to a severe misspecification of the model. The same reasoning can be straightforwardly applied to the modelling of the bid-ask spread dynamics, as well as several other important financial variables, such as the number of transactions in a given time interval or the number of traded shares per transaction, all of which are discrete-valued stochastic processes. The discreteness is not the only relevant feature of these variables, indeed they are also typically characterized by a strong intraday pattern and strong persistence (e.g. see Brownlees et al. [2011] for the volume, Groß-Klußmann and Hautsch [2013] for the spread and Andersen and Bollerslev [1997] for the magnitude of the price

[1]The material of this chapter is taken from Cattivelli and Pirino [2019].

variations).

In this chapter, we put our focus on bid-ask spreads of equity stocks. A forecasting model for the bid-ask spread of equity stocks is of interest to the large number of high frequency traders who are currently active in the market. These traders are interested in anticipating transaction costs (e.g. the quoted spread) for the purpose of minimizing execution costs. Scholars share this interest, as witnessed by the numerous contributions on optimal executions of trading orders [See, among others, Almgren and Chriss, 2001, Almgren, 2003, Alfonsi et al., 2010, Predoiu et al., 2011, Gatheral and Schied, 2011]. In this context, the bid-ask spread plays a central role because it constitutes the cost of immediate trading when the trader places a market order. For this reason, traders prefer to use limit orders when the spread is large and market orders when the spread is small [Foucault et al., 2005].

Spread predictions are relevant also in the context of risk measures. One notable example is the liquidity-adjusted intraday Value-at-Risk of Weiß and Supper [2013] in which a joint modelling of bid-ask spreads and log-returns is used for the prediction of three types of liquidity-adjusted intraday VaR's. With respect to standard risk measures, this approach has the advantage of taking into account liquidity risk, which is of great concern to both portfolio managers and investors.

In the past two decades, there have been two important contributions on bid-ask spread forecasting, one by Groß-Klußmann and Hautsch [2013] and another by Taylor [2002]. In particular, Taylor [2002] proposes the adoption of the unrestricted VAR model of Huang and Masulis [1999], while Groß-Klußmann and Hautsch [2013] introduce the Long-Memory Autoregressive Conditional Poisson (LMACP) model. The latter is particularly important in our context because it significantly outperforms bid-ask spread forecasts from AR, ARMA, ARFIMA, ACD and FIACD models [Groß-Klußmann and Hautsch, 2013]. However, it has two serious drawbacks: its unconditional mean is not finite (hence, it is not covariance stationary), and its estimation is not straightforward, given the presence of a fractional differencing operator for generating long memory in the model. In this study we overcome these issues by designing a forecasting model for

time series of bid-ask spreads (of equity stocks) which features discreteness, intraday seasonality and persistence being, simultaneously, parsimonious, accurate and fast-to-estimate.

We lay the foundations of our framework on the Autoregressive Conditional Poisson (ACP) process, introducing both a seasonal component in the model (apt to fit intraday patterns), and the heterogeneous autoregressive (HAR) specification for the dynamics over the intraday pattern. The presence of heterogeneous components generates pseudo long-memory, keeping the estimation procedure fast. This is an advantage over models, such as fractionally integrated models (e.g., the LMACP), which are nontrivial to estimate and not straightforwardly extensible in a multivariate setting. We name this new model SHARP, which is an acronym for Seasonal Heterogeneous Auto-Regressive Poisson.

In our empirical analysis, we demonstrate that the forecasting accuracy of the SHARP outperforms that of the LMACP and that of other simpler models for the dynamics of the bid-ask spread.

The SHARP, like all the discrete-time models, faces the limitation of being implemented on a time grid, which means that only a fraction of the total information generated by the historical time series is exploited to generate the forecasts. In order to overcome this limitation, we propose an extension of the SHARP model in which the information flow generated by the spread between two consecutive points of the time grid is used as an additional source of information, as in the MIDAS approach. We prove that the new extended model (that we call MIDAS-SHARP or mSHARP) shows superior forecasting accuracy with respect to all the other models considered, including the SHARP.

Finally, as an empirical application, we show how spread forecasts provided by the SHARP can be exploited to reduce the total transaction costs of a trading strategy. In particular, we prove that a trader that schedules trades according to the SHARP spread forecasts is capable of significantly (in a statistical sense) reducing execution costs with respect to other benchmark strategies.

This chapter is organized as follows. Section 1.2 provides a description of the main stylized facts for time series of bid-ask spreads, while Section 1.3 briefly reviews our benchmark model, the LMACP. Section 1.4 describes in detail our new model, the SHARP. Two possible estimation procedures (maximum likelihood and ordinary least squares) for the SHARP are presented in Section 1.5, while Section 1.6 is dedicated to the extension called MIDAS-SHARP. Using a comparative exercise, we assess the forecasting performances of the SHARP (and of the MIDAS-SHARP) in Section 1.7. In the final part of Section 1.7, we prove how spread predictions based on the SHARP model can be used to reduce the costs of a trading strategy. Finally, Section 1.8 offers a conclusion of our findings.

1.2 The dataset: stylized facts

We start our discussion with a description of the stylized facts that characterize series of bid-ask spreads of equity stocks. Our dataset contains all quote updates (tick-by-tick) of the 244 most liquid, in terms of total volume in the period 2006-2014, stocks of the NYSE[2]. To this large sample we apply a filtering procedure designed to extract ten representative stocks. First, the stocks are clustered into ten deciles according to their average quoted spread[3] in 2014. Within these ten groups, the dynamics of the bid-ask spread are quite different: quoted spreads are almost always equal to $\$0.01$ for stocks whose average spread is small (first deciles) and get more and more volatile as stocks in the last deciles are considered.

In each of the ten groups, we select the most representative stock by choosing that

[2]Prior to the empirical analysis, ask and bid prices have been corrected for the presence of outliers with the procedure proposed by Brownlees and Gallo [2006].

[3]The average quoted spread (in dollar cents) is equal to

$$\frac{100}{D \cdot J} \sum_{t=1}^{D \cdot J} (\mathrm{A}_t - \mathrm{B}_t),$$

where D is the number of trading days in the considered year, $J = 390$ is the number of intraday observations of the ask price A_t and of the bid price B_t at a one-minute frequency, and t runs over all the minutes without distinguishing between consecutive days.

with the highest median (daily traded) volume (in 2014). This filtering procedure returns the ten tickers (ordered from the first to the last average spread decile) BAC, VZ, GM, DAL, HAL, XOM, VLO, CVX, APC and IBM. Table 1.1 reports some summary statistics of the ten selected stocks computed using all the data of 2014. In particular, we report the average time (indicated as $\langle \Delta t \rangle$) between two consecutive quote updates. Not surprisingly, small tick stocks, such as IBM, show rapid (more precisely less than one second) spread updates, while large tick ones, such as BAC, are characterized by less frequent spread changes[4]. Finally, note that for the mean, the standard deviation and the maximum of the quoted spread reported in Table 1.1, we distinguish between the spread sampled every minute and every five seconds. This distinction is necessary because, in the forecasting exercise, we deal with the two frequencies separately.

We frame all models on an equispaced temporal grid, dividing each day into J equispaced periods. The sample is made of the $D = 244$ trading days of 2014. As anticipated above, we estimate the models and perform the corresponding forecasts using two frequencies: one minute and five seconds. The former case is achieved by choosing $J = 390$,[5] while the latter corresponds to $J = 390 \times 60/5 = 4\,680$. For a given J, we observe A_t and B_t, respectively, the ask and the bid price prevailing at time t, where t is the discrete-time index

$$t \in \mathcal{T} \stackrel{\text{def}}{=} \{1, ..., J, J+1, ..., 2\,J, ..., D \cdot J\}\,, \tag{1.1}$$

i.e. t runs over all the elements of the time grid $\mathcal{T}$, without distinguishing between consecutive days. We define the bid-ask spread S_t as

$$S_t \stackrel{\text{def}}{=} 100\,(\mathrm{A}_t - \mathrm{B}_t) - 1.$$

Hence S_t is defined as the number of price intervals of size $\$0.01$ between the ask price

[4]Large tick stocks are defined as stocks for which the quoted bid-ask spread is almost always equal to one dollar cent, while small tick stocks are characterized by spreads of few ticks [Eisler et al., 2012, Dayri and Rosenbaum, 2015].

[5]The NYSE trading day starts at 9:30 AM and closes at 4:00 PM, corresponding to 6 hours and 30 minutes of trading, thus $J = 6 \times 60 + 30 = 390$.

Summary Statistics

	BAC	VZ	GM	DAL	HAL	XOM	VLO	CVX	APC	IBM
Median volume ($\times 10^6$)	14.64	3.36	3.08	2.16	1.84	2.74	1.30	1.60	0.95	0.88
Mean price	16.34	48.61	34.58	37.57	58.69	97.30	51.30	120.54	94.77	182.32
$\langle \#\text{ transactions} \rangle$ ($\times 10^3$)	6.10	6.29	5.93	6.28	8.03	8.76	4.73	7.04	5.09	4.22
$\sqrt{RV \times 252} \times 100$	16.47	12.81	20.30	29.14	25.78	14.07	26.35	14.18	26.21	11.82
$\langle \Delta t \rangle$	12.99	2.88	2.73	1.16	0.95	0.74	0.96	0.58	0.82	0.91
	Quoted spread (in cents) at one-minute frequency									
Mean	1.01	1.11	1.13	1.34	1.68	1.72	2.23	2.51	4.57	6.53
Standard Deviation	0.11	0.37	0.38	0.63	1.01	1.19	1.43	1.85	3.58	4.66
Maximum	3	17	10	29	28	30	28	42	76	94
	Quoted spread (in cents) at five-second frequency									
Mean	1.01	1.11	1.12	1.31	1.67	1.71	2.12	2.50	4.54	6.47
Standard Deviation	0.11	0.36	0.36	0.61	0.97	1.15	1.39	1.78	3.48	4.58
Maximum	3	17	10	29	29	40	35	72	76	99

Table 1.1: This table reports, in order from the first to the last row, the median daily volume (in number of shares), the average daily closing price, the average number of daily transactions (indicated with $\langle \#\text{ transactions} \rangle$), the average annualized (five-minute) realized volatility, and the average time in seconds (denoted by $\langle \Delta t \rangle$) for a change in the spread. Finally, the last six rows report, in order, the average, the standard deviation and the maximum of the quoted bid-ask spread (in dollar cents) at one-minute and at five-second frequency.

A_t and the bid price B_t, minus one. When the ask and the bid are separated by the smallest possible distance, that is $\$0.01$, the variable S_t is zero. We choose to work with this quantity, instead of the quoted spread $(\mathrm{A}_t - \mathrm{B}_t)$, for modelling purposes. In fact, in Section 1.4, we model the conditional distribution of the spread with a Poisson distribution, which has support on the set of non-negative integer numbers, which includes the number zero as an element.

The high frequency dynamics of S_t is characterized by three main stylized facts, which are summarized in Figure 1.1 for the case of the ticker XOM, although they are shared by the ten assets in the sample. First, as shown in the top-left panel of Figure 1.1, S_t is an integer-valued process. Second, it shows a pronounced intraday seasonality: the top-right plot in Figure 1.1 reports, for the case of the five-second grid (i.e. $J = 4\,680$)

and as a function of the periodic intraday index $j_t = t - \left\lfloor \frac{t}{J} \right\rfloor \cdot J \in \{1, ..., J\}$, the sample mean

$$\widehat{\varphi}_{j_t} = \frac{1}{D} \sum_{d=0}^{D-1} S_{j_t + d\,J}, \tag{1.2}$$

which, as we will discuss below, is also our estimator for the unconditional expected value of the spread in the j_t-th time instant of the day (i.e. the intraday seasonal pattern). The seasonality is also evident in the periodicity of the autocorrelation of S_t (bottom-left of Figure 1.1). Finally, the bottom-right plot in Figure 1.1 reports the autocorrelation function for the de-seasonalized time series $\widetilde{S}_t = S_t / \widehat{\varphi}_{j_t}$ which clearly shows a strong persistence.

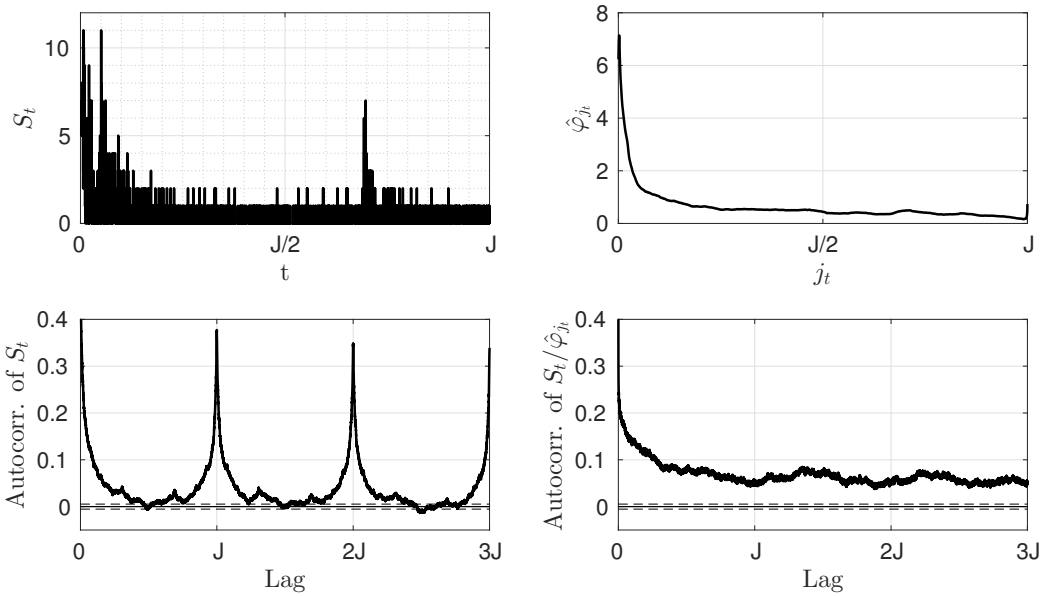

Figure 1.1: Empirical regularities of the five-second bid-ask spread for XOM. Four graphs are reported, using the first thirty trading days of the year, from January 2, 2014, to February 13, 2014: 1) (top-left) the spread series on January 8, 2014; note that it is an integer-valued stochastic process; 2) (top-right) the intraday pattern estimated as $\widehat{\varphi}_{j_t} = \frac{1}{D} \sum_{d=0}^{D-1} S_{j_t + d\,J}$ and then smoothed with a moving average filter with span equal to 201; 3) (bottom-left) the autocorrelation function of S_t; 4) (bottom-right) the autocorrelation function of the de-seasonalized time series $\widetilde{S}_t = S_t / \widehat{\varphi}_{j_t}$. Dashed lines represent 95% and 5% confidence intervals.

1.3 The Long Memory ACP (LMACP)

In our analysis, we put the focus on observation-driven models for integer-valued time series (e.g., the ACP model) that, in contrast to parameter-driven models [e.g. Zeger, 1988, Harvey and Fernandes, 1989, MacDonald and Zucchini, 1997, McKenzie, 2003], are easy to estimate even in a high-frequency setting, where the number of observations is large. In particular, in this section we will briefly describe the main features of the ACP and of its long memory version, the LMACP. Both models will be used as benchmarks to be compared with the SHARP (and the MIDAS-SHARP) in the empirical out-of-sample exercise of Section 1.7.

The Autoregressive Conditional Poisson (ACP) model. Rydberg and Shephard [2000] introduced the Autoregressive Conditional Poisson (ACP) model for modeling the high frequency dynamics of the number of trades. Discreteness is achieved through the use of a Poisson distribution for the dependent random variable S_t, whose dynamics is written as

$$\begin{aligned} \mathbb{P}_{t-1}\left[S_t = k\right] &= \lambda_t^k \, \tfrac{e^{-\lambda_t}}{k!}, \quad k \in \mathbb{N}_0, \\ \lambda_t &= c + \alpha(B) S_t + \beta(B)\lambda_t, \end{aligned} \tag{1.3}$$

where $\mathbb{P}_t\left[\cdot\right]$ is the $\mathcal{F}_t$-conditional probability[6] and $\mathbb{N}_0 = \{0, 1, 2, ...\}$ is the set of non-negative integer numbers. The parameter $c \in \mathbb{R}$ is a constant and

$$\alpha(B) = \sum_{q=1}^{a} \alpha_q B^q \quad \text{and} \quad \beta(B) = \sum_{q=1}^{b} \beta_q B^q \tag{1.4}$$

are polynomials of the backshift operator B, with $\alpha_q > 0$ for $q = 1, ..., a$ and $\beta_q > 0$ for $q = 1, ..., b$. The model is stationary if $\sum_{q=1}^{a} \alpha_q + \sum_{q=1}^{b} \beta_q < 1$. The conditions for geometric ergodicity have been studied in Fokianos et al. [2009], while the properties of the maximum likelihood estimator of the parameters α_q, $q = 1, ..., a$, and β_q, $q = 1, ..., b$, and the analytical expression of the score vector and of the Hessian of the log-likelihood are discussed in Ferland et al. [2006]. As a final remark, note that the ACP process has

[6] $\mathcal{F}_t$ is the natural filtration of the process S_t.

been studied in the literature also with different names: INGARCH in Ferland et al. [2006] and CBIN in Davis et al. [2001].

The LMACP. Groß-Klußmann and Hautsch [2013] proposed a long-memory version of the ACP model defined in equation (1.3). Persistence is introduced with a fractional differencing operator $(1-B)^d$, with $0 < d < 1$, which is a polynomial defined in terms of the hypergeometric function F

$$(1-B)^d = F(-d, 1, 1; B) = \sum_{g=0}^{\infty} \frac{\Gamma(g-d)}{\Gamma(-d)\Gamma(g+1)} B^g. \tag{1.5}$$

This operator enters in the evolution of instantaneous intensity λ_t as follows[7]:

$$(1-\phi(B))(1-B)^d S_t = \omega + (1-\beta(B))v_t, \tag{1.6}$$

where $\omega \in \mathbb{R}_0^+$, $v_t = S_t - \lambda_t$, $\phi(B) = \alpha(B) + \beta(B)$ and $\alpha(B)$ and $\beta(B)$ are defined in equation (1.4). For simplicity, Groß-Klußmann and Hautsch [2013] set $p = q = 1$. The LMACP can be re-arranged to obtain an explicit modelling for the intensity

$$\begin{aligned} (1-\phi(B))(1-B)^d S_t &= \omega + (1-\beta(B))S_t - (1-\beta(B))\lambda_t \\ (1-\beta(B))\lambda_t &= \omega - (1-\phi(B))(1-B)^d S_t + (1-\beta(B))S_t. \end{aligned} \tag{1.7}$$

[7] Actually, there are two ways in which the fractional operator can enter in defining the evolution of λ_t, nevertheless Groß-Klußmann and Hautsch [2013] found that the specification in (1.6) performs better. The other possible specification reads as

$$(1-\phi(B))(1-B)^d(S_t - \omega) = (1-\beta(B))v_t.$$

Hence

$$\begin{aligned}
\lambda_t &= \omega + \beta\lambda_{t-1} - (1-\phi(B))(1-B)^d S_t + S_t - \beta S_{t-1} \\
&= \omega + \beta\lambda_{t-1} + S_t - \beta S_{t-1} - (1-\phi(B)) \sum_{g=0}^{\infty} \frac{\Gamma(g-d)}{\Gamma(-d)\Gamma(g+1)} S_{t-g} \\
&= \omega + \beta\lambda_{t-1} + S_t - \beta S_{t-1} - \sum_{g=0}^{\infty} \frac{\Gamma(g-d)}{\Gamma(-d)\Gamma(g+1)} (S_{t-g} - \phi S_{t-g-1}) \\
&= \omega + \beta\lambda_{t-1} + \alpha S_{t-1} - \sum_{g=1}^{\infty} \frac{\Gamma(g-d)}{\Gamma(-d)\Gamma(g+1)} (S_{t-g} - \phi S_{t-g-1}),
\end{aligned} \tag{1.8}$$

and finally

$$\lambda_t = \omega + \beta\lambda_{t-1} + \alpha S_{t-1} - \sum_{g=1}^{\infty} \frac{\Gamma(g-d)}{\Gamma(-d)\Gamma(g+1)} (S_{t-g} - \phi S_{t-g-1}). \tag{1.9}$$

In order to account for both conditional as well as unconditional over-dispersion and under-dispersion, Groß-Klußmann and Hautsch [2013] propose the adoption of the double Poisson distribution $\mathcal{P}(S_t = s \,|\, \lambda_t; \gamma)$ of Efron [1986], defined as

$$\mathcal{P}(S_t = s \,|\, \lambda_t; \gamma) = c(\gamma, \lambda_t)\sqrt{\gamma}\, e^{-\gamma\lambda_t} \left(\frac{e^{-s} s^s}{s!}\right) \left(\frac{e\lambda_t}{s}\right)^{\gamma s},$$

where γ is an additional model parameter apt to capture conditional over ($\gamma < 1$) or under ($\gamma > 1$) dispersion. The normalizing factor $c(\gamma, \lambda_t)$ is usually approximated by

$$c(\gamma, \lambda_t) \approx \left(1 + \frac{1-\gamma}{12\lambda_t\gamma}\left(1 + \frac{1}{\lambda_t\gamma}\right)\right)^{-1}. \tag{1.10}$$

A seasonal component s_{j_t} modeled with the Fourier expansion

$$s_{j_t} = \delta^s \frac{j_t}{J} + \sum_{l=1}^{L} \left(\delta^s_{1,l} \cos\left(\frac{j_t}{J} 2\pi l\right) + \delta^s_{2,l} \sin\left(\frac{j_t}{J} 2\pi l\right)\right),$$

with $j_t = t - \lfloor t/J \rfloor \cdot J$, is included in the dynamics with an exponential link function, obtaining the following final specification of the LMACP:

$$\begin{aligned}
\mathbb{P}_{t-1}\left[S_t = k\right] &= c(\gamma, \lambda_t')\sqrt{\gamma}\, e^{-\gamma\lambda_t'}\left(\tfrac{e^{-k}k^k}{k!}\right)\left(\tfrac{\lambda_t' e}{k}\right)^{\gamma k}, \quad k = 0, 1, 2, \ldots, \\
\lambda_t' &= \lambda_t \exp(s_{j_t}), \\
s_{j_t} &= \delta^s \tfrac{j_t}{J} + \textstyle\sum_{l=1}^{L}\left(\delta_{1,l}^s \cos\left(\tfrac{j_t}{J} 2\pi l\right) + \delta_{2,l}^s \sin\left(\tfrac{j_t}{J} 2\pi l\right)\right), \\
\lambda_t &= \omega + (\phi - \beta) S_{t-1} + \beta\lambda_{t-1} - \textstyle\sum_{g=1}^{\infty} \tfrac{\Gamma(g-d)}{\Gamma(-d)\Gamma(g+1)}(S_{t-g} - \phi S_{t-g-1}).
\end{aligned}$$

The parameters of the model can be estimated by maximizing the log-likelihood

$$\begin{aligned}
\log \mathcal{L}(\boldsymbol{\vartheta}) &= \sum_{t=1}^{T}\Bigg\{ \log[c(\gamma, \lambda_t \exp(s_{j_t}))] + \frac{1}{2}\log(\gamma) - \gamma\lambda_t \exp(s_{j_t}) + \\
&\quad + \gamma S_t\left[1 + \log\left(\frac{\lambda_t \exp(s_{j_t})}{S_t}\right)\right]\Bigg\}, \qquad (1.11)
\end{aligned}$$

where

$$\boldsymbol{\vartheta} = (\delta^s,\, \delta_{1,1}^s,\, ..., \delta_{1,L}^s,\, \delta_{2,1}^s,\, ..., \delta_{2,L}^s,\, \omega,\, \phi,\, \beta,\, d,\, \gamma)$$

indicates the vector of model parameters. To increase the efficiency of the estimation, the parameters of the seasonal pattern (δ^s, $\delta_{i,j}^s$) can be jointly estimated (via maximum likelihood) with the parameters of the dynamics (ω, ϕ, β and d). Groß-Klußmann and Hautsch [2013] set $L = 2$ and select a truncation point of 250 observations in g, motivating that this is sufficient to obtain reliable estimates of the parameters.

The LMACP model is our benchmark since, to the best of our knowledge, it is the only model available in the past academic literature that is capable to capture the salient features of the bid-ask spread: strong autocorrelation, discreteness of observations and intraday seasonality. Moreover, Groß-Klußmann and Hautsch [2013] showed that the LMACP model generates bid-ask spread forecasts significantly more accurate than those generated by AR, ARMA, ARFIMA, ACD and FIACD models. They further proved that the spread predictions obtained with the LMACP model allow traders to reduce spread transaction costs up to 14%.

This said, the LMACP is a purely (i.e., in a strict mathematical sense) long-memory model and, hence, its estimation turns out to be cumbersome for several reasons. First, it is complicated by the fact that the log-likelihood of the model is not concave with multiple local maxima. Hence, any optimisation procedure is slowed down by the time required to verify whether the identified solution is global or not. In fact, as documented in Section 1.7 and Appendix 1.C, the average time required to estimate the LMACP is significantly larger than that of the other models considered. In particular, the estimation of the LMACP is, on average, 100 times slower than the estimation of the SHARP. This is an unwanted feature, especially in high-frequency applications such as bid-ask spread forecasting, where rapid updates of the model parameters are often required. Second, under the LMACP, the unconditional expected value of the state-variable S_t is infinite, that is $\mathbb{E}[S_t] = \infty$, hence the LMACP is not covariance-stationary. We will discuss in Section 1.7 the implications of this unfavorable feature.

In the SHARP we provide a valid and simple alternative to circumvent these issues by specifying the dynamics over the intraday pattern with an Heterogeneous Auto Regressive (HAR) structure [as in Corsi, 2009], which guarantees a fast estimation procedure, a concave log-likelihood, strong persistence and finite unconditional mean, while simultaneously improving the forecasting accuracy.

1.4 The Seasonal Heterogeneous Auto-Regressive Poisson model

In this section we formally introduce the SHARP model. Consider then an integer-valued stochastic process $(S_t)_{t\in\mathcal{T}}$ sampled on the discrete-time grid $\mathcal{T}$ defined in (1.1) and let $\mathcal{F}_t$ be its natural filtration. The definition of the SHARP process given below is generic and can be adapted to any sampling frequency, hence we leave unspecified the number J of intraday observations that appear in the definition of $\mathcal{T}$. The following is our formal definition of the SHARP model.

Definition 1. *A discrete-time process $(S_t)_{t\in\mathcal{T}}$ is a SHARP process if*

$$\begin{aligned}
\mathbb{P}_{t-1}\left[S_t = k\right] &= \lambda_t^k \, \tfrac{e^{-\lambda_t}}{k!}, \quad k \in \mathbb{N}_0, \; t \in \mathcal{T}, \\
\lambda_t &= \varphi_{j_t} \, \mu_t, \\
\mu_t &= (1 - \Sigma_\alpha) + \alpha^{(s)} \, \widetilde{S}_{t-1:t-1} + \alpha^{(m)} \, \widetilde{S}_{t-m:t-1} + \alpha^{(\ell)} \, \widetilde{S}_{t-\ell:t-1},
\end{aligned} \tag{1.12}$$

where $\varphi_{j_t} = \varphi_{t-\lfloor t/J \rfloor \cdot J}$ is a periodic and positive deterministic process, $\Sigma_\alpha = \alpha^{(s)} + \alpha^{(m)} + \alpha^{(\ell)} < 1$ with $\alpha^{(s)} > 0$, $\alpha^{(m)} > 0$, $\alpha^{(\ell)} > 0$, and where the averages $\widetilde{S}_{t_1:t_2}$, for generic indexes t_1 and t_2 in $\mathcal{T}$, are defined as

$$\widetilde{S}_{t_1:t_2} \overset{def}{=} \frac{1}{t_2 - t_1 + 1} \sum_{q=t_1}^{t_2} \frac{S_q}{\varphi_{j_q}}. \tag{1.13}$$

Before proceeding with a discussion on the estimation of the model in Definition 1, we describe step-by-step how the set of equations in (1.12) can be obtained from the Autoregressive Conditional Poisson model (ACP) of Heinen [2003] of equation (1.3). The ACP model is suited for integer-valued processes, however it is not designed to capture long-memory and seasonality patterns, as those that are described in Figure 1.1. In what follows, we separately discuss how seasonality patterns and long-memory can be included.

Seasonality. The model in (1.3) can be easily modified to include a seasonal deterministic pattern in the dynamics, more precisely to incorporate an intraday pattern for $\mathbb{E}\left[S_t\right]$. Consider for this purpose a model for the bid-ask spread S_t of the form

$$\begin{aligned}
\mathbb{P}_{t-1}\left[S_t = k\right] &= \lambda_t^k \, \tfrac{e^{-\lambda_t}}{k!}, \quad k = 0, 1, 2, \ldots, \\
\lambda_t &= \varphi_{j_t} \, \mu_t, \\
\mu_t &= c + \alpha(B) \tfrac{S_t}{\varphi_{j_t}} + \beta(B) \mu_t,
\end{aligned} \tag{1.14}$$

where the deterministic (and positive) pattern $\varphi_{j_t} = \varphi_{t-\lfloor t/J \rfloor \cdot J}$ is left unspecified and must be estimated. We will refer to this specification as seasonal ACP or sACP(a, b),

where a and b are the orders of the two polynomials $\alpha(B)$ and $\beta(B)$, respectively. The model described by the system of equations (1.14) features the separation of the intensity λ into a seasonal pattern φ and a de-seasonalized intensity μ. Under the assumption of weak-stationarity[8] for the process

$$\left(S_{j_t+d\,J}\right)_{d\in\{0,1,\ldots,D-1\}} \tag{1.15}$$

for a given $j_t \in \{1, 2, ..., J\}$, we can set the parameter c (in the last of the equations (1.14)) equal to

$$c = 1 - \sum_{q=1}^{a} \alpha_q - \sum_{q=1}^{b} \beta_q, \tag{1.16}$$

in order to satisfy the condition $\mathbb{E}[\mu_t] = 1$, which, in turn, guarantees that[9] $\mathbb{E}\left[S_t\right] = \varphi_{j_t}$, $\forall t$. By imposing (1.16), we ensure that μ describes the deviation of the intensity λ from the seasonal pattern φ. The intraday pattern φ can be estimated in several ways. We will discuss this topic further in Section 1.5.

Long-memory. The hybrid model described by equation (1.14) features an intraday seasonal pattern, but it can not fit the long-memory structure of the time-series S_t, that is the slowly decaying autocorrelation of the de-sesonalised process S_t/φ_{j_t} (see Figure 1.1). The adoption of fractionally integrated processes is a valid possibility for incorporating such a long-memory structure [see Baillie, 1996, Jasiak, 1999, Rossi and Fantazzini, 2015, Bayer et al., 2016]. Nevertheless, empirically, the estimation of these models is often problematic and time consuming. To circumvent this complexity, we take inspiration from the popular HAR model by Corsi [2009] described in Appendix 1.A. The HAR model has the advantage of reproducing slowly-decaying memory patterns, while preserving a simple structure. The definitive specification of the SHARP model is therefore

[8] The parametric restrictions needed to ensure covariance stationarity of the process (1.15) will be discussed in the proof of Theorem 1. This property guarantees that, for any given j_t, the expected value of the process (1.15) does not depend on the day d, but only on j_t, allowing the incorporation of a seasonal intraday pattern for $\mathbb{E}\left[S_t\right]$ with period J.

[9] Straightforward computations, similar to the variance targeting case for GARCH models [Engle and Mezrich, 1995], shows that in order to have $\mathbb{E}\left[S_t\right] = \varphi_{j_t}$, we need the parametric restriction (1.16).

achieved by substituting the last of the equations in (1.14) with

$$\mu_t = c + \alpha^{(s)} \widetilde{S}_{t-1:t-1} + \alpha^{(m)} \widetilde{S}_{t-m:t-1} + \alpha^{(\ell)} \widetilde{S}_{t-\ell:t-1}, \tag{1.17}$$

where the averages $\widetilde{S}_{t-1:t-1}$, $\widetilde{S}_{t-m:t-1}$ and $\widetilde{S}_{t-\ell:t-1}$ are defined in equation (1.13). The parameter c is a constant, m and ℓ are two integers with $m < \ell$ and the parameters $\alpha^{(s)}$, $\alpha^{(m)}$ and $\alpha^{(\ell)}$ correspond to, respectively, the short (s), medium (m) and long (ℓ) autoregressive components for μ_t. Note that, while the short-term component is chosen to coincide with $S_{t-1}/\varphi_{j_{t-1}}$, we allow m and ℓ to be chosen via an optimization procedure. We will discuss this point later.

Finally, in order to guarantee that the HAR structure in equation (1.17) preserves, as for the hybrid model in equations (1.14), the identity $\mathbb{E}[\mu_t] = 1$, the constant parameter c is to be chosen by imposing the constraint (1.16), which gives $c = 1 - \left(\alpha^{(s)} + \alpha^{(m)} + \alpha^{(\ell)}\right)$, whence the specification $\mu_t = (1 - \Sigma_\alpha) + \alpha^{(s)} \widetilde{S}_{t-1:t-1} + \alpha^{(m)} \widetilde{S}_{t-m:t-1} + \alpha^{(\ell)} \widetilde{S}_{t-\ell:t-1}$ with $\Sigma_\alpha = \alpha^{(s)} + \alpha^{(m)} + \alpha^{(\ell)}$ that appears in Definition 1.

1.5 Estimation of the SHARP model

The estimation of the SHARP model poses the problems of selecting the model specification (i.e. choosing the integer parameters m and ℓ) and of finding the seasonal pattern φ, as well as the parameters $\alpha^{(s)}$, $\alpha^{(m)}$ and $\alpha^{(\ell)}$.

Since the intraday pattern can be estimated separately, as formally stated below in Theorem 1, we rely on a two-step estimation procedure. The first step consists of estimating φ non-parametrically, while the second step considers the remaining parameters.

First step: the estimation of the intraday seasonal pattern. Being $(S_t)_{t\in\mathcal{T}}$ and $(S_t/\varphi_{j_t})_{t\in\mathcal{T}}$ series of non-stationary dependent random variables, the asymptotic convergence of $\widehat{\varphi}_{j_t}$ (defined in equation (1.2)) to the intraday seasonal pattern φ_{j_t} is not straightforwardly guaranteed. Nevertheless, the (weak) law of large numbers still holds

if some regularity conditions are met[10], as established in the following theorem.

Theorem 1. *Assume that the process* $(S_t)_{t\in\mathcal{T}}$ *is a SHARP process. Then, for a given* $j_t = t - \left\lfloor \frac{t}{J} \right\rfloor \cdot J \in \{1, ..., J\}$*, the parametric restrictions* $0 < \alpha^{(s)} < 1$*,* $0 < \alpha^{(m)} < 1$*,* $0 < \alpha^{(\ell)} < 1$ *and* $\Sigma_\alpha = \alpha^{(s)} + \alpha^{(m)} + \alpha^{(\ell)} < 1$ *are sufficient to guarantee that the time-series* $(S_{j_t+d\,J})_{d\in\{0,1,...,D-1\}}$ *is covariance stationary*[11] *and*

$$\widehat{\varphi}_{j_t} \overset{def}{=} \frac{1}{D}\sum_{d=0}^{D-1} S_{j_t+d\,J} \xrightarrow{p} \varphi_{j_t} \text{ as } D \to \infty. \tag{1.18}$$

Proof. See Appendix 1.B.

It is worth noting that Theorem 1 does not establish the covariance-stationarity for the SHARP process $(S_t)_{t\in\mathcal{T}}$, but for all the J processes $(S_{j_t+d\,J})_{d\in\{0,1,...,D-1\}}$, with $j_t = 1, ..., J$, which is what is required for the consistency of the estimators $\widehat{\varphi}_{j_t}$.

Let us emphasize nevertheless that, in finite sample, the use of a smoother is recommended to reduce the variance of $\widehat{\varphi}_{j_t}$. Indeed, if the number of days D in equation (1.18) is small, the variance of $\widehat{\varphi}_{j_t}$ could be high. In our empirical applications, we implement a moving average filter with span equal to 41 for the one-minute frequency and equal to 201 in the five-second case. Moreover, if the event $S_t = 0$ is frequent, it is necessary to substitute (1.18) with the robust alternative $\widehat{\varphi}^{(0)}_{j_t} = \max\{\widehat{\varphi}_{j_t}, \varphi^{\min}\}$, with $\varphi^{\min} > 0$, in order to avoid zeros in the intraday pattern[12], which could cause divergences in the averages $\widetilde{S}_{t_1:t_2}$ that appear in equation (1.13).

Second step: the estimation of m**,** ℓ**,** $\alpha^{(s)}$**,** $\alpha^{(m)}$ **and** $\alpha^{(\ell)}$**.** The vector of the parameters $\boldsymbol{\vartheta} = (m, \ell, \boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = (\alpha^{(s)}, \alpha^{(m)}, \alpha^{(\ell)})$ is estimated by maximizing the log-likelihood

[10] See, for example, Proposition 7.5 of Hamilton [1994].

[11] The parametric restrictions described in Theorem 1 are identical to the covariance-stationarity conditions of a standard ACP process. Therefore one could infer that the result in Theorem 1 are derived directly from the properties of the standard ACP process applied to the de-seasonalized process S_t/φ_{j_t}. This is not the case since the latter does not follow an ACP process. This can be easily understood by the fact that S_t/φ_{j_t} is not an integer-valued process and, therefore, its conditional distribution is not Poisson.

[12] In the empirical analysis of Section 1.7, we adopt $\varphi^{\min} = 0.1$, however out-of-sample results are largely independent on the choice of $\varphi^{\min}$.

$\mathcal{L}(\boldsymbol{\vartheta})$ of the SHARP process which is equal to

$$\sum_{t\in\mathcal{T}}\Bigg\{-\log(S_t!)-\widehat{\varphi}_{j_t}\left(1-\Sigma_\alpha+\alpha^{(s)}\,\widetilde{S}_{t-1:t-1}+\alpha^{(m)}\,\widetilde{S}_{t-m:t-1}+\alpha^{(\ell)}\,\widetilde{S}_{t-\ell:t-1}\right)+$$
$$+S_t\,\log\left[\widehat{\varphi}_{j_t}\left(1-\Sigma_\alpha+\alpha^{(s)}\,\widetilde{S}_{t-1:t-1}+\alpha^{(m)}\,\widetilde{S}_{t-m:t-1}+\alpha^{(\ell)}\,\widetilde{S}_{t-\ell:t-1}\right)\right]\Bigg\}. \quad (1.19)$$

When the integers m and ℓ in the log-likelihood (1.19) are considered as free parameters to be estimated, λ_t is no longer an affine function of the parameters. As a consequence, the log-likelihood is not concave in $\boldsymbol{\vartheta}$. The absence of concavity in the log-likelihood and the discreteness of the parameters do not guarantee the consistency of the maximum likelihood estimators. For this reason, in Appendix 1.C we study their finite sample properties. In the same section we also analyze the finite sample properties of our main competitor, the LMACP model. Here we limit to say that the finite sample performance of the ML estimator is satisfactory for sample size comparable with those used in empirical applications, e.g., when dealing with data sampled at one-minute frequency. On the contrary, if one selects a priori the parameters m and ℓ, such as in Corsi [2009], the estimation procedure is highly simplified by the concavity of the log-likelihood (1.19) as a function of $\boldsymbol{\alpha}$ [e.g, see Boyd and Vandenberghe, 2004]. In the empirical analysis presented in Section 1.7, we derive a set of recommended values for m and ℓ for different frequencies of observation. These values can be used by researchers to simplify and improve (by exploiting the concavity of the log-likelihood) the estimation of the SHARP model[13].

Standard errors and model misspecification. Standard errors for the vector of parameters $(\widehat{\boldsymbol{\varphi}},\widehat{\boldsymbol{\alpha}})=\left(\widehat{\varphi}_1,\ldots,\widehat{\varphi}_J,\widehat{\alpha}^{(s)},\widehat{\alpha}^{(m)},\widehat{\alpha}^{(\ell)}\right)$ must be robust to the misspecification of the conditional distribution and must take into account the effect of the two-step estimation procedure. For this reason, we follow the approach of Cipollini et al. [2017], rephrasing

[13] If m and ℓ are selected a-priori, the ML estimator is also robust to the misspecification of the Poisson distribution and can be considered a Quasi-Maximum Likelihood estimator. This can be seen by looking at the score vector, whose component f, with $f\in\{s,m,\ell\}$, $\frac{\partial\mathcal{L}(\boldsymbol{\vartheta})}{\partial\alpha^{(f)}}=\sum_{t\in\mathcal{T}}\frac{\partial\lambda_t}{\partial\alpha^{(f)}}\left(\frac{S_t}{\lambda_t}-1\right)$ has expected value equal to zero as long as the intensity λ_t is correctly specified (that is, if $\mathbb{E}[S_t|\mathcal{F}_{t-1}]=\lambda_t$).

the two-step procedure as a two-step GMM estimator, for which robust standard errors are known [Newey and McFadden, 1994]. In our case, both the sample average $\widehat{\varphi}_{j_t}$ and the maximum likelihood estimator used for the remaining parameters, can be seen as special cases of GMM estimators. By adapting the approach of Cipollini et al. [2017] we get for the asymptotic variance-covariance matrix (aVar) of $\sqrt{T}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ the expression

$$\text{aVar}[\sqrt{T}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})] = \left(H_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1}, H_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1} H_{\boldsymbol{\alpha}\boldsymbol{\varphi}}\right) \begin{pmatrix} \Omega_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & \Omega_{\boldsymbol{\alpha}\boldsymbol{\varphi}} \\ \Omega_{\boldsymbol{\varphi}\boldsymbol{\alpha}} & \Omega_{\boldsymbol{\varphi}\boldsymbol{\varphi}} \end{pmatrix} \left(H_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1}, H_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1} H_{\boldsymbol{\alpha}\boldsymbol{\varphi}}\right)', \tag{1.20}$$

with[14]

$$H_{\boldsymbol{\alpha}\boldsymbol{\alpha}} = \mathbb{E}[\nabla_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^2 l_t], \quad H_{\boldsymbol{\alpha}\boldsymbol{\varphi}} = \mathbb{E}[\nabla_{\boldsymbol{\alpha}\boldsymbol{\varphi}}^2 l_t], \quad \Omega = \text{Var}[(\nabla_{\boldsymbol{\alpha}} l_t', \mathbf{m}_{\lfloor t/J \rfloor}')], \tag{1.21}$$

where l_t is the logarithm of the conditional probability of observing S_t. The vector $\mathbf{m}$ is the moment (vector) function giving $\widehat{\varphi}_{j_t}$ as a GMM estimator of φ_{j_t}, defined as

$$\mathbf{m} = \sum_{d=0}^{D-1} \mathbf{m}_d = \sum_{d=0}^{D-1} \left(S_{dJ+1} - \varphi_1, S_{dJ+2} - \varphi_2, ..., S_{dJ+J} - \varphi_J\right)'.$$

The matrix Ω denotes the variance-covariance matrix of the two moment functions $\nabla_{\boldsymbol{\alpha}} l_t$ and $\mathbf{m}_{\lfloor t/J \rfloor}$. All the matrices that appear on the right hand side of equation (1.20) can be estimated by their corresponding sample counterparts, as shown in Appendix 1.D. Concerning the standard errors of the intraday pattern, the asymptotic variance-covariance matrix of $\sqrt{T}(\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})$ is equal to $\Omega_{\boldsymbol{\varphi}\boldsymbol{\varphi}}$.

OLS estimation of the SHARP. One of the main advantages of the HAR model of Corsi [2009] is that it can be simply estimated via ordinary least squares. Here we prove that the SHARP model shares this feature[15]. First, we rewrite the second of the equations in (1.12) as

$$\widetilde{S}_{t:t} = \mu_t + \varepsilon_t, \tag{1.22}$$

[14]The operators $\nabla_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^2$ and $\nabla_{\boldsymbol{\alpha}\boldsymbol{\varphi}}^2$ appearing in equation (1.21) are defined as $(\nabla_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^2)_{h,k} = \partial^2/(\partial\alpha_h \partial\alpha_k)$ and $(\nabla_{\boldsymbol{\alpha}\boldsymbol{\varphi}}^2)_{h,k} = \partial^2/(\partial\alpha_h \partial\varphi_k)$.

[15]We thank an anonymous referee for this suggestion.

where the martingale difference sequence ε_t is defined as $\varepsilon_t \stackrel{\text{def}}{=} (S_t - \lambda_t)/\varphi_{j_t}$. Then, substituting the explicit expression of μ_t from equation (1.12) into equation (1.22), we obtain an HAR-like structure in $\widetilde{S}_{t:t} - 1$, i.e.

$$(\widetilde{S}_{t:t} - 1) = \alpha^{(s)} (\widetilde{S}_{t-1:t-1} - 1) + \alpha^{(m)} (\widetilde{S}_{t-m:t-1} - 1) + \alpha^{(\ell)} (\widetilde{S}_{t-\ell:t-1} - 1) + \varepsilon_t. \quad (1.23)$$

In equation (1.23), the parameters $\boldsymbol{\alpha} = (\alpha^{(s)}, \alpha^{(m)}, \alpha^{(\ell)})'$ can now be estimated via OLS. The error term ε_t is neither Gaussian nor homoscedastic, however, robust standard errors can be calculated through equations (1.20) and (1.21), with $l_t = (\varepsilon_t)^2$, after having estimated the intraday seasonal pattern as in (1.18). We will refer to this specification of the SHARP model as olsSHARP.

1.6 MIDAS-SHARP

The estimation of the SHARP model requires, as input data, spread values sampled on the elements of the equispaced time grid $\mathcal{T}$. As a result, the spread history between any two consecutive instants of the grid $\mathcal{T}$ is ignored, leading to a loss of information. This drawback can be alleviated by exploiting the informational content of the spread observed on some time partition finer than the original partition $\mathcal{T}$, a modelling strategy that derives directly from the mixed-data sampling (MIDAS) approach of Ghysels et al. [2004, 2007], which is introduced in Appendix 1.E.

We consider then the partition $\mathcal{Q}^{(r)}$ that is obtained dividing the time interval between any two consecutive instants of $\mathcal{T}$ into r equispaced sub-intervals, in formula

$$\mathcal{Q}^{(r)} = \left\{ \frac{1}{r}, \frac{2}{r}, \ldots, D \cdot J - \frac{1}{r}, D \cdot J \right\}. \quad (1.24)$$

We indicate with $S_q^{(r)}$, $q \in \mathcal{Q}^{(r)}$, the spread prevailing at time q. For given m and ℓ, multiples of $1/r$, the SHARP model in Definition 1 is extendible into a MIDAS-like framework by specifying a dynamics for the instantaneous intensity μ_t on the coarser partition $\mathcal{T} \subset \mathcal{Q}^{(r)}$. We name this extension MIDAS-SHARP, henceforth shortened to mSHARP,

and we formalize its definition in what follows.

Definition 2. *Let $\mathcal{T}$ and $\mathcal{Q}^{(r)}$ be, respectively, the partitions defined in equations* (1.1) *and* (1.24)*. A discrete-time process $\left(S_t^{(r)}\right)_{t\in\mathcal{T}}$ is a MIDAS-SHARP process if*

$$\begin{aligned}
\mathbb{P}_{t-1}\left[S_t^{(r)}=k\right] &= \lambda_t^k\,\tfrac{e^{-\lambda_t}}{k!}, \quad k=0,1,2,\ldots, \quad t\in\mathcal{T} \\
\lambda_t &= \varphi_{j_t}^{(r)}\,\mu_t, \\
\mu_t &= (1-\Sigma_\alpha)+\alpha^{(s)}\,\widetilde{S}_{t-1:t-1}^{(r)}+\alpha^{(m)}\,\widetilde{S}_{t-m:t-1}^{(r)}+\alpha^{(\ell)}\,\widetilde{S}_{t-\ell:t-1}^{(r)},
\end{aligned} \tag{1.25}$$

where $\Sigma_\alpha=\alpha^{(s)}+\alpha^{(m)}+\alpha^{(\ell)}<1$, with $\alpha^{(s)}>0$, $\alpha^{(m)}>0$, $\alpha^{(\ell)}>0$ and $\varphi_{j_t}^{(r)}=\varphi_{t-\lfloor t/J\rfloor\cdot J}^{(r)}$ is a periodic and positive deterministic process. The averages $\widetilde{S}_{t_1:t_2}^{(r)}$ are defined as

$$\widetilde{S}_{t_1:t_2}^{(r)}=\frac{1}{(t_2-t_1)\,r+1}\sum_{q\in\{t_1,t_1+\frac{1}{r},\ldots,t_2\}}\frac{S_q^{(r)}}{\varphi_{j_q}^{(r)}},$$

and, for a given $q\in\mathcal{Q}^{(r)}\setminus\mathcal{T}$, the time-series $\left(S_{j_q+d\,J}^{(r)}\right)_{d\in\{0,1,\ldots,D-1\}}$ is non-negative and covariance stationary, with

$$\frac{1}{D}\sum_{d=0}^{D-1}S_{j_q+d\,J}^{(r)}\xrightarrow{p}\varphi_{j_q}^{(r)}=\mathbb{E}\left[S_q^{(r)}\right]<\infty. \tag{1.26}$$

There is a key difference between the SHARP and the mSHARP: both models specify the dynamics of the spread on the coarser time grid $\mathcal{T}$, nevertheless the mSHARP requires additional assumptions on the regularity of the process $S_q^{(r)}$ defined on the partition $\mathcal{Q}^{(r)}\setminus\mathcal{T}$. For this process, we assume some regularity conditions in order to guarantee that the results of Theorem 1 still hold. Moreover, the process in Definition 2 can be seen as a simple SHARP with additional covariance-stationary regressors.

It is important to note that the assumption in equation (1.26) has a double implication. First, it implicitly imposes that the unconditional expected value $\mathbb{E}\left[S_q^{(r)}\right]$ is finite and periodic, with period J, in fact

$$\mathbb{E}\left[S_{q+J}^{(r)}\right]=\varphi_{j_{q+J}}^{(r)}=\varphi_{j_q}^{(r)}=\mathbb{E}\left[S_q^{(r)}\right].$$

Second, it is needed to guarantee the consistency of the estimator $\widehat{\varphi}_{j_t}$ in the set $\mathcal{Q}^{(r)} \backslash \mathcal{T}$. In all the empirical applications that follow, the mSHARP is estimated following the same two-step procedure adopted for the SHARP[16]. Consistent standard errors for the vector of parameters $(\boldsymbol{\varphi}^{(r)}, \boldsymbol{\alpha}) = (\varphi^{(r)}_{1/r}, \varphi^{(r)}_{2/r}, ..., \varphi^{(r)}_{J-1/r}, \varphi^{(r)}_{J}, \boldsymbol{\alpha})$ can be estimated as in equation (1.20). Finally, note that, for given m and ℓ, the mSHARP is robust to misspecifications in the Poisson conditional distribution, as it happens for the SHARP.

1.7 Empirical analysis

We compare the forecasting performances of the newly defined SHARP, olsSHARP, and mSHARP models with a set of benchmark alternatives along three dimensions: accuracy [through the standard test by Diebold and Mariano, 1995], goodness-of-fit (through the Ljung-Box test statistics) and average time required to perform the estimation.

The following is the list (with some additional information) of the models involved in the horse-race exercise discussed below.

1. The SHARP model of equation (1.12).
2. The olsSHARP model of equation (1.23) (estimated with OLS).
3. The MIDAS-SHARP model of Section 1.6.
4. The Long Memory ACP (LMACP) model described in Section 1.3[17].
5. The seasonal-adjusted ACP(1,1) of equation (1.14). This model has a simple structure and provides competitive forecasts for time-series that are not particularly persistent.

[16] In more detail, in the first step the intraday pattern of the spreads $S^{(r)}_q$, with $q \in \mathcal{Q}^{(r)}$, is estimated through the estimator defined in (1.26). In the second step, the parameters $\alpha^{(s)}, \alpha^{(m)}, \alpha^{(\ell)}$ are estimated via maximum-likelihood.

[17] We implement the model following the suggestions of Groß-Klußmann and Hautsch [2013]. That is, we set $L = 2$, we approximate the normalizing factor as $1/c(\gamma, \lambda'_t) \approx 1 + (1-\gamma)(1 + 1/\lambda'_t\gamma)/12\lambda'_t\gamma$ and we truncate the infinite sum in g to 250. Other values for L have been taken into consideration, obtaining poorer performances. In fact, with $L > 2$ the number of parameters grows substantially making the estimation of the model unstable (due to the presence of several local minima in the log-likelihood) and time-consuming.

6. A purely seasonal model

$$S_t = \varphi_{j_t} + \varepsilon_t, \tag{1.27}$$

where $\varphi_{j_t} = \varphi_{t-\lfloor t/J \rfloor \cdot J}$ is a deterministic seasonal intraday pattern (estimated non-parametrically as in equation (1.2)) and the ε_t's are iid disturbances with $\mathbb{E}[\varepsilon_t] = 0$, $\text{Var}[\varepsilon_t] = \sigma^2$ (we will refer to this specification simply as "seasonal"). Including this model allows us to test whether the seasonal pattern alone delivers better or equal accurate forecasts than more sophisticated models.

7. A random walk model with seasonal adjustment

$$S_t = \begin{cases} \varphi_1 + \varepsilon_t, & \text{if } j_t = 1; \\ S_{t-1} + \varepsilon_t, & \text{otherwise;} \end{cases}$$

where the ε_t's are iid errors as in the previous model of the list and where the seasonal adjustment has been added to avoid a penalization of the forecast in the first instant of the day. This model is included in the empirical comparison because, for strongly persistent time series, S_{t-1} is informative for S_t. This model is indicated with the acronym RW.

As mentioned above, in the models 1, 2, and 3 of the previous list, the integer parameters m and ℓ are always selected from the recommended values of Table 1.2. These values are computed as averages of the estimated (using the log-likelihood (1.19)) m and ℓ from the dataset of ten stocks described in Section 1.2, using the data relative to 2013.

As an illustrative example, we report in Figure 1.2 the time series of the estimated parameters of the mSHARP model, at one-minute frequency for IBM with a rolling window of ten days. The parameter r, which defines the number of sub-intervals used in the mSHARP, is set to $r = 60$. Standard errors are computed according to equation (1.20). The paths of Figure 1.2 reveal that, at least for the case of IBM, the medium-term component $\alpha^{(m)}$ dominates the short-term and the long-term parameters, the latter being sometimes not significantly different from zero. This ranking of the auto-regressive components in the mSHARP model is confirmed for the majority of the stocks, as wit-

Average estimates of m and ℓ for the SHARP model

	0.5 sec.	1 sec.	5 sec.	10 sec.	15 sec.	30 sec.	1 min.	5 min.	10 min.	15 min.
m	5.89 (1.78)	5.29 (1.36)	6.56 (5.44)	7.96 (4.51)	8.91 (5.15)	11.78 (8.14)	9.70 (6.55)	12.14 (8.19)	9.44 (5.95)	7.97 (4.29)
ℓ	231.14 (172.42)	184.00 (119.60)	145.74 (87.44)	123.66 (70.62)	97.71 (47.81)	95.87 (47.09)	82.40 (48.46)	79.51 (42.19)	61.91 (36.98)	66.52 (37.45)

Table 1.2: The sample period from 02/01/2013 to 31/12/2013 has been divided in 25 non-overlapping intervals of 10 days and, for each interval, the parameters m and ℓ have been estimated at the corresponding frequency reported in the column label. At the end of this procedure, we are left with $25 \times 10 = 250$ estimates of each parameter for each of the sampling frequencies considered. The mean and (in parentheses) the standard deviation of these estimates are reported in the table.

nessed by the first three lines in Table 1.3 and in Table 1.4. These tables provide an overview of the magnitude of all the estimated models' parameters at, respectively, one-minute and five-second frequencies. As expected, at the frequency of five seconds, the magnitude of the short-term component $\alpha^{(s)}$ is more pronounced. The averages of the estimated coefficients confirm that bid-ask spread series are persistent, as witnessed by the fact that, typically, $\alpha^{(s)}+\alpha^{(m)}+\alpha^{(\ell)} \approx 1$. In the case of the sACP the sum $\alpha+\beta$ ranges from 0.9 for large tick stocks to 0.5 for small tick ones. On the contrary, the estimates of the fractional integration parameter d in the LMACP are not influenced by the average size of the spread but they are, on average, larger in the five-second case.

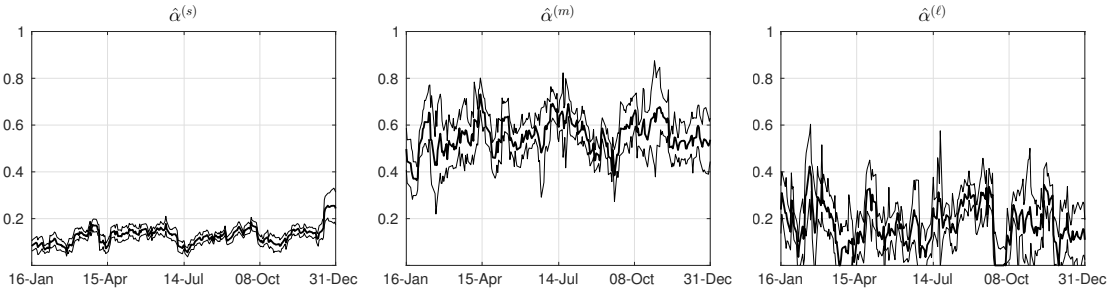

Figure 1.2: The time series of $\hat{\alpha}^{(s)}, \hat{\alpha}^{(m)}, \hat{\alpha}^{(\ell)}$, with its corresponding standard errors (1.20), for the mSHARP, estimated with the one-minute quoted spread series of IBM, in 2014. The parameter r, which defines the number of sub-intervals used in the mSHARP, is set to $r = 60$. The estimation is performed with a moving time window of ten days, which is recursively shifted by one day.

Average estimated parameters (one-minute frequency)

		BAC	VZ	GM	DAL	HAL	XOM	VLO	CVX	APC	IBM
mSHARP	$\alpha^{(s)}$	0.040	0.063	0.049	0.047	0.085	0.094	0.115	0.105	0.144	0.128
	$\alpha^{(m)}$	0.362	0.453	0.435	0.471	0.538	0.595	0.510	0.560	0.539	0.555
	$\alpha^{(\ell)}$	0.571	0.254	0.223	0.171	0.138	0.119	0.176	0.155	0.132	0.176
SHARP	$\alpha^{(s)}$	0.054	0.065	0.048	0.045	0.091	0.096	0.118	0.107	0.136	0.116
	$\alpha^{(m)}$	0.127	0.168	0.141	0.206	0.278	0.330	0.283	0.327	0.392	0.400
	$\alpha^{(\ell)}$	0.756	0.420	0.412	0.286	0.252	0.264	0.291	0.310	0.248	0.292
olsSHARP	$\alpha^{(s)}$	0.019	0.043	0.031	0.032	0.062	0.075	0.081	0.081	0.115	0.110
	$\alpha^{(m)}$	0.063	0.152	0.143	0.187	0.287	0.341	0.320	0.334	0.426	0.392
	$\alpha^{(\ell)}$	0.895	0.526	0.481	0.400	0.367	0.352	0.381	0.407	0.301	0.334
LMACP	ϕ	0.313	0.180	0.299	0.297	0.153	0.274	0.163	0.293	0.252	0.138
	β	0.689	0.076	0.152	0.158	0.186	0.235	0.351	0.280	0.188	0.232
	d	0.382	0.263	0.195	0.179	0.350	0.329	0.566	0.314	0.305	0.407
sACP	α	0.138	0.051	0.050	0.073	0.148	0.172	0.183	0.217	0.285	0.264
	β	0.850	0.886	0.870	0.741	0.507	0.554	0.559	0.329	0.142	0.212

Table 1.3: Average estimated parameters $\alpha^{(s)}$, $\alpha^{(m)}$, $\alpha^{(\ell)}$ for mSHARP, SHARP, and olsSHARP; average ϕ, β, d for LMACP and average α, β for sACP. The average is taken over a rolling window of ten days. For the mSHARP, the parameter r is equal to 60.

Computation of the forecasts. For each of the seven models listed above, we obtain forecasts of the bid-ask spread with a standard out-of-sample procedure: for a given sampling frequency, we first estimate the parameters of the considered model in a training window[18] and then we compute the one-step-ahead (i.e. either one minute or five seconds) forecasts starting from the end of the window until the end of the day that immediately follows. At the end of the day, the estimation of the parameters is renewed by shifting the estimation window of one day. The loop described thus far is iterated from the beginning to the end of 2014, for the one-minute case and from the beginning of 2014 to June 30, 2014, for the five-second case. Independently of the model used, the one-step ahead forecast is obtained as $\widehat{S}_{t+1|t} = \mathbb{E}[S_{t+1} \,|\, \mathcal{F}_t]$. The total number of forecasts varies, according to the frequency used, from $T = (244 - 10) \times 390 = 91\,260$ for

[18]We choose the length of the training window according to the frequency used: for the case of one-minute forecasts we use a window of ten days, which is shortened to five days for the case of five seconds.

Average estimated parameters (five-second frequency)

		BAC	VZ	GM	DAL	HAL	XOM	VLO	CVX	APC	IBM
mSHARP	$\alpha^{(s)}$	0.347	0.229	0.229	0.230	0.256	0.244	0.380	0.303	0.471	0.432
	$\alpha^{(m)}$	0.207	0.344	0.294	0.294	0.262	0.363	0.274	0.341	0.242	0.264
	$\alpha^{(\ell)}$	0.411	0.282	0.296	0.253	0.290	0.277	0.232	0.257	0.211	0.218
SHARP	$\alpha^{(s)}$	0.373	0.229	0.225	0.221	0.248	0.232	0.366	0.288	0.456	0.416
	$\alpha^{(m)}$	0.151	0.255	0.214	0.230	0.208	0.299	0.255	0.301	0.238	0.255
	$\alpha^{(\ell)}$	0.433	0.336	0.340	0.294	0.328	0.332	0.253	0.293	0.226	0.238
olsSHARP	$\alpha^{(s)}$	0.254	0.205	0.218	0.216	0.232	0.210	0.305	0.251	0.426	0.405
	$\alpha^{(m)}$	0.164	0.242	0.201	0.211	0.185	0.272	0.252	0.282	0.239	0.247
	$\alpha^{(\ell)}$	0.560	0.378	0.363	0.330	0.374	0.378	0.318	0.344	0.251	0.252
LMACP	ϕ	0.461	0.579	0.433	0.521	0.522	0.444	0.430	0.429	0.435	0.340
	β	0.029	0.199	0.051	0.227	0.265	0.344	0.344	0.372	0.424	0.359
	d	0.391	0.403	0.236	0.415	0.512	0.560	0.560	0.503	0.486	0.515
sACP	α	0.376	0.200	0.209	0.213	0.260	0.233	0.368	0.318	0.484	0.448
	β	0.588	0.683	0.621	0.582	0.492	0.635	0.441	0.501	0.321	0.352

Table 1.4: Average estimated parameters $\alpha^{(s)}$, $\alpha^{(m)}$, $\alpha^{(\ell)}$ for mSHARP, SHARP, and olsSHARP; average ϕ, β, d for LMACP and average α, β for sACP. The average is taken over a rolling window of five days. For the mSHARP, the parameter r is equal to 25.

the one-minute frequency to $T = (122 - 5) \times 4\,680 = 547\,560$ for the five-second case.

1.7.1 Comparative analysis of the models' performances

Forecasting accuracy. To evaluate the accuracy of the forecasting models, we use the Mean rounded Absolute Error (MrAE) and the Mean Squared Error (MSE) loss functions, defined respectively as

$$\text{MrAE} = \frac{1}{T}\sum_{t=1}^{T}\left|S_{t+1} - \text{round}\left(\widehat{S}_{t+1|t}\right)\right|, \quad \text{MSE} = \frac{1}{T}\sum_{t=1}^{T}\left(S_{t+1} - \widehat{S}_{t+1|t}\right)^2, \qquad (1.28)$$

where the sums are computed across the forecasts in 2014 and where S_{t+1} denotes the observed spread. The round function is considered because the rounded value of $\widehat{S}_{t+1|t}$ is closer, in absolute terms, to the observed spread S_{t+1} for all the considered models,

and creates coherent forecasts for an integer-valued process[19]. On the contrary, we do not round the forecasts in the squared loss function, since the mean square error is minimized by the conditional expectation (with no rounding).

Table 1.5 and Table 1.6 report, for each stock in the dataset, the loss functions for all the seven models and for the one-minute and five-second frequencies, respectively. One, two, or three stars signal that the mSHARP predictions are significantly better, according to the one-sided Diebold and Mariano [1995] test, at the 10%, 1% and 0.1% significance level, respectively[20]. The results reported in Table 1.5 and Table 1.6 indicate the mSHARP as the most performing model in terms of forecasting accuracy[21].

At this point, one may wonder whether the superior performances of the mSHARP can mostly be attributed to the specification of the seasonal component or to the HAR structure. A LMACP model with a seasonal component estimated with the non-parametric estimator $\widehat{\varphi}_{j_t}$ would be then warranted. Nevertheless, under a LMACP dynamics, the unconditional expected value of S_t is not finite and, hence, the expected value of the non-parametric estimator $\mathbb{E}[\widehat{\varphi}_{j_t}]$ would diverge to infinity and not to the intraday pattern. This said, the model comparison in Table 1.5 and Table 1.6 provides empirical evidence that the superior forecasting accuracy of the mSHARP model is not solely due to the specification of the seasonal component. In fact, for large tick stocks, the purely seasonal model can outperform the more sophisticated LMACP, as witnessed by the loss functions in Table 1.5 and Table 1.6 for the tickers BAC, VZ, GM and DAL. On the other hand, for small tick stocks, the solely contribution of the seasonal component is not sufficient to have accurate forecasts. As an example consider the case of IBM in Table 1.5: the sACP (which shares with the mSHARP/SHARP the same seasonal pattern) is outperformed by the LMACP, which is outperformed by the SHARP, beaten, in turn, by the mSHARP. In summary, both the seasonal component and the HAR structure are

[19] As an alternative, one could use the conditional median, as proposed by Freeland and McCabe [2004].

[20] Since the loss functions are characterized by heteroskedasticity, we employ, for the denominator of the Diebold and Mariano [1995] test statistics, the HAC estimator of Newey and West [1986], with a Parzen Kernel and a bandwidth estimated with the OLS procedure [Andrews, 1991]. We thank an anonymous referee for this suggestion.

[21] Results are strongly significant thanks to the large number of observations involved in the forecasting exercize ($T \sim 10^5$ for the one-minute frequency and to $T \sim 5 \times 10^5$ for the five-second one).

contributing to the superior accuracy of the mSHARP.

Goodness-of-fit. To complete the empirical comparative analysis, we report in Table 1.7 and in Table 1.8 for, respectively, the one-minute and five-second frequencies, the averages of the p-values of the Ljung-Box statistics for the Pearson's residuals

$$\eta_t = \frac{S_t - \mathbb{E}[S_t \mid \mathcal{F}_{t-1}]}{\sqrt{\text{Var}[S_t \mid \mathcal{F}_{t-1}]}}, \tag{1.29}$$

where S_t, as usual, denotes the observed spread, while $\mathbb{E}[S_t \mid \mathcal{F}_{t-1}]$ and $\text{Var}[S_t \mid \mathcal{F}_{t-1}]$ are, respectively, its conditional mean and variance under the assumed model.

The p-values in Table 1.7 reveal that, at one-minute frequency, the mSHARP, the SHARP, and the olsSHARP are well-specified at any lags and for any kind of stock, from the large (BAC) to the small (IBM) tick class, even though the p-values are quite close to a rejection for small tick stocks and for large lags. LMACP and sACP perform worse, especially for small tick stocks. This can be ascribed to the fact that, at one-minute frequency, the persistence of the time series is not particularly pronounced and so the long-memory version of the ACP does not depart much from it in term of goodness of specification. Nevertheless, as witnessed by the p-values in Table 1.7, for small tick stocks at the frequency of five seconds the LMACP is the best choice in terms of model specification. This is mainly a consequence of the fact that, at such a high frequency, the persistence of the time series is relevant, particularly for small tick stocks. The LMACP, being a (non-pseudo) long-memory process, is designed to capture such a feature, producing well-specified forecasts. Not surprisingly, the purely seasonal model and the random walk are misspecified at any frequency and for any kind of stock.

Average estimation time. In empirical applications the model that requires, *ceteris paribus*, the smallest amount of time to compute the forecast is the most convenient for practitioners and researchers. This is especially true whenever the predictions are used to determine the decisions of an optimal execution strategy and, consequently, a rapid update of the model parameters is needed. For this reason, the average time required to

Out-of-sample accuracy (one-minute frequency)

	mSHARP	SHARP	olsSHARP	LMACP	sACP	seasonal	RW
BAC	0.013	0.013	0.013	0.026***	0.015***	0.013*	0.025***
	0.013	0.013***	0.013***	0.030***	0.014***	0.021***	0.025***
VZ	0.108	0.108	0.109**	0.135***	0.109**	0.113***	0.164***
	0.108	0.109**	0.110***	0.260***	0.110***	0.113***	0.187***
GM	0.116	0.118***	0.118***	0.134***	0.118***	0.123***	0.183***
	0.116	0.116	0.117**	0.152***	0.117*	0.119**	0.206***
DAL	0.269	0.275***	0.276***	0.302***	0.275***	0.290***	0.362***
	0.269	0.275***	0.278***	0.345***	0.277**	0.296***	0.454***
HAL	0.419	0.429***	0.431***	0.465***	0.434***	0.483***	0.531***
	0.525	0.546***	0.577**	0.660***	0.538*	0.608***	0.837***
XOM	0.507	0.522***	0.522***	0.568***	0.532***	0.589***	0.624***
	0.788	0.820***	0.829***	1.021***	0.816**	0.900***	1.223***
VLO	0.665	0.680***	0.681***	0.699***	0.684***	0.742***	0.826***
	1.098	1.131***	1.163***	1.226***	1.134***	1.237***	1.757***
CVX	0.781	0.802***	0.804***	0.823***	0.820***	0.904***	0.996***
	1.753	1.840***	1.870***	1.988***	1.874***	2.098***	2.810***
APC	1.459	1.477***	1.476***	1.550***	1.547***	1.682***	1.845***
	5.741	5.787	5.885***	6.324***	5.973*	6.766***	8.521***
IBM	2.331	2.363***	2.364***	2.436***	2.483***	2.701***	2.986***
	11.693	11.943***	11.965***	12.795***	12.799***	14.595***	19.372***

Table 1.5: Average of the loss functions (for each model the first line is MrAE and the second is MSE) calculated with a moving window of ten days and a frequency of one minute. The symbols ***,**,* mean that the mSHARP delivers more accurate forecasts according to the Diebold and Mariano [1995] test (calculated with HAC standard errors) at the 0.1%, 1%, 10% significance level, respectively. Equal (in value) but statistically different loss functions are solely due to the limited number of digits shown.

Out-of-sample accuracy (five-second frequency)

	mSHARP	SHARP	olsSHARP	LMACP	sACP	seasonal	RW
BAC	0.015	0.016***	0.015	0.018***	0.016***	0.015	0.020***
	0.013	0.013***	0.013***	0.019***	0.015***	0.022***	0.020***
VZ	0.093	0.093*	0.094***	0.127***	0.094***	0.100***	0.115***
	0.080	0.080***	0.081***	0.169***	0.081***	0.099***	0.123***
GM	0.122	0.123***	0.123***	0.171***	0.124***	0.130***	0.150***
	0.101	0.102***	0.102***	0.195***	0.102***	0.119***	0.156***
DAL	0.235	0.237***	0.237***	0.272***	0.238***	0.261***	0.263***
	0.189	0.190***	0.191***	0.270***	0.191***	0.231***	0.288***
HAL	0.301	0.302***	0.304***	0.330***	0.307***	0.378***	0.325***
	0.274	0.278***	0.281***	0.343***	0.281***	0.375***	0.391***
XOM	0.413	0.416***	0.421***	0.452***	0.421***	0.547***	0.452***
	0.522	0.532***	0.542***	0.732***	0.533***	0.779***	0.746***
VLO	0.594	0.597***	0.609***	0.601***	0.603***	0.804***	0.613***
	0.864	0.870***	0.880***	0.964***	0.891***	1.384***	1.183***
CVX	0.666	0.670***	0.679***	0.673***	0.678***	0.846***	0.717***
	1.166	1.182***	1.206***	1.273***	1.199***	1.811***	1.626***
APC	1.149	1.151***	1.163***	1.134	1.159***	1.721***	1.112
	3.112	3.123***	3.142***	3.181***	3.201***	6.193***	3.936***
IBM	1.896	1.899***	1.906***	1.893	1.921***	2.706***	1.887
	8.085	8.107***	8.122***	8.111	8.287***	14.419***	10.677***

Table 1.6: Average of the loss functions (for each model the first line is MrAE and the second is MSE) calculated with a rolling window of five days and a frequency of five seconds. The symbols ***,**,* mean that the mSHARP delivers more accurate forecasts according to the Diebold and Mariano [1995] test (calculated with HAC standard errors) at the 0.1%, 1%, 10% significance level, respectively. Equal (in value) but statistically different loss functions are solely due to the limited number of digits shown.

Ljung-Box test statistics (one-minute frequency)

	mSHARP	SHARP	olsSHARP	LMACP	sACP	seasonal	RW
	0.743	0.555	0.961	0.658	0.394	0.226	0.000
BAC	0.717	0.780	0.217	0.533	0.406	0.072	0.000
	0.323	0.298	0.063	0.499	0.258	0.072	0.000
	0.697	0.577	0.873	0.187	0.355	0.006	0.000
VZ	0.379	0.449	0.333	0.037	0.380	0.001	0.000
	0.152	0.145	0.086	0.041	0.104	0.003	0.000
	0.823	0.771	0.906	0.170	0.388	0.007	0.000
GM	0.423	0.501	0.364	0.023	0.462	0.003	0.000
	0.142	0.152	0.077	0.015	0.120	0.005	0.000
	0.738	0.845	0.952	0.048	0.511	0.001	0.000
DAL	0.218	0.323	0.307	0.004	0.384	0.000	0.000
	0.138	0.121	0.105	0.000	0.073	0.001	0.000
	0.618	0.675	0.869	0.099	0.337	0.000	0.000
HAL	0.115	0.121	0.166	0.049	0.100	0.000	0.000
	0.089	0.071	0.106	0.010	0.011	0.000	0.000
	0.634	0.721	0.862	0.083	0.297	0.000	0.000
XOM	0.092	0.096	0.171	0.025	0.034	0.000	0.000
	0.057	0.036	0.041	0.002	0.001	0.000	0.000
	0.637	0.655	0.814	0.125	0.211	0.000	0.000
VLO	0.057	0.054	0.090	0.027	0.040	0.000	0.000
	0.060	0.023	0.080	0.013	0.003	0.000	0.000
	0.487	0.580	0.758	0.113	0.155	0.000	0.000
CVX	0.027	0.033	0.088	0.028	0.003	0.000	0.000
	0.041	0.026	0.045	0.000	0.000	0.000	0.000
	0.561	0.625	0.699	0.016	0.088	0.000	0.000
APC	0.030	0.041	0.080	0.000	0.000	0.000	0.000
	0.022	0.019	0.015	0.000	0.000	0.000	0.000
	0.711	0.732	0.719	0.114	0.056	0.000	0.000
IBM	0.048	0.064	0.063	0.000	0.000	0.000	0.000
	0.058	0.059	0.028	0.000	0.000	0.000	0.000

Table 1.7: The mean of the Ljung-Box test statistics for the Pearson residuals of each model calculated with 1 (first line), 10 (second line) and 390 (third line) lags and with a moving window of ten days at a frequency of one minute. The average is calculated over the iterations of the moving window.

Ljung-Box test statistics (five-second frequency)

	mSHARP	SHARP	olsSHARP	LMACP	sACP	seasonal	RW
	0.064	0.021	0.196	0.007	0.104	0.000	0.000
BAC	0.157	0.088	0.006	0.019	0.000	0.000	0.000
	0.156	0.059	0.182	0.002	0.203	0.000	0.000
VZ	0.010	0.007	0.001	0.000	0.000	0.000	0.000
	0.340	0.158	0.217	0.014	0.124	0.000	0.000
GM	0.001	0.001	0.000	0.000	0.000	0.000	0.000
	0.512	0.265	0.240	0.012	0.037	0.000	0.000
DAL	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	0.482	0.267	0.225	0.032	0.147	0.000	0.000
HAL	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.373	0.164	0.217	0.017	0.067	0.000	0.000
XOM	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.115	0.063	0.056	0.042	0.033	0.000	0.000
VLO	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.261	0.119	0.137	0.046	0.093	0.000	0.000
CVX	0.000	0.000	0.000	0.007	0.000	0.000	0.000
	0.040	0.025	0.002	0.192	0.195	0.000	0.000
APC	0.000	0.000	0.000	0.079	0.000	0.000	0.000
	0.027	0.015	0.002	0.199	0.202	0.000	0.000
IBM	0.000	0.000	0.000	0.141	0.000	0.000	0.000

Table 1.8: The mean of the Ljung-Box test statistics for the Pearson residuals of each model calculated with 1 (first line) and 10 lags (second line), with a moving window of five days and at a frequency of five seconds. The average is calculated over the iterations of the moving window. The results with the maximum number of lags, i.e. $4\,680$, are not shown because they are always smaller than 10^{-3} except for the case of the BAC with the LMACP model, for which it is equal to 0.151.

estimate the seven models discussed so far is shown in Table 1.9. They are in line with what is anticipated from the analytical properties of each model. In fact, the estimation of long memory processes is known to be quite cumbersome. On the other hand, the analytical properties of the SHARP (and of the mSHARP/olsSHARP), that is the concavity of the log-likelihood and the simplicity of the intraday pattern, translates in a rapid optimization of the likelihood function. In the case of the olsSHARP, which is estimated analytically, the estimation time is almost negligible.

Average estimation time

	mSHARP	SHARP	olsSHARP	LMACP	sACP	Seasonal	RW
1 min.	0.082	0.097	0.002	85.36	0.081	-	-
5 sec.	0.224	0.182	0.005	284.5	0.282	-	-

Table 1.9: Average time (in seconds, averaged across all ten stocks and across all the estimation windows) required to estimate each model. For the frequency of one minute, the seven models are estimated in a time window of ten days, while for the frequency of five seconds they are estimated in a time window of five days. Optimization is achieved through the Sequential Quadratic Programming (SQP) implemented in Matlab. We used an Intel Core i5-2450M CPU, 2.50GHz with 4 processors (including cores). The estimation times for the Seasonal and the Random Walk models are not reported since they are close to zero.

1.7.2 An application to optimal execution

In this section we empirically show that it is possible to use the spread predictions of the SHARP model to develop a trading schedule that (in a statstically significant way) reduces the transaction costs with respect to trading strategies based on other benchmark models. For this purpose, we design optimal execution strategies on an equispaced time grid of five seconds. In what follows, the time index t is assumed to belong to the partition $\mathcal{T}$ in (1.1) with $J = 4\,680$. We work under the assumption that the mid price process[22] M_t is a martingale, that is

$$\mathbb{E}[M_{t+h} \,|\, \mathcal{F}_t] = M_t, \quad \forall h \in \mathbb{N}_+.$$

[22]Defined as the average price between the best ask and the best bid, that is $M_t = (\mathrm{A}_t + \mathrm{B}_t)/2$.

This assumption implies that the h-step ahead forecasts at time t of the ask (A_{t+h}) and bid (B_{t+h}) prices depend on the forecast of the spread S_{t+h} as follows

$$\begin{aligned}\mathbb{E}[\mathrm{A}_{t+h}\,|\,\mathcal{F}_t] &= \mathbb{E}[M_{t+h} + 1/2(1+S_{t+h})|\,\mathcal{F}_t] = M_t + 1/2 + 1/2\,\mathbb{E}[S_{t+h}|\,\mathcal{F}_t],\\ \mathbb{E}[\mathrm{B}_{t+h}\,|\,\mathcal{F}_t] &= \mathbb{E}[M_{t+h} - 1/2(1+S_{t+h})|\,\mathcal{F}_t] = M_t - 1/2 - 1/2\,\mathbb{E}[S_{t+h}|\,\mathcal{F}_t].\end{aligned} \tag{1.30}$$

In particular, the smaller the $\mathbb{E}[S_{t+h}|\,\mathcal{F}_t]$, the smaller the $\mathbb{E}[\mathrm{A}_{t+h}\,|\,\mathcal{F}_t]$ and the higher the $\mathbb{E}[\mathrm{B}_{t+h}\,|\,\mathcal{F}_t]$. For this reason, accurate spread predictions allow for the selection of instants characterized by (statistically) smaller ask prices and (statistically) higher bid prices.

In the empirical exercise described here, we follow the approach of Taylor [2002] and Groß-Klußmann and Hautsch [2013]. Orders are assumed to be split into smaller trades and distributed over the day, a common practice used to reduce the price impact of transactions [for example, Almgren and Chriss, 2001]. More specifically, each trading day d, with $d = 0, ..., D-1$, is divided into $N_{\text{trade}} = 390$ trading intervals of one minute and, each trading interval, is divided into $N_{\text{sub}} = 12$ sub-intervals of five seconds. In this setting, a trading strategy for the d-th day consists in a collection of trading times $\{t_{1,d}, ..., t_{N_{\text{trade}},d}\}$ where, for a generic $i = 1, ..., N_{\text{trade}}$, the i-th trading time $t_{i,d}$ is chosen in the sub-grid $\mathcal{H}_i \overset{\text{def}}{=} \{(i-1)\ N_{\text{sub}} + 1, ..., i\,N_{\text{sub}}\}$ in which the i-th minute is divided. We consider four different types of strategies. In the first, which we address as the uninformed strategy, we consider the choice of a trader that cannot make any inference on the future values of the bid-ask spread. We assume that the uninformed trader selects the i-th trading time, denoted as $t_{i,d}^{(\mathrm{U})}$, through a uniform extraction from the set $\mathcal{H}_i$. In the other three strategies, the trader is allowed to forecast future bid-ask spreads through one among the SHARP, the sACP and the seasonal models described, respectively, in equations (1.12), (1.14) and (1.27). Regardless of the model used, the trading strategy is defined by the following choice rule: the trader generates, progressively from the first to the last instant in the i-th sub-grid $\mathcal{H}_i$ of the day d, the predictions of the spread in the remaining instants and executes the trade as soon as the prevailing spread is lower

than all the predicted values.

To each trading strategy it is possible to associate a stream of optimal (according to the strategy) ask and bid prices, defined as the best ask and best bid prevailing at the trading instants $\left\{t_{1,d},...,t_{N_{\text{trade}},d}\right\}$. Let, for example, be $A_{i,d}^{(\text{U})}$, with $i=1,...,N_{\text{trade}}$, be the stream of ask prices generated by the uninformed strategy during day d and let $A_{i,d}^{(\text{SHARP})}$ be the corresponding stream generated by the strategy based on SHARP predictions. The average

$$\text{G}_{\text{A}}^{(\text{U})}=\frac{100}{D}\sum_{d=0}^{D-1}\left(\frac{1}{N_{\text{trade}}}\sum_{i=1}^{N_{\text{trade}}}\frac{\text{A}_{i,d}^{(\text{U})}-\text{A}_{i,d}^{(\text{SHARP})}}{\overline{S}_{i,d}}\right) \tag{1.31}$$

with

$$\overline{S}_{i,d}=\frac{1}{N_{\text{sub}}}\sum_{k\in\mathcal{H}_i}\left(S_{Jd+k}+1\right),$$

represents the percentage average gain (in units of the average spread) of adopting, for buying the asset, the strategy that profits from SHARP predictions of the bid-ask spread instead of the uninformed strategy. For example, a $\text{G}_{\text{A}}^{(\text{U})}=1$ is indicating that, on average, the trading strategy based on SHARP predictions saves, with respect to the uninformed one, 1% of the average spread. Similarly, the quantity

$$\text{G}_{\text{B}}^{(\text{U})}=\frac{100}{D}\sum_{d=0}^{D-1}\left(\frac{1}{N_{\text{trade}}}\sum_{i=1}^{N_{\text{trade}}}\frac{\text{B}_{i,d}^{(\text{SHARP})}-\text{B}_{i,d}^{(\text{U})}}{\overline{S}_{i,d}}\right) \tag{1.32}$$

has the same interpretation of the gain $\text{G}_{\text{A}}^{(\text{U})}$ defined in (1.31), with the difference that now the trading is considered on the sell side.

Table 1.10 reports a summary of the average gains in the adoption of the strategy that profits from SHARP predictions with respect to all the other strategies considered. For small and medium tick stocks, trading based on SHARP forecasts may save up to 14% (resp. 13%) of the average spread costs in selling (resp. buying) the asset with respect to the uninformed strategy. The savings against the seasonal and the sACP strategies are smaller, even if often statistically significant, especially for small tick stocks.

Percentage average gain of the SHARP-based trading schedule

	BAC	VZ	GM	DAL	HAL	XOM	VLO	CVX	APC	IBM
$G_A^{(U)}$	0.43^{**}	3.63^{**}	3.43^{**}	7.96^{**}	10.17^{**}	13.05^{**}	12.60^{**}	12.35^{**}	10.98^{**}	11.76^{**}
$G_B^{(U)}$	0.64^{**}	3.55^{**}	4.85^{**}	9.51^{**}	11.65^{**}	12.64^{**}	13.66^{**}	14.31^{**}	11.73^{**}	12.93^{**}
$G_A^{(\varphi)}$	0.00^{**}	0.05	0.08^{**}	0.06	0.36^{**}	0.21	0.52^{**}	1.04^{**}	1.04^{**}	0.88^{**}
$G_B^{(\varphi)}$	0.00^{**}	0.02	-0.02	0.19^{**}	0.30^{**}	0.26^{*}	0.80^{**}	1.12^{**}	0.88^{**}	1.00^{**}
$G_A^{(sACP)}$	0.00^{**}	0.04	0.09^{**}	0.00	0.16	-0.07	0.06	0.53^{**}	0.60^{**}	0.40^{**}
$G_B^{(sACP)}$	0.00^{**}	0.01	-0.03	0.11	0.04	0.05	0.46^{**}	0.24	0.56^{**}	0.64^{**}

Table 1.10: The rows $G_A^{(U)}$, $G_A^{(\varphi)}$ and $G_A^{(sACP)}$ report the percentage average gains (defined as in equation (1.31)) in adopting a trading strategy based on SHARP forecasts with respect to, respectively, the uniformed, the seasonal and the sACP trading strategies. Similarly, the rows $G_B^{(U)}$, $G_B^{(\varphi)}$ and $G_B^{(sACP)}$, report the percentage average gains (defined as in equation (1.32)) in adopting a trading strategy based on SHARP forecasts with respect to, respectively, the uniformed, the seasonal and the sACP trading strategies. The gain G_A (resp. G_B) is positive when the average ask (resp. bid) price paid with the SHARP-based strategy is smaller (resp. higher) than that paid with the other one. The superscripts $**$ and $*$ indicate that the null of zero percentage average gain is rejected with, respectively, 5% and 10% significance level.

1.8 Conclusions

In this study we propose a parsimonious and accurate discrete-time forecasting model for integer-valued time series. The model, named SHARP, features seasonality and pseudo-long-memory patterns. On a dataset of ten NYSE stocks, representative of different bid-ask spread dynamics, both the SHARP and its MIDAS extension (named mSHARP) show, despite their simple structure, superior forecasting performances with repsect to the benchmark competitor, i.e. the LMACP, which is a (non-pseudo) long-memory model. The results of the comparison confirm that it is possible to have competitive and reliable forecasts of the bid-ask spread without necessarily resorting to long-memory processes, whose estimation procedure is often quite cumbersome and could be up to a thousand times slower. Finally, as an empirical application, we show how bid-ask spread forecasts based on the SHARP model provide insightful information for reducing the total costs of transacting, especially for medium and small tick stocks.

Acknowledgments

We are grateful to Giampiero Gallo, Luca Trapin and to the seminar participants at the "Recent developments in econometric methodologies" (Bergamo, November 25-26, 2016) and at the "Seventh Italian Congress of Econometrics and Empirical Economics" (Messina, January 25-27, 2017) for discussions. A Matlab package with the main routines for the estimation, simulation and forecasting of the SHARP model and its extensions are available at `https://sites.google.com/a/sns.it/lucacattivelli`. Minor routines are available upon request.

Appendix 1.A Pseudo Long Memory and the HAR

In the LMACP model, long-memory is obtained with the introduction of a fractional difference operator. However, fractionally integrated models are often problematic, since the estimation procedure is quite cumbersome and they not easily extendible to multivariate processes. The Heterogeneous Auto Regressive (HAR) model by Corsi [2009] provides a valid and simple alternative to circumvent these issues.

The HAR is an autoregressive model with the feature of considering the variable of interest realized over intervals of size m and ℓ, corresponding to the medium (m) and long (ℓ) term autoregressive components.[23] For example, for the case of the deseasonalized time series $\widetilde{S}_t = S_t/\widehat{\varphi}_{j_t}$, the HAR model would be written as

$$\widetilde{S}_t = c + \alpha^{(s)}\,\widetilde{S}_{t-1} + \alpha^{(m)} \sum_{q=t-m}^{t-1} \frac{\widetilde{S}_q}{m} + \alpha^{(\ell)} \sum_{q=t-\ell}^{t-1} \frac{\widetilde{S}_q}{\ell} + \epsilon_t, \tag{1.33}$$

where the ϵ_t's are i.i.d. errors. Corsi [2009] selected $m = 5$ and $\ell = 22$, however different choices for m and ℓ are possible[24]. Having selected a value for m and ℓ, the parameters $\alpha^{(s)}$, $\alpha^{(m)}$ and $\alpha^{(\ell)}$ can be estimated by applying simple OLS.

[23] The parameters m and ℓ are two integers with $m < \ell$.

[24] The HAR model was proposed in Corsi [2009] to describe the evolution of the daily realized volatility, being the medium term component a weakly component ($m = 5$) and the long term autoregressive component a monthly term ($\ell = 22$).

Even if from a mathematical point of view the HAR specification (1.33) does not define a long-memory process, it produces an extremely persistent process, with slowly decaying memory patterns. In spite of the simplicity of its structure, the simulation results of Corsi [2009] show that the HAR model successfully achieves the purpose of reproducing the main empirical features of long memory processes in a tractable and parsimonious way.

As a final remark, note that the model in equation (1.33) would be able to capture the strong persistence of the spread through the HAR specification, and the intraday seasonality of S_t thanks to the use of the seasonal pattern $\widehat{\varphi}_{j_t}$. However, it is not suited to describe integer-valued series, differently from the SHARP model of equation (1.12).

Appendix 1.B Proof of Theorem 1.

From the last two lines of equation (1.12), we obtain that

$$\frac{\lambda_t}{\varphi_{j_t}} = \mu_t = (1 - \Sigma_\alpha) + \alpha_H(B)\frac{S_t}{\varphi_{j_t}}, \tag{1.34}$$

with $\alpha_H(B) = \alpha^{(s)}B + \frac{\alpha^{(m)}}{m}\sum_{k=1}^{m} B^k + \frac{\alpha^{(\ell)}}{\ell}\sum_{k=1}^{\ell} B^k$. Now we define the martingale difference sequence v_t as $v_t = S_t - \lambda_t$. Substituting v_t into equation (1.34), we get that

$$\left(1 - \alpha_H\left(B\right)\right)\frac{S_t}{\varphi_{j_t}} = (1 - \Sigma_\alpha) + \frac{v_t}{\varphi_{j_t}}. \tag{1.35}$$

The operator $(1 - \alpha_H(B))$ that appears in equation (1.35) is invertible if and only if its roots lie outside of the unit circle, that is, if $\alpha^{(s)} + \alpha^{(m)} + \alpha^{(\ell)} < 1$. The invertibility of $(1 - \alpha_H(B))$ and equation (1.35) imply that

$$S_t = \varphi_{j_t} + \varphi_{j_t}(1 - \alpha_H(B))^{-1}\frac{v_t}{\varphi_{j_t}}, \tag{1.36}$$

so that, for a given $h \in \{1, ..., J\}$, consider the process

$$Y_d^{(h)} = S_{h+d\,J} - \varphi_h, \;\; d = 0, 1, ..., D-1 \tag{1.37}$$

whose covariance at lag k is written as

$$\begin{aligned}
\mathbb{E}\left[Y_d^{(h)}\, Y_{d+k}^{(h)}\right] &= \mathbb{E}\left[\left(\varphi_h(1-\alpha_H(B))^{-1}\frac{v_{h+d\,J}}{\varphi_h}\right)\left(\varphi_h(1-\alpha_H(B))^{-1}\frac{v_{h+(d+k)\,J}}{\varphi_h}\right)\right] \\
&= \mathbb{E}\left[\left(\varphi_h\sum_{s=0}^{\infty}\delta_s B^s\frac{v_{h+d\,J}}{\varphi_h}\right)\left(\varphi_h\sum_{s'=0}^{\infty}\delta_{s'} B^{s'}\frac{v_{h+(d+k)\,J}}{\varphi_h}\right)\right] \\
&= \mathbb{E}\left[\left(\varphi_h\sum_{s=0}^{\infty}\delta_s\frac{v_{h+d\,J-s}}{\varphi_{j_{h-s}}}\right)\left(\varphi_h\sum_{s'=0}^{\infty}\delta_{s'}\frac{v_{h+(d+k)\,J-s'}}{\varphi_{j_{h-s'}}}\right)\right] \\
&= \varphi_h^2\sum_{s=0}^{\infty}\delta_s\,\delta_{s+k\,J}\,\mathbb{E}\left[\frac{v_{j+d\,J-s}^2}{\varphi_{j_{h-s}}^2}\right] \\
&= \varphi_h^2\sum_{s=0}^{\infty}\frac{\delta_s\,\delta_{s+k\,J}}{\varphi_{j_{h-s}}},
\end{aligned} \tag{1.38}$$

with δ_q such that $\sum_{q=0}^{\infty}\delta_q\,B^q = (1-\alpha_H\,(B))^{-1}$. From equality (1.38) we see that, for a given h, the process $(S_{h+d\,J})_{d\in\{0,1,...,D-1\}}$ is covariance stationary, that is the coefficients

$$\gamma_k^{(h)} \stackrel{\text{def}}{=} \mathbb{E}[(S_{h+d\,J}-\varphi_h)(S_{h+(d+k)\,J}-\varphi_h)] = \mathbb{E}[Y_d^{(h)}\,Y_{d+k}^{(h)}],$$

depend only on the lag k, and not on d. Moreover, the process $(S_{h+d\,J})_{d\in\{0,1,...,D-1\}}$ is not a long memory process, therefore the autocorrelation is absolutely summable, whence the consistency result (1.18)

$$\widehat{\varphi}_h \stackrel{\text{def}}{=} \frac{1}{D}\sum_{d=0}^{D-1} S_{h+d\,J} \xrightarrow{p} \varphi_h \text{ and as } D \to \infty,$$

follows from Proposition 7.5 in Hamilton [1994].

Appendix 1.C Finite sample properties

In this Appendix we analyze the finite sample properties of the SHARP model, checking whether the absence of concavity in the log-likelihood and the discreteness of the parameters m and ℓ affect the performance of the estimators when m and ℓ are considered as free parameters to be estimated. Moreover, we control the effectiveness of the proposed two-step procedure for the estimation of the SHARP model comparing its finite sample properties with that of the LMACP.[25]

In order to compare the maximum likelihood estimators, we compute the relative bias and relative standard deviation defined as follows: consider the ratio $\widehat{\theta}/\theta_0$, where $\widehat{\theta}$ is the estimator of a given model parameter and θ_0 is its true value (used to generate the artificial sample). If the estimator $\widehat{\theta}$ is unbiased, then $\mathbb{E}\left[\widehat{\theta}/\theta_0\right] = 1$. We define the relative bias as

$$\text{RBIAS} = \mathbb{E}\left[\widehat{\theta}/\theta_0\right] - 1 = \left(\mathbb{E}\left[\widehat{\theta}\right] - \theta_0\right)/\theta_0.$$

We also define the relative standard deviation of $\widehat{\theta}$ as

$$\text{RSTD} = \sqrt{\text{VaR}\left[\widehat{\theta}/\theta_0\right]} = \sqrt{\text{VaR}\left[\widehat{\theta}\right]}/\left|\theta_0\right|$$

and the relative root mean square error as $\text{RRMSE} = \sqrt{\text{RBIAS}^2 + \text{RSTD}^2}$. These quantities can be estimated with their finite-sample counterparts

$$\widehat{\text{RBIAS}} = \frac{\bar{\theta} - \theta_0}{\theta_0}, \quad \widehat{\text{RSTD}} = \frac{1}{|\theta_0|}\sqrt{\frac{\sum_{r=1}^{R}(\widehat{\theta}_r - \bar{\theta})^2}{R}}, \quad \widehat{\text{RRMSE}} = \sqrt{\widehat{\text{RBIAS}}^2 + \widehat{\text{RSTD}}^2}, \tag{1.39}$$

where $\bar{\theta}$ is the average of parameters $\widehat{\theta}_r$ estimated in the $r-$th replication with $r = 1, ..., R$, that is $\bar{\theta} = \frac{1}{R}\sum_{r=1}^{R}\widehat{\theta}_r$, being R the total number of replications.

The results of the simulations are reported in Table 1.C.1 for $R = 2000$ replications using a length of 3900 observations for each replica[26]. Along with the estimated bias

[25] In the LMACP, we truncate the fractional operator $(1-B)^d$ of equation (1.5) at lag 250, this truncation being suggested by Groß-Klußmann and Hautsch [2013].

[26] The length of 3900 observations has been selected to analyze the finite sample properties of the estima-

and standard deviation we report the true value of each parameter adopted in the simulation. In order to have realistic values for such parameters we used the estimates for the one-minute bid-ask spread series obtained from the first 10 trading days of January 2014 for IBM.

The simulation results show that the approximation in equation (1.10) brings some bias in the estimation of dispersion parameter γ,[27] while the remaining estimates are unbiased. Moreover, the variance of the estimator is quite large for the parameters ϕ, β, $\delta^s_{1,2}$, $\delta^s_{2,1}$ of the LMACP and for the integer parameter ℓ of the SHARP, and small for the others.

Finally, in Figure 1.C.1 and Figure 1.C.2, we report the kernel (smoothed) distribution of the estimated parameters across the 2000 replications of the model (black continuous line) along with the asymptotic density of the maximum likelihood estimator[28] for the non-integer parameters (blue dotted line).

The computational time requested to perform the optimisations is quite different for the two models. In particular, the estimation of the LMACP model results to be quite complicated, due to the presence of many local maxima (this can be mostly attributed to the fact that the log-likelihood is not concave). In fact, in order to find the global minimum it is necessary to use many different initial conditions for the optimization procedure. In addition, the estimate is quite long, taking an average time of approxi-

tors when the sample size consists of 10 days of trading with 390 observations per day, which corresponds to observing the process S_t every minute.

[27] If we replace the approximation (1.10) with a (computationally intensive) numerical approximation of the normalisation term

$$c(\gamma, \lambda_t) \sim \left(\sum_{s=0}^{\infty} \sqrt{\gamma}\, e^{-\gamma \lambda_t} \left(\frac{e^{-s} s^s}{s!} \right) \left(\frac{e \lambda_t}{s} \right)^{\gamma s} \right)^{-1},$$

the bias disappears.

[28] This is a Gaussian variable with mean equal to the corresponding true value used for the simulation and covariance matrix given by the information matrix

$$I(\Theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial \Theta} \log \mathcal{L}(\Theta) \right)^2 \right].$$

The information matrix has been estimated with the empirical Hessian estimator $\widehat{I(\Theta)} = -H^{-1}\left(\widehat{\Theta}\right)$, where $\widehat{\Theta}$ is the MLE of the vector of model parameters and where H is the Hessian matrix defined in equation (1.40) [see Davidson and MacKinnon, 2004].

Overview of the models.

LMACP:
$$\begin{cases} \mathbb{P}_{t-1}\left[S_t = k\right] = c(\gamma, \lambda'_t)\sqrt{\gamma}\, e^{-\gamma\lambda'_t}\left(\frac{e^{-k}k^k}{k!}\right)\left(\frac{\lambda'_t\, e}{k}\right)^{\gamma\, k}, \quad k = 0, 1, 2, \ldots, \\ \lambda'_t = \lambda_t\, \exp(s_{j_t}), \\ s_{j_t} = \delta^s\, j_t/J + \sum_{l=1}^{2}\left(\delta^s_{1,l}\, \cos\left(2\,\pi\, l\, j_t/J\right) + \delta^s_{2,l}\, \sin\left(2\,\pi\, l\, j_t/J\right)\right), \\ \lambda_t = \omega + (\phi - \beta)S_{t-1} + \beta\lambda_{t-1} - \sum_{g=1}^{\infty}\frac{\Gamma(g-d)}{\Gamma(-d)\Gamma(g+1)}(S_{t-g} - \phi S_{t-g-1}). \end{cases}$$

SHARP:
$$\begin{cases} \mathbb{P}_{t-1}\left[S_t = k\right] = \lambda_t^k\, \frac{e^{-\lambda_t}}{k!}, \quad k = 0, 1, 2, \ldots, \\ \lambda_t = \varphi_{j_t}\, \mu_t,\ \mathbb{E}\left[S_t\right] = \varphi_{j_t}, \\ \mu_t = (1 - \Sigma_\alpha) + \alpha^{(s)}\, \widetilde{S}_{t-1:t-1} + \alpha^{(m)}\, \widetilde{S}_{t-m:t-1} + \alpha^{(\ell)}\, \widetilde{S}_{t-\ell:t-1}, \\ \Sigma_\alpha = \alpha^{(s)} + \alpha^{(m)} + \alpha^{(\ell)}. \end{cases}$$

Parameters of the conditional intensities.

	SHARP					LMACP				
	$\alpha^{(s)}$	$\alpha^{(m)}$	$\alpha^{(\ell)}$	m	ℓ	ω	ϕ	β	d	γ
True	0.120	0.305	0.318	9	60	5.133	0.420	0.530	0.450	0.569
$\widehat{\text{RBIAS}}$	-0.001	0.004	-0.051	-0.004	0.039	0.024	-0.024	-0.008	0.006	0.143
$\widehat{\text{RSTD}}$	0.146	0.138	0.185	0.098	0.240	0.083	0.583	0.459	0.219	0.022
$\widehat{\text{RRMSE}}$	0.146	0.138	0.192	0.099	0.243	0.086	0.583	0.459	0.219	0.144

Parameters of the seasonal components.

	SHARP	LMACP				
	φ_{j_t}	δ^s	$\delta^s_{1,1}$	$\delta^s_{1,2}$	$\delta^s_{2,1}$	$\delta^s_{2,2}$
True	-	-1.2285	-0.0235	-0.0268	-0.3002	-0.1111
$\widehat{\text{RBIAS}}$	0.0058	0.0130	0.0171	-0.0003	0.0131	0.0138
$\widehat{\text{RSTD}}$	0.0851	0.0608	0.5455	0.4886	0.0809	0.1451
$\widehat{\text{RRMSE}}$	0.0853	0.0621	0.5458	0.4886	0.0820	0.1457

Table 1.C.1: The parameters inputted in the simulation (row labelled as "True") and the statistics of the estimated parameters, i.e., the relative bias (RBIAS), the relative standard deviation (RSTD) and the relative root mean square error (RRMSE) defined in equations (1.39). The normalizing factor $c(\gamma, \lambda_t)$ in the LMACP model is approximated by $c(\gamma, \lambda_t) \approx \left(1 + \frac{1-\gamma}{12\lambda_t\gamma}\left(1 + \frac{1}{\lambda_t\gamma}\right)\right)^{-1}$. For the seasonal component φ_{j_t} we report the mean over all replicas of the sample averages of the three loss functions over the entire sample.

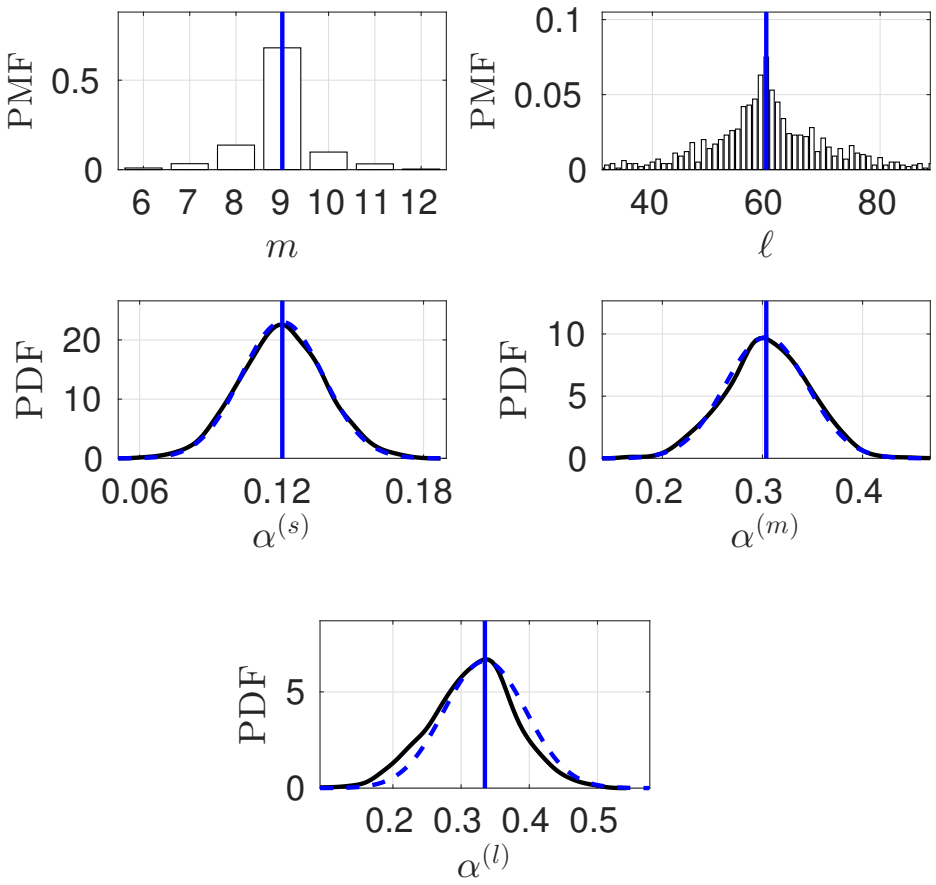

Figure 1.C.1: Finite sample distribution of the estimators of the parameters of the SHARP model. We report, in black, the distribution (Probability Distribution Function, PDF, for the continuous parameters and Probability Mass Function, PMF, for the discrete ones) of the estimated parameters (indicated in the horizontal axis of each sub-figure) across 2000 replications of the model. We show also the theoretical distribution (blue dotted line), when known, of the corresponding maximum likelihood estimator and the true simulated value with a vertical blue line.

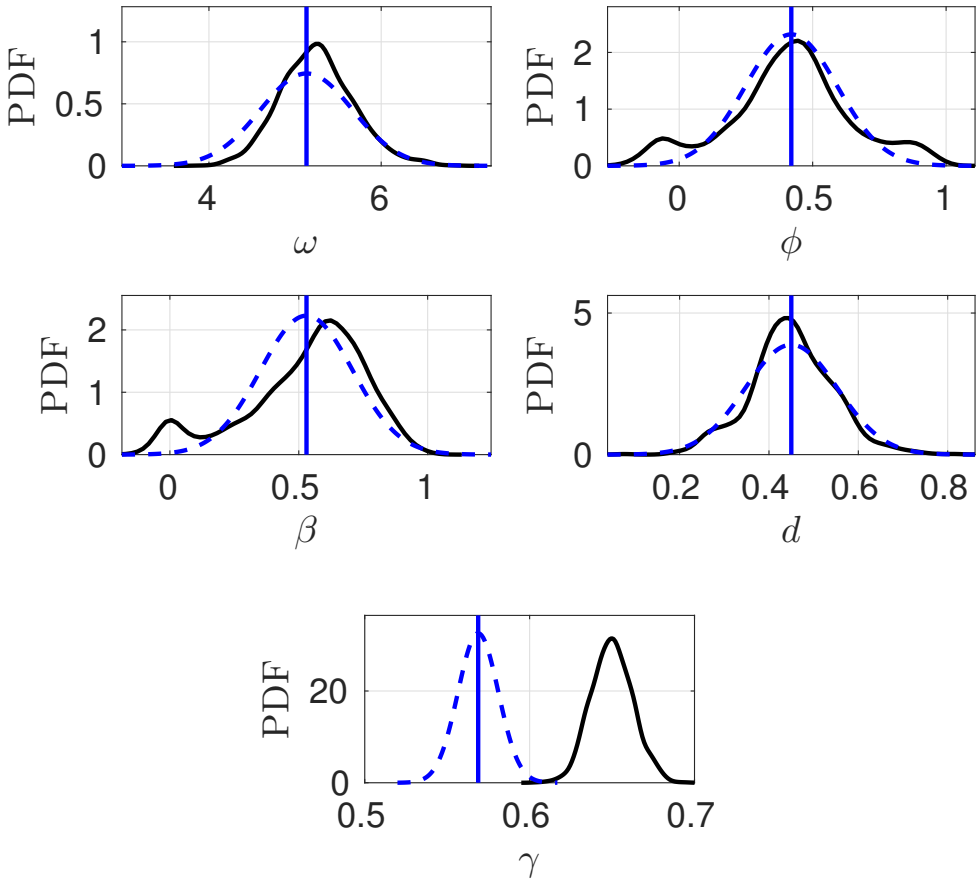

Figure 1.C.2: Finite sample distribution of the estimators of the main parameters of the LMACP model. We report, in black, the Probability Distribution Function (PDF) of the estimated parameters (indicated in the horizontal axis of each sub-figure) across 2000 replications of the model. We show also the theoretical distribution (blue dotted line) of the corresponding maximum likelihood estimator and the true simulated value with a vertical blue line.

mately 70 seconds[29].

The optimization procedure is quite long also in the estimation of the SHARP model with free parameters m and l. Indeed, in this case, the log-likelihood is not concave and several local minima can be found. Conversely, the algorithm has no trouble converging for the SHARP with fixed m and ℓ and the results are independent of the initial values imposed. Indeed, even if the initial values are far from the true ones, the algorithm converges always toward the same set of parameters. In this case, the optimization procedure converges in an average time of 0.073 seconds having selected random initial values for the parameters. In summary, the maximum likelihood estimation provides good estimates of the parameters for the SHARP model (both with fixed an free m and ℓ) and for the LMACP model, even if the SHARP with fixed m and ℓ presents the appealing property of being fast and simple to be estimated.

Appendix 1.D Standard Errors

Here we report the sample counterparts of Ω and H (defined in equation (1.21)) for the calculation of the standard errors though equation (1.20):

$$\left(\widehat{\Omega}_{\boldsymbol{\alpha\varphi}}\right)_{h,k} = \frac{1}{D}\sum_{d=0}^{D-1}(\nabla_{\boldsymbol{\alpha}} l_{dJ+k})_h\ (\mathbf{m}_d)_k = \frac{1}{D}\sum_{d=0}^{D-1}(\nabla_{\boldsymbol{\alpha}} l_{dJ+k})_h\,(S_{dJ+k} - \widehat{\varphi}_k),$$

with $h \in \{1,2,3\}$ and $k \in \{1,...,J\}$,

$$\left(\widehat{\Omega}_{\boldsymbol{\varphi\varphi}}\right)_{h,k} = \frac{1}{D}\sum_{d=0}^{D-1}(\mathbf{m}_d)_h(\mathbf{m}_d)_k = \frac{1}{D}\sum_{d=0}^{D-1}(S_{dJ+h} - \widehat{\varphi}_h)\,(S_{dJ+k} - \widehat{\varphi}_k),$$

with both h and k integers in $\{1,...,J\}$, and[30]

$$\widehat{H}_{\boldsymbol{\alpha\alpha}} = \frac{1}{T}\sum_{t\in\mathcal{T}}(\nabla^2_{\boldsymbol{\alpha\alpha}} l_t), \quad \widehat{H}_{\boldsymbol{\alpha\varphi}} = \frac{1}{T}\sum_{t\in\mathcal{T}}(\nabla^2_{\boldsymbol{\alpha\varphi}} l_t), \quad \widehat{\Omega}_{\boldsymbol{\alpha\alpha}} = \frac{1}{T}\sum_{t\in\mathcal{T}}(\nabla_{\boldsymbol{\alpha}} l_t)(\nabla_{\boldsymbol{\alpha}} l_t)',$$

[29]The optimization algorithm is the Sequential Quadratic Programming (SQP) and the computer is a Intel Core i5 −2450M CPU, 2.50GHz with 4 processors (including cores).

[30]Notice that $\left(\widehat{\Omega}_{\boldsymbol{\varphi\alpha}}\right) = \left(\widehat{\Omega}_{\boldsymbol{\alpha\varphi}}\right)'$.

with $T = J \cdot D$. We recognize the Hessian of the log-likelihood in $\widehat{H}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}$ and the outer-product $(\nabla_{\boldsymbol{\alpha}} l_t)(\nabla_{\boldsymbol{\alpha}} l_t)'$ of the gradient in $\widehat{\Omega}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}$.

A notable simplification comes from the fact that the first and the second derivatives of the log-likelihood as a function of the parameters $\alpha^{(s)}$, $\alpha^{(m)}$ and $\alpha^{(\ell)}$ are analytically tractable. In fact, for every couple of indexes f and l, both chosen in $\{s, m, \ell\}$, the corresponding element of the Hessian matrix is computed as

$$-\frac{\partial^2 \mathcal{L}(\boldsymbol{\vartheta})}{\partial \alpha^{(f)} \partial \alpha^{(l)}} = \sum_{t \in \mathcal{T}} \frac{S_t}{\lambda_t^2} \frac{\partial \lambda_t}{\partial \alpha^{(f)}} \frac{\partial \lambda_t}{\partial \alpha^{(l)}}, \tag{1.40}$$

with

$$\frac{\partial \lambda_t}{\partial \alpha^{(f)}} = \varphi_{j_t} \left(-1 + \widetilde{S}_{t-f:t-1} \right).$$

Besides, the score vector and the Hessian matrix can be used to simplify the numerical optimization of the likelihood. Indeed the optimization algorithms are faster and more robust with the inclusions of the derivatives.

Appendix 1.E Mixed Data Sampling (MIDAS)

The SHARP and the LMACP models are both based on values sampled on the fixed equispaced time grid $\mathcal{T}$, ignoring the values taken by S_t between two consecutive instants of $\mathcal{T}$. This inevitably brings a loss of information, since the information of the dynamics of the spread between consecutive instants of the grid is not included in the model in any way. Consider, as an example, Figure 1.E.1, where a fraction of the continuous time dynamics of the spread is shown. The red dots that appear in Figure 1.E.1 are the values taken by the spread every minute, that is, on the grid $\mathcal{T}$ with $J = 390$. It is evident that a large part of the dynamics of the spread is discarded when only one-minute data are considered.

A possible solution is to construct models that combine data with different sampling frequencies, as in the Mi(xed) Da(ta) S(ampling) regression (MIDAS) framework of Ghysels et al. [2004, 2007]. MIDAS regressions are tightly parameterized regressions

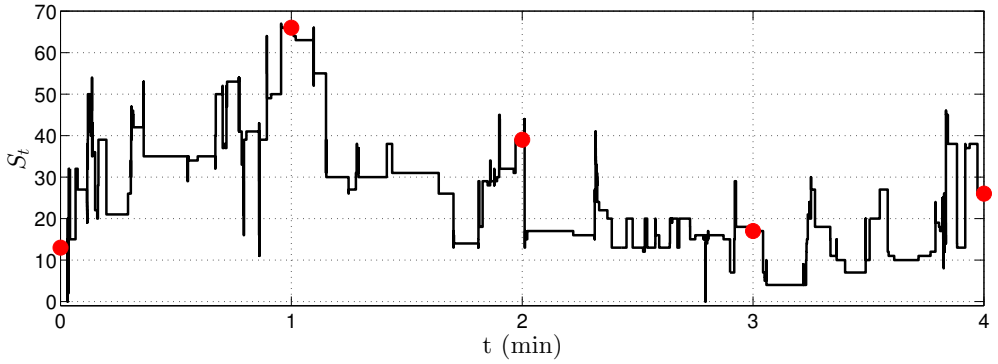

Figure 1.E.1: Spread dynamics of IBM, considering the first four minutes of January 2nd, 2014. The values taken by the spread every minute are highlighted with a red circle.

involving processes sampled at different frequencies. We introduce MIDAS regressions supposing that the de-seasonalized spread $\widetilde{S}_t$ is sampled at some fixed, say every one minute ($J = 390$), sampling frequency. Then, we consider the set of past values of the de-seasonalized spread $\widetilde{S}_t^{(r)}$ sampled r times faster, for example every 1 second ($r = 60$). Using this notation, a simple linear MIDAS regression for the de-seasonalized spread $\widetilde{S}_t$ can be written as

$$\widetilde{S}_t = c + \delta(B^{1/r})\, \widetilde{S}_{t-1}^{(r)} + \epsilon_t \tag{1.41}$$

where $\delta(B^{1/r}) = \sum_{g=0}^{g^{max}} \delta_g B^{g/r}$ is a polynomial of length g^{max} in the $B^{1/r}$ operator and $B^{g/r}\widetilde{S}_{t-1}^{(r)} = \widetilde{S}_{t-1-g/r}^{(r)}$. Equation (1.41) can be seen as a projection of the lower frequency data $\widetilde{S}_t$ onto the rich information set given by the high-frequency dynamics of the bid-ask spread.

The key advantage of the estimation of the MIDAS regression (1.41) is the fact that it is more efficient than the estimation of the typical approach of aggregating all series to the least frequent sampling [Ghysels et al., 2004], such as in the SHARP model.

Chapter 2

Adaptive Lasso for vector Multiplicative Error Models[1]

2.1 Introduction

Non-negative-valued processes are often encountered in financial time series analysis (e.g. realized volatilities, daily ranges, bid-ask spreads, order-book depths, transaction volumes, durations, etc.). The Multiplicative Error Model (MEM) [Engle, 2002] is designed to capture some key stylized facts of these time series, such as persistence and clustering. These features are addressed through a GARCH–type structure, being the MEM a product of the conditional expectation of the variable of interest times an independent and identically distributed (i.i.d.) innovation term. Apart from GARCH, this multiplicative structure encompasses several notable univariate models [e.g. Engle and Russell, 1998, Chou, 2005] and was extended to the multivariate case, called vector MEM (vMEM) [Cipollini et al., 2006, Hautsch, 2011, Cipollini et al., 2013, 2017]. In this context, the variables of interest are modeled as the Hadamard product of a vector of conditional expectations times a vector of innovations with a positive definite covariance matrix. For the latter, several choices are available: e.g. the Gamma-copula approach

[1]The material of this chapter is taken from Cattivelli and Gallo [2018].

of Cipollini et al. [2017], and the log-normal distribution introduced by Taylor and Xu [2013, 2017] in their log-vMEM specification, and adopted by Cipollini et al. [2017] in a regular vMEM. A parametric specification is avoidable, following the semiparametric GMM–based approach of Cipollini et al. [2013].

Even if the vMEM is conceptually simple, its estimation may become problematic when the dimensionality N of the vector of interest is large: as a matter of fact, leaving the model unrestricted implies estimating zero–parameters, leading to inefficient parameter estimates, and poor forecasting performances. Imposing some parameter constraints is often arbitrary and may lead to model misspecification: a General-to-Specific strategy (henceforth GtS) can be adopted as in Engle and Gallo [2006] and Cipollini and Gallo [2010], where statistically insignificant coefficients are removed over repeated estimations, in order to retain just the relevant regressors. However, this strategy is not guaranteed to find the correct set of significant covariates, since it does not explore all possible subsets of predictors, and may lead to questions about the overall size of the test.

In order to avoid model misspecification and to simplify the variable selection procedure, in this study we pursue a different approach, adopting Lasso techniques [Tibshirani, 1996], a popular shrinkage method which achieves both estimation efficiency and variable selection accuracy. We are motivated in this direction by the fact that the Lasso usually manages to have better performance in forecasting than other parameter selection techniques, when the number of regressors is large [Kock and Callot, 2014]. Compared to sequential top-down or down-top strategies, such as the GtS approach, the Lasso method is also more computationally efficient[2]. The family of Lasso procedures is nowadays extremely large and includes the renowned Adaptive Lasso, with interesting asymptotic properties [see Zou, 2006]. One of the theoretical results of this study is that the oracle property of the Adaptive Lasso (consistent in variable selection and performing as well as if the true underlying model was known in advance) is ex-

[2]With a GtS strategy the specification search is based on deleting coefficients one by one, thus envisaging a number of re-estimations in the order of $O(N^2)$ [Cipollini and Gallo, 2010], while the number of Lasso estimations is set by the researcher with no need for it to depend on N.

tended to the vMEM. Similar examples of Lasso techniques for high dimensional time series can be found in Hsu et al. [2008], Kock and Callot [2015], Audrino and Knaus [2016] and in Caner [2009] within a GMM framework.

We explore the finite sample properties of the Adaptive Lasso estimator, by running a comprehensive Monte Carlo simulation exercise: its results indicate that the Adaptive Lasso techniques achieve an interesting reduction of the Root Mean Square Error (RMSE) of the estimates and a greater precision in variable selection, while keeping low the number of estimations, relative to the GtS approach. An interesting outcome is that the procedure works better on relatively short samples, an issue which is relevant when parameter stability is at stake.

This Chapter is organized as follows. Section 2.2 briefly reviews several univariate and multivariate MEM specifications. Section 2.3 describes possible approaches for variable selection with a focus on the Adaptive Lasso technique for estimating vector Multiplicative Error Models, while Section 2.4 considers the specification of the vMEM with a log-normal conditional distribution. In Section 2.5 we investigate the finite sample properties of the Adaptive Lasso estimator in comparison with the GtS approach. Conclusions follow in Section 2.6. Finally, some technical material on the Adaptive Lasso estimator for the vMEM is presented in the Appendices 2.A and 2.B.

2.2 Multiplicative Error Models (MEM)

The univariate specification The Multiplicative Error Model specifies the evolution of a non-negative valued process y_t with a multiplicative structure, as an extension of the GARCH–type approach where the conditional expectation of a variable of interest is modeled as a combination of past observations and past conditional expectations.

The stochastic process y_t is hence described in terms of the product of a time varying conditional mean μ_t (which depends on the past values of the series) and an i.i.d. series

of non-negative valued random variables ϵ_t with unit mean ($E[\epsilon_t] = 1$), such that

$$y_t = \mu_t \, \epsilon_t. \tag{2.1}$$

The conditional mean μ_t can be specified as

$$\mu_t = \omega + \alpha \, y_{t-1} + \beta \, \mu_{t-1}, \tag{2.2}$$

with the stationarity condition $\alpha+\beta < 1$ [Engle, 2002]. Further refinements are possible, such as longer lags or terms related to asymmetric behavior in response to the sign of returns or to the direction of a trade, being a buy or a sell [Cipollini et al., 2013]. This specification has been adopted by Engle and Russell [1998] for inter-trade durations, by Manganelli [2005] for transaction volumes, by Chou [2005] for high–low range and by Engle and Gallo [2006] for the dynamic interaction of different volatility measures.

A robust specification for the density of the error term ϵ_t is given by the family of gamma densities:

$$f(\epsilon_t|\mathcal{F}_{t-1}) = \frac{1}{\Gamma(a)b^a}\epsilon_t^{a-1} \exp\left(-\frac{\epsilon_t}{b}\right). \tag{2.3}$$

Under the gamma specification, the condition $E[\epsilon_t] = 1$ is guaranteed by imposing $b = 1/a$, while the conditional expectation of the process y_t is equal to $\mathbb{E}[y_t|\mathcal{F}_{t-1}] = \mu_t$ and its conditional variance is given by $\mathbb{V}[y_t|\mathcal{F}_{t-1}] = \mu_t^2/a$. The maximum likelihood estimator for the error density function (2.3) is a quasi maximum likelihood estimators, hence consistent and asymptotically normal [Engle, 2002, Engle and Gallo, 2006].

Alternative specifications for the probability distribution of the innovations have been adopted in the literature. For example the exponential and the Weibull distributions [Engle and Russell, 1998], the generalized gamma [Dufour et al., 2000], and the lognormal [Taylor and Xu, 2013]. Also discrete-continuous mixture distributions can be adopted, as in Hautsch et al. [2013], in order to capture a clustering of observations at zero.

vector MEM (vMEM) Cipollini et al. [2007, 2013] extended the univariate MEM to a

multivariate specification (vector MEM or vMEM) introducing a contemporaneous correlation among the errors and dynamic interdependencies among the univariate series. Hence, the vMEM generalizes the univariate MEM to situations in which the process of interest is a vector of such processes. A baseline N-dimensional vMEM for the vector $\{\mathbf{y}_t : t = 1, 2, ..., T\}$ is given by:

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\mu}_t \odot \boldsymbol{\epsilon}_t \\ \boldsymbol{\mu}_t &= \boldsymbol{\omega} + A\mathbf{y}_{t-1} + B\boldsymbol{\mu}_{t-1}, \end{aligned} \tag{2.4}$$

for some initial values $\mathbf{y}_0, \boldsymbol{\mu}_0$, where A, B are $N \times N$ coefficient matrices, $\boldsymbol{\omega}$ is a $N \times 1$ vector of constant terms and $\boldsymbol{\epsilon}_t$ is a i.i.d. series of random vectors with unit mean and positive definite covariance matrix Σ. The parameters describing the interdependences among the series are stacked in the vector $\boldsymbol{\beta} = \text{vec}(D)$, with $D = (A, B)$ and $d = \#\boldsymbol{\beta} = 2N^2$.

In the current framework, it has always been a concern that the model does not produce negative forecasts: this is the main motivation behind the EGARCH [Nelson, 1991], and the log-vMEM [Taylor and Xu, 2017]. Although in the presence of negative off-diagonal values in A and B there exists an abstract risk of generating negative forecasts, in practice, rather than imposing nonnegativity constraints[3], we can allow for some moderate negative off-diagonal coefficient values capable of generating some interesting short to medium run dynamic interdependence without causing any trouble into negative territory [e.g. see Hautsch, 2008, Cipollini et al., 2017].

Two different estimation approaches for the elements in Σ, $\boldsymbol{\omega}$ and D have been developed.

First, Cipollini et al. [2007] suggested a parametric approach where the joint distribution of innovations is specified, for example through Gamma marginal probability density functions and a copula function (Normal or Student's t). After having specified a particular distribution function for the innovation term $\boldsymbol{\epsilon}_t$, the maximum likelihood

[3] Cf. the discussion in Cipollini et al. [2006].

estimates can be obtained by minimizing the negative log-likelihood $-\ell(\tilde{\boldsymbol{\beta}}, \tilde{\Sigma})$:

$$\left(\hat{\boldsymbol{\beta}}(\text{MLE}), \hat{\Sigma}(\text{MLE})\right) = \underset{\tilde{\boldsymbol{\beta}}, \tilde{\Sigma}}{\operatorname{argmin}} \left\{ -\frac{1}{T} \ell(\tilde{\boldsymbol{\beta}}, \tilde{\Sigma}) \right\}. \tag{2.5}$$

A brief discussion about possible choices of the innovation term distribution is presented in Section 2.4.

Then, in Cipollini et al. [2013] a semiparametric specification has been proposed, based on the Generalized Method of Moments, avoiding the need of choosing a density function for the innovations.

In general, the estimation procedure for the vector $\boldsymbol{\omega}$ can be greatly simplified if the underlying model is assumed to be stationary. With expectation targeting [mirroring the variance targeting of Engle and Mezrich [1996] for GARCH; for details see Cipollini et al., 2013] we can express $\boldsymbol{\omega}$ in equation (2.4) in terms of the unconditional mean $\boldsymbol{\mu} = \mathbb{E}[\mathbf{y}_t]$ as

$$\boldsymbol{\omega} = (\mathbb{I} - A - B)\boldsymbol{\mu},^{4}$$

where the sample average $\hat{\boldsymbol{\mu}} = \sum_{t=1}^{T} \mathbf{y}_t / T$ can be substituted for the unknown $\boldsymbol{\mu}$.

2.3 Variable selection for the vMEM

The matrices A and B are the main focus of our study. These matrices describe how information at time $t-1$ influences the conditional expectations of $\mathbf{y}$ at time t. Possible spillovers effects are described by the off-diagonal elements of A and B: a non-zero element ($A_{i,j}$ and/or $B_{i,j}$) denotes the presence of a link from the series j to series i, where the former denotes the impact of the most recent observation, while the latter measures the inertia in the evolution of the conditional expectation.

However, the number of parameters in the matrices A and B increases quadratically in the dimension N of the process $\boldsymbol{y}_t$ and the the data-generating-process often contains zero parameters in these matrices which leads to inefficient parameter estimates and

[4] The symbol $\mathbb{I}$ denotes the identity matrix.

poor forecasting performances, if these parameters are not correctly estimated as being zeros. This issue has been addressed by Cipollini and Gallo [2010], who proposed an algorithm for choosing the relevant variables in a vMEM model. Starting from a "large" formulation where the matrices A and B are full, they check which coefficients are the least significant (according to the t test statistic) and repeat the estimation procedure deleting them one at a time. The algorithm stops when only significant variables are included in the formulation.

However, this General-to-Specific (GtS) strategy has a drawback: it is not guaranteed to find the correct set of significant covariates since it does not explore all possible subsets of predictors. Indeed, there are several controversies surrounding this type of top-down model selection procedures [Phillips, 2005].

A valid alternative is given by Lasso techniques [Tibshirani, 1996]. The Lasso is a shrinkage method that selects the model and estimates the parameters simultaneously. The Lasso estimates can be found by minimizing the negative log-likelihood $-\frac{1}{T}\ell(\tilde{\boldsymbol{\beta}}, \tilde{\Sigma})$ with a (possibly weighted) Lasso penalty term $\lambda_T \sum_{j=1}^{d} |\tilde{\beta}_j|$, that is:

$$\left(\hat{\boldsymbol{\beta}}(\lambda_T), \hat{\Sigma}(\lambda_T)\right) = \underset{\tilde{\boldsymbol{\beta}}, \tilde{\Sigma}}{\operatorname{argmin}} \left\{ -\frac{1}{T}\ell(\tilde{\boldsymbol{\beta}}, \tilde{\Sigma}) + \lambda_T \sum_{j=1}^{d} |\tilde{\beta}_j| \right\}. \tag{2.6}$$

The penalty term has the ability to shrink the coefficients $\hat{\beta}_j$ toward zero,[5] as shown in figure 2.1, and hence gives interpretable results [Tibshirani, 1996]. Another important element of this procedure is the parameter λ_T, which selects the amount of shrinkage. The simple case in which λ_T is equal to zero is completely equivalent to a standard maximum likelihood approach, while the asymptotic case $\lambda_T \to \infty$ produces zero estimates for all parameters, because of the unique minimum of the penalisation term $\sum_{j=1}^{d} |\tilde{\beta}_j|$ in $\tilde{\boldsymbol{\beta}} = \mathbf{0}$. Usually, the shrinkage parameter λ_T is selected through cross validation.

However, the Lasso is consistent in variable selection only under strong assumptions [Zou, 2006], producing also biased (downward) estimates for large coefficients. The

[5] In the empirical exercises that follow, we will consider a parameters to be a zero if its estimate is smaller, in absolute value, than two times the numerical tolerance of the optimizer of (2.6)

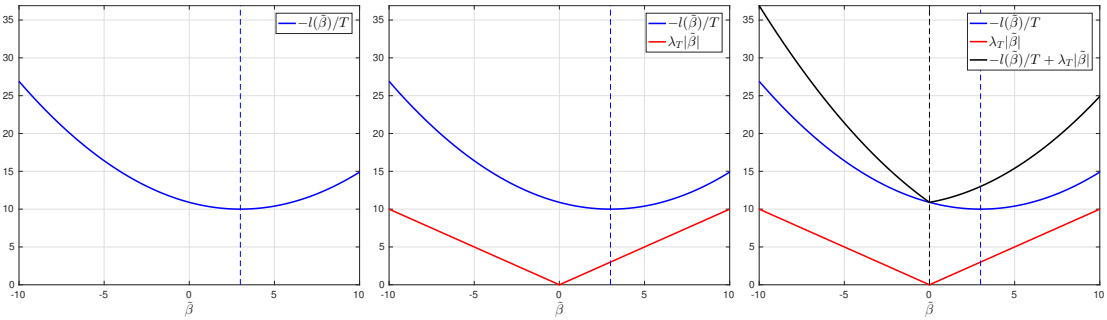

Figure 2.1: On the left: a log-likelihood function $\frac{1}{T}\ell(\tilde{\beta})$, multiplied by -1, with minimum in $\tilde{\beta} = 3$. In the middle: the log-likelihood and the Lasso term $\lambda_T\tilde{\beta}$, with $\lambda_T = 1$. On the right: the log-likelihood plus the Lasso term, whose minimum is now in $\tilde{\beta} = 0$.

Adaptive Lasso estimator addresses these issues by introducing the following weighted penalisation

$$\left(\hat{\boldsymbol{\beta}}(\lambda_T), \hat{\Sigma}(\lambda_T)\right) = \underset{\tilde{\boldsymbol{\beta}}, \tilde{\Sigma}}{\operatorname{argmin}} \left\{ -\frac{1}{T}\ell(\tilde{\boldsymbol{\beta}}, \tilde{\Sigma}) + \lambda_T \sum_{j=1}^{d} \hat{w}_j |\tilde{\beta}_j| \right\}, \tag{2.7}$$

where the weights $\hat{w}_j = 1/|\hat{\beta}_j(\text{MLE})|^\gamma$, for some $\gamma > 0$, are selected in such a way to obtain large shrinkage for small parameters (that is, for small $|\hat{\beta}_j(\text{MLE})|$) and small shrinkage for large coefficients. Insignificant parameters are rapidly narrowed to zero while the estimation precision of significant coefficients is not affected. Zou [2006] shows that the Adaptive Lasso for linear models and generalized linear models possesses the oracle property, that is, it is consistent in variable selection and performs as well as if the true underlying model were given in advance. The shrinkage parameter λ_T can be selected with a particular cross-validation approach, as shown in Section 2.5.

In the following theorem we show under which conditions the Adaptive Lasso for vMEM (AL-vMEM) is consistent in variable selection and performs as well as if the true underlying model were given in advance. Since we are dealing with a variable selection issue, we assume that the true model depends only on a subset of the predictors $\mathcal{D} = \{j : \beta_j \neq 0\}$, with $\#\mathcal{D} = d_{\mathcal{D}} \leq d$. For clarity, let us state first some technical assumptions needed for Theorem 2 below:

Assumption 1. *For all $i, j, k \in \{1, 2, ..., d\}$ and all $t = 1, ..., T$, we assume the following:*

- *the function $\tilde{\boldsymbol{\beta}} \to \log f(\mathbf{y}_t \,|\, \mathcal{F}_{t-1}, \tilde{\boldsymbol{\beta}})$ is of class C^3, where $f(\mathbf{y}_t \,|\, \mathcal{F}_{t-1}, \tilde{\boldsymbol{\beta}})$ is the conditional distribution of $\mathbf{y}_t$. Hence, the log-likelihood function $\ell(\tilde{\boldsymbol{\beta}})$ is given by $\ell(\tilde{\boldsymbol{\beta}}) = \sum_{t=1}^{T} \log\Big(f(\mathbf{y}_t \,|\, \mathcal{F}_{t-1}, \tilde{\boldsymbol{\beta}})\Big)$;*
- *the Fisher information matrix $\mathcal{I}(\boldsymbol{\beta})$, defined as $(\mathcal{I}(\boldsymbol{\beta}))_{i,j} = -\mathbb{E}[\frac{\partial^2 \ell(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\beta}_i \partial \tilde{\beta}_j}|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}}]/T = -\mathbb{E}[\frac{\partial^2 f(\mathbf{y}_t \,|\, \tilde{\boldsymbol{\beta}})}{\partial \tilde{\beta}_i \partial \tilde{\beta}_j}|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}}]$, is finite and positive definite;*
- *there exist some function Z_{ijk} and an open set Ω that contains the true parameter $\boldsymbol{\beta}$ such that $\forall \tilde{\boldsymbol{\beta}} \in \Omega$*

$$\left| \frac{\partial^3 \log f(\mathbf{y}_t \,|\, \mathcal{F}_{t-1}, \tilde{\boldsymbol{\beta}})}{\partial \tilde{\beta}_i \partial \tilde{\beta}_j \partial \tilde{\beta}_k} \right| \leq Z_{ijk}(\mathbf{y}_t) < \infty,$$

with $\mathbb{E}[Z_{ijk}(\mathbf{y}_t)] < \infty$.

Theorem 2. *Assume that the process $\mathbf{y}_t$ follows the stationary vMEM of equation* (2.4)*. Let $\boldsymbol{\epsilon}_t$ be a sequence of unit mean i.i.d. random vectors with finite and positive definite covariance matrix Σ. Under Assumptions 1, if $\lambda_T \sqrt{T} \to 0$ and $\lambda_T T^{(\gamma+1)/2} \to \infty$, then the Adaptive Lasso estimator $\hat{\boldsymbol{\beta}}(\lambda_T)$ is*

- *asymptotically normal in $\mathcal{D}$: $\sqrt{T}(\hat{\boldsymbol{\beta}}_{\mathcal{D}}(\lambda_T) - \boldsymbol{\beta}_{\mathcal{D}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\beta}_{\mathcal{D}}))$,*
- *consistent in variable selection: $\lim_{T\to\infty} \mathbb{P}[\hat{\mathcal{D}}(\lambda_T) = \mathcal{D}] = 1$,*

where $\mathcal{I}(\boldsymbol{\beta}_{\mathcal{D}})$ is the Fisher Information matrix for the vector of non-zero parameters $\boldsymbol{\beta}_{\mathcal{D}} = \{\beta_j : j \in \mathcal{D}\}$, with $\hat{\boldsymbol{\beta}}_{\mathcal{D}}(\lambda_T) = \{\hat{\beta}_j(\lambda_T) : j \in \mathcal{D}\}$ and $\hat{\mathcal{D}}(\lambda_T) = \{j : \hat{\beta}_j(\lambda_T) \neq 0\}$.

Proof. See Appendix 2.A.

As T goes to infinity, the shrinkage parameter λ_T must go to zero rapidly ($\lambda_T \sqrt{T} \to 0$) but not too fast ($\lambda_T T^{(\gamma+1)/2} \to \infty$). For example, in the important case $\gamma = 1$, we have that $\lambda_T \sim T^{-\delta}$, with $0.5 < \delta < 1$.

Finally, let us emphasize that Theorem 2 can be easily extended to the case in which the specification of the conditional mean $\boldsymbol{\mu}_t$ (in the second line of (2.4)) is generalized

with the adoption of a monotonic differentiable link function $g(\boldsymbol{\mu}_t)$ and a larger set of predictors, as in the logarithmic specification of Bauwens and Giot [2000], adopted also in Taylor and Xu [2017], where $\log \boldsymbol{\mu}_t = \boldsymbol{\omega} + \sum_{i=1}^{p} A_i \log \mathbf{y}_{t-1} + \sum_{j=1}^{q} B_q \log \boldsymbol{\mu}_{t-1}$.

2.4 Multivariate log-normal distribution

Only a few joint probability distributions for non-negative valued random vectors are available. This fact complicates the effort in finding a good specification for the multivariate distribution of the error terms $\boldsymbol{\epsilon}_t$. In this context, Cipollini et al. [2017] suggest the adoption of copula functions together with Gamma marginals. As an alternative, they also explore the case in which the innovation terms $\boldsymbol{\epsilon}_t$ are log-normally distributed as initially introduced by Taylor and Xu [2017] in their log-vMEM specification.

We further develop this approach, starting from the general definition

$$\mathbf{y}_t \oslash \boldsymbol{\mu}_t \equiv \boldsymbol{\epsilon}_t \,|\, \mathcal{F}_{t-1} \sim \text{LN}(\mathbf{m}, V), \tag{2.8}$$

where $\mathbf{m}$ is a location vector and V is a scale matrix. Under the log-normal specification (2.8) and the need for a unit expected value, the mean $\mathbb{E}[\boldsymbol{\epsilon}_t] = \exp\left(\mathbf{m} + 0.5\text{diag}(V)\right)$ (exp, the exponential function, is applied element by element) and the variance-covariance matrix $\mathbb{V}[\boldsymbol{\epsilon}_t] = \Sigma$ of the error term $\boldsymbol{\epsilon}_t$ are given by

$$\mathbb{E}[\boldsymbol{\epsilon}_t] = \mathbf{1} \Leftrightarrow \mathbf{m} = -0.5\,\text{diag}(V), \qquad \Sigma = (\exp(V) - \mathbf{1}\mathbf{1}'). \tag{2.9}$$

The resulting log-likelihood function turns out to be equal to

$$\begin{aligned} \ell(\tilde{\boldsymbol{\beta}},\tilde{V}) = &- G - \frac{T}{2}\log\left(\det(\tilde{V})\right) - \\ &\frac{1}{2}\sum_{t=1}^{T}\left(\log \boldsymbol{\epsilon}_{t,\tilde{\boldsymbol{\beta}}} + 0.5\,\text{diag}(\tilde{V})\right)' \tilde{V}^{-1}\left(\log \boldsymbol{\epsilon}_{t,\tilde{\boldsymbol{\beta}}} + 0.5\,\text{diag}(\tilde{V})\right), \end{aligned} \tag{2.10}$$

where

$$G = \sum_{t=1}^{T} \sum_{n=1}^{N} y_{n,t} - \frac{T\,N}{2} \log(2\pi)$$

and $\log \boldsymbol{\epsilon}_{t,\tilde{\boldsymbol{\beta}}} = \log \mathbf{y}_t - \log \boldsymbol{\mu}_{t,\tilde{\boldsymbol{\beta}}}$ is the log (again, element by element) of the error vector for a given value of $\tilde{\boldsymbol{\beta}}$. Consistency and asymptotic normality of the maximum likelihood estimator based on the log-likelihood function (2.10) are discussed in Taylor and Xu [2017], in which the explicit expressions of the score vector and of the Hessian matrix are given.[6] However, as for GARCH models, given the recursive nature of the vMEM of equation (2.4), the log-likelihood (2.10) depends also on all the past values of the process $\mathbf{y}_t$ with $t \leq 0$. For practical purposes, some initial values for $\boldsymbol{\mu}_0$ have to be assumed: here we rely on the customary initialization with the sample mean, that is, $\boldsymbol{\mu}_0 = \hat{\boldsymbol{\mu}}$, informally extending the results in the GARCH literature [e.g. Lee and Hansen, 1994, Francq and Zakoïan, 2004, Straumann and Mikosch, 2006] that the effects of such assumptions are asymptotically negligible, given that the estimation period is usually long enough to make the initial observations ininfluent on the overall estimates.

The optimization of the log-likelihood (2.10) is complicated by its explicit dependence on the variance-covariance matrix $\tilde{V}$. Cipollini et al. [2017] circumvented this issue by proposing an iterative procedure based on a method of moment estimator, adopting as moments the score function in $\boldsymbol{\beta}$ and the variance-covariance matrix of the log-residuals. Here we propose a different approach which does not make use of iterative solutions, but is rather based on the following maximum likelihood estimator

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\text{MLE}) = \underset{\tilde{\boldsymbol{\beta}}}{\operatorname{argmin}} \Bigg\{ & \frac{G}{T} + \frac{1}{2} \log\left(\det(\hat{V}_{\tilde{\boldsymbol{\beta}}})\right) + \\ & \frac{1}{2T} \sum_{t=1}^{T} \left(\log \boldsymbol{\epsilon}_{t,\tilde{\boldsymbol{\beta}}} - \hat{\mathbf{m}}_{\tilde{\boldsymbol{\beta}}}\right)' \hat{V}_{\tilde{\boldsymbol{\beta}}}^{-1} \left(\log \boldsymbol{\epsilon}_{t,\tilde{\boldsymbol{\beta}}} - \hat{\mathbf{m}}_{\tilde{\boldsymbol{\beta}}}\right) \Bigg\}, \end{aligned} \tag{2.11}$$

with (for details cf. Appendix 2.B)

$$\left(\hat{V}_{\tilde{\boldsymbol{\beta}}}\right)_{ij} = \left(S_{\tilde{\boldsymbol{\beta}}}\right)_{ij} - \left(\sqrt{1 + \left(S_{\tilde{\boldsymbol{\beta}}}\right)_{ii}} - 1\right) \left(\sqrt{1 + \left(S_{\tilde{\boldsymbol{\beta}}}\right)_{jj}} - 1\right), \tag{2.12}$$

[6] Our specification and theirs differ in how the term $\log(\boldsymbol{\mu}_t)$ is treated.

and

$$S_{\tilde{\boldsymbol{\beta}}} = \frac{1}{T}\sum_{t=1}^{T} \log \boldsymbol{\epsilon}_{t,\tilde{\boldsymbol{\beta}}} \left(\log \boldsymbol{\epsilon}_{t,\tilde{\boldsymbol{\beta}}}\right)', \quad \hat{\mathbf{m}}_{\tilde{\boldsymbol{\beta}}} = -0.5\,\mathrm{diag}(\hat{V}_{\tilde{\boldsymbol{\beta}}}) \tag{2.13}$$

The minimum in (2.11) can be obtained thanks to the explicit dependence of $\hat{V}_{\tilde{\boldsymbol{\beta}}}$ on $\tilde{\boldsymbol{\beta}}$, as described in equations (2.12)-(2.13).[7] Then, the estimate of V is found from (2.12) substituting $\tilde{\boldsymbol{\beta}}$ with its MLE estimate $\hat{\boldsymbol{\beta}}(\mathrm{MLE})$, that is

$$\left(\hat{V}_{\hat{\boldsymbol{\beta}}(\mathrm{MLE})}\right)_{ij} = \left(S_{\hat{\boldsymbol{\beta}}(\mathrm{MLE})}\right)_{ij} - \left(\sqrt{1+\left(S_{\hat{\boldsymbol{\beta}}(\mathrm{MLE})}\right)_{ii}} - 1\right)\left(\sqrt{1+\left(S_{\hat{\boldsymbol{\beta}}(\mathrm{MLE})}\right)_{jj}} - 1\right).$$

This estimator is asymptotically consistent, as discussed in Appendix 2.B and exact in the one-dimensional case $N = 1$.

The corresponding Adaptive Lasso estimator is obtained substituting (2.11) with[8]

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\lambda_T) = \underset{\tilde{\boldsymbol{\beta}}}{\mathrm{argmin}} \Bigg\{ & \frac{G}{T} + \frac{1}{2}\log\left(\det(\hat{V}_{\tilde{\boldsymbol{\beta}}})\right) + \\ & \frac{1}{2T}\sum_{t=1}^{T}\left(\log \boldsymbol{\epsilon}_{t,\tilde{\boldsymbol{\beta}}} - \hat{\mathbf{m}}_{\tilde{\boldsymbol{\beta}}}\right)' \hat{V}_{\tilde{\boldsymbol{\beta}}}^{-1}\left(\log \boldsymbol{\epsilon}_{t,\tilde{\boldsymbol{\beta}}} - \hat{\mathbf{m}}_{\tilde{\boldsymbol{\beta}}}\right) + \\ & \lambda_T \sum_{j=1}^{d} \hat{w}_j |\tilde{\beta}_j| \Bigg\}.\end{aligned} \tag{2.14}$$

Now, we discuss the computation of the standard errors of the Adaptive Lasso estimator. Following Fan and Li [2001] and Zou [2006], we estimate the variance-covariance matrix for the non-zero components of $\hat{\boldsymbol{\beta}}(\lambda_T)$ with the sandwich formula

$$\begin{aligned}\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_{\mathcal{D}}(\lambda_T)) = & \left\{\nabla^2\ell(\hat{\boldsymbol{\beta}}_{\mathcal{D}}(\lambda_T)) + T\,\Gamma(\hat{\boldsymbol{\beta}}_{\mathcal{D}}(\lambda_T))\right\}^{-1} \widehat{\mathrm{cov}}(\nabla\ell(\hat{\boldsymbol{\beta}}_{\mathcal{D}}(\lambda_T))) \\ & \left\{\nabla^2\ell(\hat{\boldsymbol{\beta}}_{\mathcal{D}}(\lambda_T)) + T\,\Gamma(\hat{\boldsymbol{\beta}}_{\mathcal{D}}(\lambda_T))\right\}^{-1},\end{aligned} \tag{2.15}$$

[7] The analytical expression of the score and of the Hessian matrix of the log-likelihood (2.11) are complicated by the presence of the derivative of the term $\hat{V}_{\tilde{\boldsymbol{\beta}}}^{-1}$, so that it is more expedient to rely on numerical derivatives in the optimization of the objective functions.

[8] For practical purposes, as we will see later, for the optimization of the L-1 regularized objective function (2.14), we employ the SQP algorithm which is shown to minimize the iteration costs among a large number of optimization algorithms [Schmidt et al., 2007].

where $\Gamma(\hat{\boldsymbol{\beta}}(\lambda_T)) = \lambda_T \, \text{diag}(\hat{w}_1/\hat{\beta}_1(\lambda_T), ..., \hat{w}_d/\hat{\beta}_d(\lambda_T))$ and the covariance $\widehat{\text{cov}}(\nabla \ell(\hat{\boldsymbol{\beta}}_{\mathcal{D}}(\lambda_T)))$ can be estimated using the outer product of the gradient. For the zero components $\hat{\beta}_j(\lambda_T) = 0$, the estimated standard errors are equal to zero. Fan et al. [2004] showed that this sandwich formula is a consistent estimator of the variance-covariance matrix $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}_{\mathcal{D}}(\lambda_T))$. Being a sandwich estimator, it is also robust to the misspecification of the conditional distribution of the vector of interest.

2.5 Monte Carlo simulations

While the oracle property is an asymptotic property, it is of interest to verify the behavior of the Adaptive Lasso estimator (2.14) with respect to the simpler *ML* estimator (2.11) in finite samples. By the same token, the possible presence of zeros makes the GtS approach another useful benchmark to evaluate our proposal. With this aim, we design a simulation exercise to compare the performance of these estimators in realistic situations in which only a finite number of observations is available. We focus on the sparse case in which the matrices A and B are characterized by the presence of many off-diagonal zero-elements. We simulate 500 time series of length $T = 736$ with a 7-dimensional vector $\mathbf{y}_t$ ($N = 7$)[9] modeled as a vMEM (equation (2.4)) with log-normal error terms $\boldsymbol{\epsilon}_t$. The estimation procedure consists of three steps:

1. *Maximum likelihood estimation*: we get the estimates $\hat{\boldsymbol{\beta}}(\text{MLE})$ from (2.11) using the full window of 736 observations, obtaining the weights $\hat{w}_j$. We focus on the case $\gamma = 1$, as in Audrino and Camponovo [2013], Audrino and Knaus [2016].[10]

2. *Evaluation of the optimal shrinkage parameter* $\hat{\lambda}_T$: following equation (2.14), and using only the first 600 observations, we scan candidate Adaptive Lasso parameters over a grid of $n_\lambda = 30$ different values of λ_T. Among these, the optimal shrinkage

[9] T was chosen equal to 736 in order to reproduce the length of the realized volatility series analyzed in Section 3.2.

[10] To avoid numerical instabilities and possible divergences in the weights $\hat{w}_j$, when $|\hat{\beta}_j(\text{MLE})|$ is smaller than a tiny threshold β_{min}, the weights are calculated with β_{min} in place of the ML estimate, that is $\hat{w}_j = 1/\max(|\hat{\beta}_j(\text{MLE})|, \beta_{min})$. In our empirical applications, we impose $\beta_{min} = 0.005$.

parameter $\hat{\lambda}_T$ is chosen to be the one delivering the smallest distance between the remaining 136 observations and the one-step ahead forecasts $\{\hat{\mathbf{y}}_{t|t-1}(\lambda_T)\}$ according to the following log-normal loss function:

$$L_{LN}(\mathbf{y}_t, \hat{\mathbf{y}}_{t|t-1}) = \log\left(\det(V)\right) + \left(\log \mathbf{y}_t \oslash \hat{\mathbf{y}}_{t|t-1} + \frac{1}{2}\text{diag}V\right)' V^{-1} \left(\log \mathbf{y}_t \oslash \hat{\mathbf{y}}_{t|t-1} + \frac{1}{2}\text{diag}V\right). \tag{2.16}$$

Given that the true value of the variance covariance matrix V cannot be known in empirical application, we use the estimated $\hat{V}_{\hat{\boldsymbol{\beta}}(\lambda_T)}$.

3. *Adaptive Lasso estimation*: we get the estimates $\hat{\boldsymbol{\beta}}(\hat{\lambda}_T)$ from (2.14) using the full window and fixing the optimal shrinkage parameter $\hat{\lambda}_T$ at the value obtained in the second step.

In the second step, for the choice of the grid of possible values of λ_T we adopt a heuristic rule, fixing a maximum value λ_{max}, and then taking a equally spaced grid between $\lambda_{\text{max}}/n_\lambda$ and λ_{max}. The maximum value λ_{max} is taken to be proportional to the log-likelihood calculated in the first step and inversely proportional to the number of parameters d and to the number of observations T:

$$\lambda_{\text{max}} = \frac{|\ell(\hat{\boldsymbol{\beta}}(\text{MLE}))|}{T\, d\, v}, \tag{2.17}$$

with a granularity parameter v selected to be small enough as to deliver a U-shape loss function (2.16) (for an illustration see Figure 3.2). Such a choice of the upper bound in equation (2.17) makes the Lasso term $\lambda_T \sum_{j=1}^{d} \hat{w}_j |\tilde{\beta}_j|$ (which is of order $\sim \lambda_T\, d$ when $\gamma = 1$) proportional to the likelihood term $\frac{1}{T}\ell(\tilde{\boldsymbol{\beta}}, \tilde{\Sigma})$ in equation (2.7).

Two Monte Carlo exercises are carried out,[11] respectively designed to

[11] In the following Monte Carlo exercises, we impose some standard constraints in order to improve the performances of all the estimators we are considering. In particular, the elements of A and B are assumed to be smaller than one and bigger than -0.5. The diagonal elements are assumed to be non-negative. The same constraints will be also applied to the empirical results of Sections 3.2 and 3.4.

- investigate the Lasso estimator's capability to recover the situation of complete lack of interdependence, i.e., when the matrices A and B are diagonal;

- verify whether the Lasso estimator is able to identify a sparse representation of the matrices A and B, having 63 zeros over the $d = 2 \times N^2 = 98$ parameters.

The values of the coefficients are selected to be in line with previous studies in volatility clustering (relatively low A coefficients and larger values on the main diagonals, especially of B) in the presence of a strong contemporaneous correlation V.[12] The variable selection procedure is confined within the off-diagonal parameters [as in Cipollini and Gallo, 2010]. This assumption will be particularly important in the empirical out-of-sample exercise of Section 3.4, where the diagonal elements are known to be the most important regressors and can not be removed if an optimal forecasting accuracy is required. The parameter v of equation (2.17) is equal to 1 for the diagonal case and equal to 5 for the sparse one. The estimation statistics for the matrix of coefficients A used in the simulation exercise can be found, for the two cases, respectively in Tables 2.1 and 2.3.[13]

In these tables we also provide two important comparisons by repeating the same scheme to explore:

- the parallel performance of the General-to-Specific approach of Cipollini and Gallo [2010], making it start from a "large" formulation where the matrices A and B are full, and then deleting, iteration by iteration, the one which has the smallest t test statistic. The algorithm stops when only significant variables (at $\alpha = 5\%$) are left.

- the effects of a misspecified conditional distribution on the finite sample properties of the Adaptive Lasso estimator (2.14): here we extract the vector $\boldsymbol{\epsilon}_t$ from a

[12] To be realistic, we take values for V from the corresponding estimate found later in Section 3.2 for the case of seven European indices in the period 2010-2012.

[13] The statistics for the matrix B are not shown because they are very similar to these of A, but are available from the authors upon request.

multivariate distribution with Gamma marginals and a Student's t copula.[14]

To complement the analysis, we build two parallel tables (Tables 2.2 and 2.4), where we suggest an evaluation based on the percentage of false and true zeros detected by individual t-tests at various significance levels, with those delivered by the GtS, the Adaptive Lasso and the Adaptive Lasso where the error distribution is misspecified.

In the diagonal case, the t-tests at the 5% level[15], the General-to-Specific approach, the Adaptive Lasso estimator and its misspecified version have a large ($\sim 92 - 94\%$) probability of detecting the zeros on the off-diagonal elements (see Table 2.2). However, there is a notable difference in the average RMSE of these approaches. Indeed, the average RMSE gain over the *MLE*, defined as

$$\text{Gain}_i = 100 \times \frac{\text{RMSE}_{MLE} - \text{RMSE}_i}{\text{RMSE}_{MLE}}, \tag{2.18}$$

with i equal to a considered method, is equal to 79.22 for the GtS approach, 92.87 for the Adaptive Lasso estimator and 92.84 for the misspecified AL.

In the sparse configuration, the Adaptive Lasso detects the 63 zeros with an average success rate of 67.39%, while, only the smaller parameters (especially the negative ones) are wrongly classified as being zeros. Individual t-tests with critical values at the 40% perform significantly worse (see Table 2.4). If larger critical values are considered, the frequency of correctly detected zeros increases, outperforming the Adaptive Lasso. However, also the frequency of wrong zeros increases, rapidly reaching and exceeding 50%. The GtS approach has a larger probability of finding both the true zeros and the false zeros compared to the Adaptive Lasso approach, making the comparison difficult. Hence, we consider also a non-standard alternative with a different significance level ($\alpha = 10\%$), called GtS 10%, which is out-performed by the Adaptive Lasso and by the misspecified AL, as shown in the summary results of Table 2.4. Furthermore, the mean squared errors of the estimated parameters of the Adaptive Lasso estimator are signif-

[14] In the t copula, the number of degrees of freedom ν was set equal to 8, as in the empirical results of Cipollini et al. [2017].The covariance matrix Σ has been set equal to the one of the log-normal case.

[15] The t-stats have been evaluated considering the value of the parameters $\hat{\boldsymbol{\beta}}(\text{MLE})$ divided by their standard errors.

Matrix A

True values

0.258	0	0	0	0	0	0
0	0.278	0	0	0	0	0
0	0	0.244	0	0	0	0
0	0	0	0.291	0	0	0
0	0	0	0	0.253	0	0
0	0	0	0	0	0.245	0
0	0	0	0	0	0	0.264

RMSE: *MLE*

0.076	0.088	0.084	0.088	0.072	0.096	0.080
0.065	0.085	0.075	0.077	0.066	0.085	0.072
0.063	0.081	0.076	0.078	0.066	0.083	0.072
0.052	0.067	0.062	0.065	0.053	0.066	0.059
0.058	0.078	0.072	0.069	0.063	0.077	0.063
0.048	0.061	0.058	0.059	0.049	0.069	0.055
0.076	0.092	0.085	0.090	0.076	0.098	0.088

Percentage of zeros: GtS

0	95.2	96.2	96.0	96.0	96.2	96.0
94.0	0	95.4	94.6	94.4	96.6	95.6
95.4	94.8	0	96.8	93.4	95.0	93.8
94.4	95.0	94.0	0	95.2	96.0	92.8
94.8	95.8	95.2	95.2	0	96.4	94.6
95.4	94.6	95.4	95.4	94.4	0	95.6
95.2	95.8	96.0	95.8	94.2	94.0	0

GtS percentage gain over *MLE*

69.6	82.7	80.8	79.5	79.5	84.7	81.4
82.8	74.0	79.9	80.7	79.0	87.9	83.0
82.3	83.7	69.4	81.2	78.4	84.0	80.3
79.2	82.3	78.3	56.6	79.8	78.8	74.9
80.1	84.1	81.8	80.6	61.3	81.2	77.9
83.2	84.8	85.4	81.1	75.6	67.5	80.6
82.5	86.7	81.0	79.0	80.3	80.9	71.2

Percentage of zeros: Adaptive Lasso

0	93.0	94.0	93.0	95.2	94.0	95.8
93.8	0	96.0	93.6	95.8	93.2	91.6
94.8	93.6	0	93.0	94.2	93.8	93.4
95.8	93.6	95.8	0	94.0	95.8	96.2
95.6	93.4	94.6	93.8	0	95.4	96.0
95.4	93.0	95.6	93.0	95.4	0	95.4
94.4	93.4	96.2	93.6	95.8	94.4	0

A. Lasso percentage gain over *MLE*

79.3	96.8	96.1	94.8	96.5	96.1	96.3
95.3	81.0	96.5	95.7	96.7	96.1	95.5
94.3	96.0	77.3	96.9	94.5	94.8	93.7
95.3	95.0	96.4	71.4	95.3	96.0	96.1
96.1	94.7	94.0	95.1	70.0	96.5	95.4
92.9	96.3	95.8	95.2	92.8	75.5	96.7
96.1	96.9	97.1	95.5	96.4	95.7	80.3

Percentage of zeros: misspecified Adaptive Lasso

0	92.4	95.6	96.4	95.4	94.8	96.8
95.2	0	94.4	95.6	95.2	95.0	95.0
94.8	95.0	0	95.2	96.2	95.0	94.6
96.2	93.8	96.4	0	96.0	95.4	95.4
96.2	95.4	96.0	97.6	0	97.8	95.6
95.6	93.0	97.4	94.6	94.8	0	96.6
96.4	92.0	97.2	97.0	95.4	94.2	0

Missp. A. Lasso percentage gain over *MLE*

76.9	95.3	95.4	93.3	94.7	96.3	96.9
94.7	80.7	95.6	97.1	97.2	96.9	94.8
95.1	95.9	77.0	95.6	95.6	94.5	95.7
96.4	96.9	94.2	71.6	93.4	94.5	95.6
94.8	96.3	97.1	97.0	69.7	93.7	94.7
96.2	96.1	97.6	96.8	95.8	74.0	96.8
96.8	95.2	96.6	96.7	95.8	94.3	79.6

Table 2.1: Monte Carlo exercise. Diagonal case, $T = 736$. Top-left panel: the elements of the matrix A used for Monte Carlo simulations. Top-right panel: the Root Mean Square Error (RMSE) of the *MLE*. Remaining panels on the left: the percentage of times in which the estimator finds a zero for the corresponding parameter. Remaining panels on the right: percentage gain in RMSE over the *MLE* (2.18).

t-test,40%	t-test,20%	t-test,10%	t-test,5%	GtS	A. Lasso	Missp. A. Lasso
56.42%	76.24%	87.19%	92.96%	94.83%	94.07%	94.91%

Table 2.2: The percentage of true zeros (that is, zeros detected in the off-diagonal elements of A and B) identified by individual t tests, by the General-to-Specific approach, by the Adaptive Lasso estimator and by the Adaptive Lasso estimator with a misspecified conditional distribution. A and B are diagonal matrices (see Table 2.1) and $T = 736$.

icantly smaller with respect to the *ML* and the GtS approaches. Indeed, the average gain in RMSE is equal to 43.29 for the GtS approach, 27.32 for the GtS 10%, 58.72 for the Adaptive Lasso estimator and 57.71 for the misspecified AL.

In order to check the performance of the Adaptive Lasso estimator with larger sample sizes, we simulate 500 time series of length $T = 1500$. For the evaluation of the optimal shrinkage parameter $\hat{\lambda}_T$, the model is estimated using the first 1300 observations with 30 different values of λ_T[16]. The optimal shrinkage parameter $\hat{\lambda}_T$ is the value of λ_T with the smallest log-normal distance (2.16) between the remaining 200 observations and the forecasts $\{\hat{\mathbf{y}}_{t|t-1}(\lambda_T)\}$.

Also in this case the MSE of the Adaptive Lasso approach are considerably smaller compared to the other approaches (see Table 2.5), more specifically the average RMSE gain is equal to 30.26 for the GtS approach, 21.45 for the GtS 10%, 45.36 for the Adaptive Lasso estimator and 43.65 for the misspecified AL, while its performance in variable selection is comparable to the one of the GtS 10% algorithm (see Table 2.4 and 2.5).

In conclusion, we enter the empirical illustration of our proposal with a strong evidence that the Adaptive Lasso estimator has the capability to discriminate between a diagonal and a sparse configuration for A and B, especially in the case of a limited number of observations, even in the presence of a misspecified error distribution. With more observations, the gap is reduced but the main advantage of the Adaptive Lasso is still to reduce the MSE of the estimates by comparison with the other existing approaches.

The comparison between the Adaptive Lasso and the GtS approaches is also related to the issue of the computational efficiency. In fact, GtS needs to perform a search by deleting coefficients one by one with a potential number of re-estimations of the order of $N(N-1)$ for each matrix (A and B), so for even small N, say $N = 7$, that would be 84 times. There are also present the usual complications arising with a GtS strategy, in terms of the actual size of the sequential testing and consistency in variable selection. Indeed, the results of these top-down strategies are typically suboptimal, search path dependent and affected by the order of the variables included into the system [Hsu

[16] In this case, we have imposed $v = 10$.

Matrix A

True values

0.107	0	0	0	0	0	0.107
0	0.128	0	0	0.178	0	0
0	0	0.184	0	0	0	-0.046
0.081	0	0	0.196	0	0	0
0	-0.023	0	0	0.226	-0.011	0
0	0	0.071	0	-0.063	0.303	0.023
0	0.022	0	0.202	0	0	0.123

RMSE: *MLE*

0.066	0.080	0.077	0.099	0.071	0.091	0.077
0.054	0.073	0.069	0.086	0.066	0.082	0.065
0.062	0.066	0.073	0.079	0.060	0.071	0.062
0.051	0.066	0.058	0.067	0.054	0.064	0.053
0.060	0.072	0.061	0.081	0.077	0.075	0.060
0.050	0.057	0.058	0.066	0.051	0.074	0.054
0.078	0.084	0.074	0.086	0.075	0.090	0.070

Percentage of zeros: GtS

0.4	86.6	82.6	79.4	83.8	79.0	21.0
86.0	0.2	81.8	84.0	2.0	86.2	75.4
85.6	83.8	0	82.0	84.8	86.6	55.2
30.0	77.2	86.4	0.2	82.0	86.0	80.6
86.2	85.4	81.4	82.4	0	89.6	82.8
86.0	80.4	31.8	86.4	31.6	0	71.6
81.2	78.4	79.8	5.0	80.6	85.8	0.2

GtS percentage gain over *MLE*

25.9	52.1	53.8	41.1	42.8	62.3	18.2
48.6	41.4	54.8	52.2	35.7	68.8	47.7
57.3	43.8	38.8	47.4	41.9	65.6	31.3
-0.6	41.6	52.2	23.3	44.2	61.2	27.8
51.4	45.0	47.6	49.3	28.6	66.0	45.7
50.1	39.8	22.3	45.7	13.3	44.4	33.3
49.8	50.8	50.3	24.3	45.8	69.7	27.0

Percentage of zeros: Adaptive Lasso

0.6	75.6	71.6	69.0	74.8	57.8	17.6
73.8	0	73.8	74.2	0.8	71.8	68.4
69.6	73.2	0	74.0	72.2	70.0	44.6
19.0	65.4	75.8	0	74.4	76.6	71.6
77.4	72.6	67.6	76.2	0	73.2	76.4
72.4	73.4	18.8	78.4	21.6	0	64.2
67.4	63.6	67.8	1.2	70.6	71.4	0.2

A. Lasso percentage gain over *MLE*

40.0	76.6	73.4	63.3	73.1	78.3	-0.5
70.7	53.0	70.6	74.7	40.5	82.5	61.8
69.0	66.9	33.6	71.1	70.7	82.2	18.2
17.6	67.4	73.4	23.0	71.7	82.0	55.1
69.3	57.6	66.6	68.6	31.3	74.6	63.2
72.7	71.1	31.4	74.2	22.5	49.1	44.0
71.1	64.8	73.7	32.8	70.7	80.2	27.6

Percentage of zeros: misspecified Adaptive Lasso

0.2	78.0	72.8	70.4	75.4	65.6	15.2
75.2	0	77.2	74.6	1.4	78.4	72.8
73.2	74.0	0	73.2	76.2	67.2	45.6
25.4	65.2	76.8	0.2	77.2	75.0	69.8
78.0	72.4	63.8	79.2	0	76.4	77.0
76.2	76.0	20.8	80.4	26.6	0	64.2
74.0	71.6	72.2	3.6	69.8	72.4	0.4

Missp. A. Lasso percentage gain over *MLE*

41.9	74.4	74.7	62.4	69.1	81.2	0.8
75.0	51.6	68.5	71.0	39.7	85.9	66.9
71.2	66.4	25.0	69.5	76.0	84.4	15.9
12.3	69.7	76.0	17.4	73.5	80.2	52.1
69.7	58.6	65.5	66.5	29.0	78.2	68.0
70.4	70.6	29.4	69.2	17.5	49.7	46.1
70.9	65.3	70.8	9.8	68.1	81.3	20.4

Table 2.3: Monte Carlo exercise. Sparse case, $T = 736$. Top-left panel: the elements of the matrix A used for Monte Carlo simulations. Top-right panel: the Root Mean Square Error (RMSE) of the *MLE*. Remaining panels on the left: the percentage of times in which the estimator finds a zero for the corresponding parameter. Remaining panels on the right: percentage gain in RMSE over the *MLE* (2.18).

	$T = 736$		$T = 1500$	
Technique	False zeros	True zeros	False zeros	True zeros
t-test,40%	39.61%	63.78%	30.64%	58.89%
t-test,20%	53.14%	80.62%	37.74%	70.26%
t-test,10%	62.56%	89.47%	43.31%	77.74%
t-test,5%	68.08%	94.03%	52.24%	88.04%
GtS	33.11%	79.48%	26.09%	75.76%
GtS, 10%	29.75%	68.95%	21.54%	69.61%
A. Lasso	25.96%	67.39%	20.77%	68.42%
Missp. A. Lasso	27.09%	69.22%	21.59%	68.73%

Table 2.4: The percentage of false zeros (that is, zeros detected in the non-zero elements of A and B) and of true zeros (that is, zeros detected in the zero elements of A and B) identified by individual t tests, by the GtS approach with 5% and 10% significance levels, by the Adaptive Lasso estimator and by the Adaptive Lasso estimator with a misspecified conditional distribution. The first two columns refers to the case $T = 736$, while the last two to the case $T = 1500$. A and B are sparse matrices (see Table 2.3). and 2.5).

et al., 2008]. So, we start from the result, shown in other contexts, that, compared to the existing subset selection methods with parameter constraints such as the GtS strategy, the Lasso method is computationally efficient and its result is robust to the order of the series included in the autoregressive model.

2.6 Conclusions

When multivariate systems of dynamic equations grow larger, manageability of parameter estimation becomes an issue, accompanied by the problem of identification of which dynamic links from one component to another are insignificant. In this respect, the vector Multiplicative Error Models applied to studying the dynamic interdependence of non negative valued processes are no exception: reconstructing the interactions among several univariate time series crucially relies on the isolation of zeros in the corresponding adjacency matrix. Such models are affected by the well-known curse of dimensionality, not only because the order of magnitude of parameters to be estimated is proportional to N^2, but also by an increasing number of zeros which makes their estimation less precise given the large standard errors of the estimators in an unrestricted model. Some form of a pruning strategy, such as a GtS approach based on a potentially high number

Matrix A

True values

0.107	0	0	0	0	0	0.107
0	0.128	0	0	0.178	0	0
0	0	0.184	0	0	0	-0.046
0.081	0	0	0.196	0	0	0
0	-0.023	0	0	0.226	-0.011	0
0	0	0.071	0	-0.063	0.303	0.023
0	0.022	0	0.202	0	0	0.123

RMSE: *MLE*

0.062	0.075	0.072	0.073	0.076	0.070	0.063
0.056	0.066	0.067	0.062	0.065	0.064	0.055
0.055	0.072	0.062	0.062	0.062	0.059	0.053
0.049	0.064	0.055	0.059	0.059	0.051	0.050
0.048	0.068	0.059	0.054	0.066	0.053	0.052
0.043	0.059	0.056	0.051	0.050	0.057	0.039
0.051	0.072	0.058	0.061	0.067	0.063	0.054

Percentage of zeros: GtS

0	76.2	74.5	77.4	74.5	76.8	22.0
80.4	0	77.2	79.8	4.8	82.4	71.5
80.8	74.1	0.2	79.2	78.4	82.4	51.3
31.3	73.7	74.9	0.4	83.4	82.6	77.0
77.8	78.2	74.7	80.2	0.2	79.4	76.2
87.1	74.3	35.6	83.6	40.6	0.2	73.3
80.6	70.1	76.6	2.4	78.2	78.2	0

GtS percentage gain over *MLE*

40.0	45.9	30.0	45.8	26.9	46.8	14.6
55.1	38.9	33.3	49.0	21.3	57.5	31.7
54.2	33.3	19.1	45.8	22.7	54.1	13.2
10.3	35.6	28.2	17.9	29.7	43.8	32.4
47.4	33.3	31.7	50.0	17.4	45.7	25.7
57.1	31.0	5.4	44.1	0	9.1	0
42.6	35.8	29.5	24.5	33.3	52.5	19.5

Percentage of zeros: Adaptive Lasso

0	77.2	78.2	70.8	76.5	77.6	5.6
70.5	0	72.8	71.3	0.7	77.2	67.1
70.1	69.0	0	71.6	72.0	77.9	26.5
6.7	62.7	77.2	0	78.9	81.9	72.0
79.0	64.9	69.8	74.3	0	66.8	75.0
78.4	73.1	6.0	81.0	14.6	0	59.8
77.2	46.7	78.7	0.4	78.7	75.4	0

A. Lasso percentage gain over *MLE*

56.4	77.0	74.0	62.7	67.3	66.0	25.0
77.6	64.8	71.1	74.5	51.1	77.5	61.0
77.1	70.4	40.4	70.8	70.5	81.1	28.9
17.9	57.8	66.7	28.2	64.9	71.9	58.8
71.1	52.1	61.0	69.0	34.8	62.9	62.9
68.6	59.5	24.3	67.6	13.9	36.4	27.6
70.2	60.4	75.0	49.0	64.4	75.0	43.9

Percentage of zeros: misspecified Adaptive Lasso

0.4	74.6	76.9	71.3	76.9	68.7	15.2
81.8	0	76.6	74.3	2.3	73.9	68.0
73.9	71.6	0	76.2	74.3	76.6	44.6
24.1	70.0	76.2	0	80.5	77.9	71.6
76.6	70.0	69.7	80.9	0	67.7	75.6
79.9	72.3	23.8	83.8	37.6	0	63.6
77.9	68.3	77.2	0.3	75.2	75.2	0

Missp. A. Lasso percentage gain over *MLE*

50.9	75.4	66.0	71.2	59.6	63.8	22.9
71.4	57.4	62.2	76.5	42.6	75.0	56.1
70.8	61.1	31.9	75.0	61.4	67.6	21.1
10.3	55.6	61.5	20.5	56.8	71.9	52.9
65.8	50.0	53.7	73.8	30.4	60.0	54.3
71.4	61.9	21.6	73.5	2.8	9.1	27.6
72.3	60.4	68.2	38.8	64.4	70.0	31.7

Table 2.5: Monte Carlo exercise. Sparse case, $T = 1500$. Top-left panel: the elements of the matrix A used for Monte Carlo simulations. Top-right panel: the Root Mean Square Error (RMSE) of the *MLE*. Remaining panels on the left: the percentage of times in which the estimator finds a zero for the corresponding parameter. Remaining panels on the right: percentage gain in RMSE over the *MLE*, defined as $100 \times (\text{RMSE}_{MLE} - \text{RMSE}_i)/\text{RMSE}_{MLE}$, with i equal to the GtS, Adaptive Lasso and misspecified Adaptive Lasso estimators.

of successive estimations with repercussions on the actual size of the sequential testing procedure, is a feasible form of model selection but potentially cumbersome.

In this study we have discussed the merits of adopting Adaptive Lasso techniques within the vMEM class, to alleviate some of the previously mentioned problems. We have adopted a parametric log–normal specification for the innovation term which can rely on some previously derived analytic results in the literature [Taylor and Xu, 2017]. Under suitable assumptions, the Adaptive Lasso is shown to enjoy the oracle property and allows for a parsimonious specification of the vMEM, improving the estimation performances and the interpretability of the results.

We have run a comprehensive Monte Carlo simulation in order to assess the properties of our procedure along several dimensions: sample period length; diagonal versus sparse coefficient matrix; error distribution different from the log-normal; comparison with a General to Specific model search. On all of these dimensions we show that the Adaptive Lasso vMEM performs well, especially (relative to the GtS) with a shorter sample period.

More specifically, simulation results show a significant reduction of the RMSE of the estimates and better results in variable selection for the Adaptive Lasso estimator compared to the GtS approach when the shorter window of 736 observations is considered. Concerning the larger window ($T = 1500$), the Lasso outperforms the GtS approach in terms of the RMSE of the estimates, while the results for the precision in variable selection are comparable between the two approaches.

Acknowledgements

We are grateful to Fabrizio Lillo, Luca Trapin, Fabrizio Cipollini, Matteo Barigozzi and to the seminar participants at the “41st Annual Meeting of the Association for Mathematics Applied to Social and Economic Sciences” (Cagliari, 2017) and at the “Rimini Conference in Economics and Finance” (Rimini, 2018) for discussions. All errors remain our own.

Appendix 2.A Proof of Theorem 2

We now provide the proof of Theorem 2 as stated in the main text.

Asymptotic Normality. Let $\mathbf{u} = \sqrt{T}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ such that $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{u}/\sqrt{T}$. Define

$$U_T(\mathbf{u}) \stackrel{\text{def}}{=} -\ell(\boldsymbol{\beta} + \mathbf{u}/\sqrt{T}) + T\lambda_T \sum_{j=1}^{d} \hat{w}_j |(\boldsymbol{\beta} + \mathbf{u}/\sqrt{T})_j|. \tag{2.19}$$

The Taylor expansion of $U_T(\mathbf{u})$ around $\mathbf{u} = \mathbf{0}$ is denoted as

$$U_T(\mathbf{u}) = U_T(\mathbf{0}) + M_1(T) + M_2(T) + M_3(T) + M_\lambda(T),$$

where $M_1(T)$ is the first order term, $M_2(T)$ is the second order term, $M_3(T)$ contains terms of order larger or equal to three and, finally, $M_\lambda(T)$ is the Adaptive Lasso term. We start analysing the first order term $M_1(T)$:

$$\begin{aligned} M_1(T) = -\boldsymbol{\nabla}_{\boldsymbol{u}} \ell(\boldsymbol{\beta} + \mathbf{u}/\sqrt{T})\big|'_{\mathbf{u}=\mathbf{0}} \mathbf{u} &= -\frac{1}{\sqrt{T}} \boldsymbol{\nabla}_{\tilde{\boldsymbol{\beta}}} \ell(\tilde{\boldsymbol{\beta}})\big|'_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}} \boldsymbol{u} \\ &= -\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \boldsymbol{\nabla}_{\tilde{\boldsymbol{\beta}}} \log f(\mathbf{y}_t \,|\, \mathcal{F}_{t-1}, \tilde{\boldsymbol{\beta}})\big|'_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}} \boldsymbol{u}. \end{aligned} \tag{2.20}$$

It follows that:

$$\begin{aligned} &\mathbb{E}[M_1(T)] = \mathbf{0}, \\ &\mathbb{V}[M_1(T)] = \boldsymbol{u}' \mathcal{I}(\boldsymbol{\beta}) \boldsymbol{u}, \end{aligned} \tag{2.21}$$

where we have used the fact that $\mathbb{E}[\boldsymbol{\nabla}_{\tilde{\boldsymbol{\beta}}} \ell(\tilde{\boldsymbol{\beta}})|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}}] = \mathbf{0}$ and the variance-covariance matrix of the gradient $\boldsymbol{\nabla}_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$ is:

$$\mathbb{V}[\boldsymbol{\nabla}_{\tilde{\boldsymbol{\beta}}} \ell(\tilde{\boldsymbol{\beta}})|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}}] = T\,\mathcal{I}(\boldsymbol{\beta}). \tag{2.22}$$

Being $M_1(T)$ the sum of a finite variance martingale difference sequence divided by $\sqrt{T}$ (see equation (2.20)), it converges to a normal random variable for $T \to \infty$ thanks to the central limit theorem, that is $M_1(T) \stackrel{d}{\to} -\boldsymbol{u}' W$, where $W \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}))$.

The second order term $M_2(T)$ is given by

$$M_2(T) = -\frac{1}{2T}\boldsymbol{u}'\, H_{\tilde{\boldsymbol{\beta}}}(\ell(\tilde{\boldsymbol{\beta}}))|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}}\, \boldsymbol{u} = -\frac{1}{2T}\sum_{t=1}^{T} \boldsymbol{u}'\, H_{\tilde{\boldsymbol{\beta}}}\left(\log f(\mathbf{y}_t \,|\, \mathcal{F}_{t-1}, \tilde{\boldsymbol{\beta}})\right)\Big|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}} \boldsymbol{u}, \quad (2.23)$$

where $H_{\tilde{x}}(g(\tilde{x}))|_{\tilde{x}=x}$ is the Hessian of the function $g(\tilde{x})$ calculated in $\tilde{x} = x$. We observe that

$$\mathbb{E}[M_2(T)] = \frac{1}{2}\boldsymbol{u}'\mathcal{I}(\boldsymbol{\beta})\boldsymbol{u}.$$

Furthermore, $M_2(T)$ is the average of a sequence of random variables with finite mean and finite variance (see equation (2.23)), then, thanks to the Law of Large Numbers, it converges to its unconditional mean: $M_2(T) \xrightarrow{p} \boldsymbol{u}'\mathcal{I}(\boldsymbol{\beta})\boldsymbol{u}/2$. The asymptotic distribution of the $M_\lambda(T)$ is described in the proof of Theorem 2 of Zou [2006] and can be summarized as:

$$T\lambda_T\hat{w}_j\left(|\beta_j + u_j/\sqrt{T}| - |\beta_j|\right) \xrightarrow{p} \begin{cases} 0 & \text{if } j \in \mathcal{D}, \\ 0 & \text{if } j \in \mathcal{D}^0 \text{ and } u_j = 0, \\ \infty & \text{if } j \in \mathcal{D}^0 \text{ and } u_j \neq 0, \end{cases} \quad (2.24)$$

where $\mathcal{D}^0 = \{j : \beta_j = 0\}$. Note that the Adaptive Lasso term $M_\lambda(T)$ is asymptotically irrelevant in the set of non-zero parameters $\mathcal{D}$. Meanwhile, in the set of zero parameters $\mathcal{D}^0$, it is zero when $\tilde{\beta}_j$ is equal to the true parameter β_j and infinite elsewhere. This is the reason why the Adaptive Lasso estimator is able to discriminate between zero and non-zero parameters.

Thanks to the third regularity condition of Assumption 1, the term $M_3(T)$ converges to zero asymptotically, as in the case of ordinary maximum likelihood estimators [Lehmann and Casella, 2006]. Now, summarising the limiting behaviour of $M_1(T)$, $M_2(T)$, $M_\lambda(T)$ and $M_3(T)$, we have that:

$$V_T(\mathbf{u}) \overset{\text{def}}{=} U_T(\mathbf{u}) - U_T(\mathbf{0}) \xrightarrow{d} V(\mathbf{u}) = \begin{cases} \frac{1}{2}\boldsymbol{u}'_{\mathcal{D}}\,\mathcal{I}(\boldsymbol{\beta}_{\mathcal{D}})\,\boldsymbol{u}_{\mathcal{D}} - \boldsymbol{u}'_{\mathcal{D}}W_{\mathcal{D}} & \text{if } u_j = 0 \ \ \forall j \notin \mathcal{D}, \\ \infty & \text{otherwise.} \end{cases} \quad (2.25)$$

It follows that the minimum of $V(\mathbf{u})$ is at $(\mathcal{I}(\boldsymbol{\beta}_{\mathcal{D}})^{-1}W_{\mathcal{D}}, \mathbf{0}_{\mathcal{D}^0})$. Now let $\hat{\mathbf{u}}(T)$ be equal to $\operatorname{argmin} V_T(\mathbf{u})$, then

$$\hat{\mathbf{u}}(T)_{\mathcal{D}} \xrightarrow{d} \mathcal{I}(\boldsymbol{\beta}_{\mathcal{D}})^{-1}W_{\mathcal{D}} \quad \text{and} \quad \hat{\mathbf{u}}(T)_{\mathcal{D}^0} \xrightarrow{d} \mathbf{0}_{\mathcal{D}^0}.$$

$W_{\mathcal{D}}$ is normally distributed, hence $\hat{\boldsymbol{\beta}}(\lambda_T) = \boldsymbol{\beta} + \hat{\mathbf{u}}(T)/\sqrt{T}$ is asymptotically normal and $\hat{\boldsymbol{\beta}}(\lambda_T)$ converges to the true parameter $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{D}}, \mathbf{0}_{\mathcal{D}^0})$, that is

$$\begin{aligned} \sqrt{T}\left(\hat{\boldsymbol{\beta}}_{\mathcal{D}}(\lambda_T) - \boldsymbol{\beta}_{\mathcal{D}}\right) &\xrightarrow{d} \mathcal{I}_{\mathcal{D}}(\boldsymbol{\beta})^{-1}W_{\mathcal{D}}, \\ \hat{\boldsymbol{\beta}}_{\mathcal{D}^0} &\xrightarrow{d} \mathbf{0}_{\mathcal{D}^0}. \end{aligned} \tag{2.26}$$

Consistency in variable selection. From equation (2.26), it follows that, if $j' \in \mathcal{D}$, then $\lim_{T\to\infty} \mathbb{P}[j' \in \hat{\mathcal{D}}^0(T)] = 0$. In order to show that $\lim_{T\to\infty} \mathbb{P}[j \in \hat{\mathcal{D}}(T)] = 0$ for $j \in \mathcal{D}^0$, we recall the Karush-Kuhn-Tucker (KKT) necessary conditions for $\hat{\beta}_j$ to be an optimum[17]:

$$\left(\boldsymbol{\nabla}_{\tilde{\boldsymbol{\beta}}}\ell(\tilde{\boldsymbol{\beta}})|_{\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}}\right)_j = T\lambda_T\hat{w}_j. \tag{2.27}$$

Being condition (2.27) necessary, it follows that

$$\begin{aligned} \mathbb{P}\left[j \in \hat{\mathcal{D}}(T)\right] &\leq \mathbb{P}\left[\left(\boldsymbol{\nabla}_{\tilde{\boldsymbol{\beta}}}\ell(\tilde{\boldsymbol{\beta}})|_{\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}}\right)_j = T\lambda_T\hat{w}_j\right] = \\ &= \mathbb{P}\left[\frac{1}{\sqrt{T}}\left(\boldsymbol{\nabla}_{\tilde{\boldsymbol{\beta}}}\ell(\tilde{\boldsymbol{\beta}})|_{\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}}\right)_j = \sqrt{T}\lambda_T\hat{w}_j\right]. \end{aligned} \tag{2.28}$$

In what follows, we study the asymptotic behaviour of the two terms appearing in the last line of equation (2.28). We start with the term on the right. Since $\hat{\beta}_j(\text{mle})\sqrt{T}$ is asymptotically finite and $\lambda_T\, T^{(\gamma+1)/2} \to \infty$, we obtain that

$$\sqrt{T}\lambda_T\hat{w}_j = \frac{\sqrt{T}\lambda_T}{|\hat{\beta}_j(\text{mle})|^\gamma}\frac{T^{\gamma/2}}{T^{\gamma/2}} = \frac{\lambda_T\, T^{(\gamma+1)/2}}{|\hat{\beta}_j(\text{mle})\sqrt{T}|^\gamma} \xrightarrow{p} \infty.$$

[17] We are implicitly assuming that $\hat{\beta}_j$ is positive. The same results holds for $\hat{\beta}_j \leq 0$.

Now we focus on the term on the left, studying the Taylor expansion of $\left(\boldsymbol{\nabla}_{\tilde{\boldsymbol{\beta}}}\ell(\tilde{\boldsymbol{\beta}})|_{\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}}\right)_j$ around β_j:

$$\frac{1}{\sqrt{T}}\left(\boldsymbol{\nabla}_{\tilde{\boldsymbol{\beta}}}\ell(\tilde{\boldsymbol{\beta}})|_{\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}}\right)_j = K_0(T) + K_1(T) + K_2(T),$$

with

$$\begin{aligned}
K_0(T) &= \frac{1}{\sqrt{T}}\left(\boldsymbol{\nabla}_{\tilde{\boldsymbol{\beta}}}\ell(\tilde{\boldsymbol{\beta}})|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}}\right)_j, \\
K_1(T) &= \frac{1}{T}\left(H_{\tilde{\boldsymbol{\beta}}}(\ell(\tilde{\boldsymbol{\beta}}))|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}}\right)_{j,j}\sqrt{T}(\hat{\beta}_j-\beta_j), \\
K_2(T) &= \frac{1}{T}\sum_{t=1}^{T}\frac{\partial^3 \log f(\mathbf{y}_t \mid \mathcal{F}_{t-1},\tilde{\boldsymbol{\beta}})}{\partial\tilde{\beta}_j^3}\bigg|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}^*}\left(\sqrt{T}(\hat{\beta}_j-\beta_j)\right)^2\frac{1}{\sqrt{T}},
\end{aligned} \tag{2.29}$$

where $\boldsymbol{\beta}^*$ is between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. Following the same reasoning done for $M_1(T)$ and $M_2(T)$, we obtain that $K_0(T) \xrightarrow{d} \mathcal{N}(0,(\mathcal{I}(\boldsymbol{\beta}))_{j,j})$ and $H_{\tilde{\boldsymbol{\beta}}}(\ell(\tilde{\boldsymbol{\beta}}))|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}} \xrightarrow{d} T\,(\mathcal{I}(\boldsymbol{\beta}))_{j,j}$. From equation (2.26), it follows that: *i.* $K_1(T)$ converges to a normal random variable and *ii.* $K_2(T)$ is of order $O_p(1/\sqrt{T})$. Therefore, the probability that $K_0(T)+K_1(T)+K_2(T)$ is equal to $\sqrt{T}\lambda_T\hat{w}_j$ goes to zero asymptotically, that is $\mathbb{P}[j\in\hat{\mathcal{D}}(T)] \xrightarrow{p} 0$. □

Appendix 2.B Quasi-*MLE* estimator of V

In this appendix, we discuss the estimation of the variance-covariance matrix V of equation (2.12) in the case in which the innovation terms $\boldsymbol{\epsilon}_t$ are log-normally distributed. The maximum likelihood estimator for V can be obtained in closed form in the univariate case ($N=1$) as follows:

$$\underset{\tilde{V}}{\operatorname{argmin}}\left\{-\frac{1}{T}\ell(\tilde{\beta},\tilde{V})\right\} = 2\,\left(\sqrt{1+s_{\tilde{\beta}}}-1\right), \tag{2.30}$$

where

$$s_{\tilde{\beta}} \overset{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T}(\log\epsilon_{t,\tilde{\beta}})^2 = \frac{1}{T}\sum_{t=1}^{T}(\log(y_t/\mu_{t,\tilde{\beta}}))^2$$

and $\mu_{t,\tilde{\beta}}$ is the value of μ_t for a given value of $\tilde{\beta}$. Equation (2.30) has a simple interpretation in terms of moment conditions. Indeed, first note that $\mathbb{E}[(\log\epsilon_t)^2] = \mathbb{V}[\log\epsilon_t] +$

$(\mathbb{E}[\log \epsilon_t])^2 = V + V^2/4$. Substituting the expectation $\mathbb{E}[(\log \epsilon_t)^2]$ with its sample counterpart s, we obtain the quadratic equation $\hat{V}^2 + 4\hat{V} - 4s = 0$ with the unique (positive) solution $\hat{V} = 2\,(\sqrt{1-s} - 1)$.

In the multivariate case ($N > 1$), after some algebra, it is possible to obtain the following first order condition in V[18]:

$$\begin{aligned}\hat{V} - S - \frac{1}{2}\,\boldsymbol{\gamma}\,\mathrm{diag}(\hat{V})' - \frac{1}{2}\,\mathrm{diag}(\hat{V})\,\boldsymbol{\gamma}' - \frac{1}{4}\,\mathrm{diag}(\hat{V})\,\mathrm{diag}(\hat{V})' + \\ + \hat{V}\,\mathrm{Diag}(\hat{V}^{-1}\boldsymbol{\gamma})\,\hat{V} + \frac{1}{2}\,\hat{V}\,\mathrm{Diag}(\hat{V}^{-1}\,\mathrm{diag}(\hat{V}))\,\hat{V} = 0,\end{aligned} \tag{2.31}$$

with $S = \frac{1}{T}\sum_{t=1}^{T} \log \boldsymbol{\epsilon}_t\,(\log \boldsymbol{\epsilon}_t)'$ and $\boldsymbol{\gamma} = \frac{1}{T}\sum_{t=1}^{T} \log \boldsymbol{\epsilon}_t$[19]. Unfortunately no closed form solutions for $\hat{V}$ are available. However, a simple asymptotically consistent approximation is proposed. Indeed, from the equality $\mathbb{E}[\boldsymbol{\gamma}] = -0.5\,\mathrm{diag}(V)$, it follows that, for $T \to \infty$, $\boldsymbol{\gamma} \xrightarrow{p} -0.5\,\mathrm{diag}(V)$, and

$$\begin{cases} \hat{V}\,\mathrm{Diag}(\hat{V}^{-1}\boldsymbol{\gamma})\,\hat{V} + \frac{1}{2}\,\hat{V}\,\mathrm{Diag}(\hat{V}^{-1}\,\mathrm{diag}(\hat{V}))\,\hat{V} & \longrightarrow_p \quad 0 \\ -\frac{1}{2}\,\boldsymbol{\gamma}\,\mathrm{diag}(\hat{V})' - \frac{1}{2}\,\mathrm{diag}(\hat{V})\,\boldsymbol{\gamma}' - \frac{1}{4}\,\mathrm{diag}(\hat{V})\,\mathrm{diag}(\hat{V})' & \longrightarrow_p \quad \frac{1}{4}\,\mathrm{diag}(\hat{V})\,\mathrm{diag}(\hat{V})'. \end{cases} \tag{2.32}$$

Introducing the asymptotic limits (2.32) into equation (2.31), we obtain the following (asymptotic) first order condition in V:

$$\hat{V} - S + \frac{1}{4}\,\mathrm{diag}(\hat{V})\,\mathrm{diag}(\hat{V})' \xrightarrow{p} 0. \tag{2.33}$$

The diagonal elements of $\hat{V}$ satisfy (asymptotically) the equation $\hat{V}_{ii} - S_{ii} + \frac{1}{4}\,\hat{V}_{ii}\,\hat{V}_{ii} = 0$ with positive solution

$$\hat{V}_{ii} = 2\,(\sqrt{1 + S_{ii}} - 1). \tag{2.34}$$

The off-diagonals terms satisfies (asymptotically) the condition $\hat{V}_{ij} - S_{ij} + \frac{1}{4}\,\hat{V}_{ii}\,\hat{V}_{jj} = 0$, such that

$$\hat{V}_{ij} = S_{ij} - \frac{1}{4}\,\hat{V}_{ii}\,\hat{V}_{jj}. \tag{2.35}$$

[18] The first order conditions are obtained maximising the log-likelihood (2.10) w.r.t. $\tilde{V}$.

[19] Here, we are not showing the explicit dependence on $\tilde{\boldsymbol{\beta}}$ for not burdening the notation.

Equations (2.34) and (2.35) can be rewritten in an unified notation as

$$\hat{V}_{ij} = S_{ij} - \left(\sqrt{1 + S_{ii}} - 1\right)\left(\sqrt{1 + S_{jj}} - 1\right), \tag{2.36}$$

which is our final estimator. As in the one dimensional case, it is possible to understand this result from the moment condition $\mathbb{E}[\log \boldsymbol{\epsilon_t} \log \boldsymbol{\epsilon_t}'] = V + \mathbf{m}\mathbf{m}' = V + \text{diag}(V)\,\text{diag}(V)'/4$. Replacing $\mathbb{E}[\log \boldsymbol{\epsilon_t} \log \boldsymbol{\epsilon_t}']$ with its sample counterpart S, we easily obtain equation (2.33).

Actually, $\hat{V}$ depends on S, which in turn depends on $\tilde{\boldsymbol{\beta}}$. Therefore, in this study, we use the notation $\hat{V}_{\tilde{\boldsymbol{\beta}}}$ in order to indicate the estimator defined by equation (2.36). Note that the exact maximum likelihood estimator of V for the univariate case (2.30) is a particular case of $\hat{V}_{\tilde{\boldsymbol{\beta}}}$.

Chapter 3

Detection of Volatility spillovers among European financial markets[1]

3.1 Introduction

In this chapter, we analyse the existence and the strength of volatility spillovers among European financial markets, during and after the European sovereign debt crisis, reconstructing the network of interdependencies through which shocks spread among these countries.

Specifically, we focus on two issues:

- *Network stability* We find noticeable different propagation mechanisms among a vector of realized volatilities from the height of the crisis (2010-2012, stronger spillover effects) to the absorption of fears (2013-2015, following the "whatever it takes" speech by Mario Draghi in July 2012). During the financial crisis, the European markets tightened their interrelatedness, with several channels through which shocks propagated throughout the nodes of the network. As with other studies about network stability [Billio et al., 2012, Engle et al., 2012], linkages appear also here to be time-varying and extremely dependent on prevailing mar-

[1] The material of this chapter is taken from Cattivelli and Gallo [2018].

ket conditions. In order to identify the presence, the size and the direction of the spillovers, we model the volatility series through a vector Multiplicative Error model with log-normal innovations and with the adoption of the Adaptive Lasso technique developed in Chapter 2.

- *Robustness to common movements* Given that the volatility series exhibit some slow–moving common behavior, we investigate whether interdependence is modified by the explicit consideration of a common low-frequency component among the series. Extending the SeminonParametric vMEM (SPvMEM) of Barigozzi et al. [2014], through the use of the Adaptive Lasso approach, we allow for full or partial interdependence in the dynamics of the volatility series around a common component. We then check the robustness of our spillover network by disentangling the common component term from specific volatility spillovers effects. We find that for the European financial markets the network of spillovers does not change.

In the empirical application, we employ several detailed estimation diagnostics tests, reporting the residual autocorrelation properties and several tests for the assumption of a log-normal distribution for the error terms. We found that the assumed distribution performs fairly well in practice, especially in capturing the marginal distribution of the innovations.

Finally, with the aim of testing the goodness of Adaptive Lasso techniques for vMEM, we also show that one-day-ahead forecasts obtained by these models deliver superior forecasting accuracy compared to standard vMEMs for the dynamics of the realized volatility among European financial markets. In particular, the Adaptive Lasso method is shown to perform better compared to the General to Specific (GtS) approach of Cipollini and Gallo [2010], especially when the number of observations is small relative to the number of parameters.

This Chapter is organized as follows. First we study volatility spillover effects among European financial markets without (Section 3.2) or with (Section 3.3) the "low-

frequency common component" of Barigozzi et al. [2014]. Then, in Section 3.4, we compare the forecasting performances of several MEM, vMEM and SPvMEM specifications for the realized volatility. Conclusions follow in Section 3.5.

3.2 Volatility Spillovers in European Markets during and after the Debt Crisis

In this Section we adopt the Adaptive Lasso estimation technique for vMEM to a relevant empirical application on realized volatility series calculated from market indices across major European financial markets. In this context, properly selecting zeros parameters in the matrices A and B of equation (2.4) allows one to reconstruct the network of interdependencies in an area battered by acute episodes of turmoil in correspondence with the sovereign debt crisis (2010-2012), later followed by a more tranquil period. This issue is particular important since the topology of the network, through which shocks spread among countries, has relevant impact for risk managers, policy makers, regulators and asset managers [Billio and Pelizzon, 2003]. We are working under the anticipation that this network may be time dependent and, in this respect, we feel supported by our Monte Carlo results (Section 2.5) that our Adaptive Lasso vMEM is capable of better detecting the existence and the direction of these interactions even with three years of daily data (a relatively short sample of 736 observations) over a GtS approach, even in the possible presence of a vector of innovations not log-normally distributed.

We identify relevant links using the non-zero estimates of our Adaptive Lasso vMEM developed in Section 2.4 to identify the interdependences across the realized volatilities of the following indices: DAX (Germany), CAC40 (France), AEX (Netherlands), SMI (Switzerland), IBEX (Spain), FTSEMIB (Italy) and FTSE100 (Great Britain). Among the seven, there are the main five economies for the Euro area (Germany, France, Italy, Spain and the Netherlands) and two relevant financial indices in Europe (SMI and FTSE100). These markets constitute a large share of the European financial market and can be thought of as a proxy for the European financial system. We chose to repeat the analysis

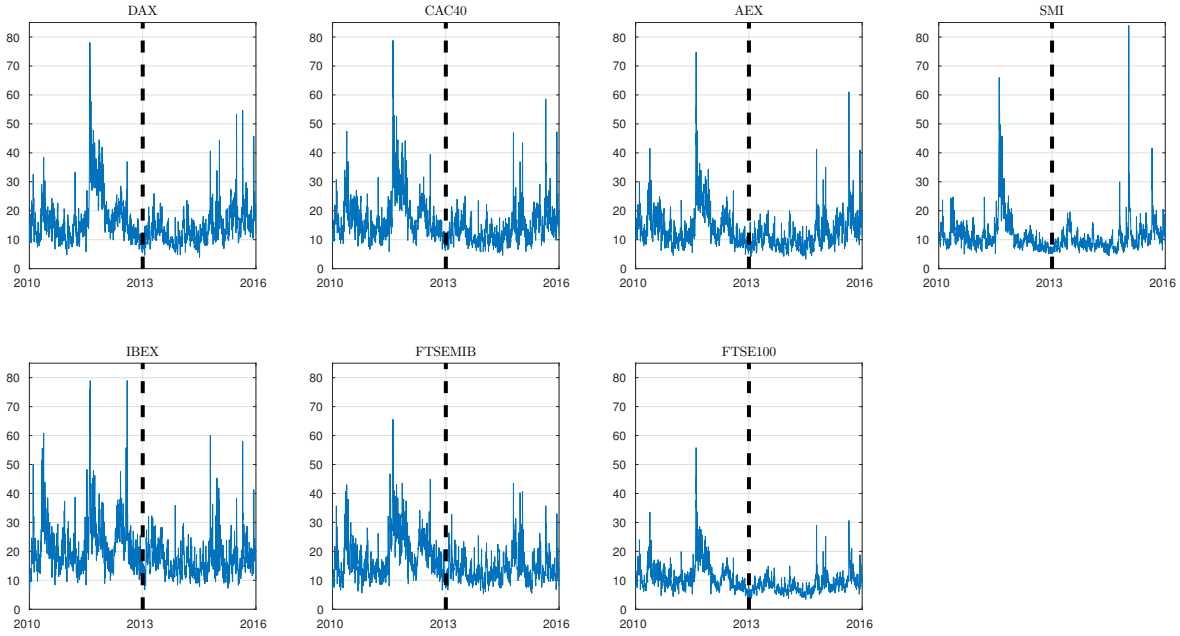

Figure 3.1: The dynamics of the realized volatility in percentage annualized terms [$\sqrt{RV_t \times 252} \times 100$, where RV_t is the realized Kernel Variance at day t obtained from the Oxford Man Realized Volatility Library, Heber et al., 2009].

over two sub-samples, the first in the most severe period of the crisis (2010-2012), and the second in its concluding period (2013-2015), as shown in Figure 3.1.

The values of the realized kernel volatilities [Barndorff-Nielsen et al., 2008] are derived from the Oxford Man Realized Volatility Library [Heber et al., 2009]. The realized kernel is robust to microstructure noise and has the fastest possible convergence rate across high frequency volatility measures.

As we did in Chapter 2, each estimation window is divided in two intervals of 600 and 136 observations, respectively[2]. The first interval is used for the evaluation of the optimal shrinkage parameters $\hat{\lambda}_T$, which is selected minimizing the loss function (2.16) (see the two graphs on the top of Figure 3.2). Then, the Adaptive Lasso estimates $\hat{\boldsymbol{\beta}}(\hat{\lambda}_T)$ are obtained using the observations of the entire window. In Table 3.1 we report the obtained estimates and their standard errors.

As notable in Table 3.1, in the estimation procedure, we allow for the presence of

[2]The period 2010-2012 counts 736 observations. The same number of observations is present in the second time period (2013-2015).

AL-vMEM 2010-2012

Matrix A:

	DAX	CAC40	AEX	SMI	IBEX	FTSEMIB	FTSE100
DAX	**0.321** (0.063)	**0.007** (0.029)	**0.033** (0.073)	0	0	**0.029** (0.068)	0
CAC40	**0.009** (0.039)	**0.233** (0.030)	**0.025** (0.067)	0	**0.041** (0.041)	0	**0.110** (0.135)
AEX	0	**-0.014** (0.030)	**0.276** (0.055)	**0.006** (0.029)	**0.029** (0.030)	0	**0.087** (0.095)
SMI	**-0.018** (0.246)	0	**0.027** (0.050)	**0.289** (0.133)	0	0	**0.069** (0.188)
IBEX	**-0.014** (0.030)	**0.058** (0.104)	0	0	**0.356** (0.205)	**0.028** (0.330)	0
FTSEMIB	0	0	0	0	**0.097** (0.068)	**0.304** (0.116)	**-0.056** (0.066)
FTSE100	**-0.019** (0.026)	0	**0.030** (0.070)	**0.035** (0.026)	**0.018** (0.021)	**-0.005** (0.024)	**0.332** (0.047)

Matrix B:

DAX	**0.630** (0.097)	0	**-0.016** (0.049)	**0.016** (0.028)	**-0.027** (0.106)	**-0.015** (0.077)	0
CAC40	0	**0.622** (0.128)	**-0.003** (0.010)	0	**-0.059** (0.137)	**0.025** (0.309)	0
AEX	**-0.001** (0.010)	0	**0.694** (0.053)	0	**-0.053** (0.103)	**0.009** (0.213)	**-0.056** (0.072)
SMI	**0.006** (0.362)	0	**-0.017** (0.130)	**0.701** (0.157)	**-0.007** (0.016)	0	**-0.079** (0.208)
IBEX	**-0.012** (0.028)	**-0.091** (0.120)	0	**-0.015** (0.023)	**0.562** (0.074)	0	**0.104** (0.132)
FTSEMIB	0	**0.090** (0.068)	**-0.026** (0.052)	0	**-0.114** (0.076)	**0.636** (0.083)	0
FTSE100	0	**0.031** (0.158)	0	0	**-0.030** (0.033)	0	**0.560** (0.194)

AL-vMEM 2013-2015

Matrix A:

DAX	**0.293** (0.036)	0	**0.003** (0.006)	0	0	0	0
CAC40	**0.024** (0.049)	**0.231** (0.036)	0	0	0	**0.017** (0.014)	0
AEX	**0.031** (0.040)	0	**0.243** (0.025)	0	0	0	**0.010** (0.007)
SMI	0	0	**0.000** (0.004)	**0.366** (0.027)	0	0	0
IBEX	0	0	0	0	**0.268** (0.029)	**0.038** (0.034)	**0.005** (0.005)
FTSEMIB	0	0	0	0	0	**0.335** (0.034)	0
FTSE100	0	0	**0.010** (0.039)	0	0	0	**0.318** (0.048)

Matrix B:

DAX	**0.660** (0.037)	**-0.001** (0.001)	0	**-0.001** (0.001)	0	0	**-0.008** (0.011)
CAC40	0	**0.680** (0.033)	0	**-0.000** (0.000)	0	0	0
AEX	0	0	**0.674** (0.031)	0	0	0	0
SMI	0	0	0	**0.556** (0.237)	0	0	**0.025** (0.424)
IBEX	**-0.017** (0.037)	0	0	0	**0.638** (0.041)	0	0
FTSEMIB	0	**0.010** (0.040)	**-0.034** (0.066)	0	0	**0.610** (0.033)	**-0.001** (0.001)
FTSE100	0	0	0	0	0	0	**0.609** (0.040)

Table 3.1: In bold, the estimated matrices A and B with the Adaptive Lasso procedure (2.14). On the right of each element, the standard error of the estimates (2.15).

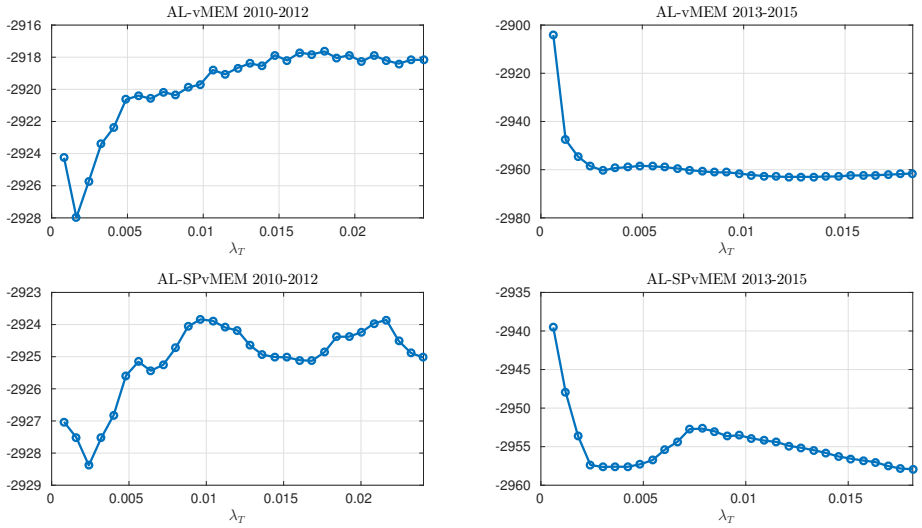

Figure 3.2: The sum of the log-normal loss function (2.16) $\sum_{t=601}^{736} L_{LN}(\mathbf{y}_t, \hat{\mathbf{y}}_{t|t-1}(\lambda_T))$ for different values of λ_T. On the left (right) we show the results for the interval 2010-2012 (respectively, 2013-2015). The first (second) line shows the results for the AL-vMEM (AL-SPvMEM). The optimal shrinkage parameter $\hat{\lambda}_T$ is the value of λ_T at the minimum of these curves.

off–diagonal negative values in the matrices A and B, without imposing non-negativity constraints. Indeed, in practical applications, it is customary not to impose constraints (cf., for example, Hautsch [2008] who also has negative values off-diagonal) at estimation stage. Moderate negative values in, say, the $\mathbf{A}$ matrix can, in fact, find an economic theoretical justification. Think of a news generating a sale of an asset or the exit from one market in favor of another. If we think of volatility increasing for *bad* news and decreasing for *good* ones, this is exactly an effect we want to capture in a multivariate context. In keeping with these empirical studies, in our design, a heuristic rule is implicitly followed for the order of magnitude of the negative values to be low relative to the main diagonal for the sake of keeping some interesting data structure (a similar approach was followed by Cipollini et al. [2013] where some negative off-diagonal values were inserted in the simulation design). Imposing nonnegativity (sufficient) conditions on the parameters, such as those derived in Cipollini et al. [2006], often turns out to be counterproductive, as it results in corner solutions at zero (absence of the interdependence) with complications for statistical inference. In this, we share the view of Bauwens et al.

[2006] in an MGARCH context when they state that *...negative values of the coefficients are not incompatible with a positive conditional variance. If one imposes positivity restrictions to facilitate estimation, one incurs the risk of rejecting θ_0 from the parameter space.*

Now, in order to visualize the system of interactions, we rely on a network structure, based on the estimates of Table 3.1. Spillover effects from index j to index i can be found looking at the element i, j of the matrices A and B. By analogy to the GARCH approach [cf. Hautsch, 2008], A measure the *impact* of the most recent observations in determining the adjustment of the conditional expectation to the new information while B represents the *inertia* of how conditional expectations based on an information set carry over to the next period. This information is summarized in the persistence matrix Γ, defined as $\Gamma = A + B$. This matrix is particularly important since it contains all the information about the stationarity of the process and about the long term dynamics through the equation:

$$\hat{\mathbf{y}}_{t+\tau|t} = \boldsymbol{\omega} + \Gamma \hat{\mathbf{y}}_{t+\tau-1|t}, \tag{3.1}$$

which holds for $\tau > 1$. As in Gallo and Velucchi [2009], we associate a directed network to this matrix with an arrow from j to i if $\Gamma_{ij} \neq 0$. Independent markets are characterized by a row and a column of zeros in Γ (except for the diagonal term), while dominant (dominated) markets have only out-going (in-going) links. The obtained networks of interactions are shown on top of Figure 3.3. As expected, the number of significant links is substantially larger during the most severe period of the crisis. In this period, the network is dense, showing a strong volatility interdependence. A strong interaction can be found among Italy, France and Great Britain. In this triangle, Great Britain appears dominant, while Italy seems to be dominated. In the second subperiod, Great Britain and Germany are strongly connected markets, while Spain is dominated by the other markets. In general, changes in the network seem to be dramatic, even if, in both intervals, there are no completely independent markets, suggesting a continuous interactions among European countries.

Some diagnostic tests are needed in order to check the goodness-of-fit of the as-

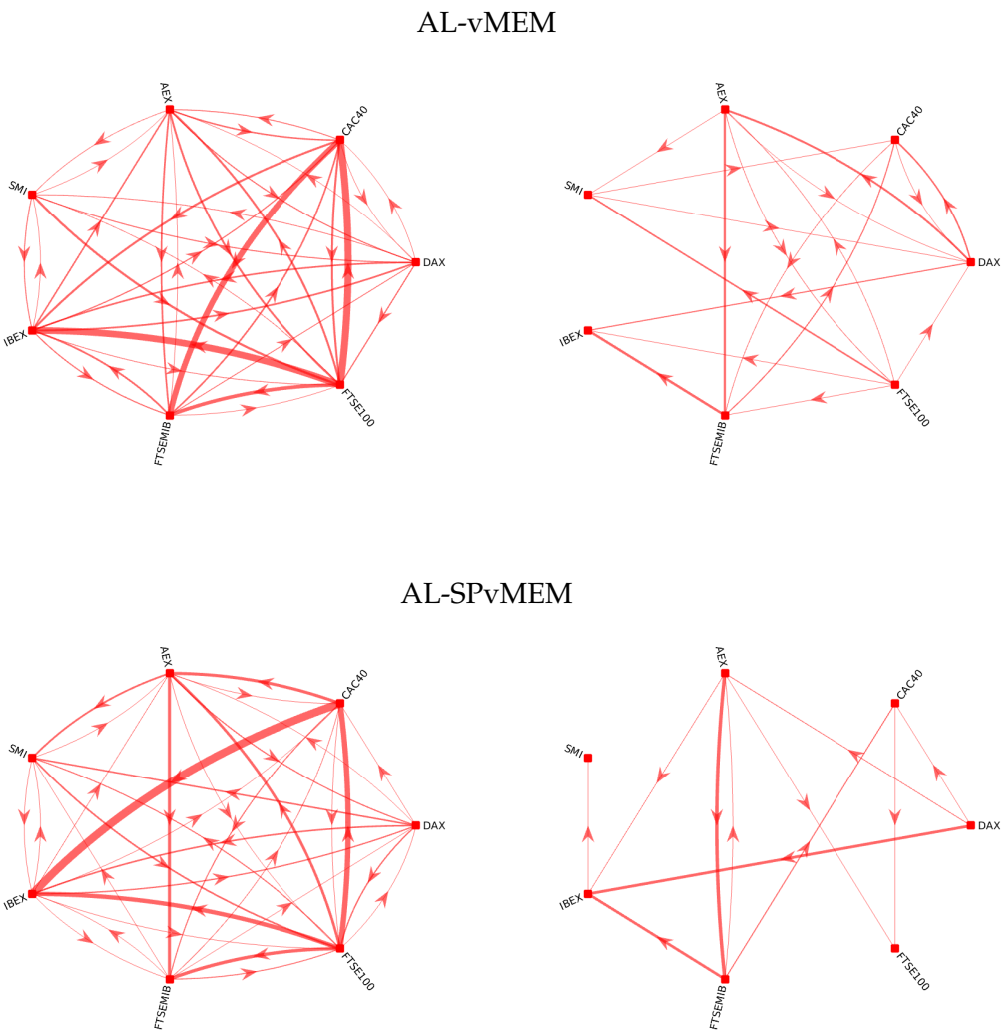

Figure 3.3: The network of interactions. On the left (right) we show the results for the interval 2010-2012 (2013-2015). The first (second) line shows the results for the AL-vMEM (AL-SPvMEM). The thickness of the links is proportional to the strength of the interaction.

model	years	residuals	DAX	CAC40	AEX	SMI	IBEX	FTSEMIB	FTSE100
AL-vMEM	2010-2012	$\hat{\epsilon}_t$	0.029	0.006	0.010	0.201	0.037	0.014	0.023
		$\hat{\epsilon}_t^2$	0.062	0.054	0.048	0.280	0.110	0.015	0.096
	2013-2015	$\hat{\epsilon}_t$	0.027	0.050	0.093	0.109	0.092	0.030	0.231
		$\hat{\epsilon}_t^2$	0.249	0.467	0.618	1.000	0.453	0.356	0.377
AL-SPvMEM	2010-2012	$\hat{\epsilon}_t$	0.062	0.012	0.012	0.172	0.030	0.028	0.021
		$\hat{\epsilon}_t^2$	0.158	0.109	0.049	0.433	0.109	0.049	0.131
	2013-2015	$\hat{\epsilon}_t$	0.029	0.043	0.083	0.108	0.028	0.010	0.280
		$\hat{\epsilon}_t^2$	0.256	0.456	0.656	1.000	0.287	0.183	0.457

Table 3.2: The p-value of the Ljung-Box test for the null hypothesis of absence of autocorrelation in the first 22 residuals.

sumed model since the correct functional form of the innovations distribution and of the conditional mean $\boldsymbol{\mu}_t$ are essential for the validity of Theorem 2. First we consider univariate tests, focusing on possible misspecifications in the conditional mean of the seven series. If $\boldsymbol{\mu}_t$ is misspecified, then also the estimated ω, A and B (which completely determine the dynamics of $\boldsymbol{\mu}_t$) and the resulting networks of interactions in Figure 3.3 may be misrepresented.

In Table 3.2 we present the p-values of the Ljung Box test for the null hypothesis of absence of autocorrelation in the first 22 residuals. Only small (usually insignificant) autocorrelation is present, in contrast with commonly used time-series models for realized volatility [Corsi et al., 2008]. Then, in Figure 3.4, we show the cross-correlations of the residuals $\hat{\boldsymbol{\epsilon}}_t$ for the interval 2010-2012.[3] Only mild correlations at lag 1 are still present in the residuals, meaning that the conditional mean $\boldsymbol{\mu}_t$ has a good specification and the estimated A and B are reliable proxies for the network of interactions.

Then we consider the standardized log-residuals

$$\mathbf{e}_t = \hat{V}^{-1/2}(\log \hat{\boldsymbol{\epsilon}}_t - \hat{\mathbf{m}}),$$

which are independently normally distributed with zero mean and the covariance ma-

[3]The results for the interval 2013-2015 are very similar and available upon request.

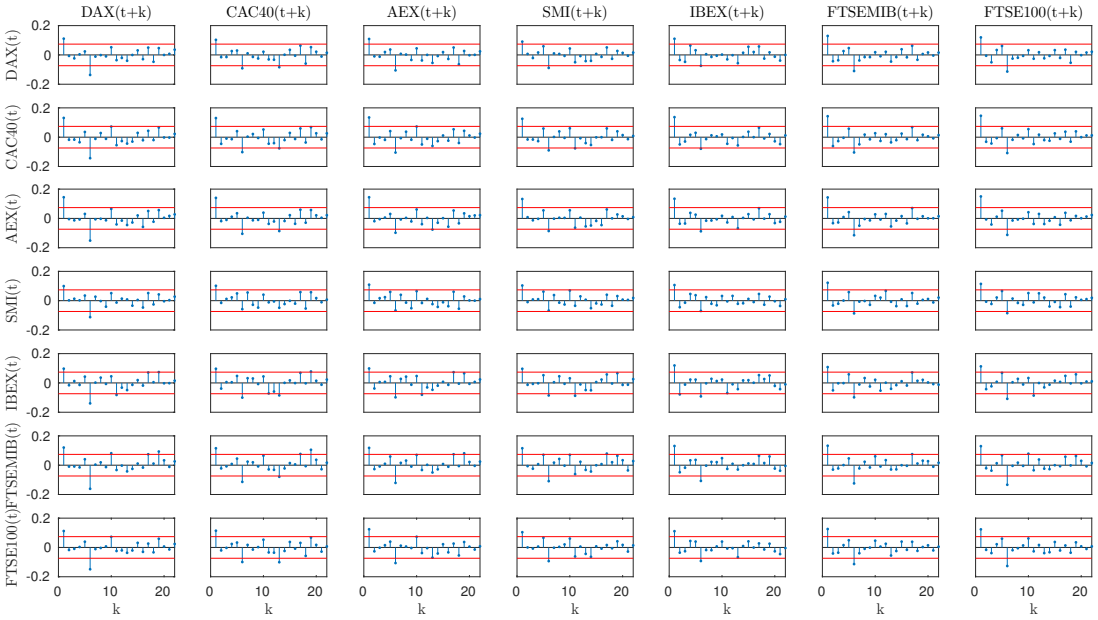

Figure 3.4: Auto and cross-correlogram of the residuals $\hat{\epsilon}_t$ for the vMEM with Adaptive Lasso penalization in the interval 2010-2012.

trix equal to an identity matrix if the conditional distribution is correctly specified. The one-sample Kolmogorov-Smirnov (KS) test, the Lilliefors [1967] (Lillie) test and the Anderson and Darling [1952] (AD) test[4] for the null hypothesis that the standardized log-residuals come from a normal distribution are present in Table 3.3. These tests indicate that the log-normality assumption for the univariate distributions is fairly appropriate, as also verified by the visual inspection of QQ-plot of the residuals in Figure 3.5, where the deviations from log-normality are concentrated in the right tails.

Following Bauwens et al. [2006], we then evaluate the multivariate normality of the log-residual testing whether $\text{cov}(e_{i,t}^2, e_{j,t}^2) = 0, \ \forall i \neq j$, as implied in the normality assumption. The corresponding test, first developed by Ding and Engle [2001], considers the following Lagrange Multiplier test statistic $T\,R^2$, where R^2 is the uncentered R-squared from a regression of 1 on $[\boldsymbol{k}_t', \boldsymbol{s}_t]$ and $\boldsymbol{k}_t$ is a $N(N-1)/2 \times 1$ vector with el-

[4]In the KS and AD tests, the reference probability distribution is the standard normal distribution. Alternatively, one could adopt the more sophisticated tests introduced in Hong and Lee [2011], Perera and Silvapulle [2017], Kheifets [2015] for checking the goodness-of-fit of the assumed distribution.

model	years	test	DAX	CAC40	AEX	SMI	IBEX	FTSEMIB	FTSE100
AL-vMEM	2010-2012	KS	0.588	0.502	0.536	0.059	0.308	0.082	0.139
		Lillie	0.208	0.073	0.156	0.023	0.019	0.001	0.004
		AD	0.324	0.215	0.538	0.065	0.106	0.128	0.156
	2013-2015	KS	0.115	0.054	0.337	0.008	0.003	0.699	0.380
		Lillie	0.001	0.001	0.057	0.001	0.001	0.222	0.074
		AD	0.045	0.067	0.214	0.000	0.005	0.435	0.269
AL-SPvMEM	2010-2012	KS	0.405	0.429	0.406	0.127	0.263	0.134	0.156
		Lillie	0.081	0.068	0.085	0.012	0.011	0.004	0.004
		AD	0.240	0.173	0.524	0.094	0.084	0.110	0.163
	2013-2015	KS	0.067	0.032	0.335	0.015	0.010	0.327	0.278
		Lillie	0.001	0.001	0.041	0.001	0.001	0.027	0.036
		AD	0.045	0.059	0.152	0.000	0.004	0.422	0.295

Table 3.3: The p-value of one-sample Kolmogorov-Smirnov (KS), Lilliefors (Lillie) and the Anderson-Darling (AD) tests for the null hypothesis that the standardized log-residuals $e_{i,t}$, with $i = 1, ..., 7$ come from a normal distribution.

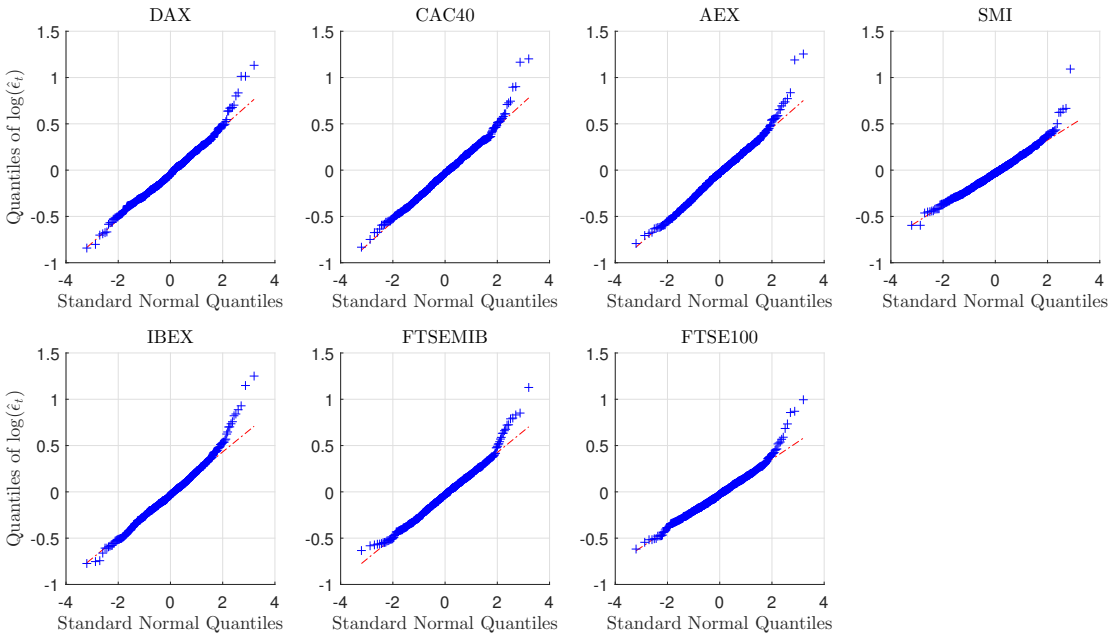

Figure 3.5: The QQ-plot of the residuals $\hat{\epsilon}_t$ for the vMEM with Adaptive Lasso penalisation in the interval 2010-2012.

ements $(e_{i,t}^2 - 1)(e_{j,t}^2 - 1)$, $i \neq j$, while $\boldsymbol{s}_t$ is the score vector of the model under the assumption of a log-normal distribution mentioned before. Such a test statistic has an asymptotic χ^2 distribution with $N(N+1)/2$ degrees of freedom. The null hypothesis is rejected at the 1% level, a result confirmed also by the Mardia [Mardia, 1970] and the Henze-Zirkler [Henze and Zirkler, 1990] multivariate normality tests on the standardized log-residuals.

Since, as stated in the Sklar's theorem every multivariate cumulative distribution function of a random vector can be expressed in terms of its marginals and a copula, the acceptable behavior of the marginals coupled with the rejection of the null of multivariate log-normality shifts the focus on the difficulties by the Normal copula (implied by the log–normal joint distribution of the innovations) in capturing some characteristics of the contemporaneous dependence (presumably some common extreme events affecting the tails).[5] In this respect, also Cipollini et al. [2017] found the Gaussian copula to be outperformed by the Student t copula, which allows for asymptotically dependent tails and generates in the estimation results significantly smaller information criteria values and more precise out-of-sample forecasts. However, such a misspecification seems to be of secondary importance in light of the Monte Carlo simulation results of Section 2.5 which gave us reliable estimates, especially in variable selection. The Monte Carlo results show, indeed, that the effects of a misspecification in the conditional distribution of the innovations are in practice almost negligible when compared to the correctly-specified case.

3.3 Volatility spillovers in the presence of a low-frequency common component

By looking at the realized volatilities in Figure 3.1, one sees that they are characterized by the presence of a common secular trend (even more strikingly so, superimposing the

[5] Extending our framework to other copulas seems to be beyond the scopes of this paper, as it could lead to problematic estimation procedures, so it is left for future research.

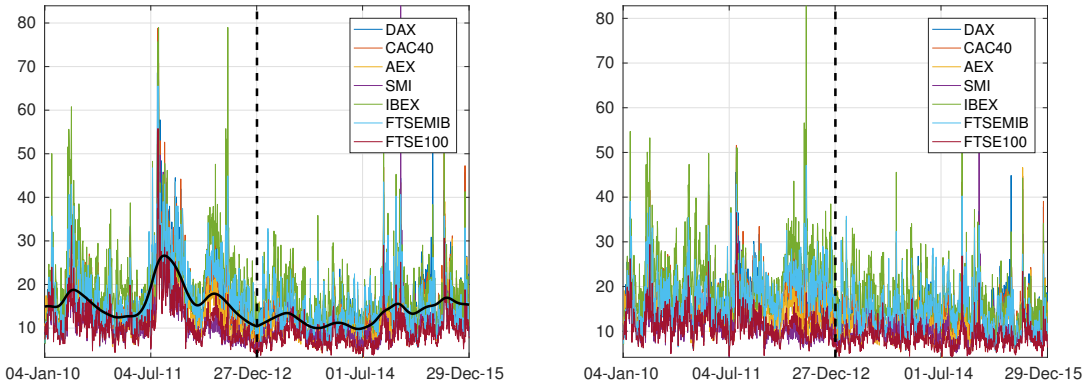

Figure 3.6: On the left: RV series $\mathbf{y}_t$ (annualized and in percentage terms) with their common trend $\hat{\phi}(z_t)$ in black. On the right: the de-trended RV series $\tilde{\mathbf{y}}_t = \mathbf{y}_t/\hat{\phi}(z_t)$, having selected a Gaussian Kernel K (see equation (3.3)) with a bandwidth h of one month. The common trend $\hat{\phi}(z_t)$ of the first period (that is, until 27-Dec-2012) and of the second period (that is, until 29-Dec-2015) have been estimated independently.

series to each other as in Figure 3.6, on the left). This common average level could be related to an undiversifiable level of risk in the considered markets. In order to disentangle the effects of this systematic trend and of more idiosyncratic components around it, Barigozzi et al. [2014] introduced a SeminonParametric extension of the vector Multiplicative Error Model, called SPvMEM. In this model, the common low–frequency component is estimated nonparametrically, while the market specific components are assumed to follow dynamics described by univariate MEMs.

This model is particularly relevant for our analysis, because, in a forecasting horse race against a set of competing specifications conducted by Barigozzi et al. [2014], the SPvMEM has shown to deliver the best out-of-sample performance in two Realized volatility panels, the first consisting in nine SPDR Sectoral Indices of the S&P500, and the second containing ninety constituents of the S&P100.

Formally, the SPvMEM specifies the dynamics of the trend adjusted processes $\tilde{y}_{n,t} = y_{n,t}/\phi(z_t)$, with $n = 1, .., N$, via N univariate MEMs (2.1)-(2.2) as

$$\begin{aligned} \tilde{y}_{n,t} &= \mu_{n,t} \odot \epsilon_{n,t} \\ \mu_{n,t} &= \omega_n + \alpha_n \, \tilde{y}_{n,t-1} + \beta_n \, \mu_{n,t-1}, \end{aligned} \tag{3.2}$$

where $z_t = t/T$ denotes the (rescaled) time index and $\phi(z_t)$ the deterministic common

trend, which is a scalar smooth function which captures the low frequency common trend, under the assumption that $\phi : [0,1] \to \mathcal{P} \in \mathbf{R}^+$ and that $\int_0^1 \phi(u)\, du = 1$. These assumptions permit to estimate $\phi(z_t)$ with a Nadaraya–Watson estimator applied to the weighted average of the rescaled series $y_{n,t}/\mu_{n,t}$. That is, for any $z_\tau \in [0,1]$,

$$\hat{\phi}(z_t) = \frac{\sum_{\tau=1}^{T} K\left(\frac{z_\tau - z_t}{h}\right) \sum_{n=1}^{N} \frac{y_{n,t}}{\mu_{n,t}} \frac{\nu_n}{\sum_{i=1}^{N} \nu_i}}{\sum_{\tau=1}^{T} K\left(\frac{z_\tau - z_t}{h}\right)}, \tag{3.3}$$

where h is a bandwidth of the kernel K and $\nu_i = \mathbb{V}[\epsilon_{i,t}]^{-1}$. The system of equation (3.2) can be considered a particular case of equation (2.4) applied to the vector $\tilde{\mathbf{y}}_t = \mathbf{y}_t/\phi(z_t)$, where the matrices A and B are assumed to be diagonal. Hence the model can be estimated iteratively applying (3.3) and the maximum likelihood estimator (2.11) (with A and B diagonal) applied to the trend adjusted processes $\tilde{\mathbf{y}}_t$, until convergence.[6]

From an economic poit of view, the aim of Barigozzi et al. [2014] is to understand which movements are due to common ($\phi(z_t)$) and individual sources ($\mu_{n,t}$). Here we further explore this line checking whether the interaction among volatility series may or may not be completely captured by the common trend. Said differently, we extend the original SPvMEM considering a trend adjusted process $\tilde{\mathbf{y}}_t = \mathbf{y}_t/\phi(z_t)$ modelled as a vector MEM of equation (2.4)[7], relaxing the assumption on the matrices A and B, in formula:

$$\begin{aligned} \tilde{\mathbf{y}}_t &= \boldsymbol{\mu}_t \odot \boldsymbol{\epsilon}_t \\ \boldsymbol{\mu}_t &= \boldsymbol{\omega} + A\tilde{\mathbf{y}}_t + B\boldsymbol{\mu}_{t-1}. \end{aligned} \tag{3.4}$$

Then we examine if volatility spillovers can be detected (with the Adaptive Lasso approach of Section 2.3) even if the common trend has been previously removed from the series. This allows us to check if the networks of spillover effects obtained in Section 3.2 are robust to the removal of the secular common trend among the series. The model

[6] For the sake of reducing the computational burden, the number of iterations is always equal to 2 in this study.

[7] We assume a log-normal distribution for the error terms.

can be estimated iteratively applying (3.3) and (2.14)[8] until convergence. We will refer to this model specification as the AL-SPvMEM (Adaptive Lasso - SeminonParametric vMEM).

In the empirical analysis, we apply this estimation procedure to the RV series of Figure 3.1. The resulting trend adjusted process $\mathbf{y}_t/\hat{\phi}(z_t)$ is shown in Figure 3.6 while the estimated parameters can be found in Table 3.4. The network of spillover effects can be obtained as a by-product of the estimation procedure (see the second line of Figure 3.3). Interestingly, the structures of the networks are not appreciably altered by the introduction of the trend. The number of links and their strength are not substantially changed, with a larger number of links during the crisis. Actually, some differences with the previous results are present. In particular, in this first period, the triangle Italy-France-Great Britain is included in a larger system with Spain and Netherlands, while in the second period Great Britain loses its central role. The overall results indicate that the introduction of the common trend does not completely capture the possible interactions among individual indices.

Diagnostic tests (residual autocorrelation and normality tests) on the estimated AL-SPvMEM are available from Tables 3.2–3.3. Finally, we check the goodness of the fit for the DAX index, showing, in Figure 3.7, the autocorrelation of the residuals $\hat{\epsilon}_t$ and of the squared residuals $\hat{\epsilon}_t^2$. The results for the other series are summarised in the lower part of Table 3.2. In Figure 3.7 we show also the histogram of $\hat{\epsilon}_t$ and its QQ-plot. In general, the results are similar to the ones obtained with the vMEM without the common trend.[9]

3.4 Volatility forecasting

The predictive ability of the Adaptive Lasso technique applied to the vMEM needs to be brought forward within an out–of–sample forecasting exercise. For our $N = 7$ series, we choose eight possible MEM specifications, each involving just one lag (possible improvements could be obtained with a finer specification search) with log-normal

[8] Equation (2.14) must be applied to the trend adjusted process $\tilde{\mathbf{y}}_t$.

[9] Also in this case, the multivariate log-normality tests have been rejected at the 1% level.

AL-SPvMEM 2010-2012

Matrix A:

	DAX	CAC40	AEX	SMI	IBEX	FTSEMIB	FTSE100
DAX	**0.320** (0.060)	0	**0.042** (0.283)	0	0	**0.004** (0.035)	**-0.009** (0.114)
CAC40	**0.001** (0.023)	**0.235** (0.047)	**0.031** (0.106)	0	**0.028** (0.026)	0	**0.083** (0.160)
AEX	**-0.003** (0.057)	0	**0.282** (0.084)	0	**0.017** (0.026)	0	**0.041** (0.181)
SMI	**-0.024** (0.100)	0	**0.045** (0.099)	**0.293** (0.046)	**-0.003** (0.005)	**-0.000** (0.001)	**0.017** (0.133)
IBEX	**-0.025** (0.170)	**0.064** (0.068)	0	**-0.004** (0.026)	**0.340** (0.131)	**0.005** (0.102)	0
FTSEMIB	0	0	0	0	**0.091** (0.196)	**0.269** (0.684)	**-0.052** (1.127)
FTSE100	**-0.020** (0.195)	0	**0.040** (0.136)	**0.025** (0.027)	**0.016** (0.037)	**-0.015** (0.047)	**0.321** (0.319)

Matrix B:

DAX	**0.618** (0.240)	0	**-0.060** (0.581)	0	**-0.020** (0.052)	0	**0.007** (0.399)
CAC40	0	**0.599** (0.217)	**-0.034** (0.318)	0	**-0.028** (0.193)	**0.010** (0.224)	0
AEX	**-0.001** (0.018)	**-0.047** (0.161)	**0.658** (0.244)	**-0.001** (0.001)	**-0.026** (0.053)	0	0
SMI	0	0	**-0.073** (0.229)	**0.678** (0.078)	**-0.003** (0.013)	**-0.002** (0.008)	0
IBEX	0	**-0.188** (1.395)	0	0	**0.624** (0.260)	0	**0.068** (0.734)
FTSEMIB	0	**0.017** (2.850)	**-0.047** (1.133)	0	**-0.085** (0.032)	**0.663** (1.362)	0
FTSE100	0	**0.008** (0.303)	**-0.034** (0.282)	0	**-0.016** (0.024)	0	**0.581** (0.529)

AL-SPvMEM 2013-2015

Matrix A:

DAX	**0.283** (0.025)	0	0	0	0	0	0
CAC40	**0.003** (0.011)	**0.238** (0.028)	0	0	0	**0.016** (0.010)	0
AEX	**0.012** (0.053)	0	**0.256** (0.052)	0	0	0	0
SMI	0	0	0	**0.359** (0.022)	0	0	0
IBEX	0	0	0	0	**0.257** (0.026)	**0.042** (0.025)	0
FTSEMIB	0	0	0	0	0	**0.326** (0.025)	0
FTSE100	0	0	0	0	0	0	**0.321** (0.022)

Matrix B:

DAX	**0.634** (0.033)	0	0	0	0	0	0
CAC40	0	**0.670** (0.028)	0	0	0	0	0
AEX	0	0	**0.664** (0.024)	0	0	**-0.005** (0.005)	0
SMI	0	0	0	**0.563** (0.026)	**0.001** (0.002)	0	0
IBEX	**-0.044** (0.014)	0	**-0.010** (0.006)	0	**0.668** (0.036)	0	0
FTSEMIB	0	0	**-0.054** (0.011)	0	0	**0.636** (0.029)	0
FTSE100	0	**-0.003** (0.002)	**-0.001** (0.001)	0	0	0	**0.598** (0.030)

Table 3.4: In bold, the estimated matrices A and B with the Adaptive Lasso procedure (2.14). On the right of each element, the standard error of the estimates (2.15). The log-normal loss functions for the evaluation of the optimal shrinkage parameters $\hat{\lambda}_T$ are shown at bottom of Figure 3.2.

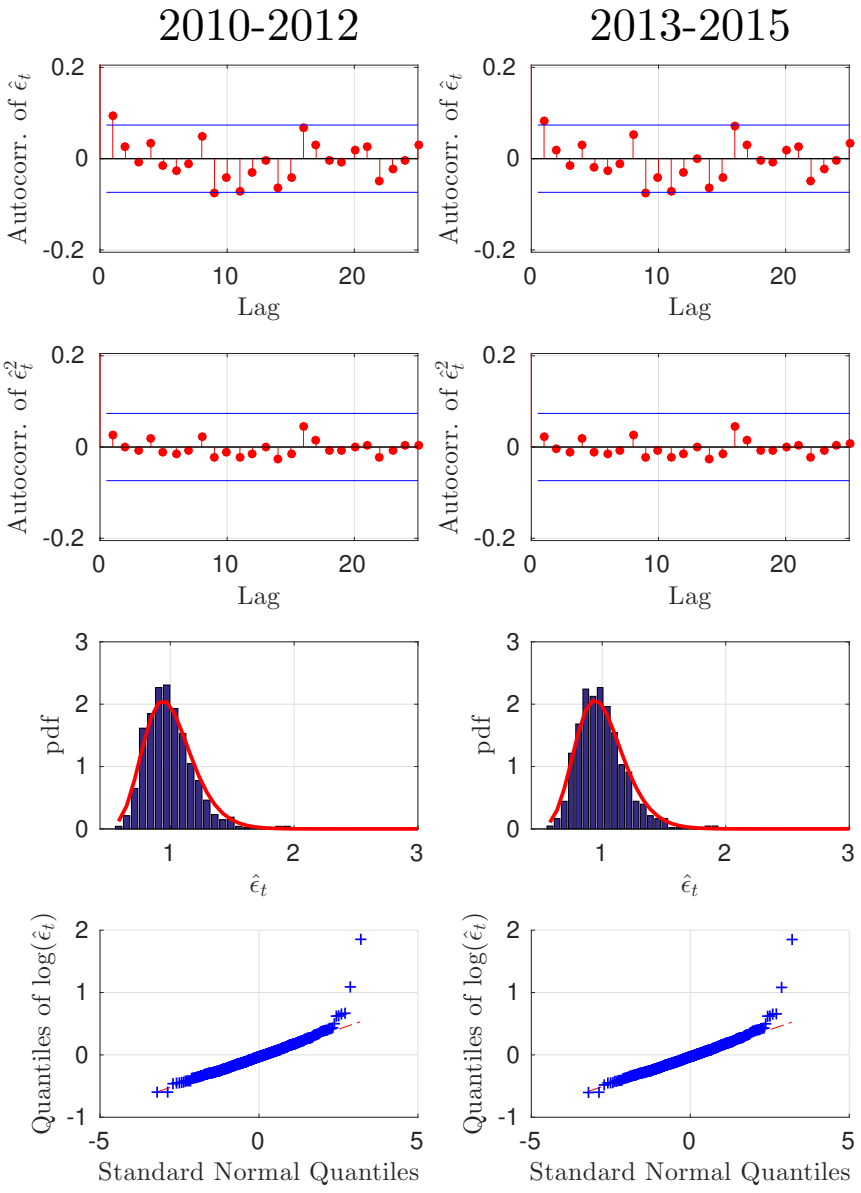

Figure 3.7: Diagnostic tests for the SPvMEM with Adaptive Lasso penalisation on the DAX index. The interval 2010-2012 in shown on the left, while the interval 2013-2015 on the right. We show the autocorrelation of the residuals $\hat{\epsilon}_t$ and of the squared residuals $\hat{\epsilon}_t^2$, then the histogram of the residuals $\hat{\epsilon}_t$ and its QQ-plot.

	Acronym	Model Description
		Without Low–frequency Component
i.	MEM	N univariate MEMs estimated with the standard *MLE* approach.
ii.	vMEM	a N-dimensional vMEM, whose log-likelihood is shown in equation (2.11).
iii.	GtS-vMEM	a N-dimensional vMEM where zeros are identified with the GtS approach of Cipollini and Gallo [2010].
iv.	AL-vMEM	an AL-vMEM estimated with the Adaptive Lasso approach of equation (2.14). This is the *benchmark model* for the case with the shortest window (T=736).
		With Low–frequency Component
v.	SPvMEM	the N-dimensional base SPvMEM [Barigozzi et al., 2014] in equation (3.4) where the matrices A and B are assumed to be diagonal and the parameters are estimated equation-by-equation.
vi.	Full SPvMEM	the N-dimensional full (multivariate) SPvMEM in equation (3.4) with full matrices A and B, and with the parameters estimated adopting the iterative *MLE* approach.
vii.	GtS-SPvMEM	a N-dimensional (multivariate) SPvMEM estimated iteratively as described in Section 3.3 but with the GtS algorithm in place of the Adaptive Lasso estimator.
viii.	AL-SPvMEM	a N-dimensional (multivariate) SPvMEM estimated iteratively with the Adaptive Lasso approach, as described in Section 3.3. This is the *benchmark model* for the case with the longest window (T=1500).

Table 3.5: List of models used in out of sample forecasting, with the acronyms used in subsequent tables.

distributions for the error terms: their description with the acronyms used is listed in Table 3.5.[10] It is a mix of models without and with the low–frequency component of the SPvMEM, possibly contemplating specification searches according to GtS and Adaptive Lasso strategies.

Starting from January 5th, 2012, to June 7, 2017, we compare the series of one-step-ahead forecasts for each model using a rolling window setup with the parameters updated every month, that is, every 22 days. In our four specifications with the low–frequency component (that is, v-vi-vii-viii), the forecasts are obtained keeping constant the last estimate of the nonparametric trend $\phi(z_t)$ for the considered forecast horizon.

[10] A Matlab package with the routines for the estimation and forecasting of the eight specifications are available at `https://sites.google.com/a/sns.it/lucacattivelli`. Minor routines are available upon request.

The comparison among the eight models is conducted comparing the sum of the absolute value (ABS) loss functions and of the Mean Squared Error (MSE) loss functions. The exercise is repeated twice, first with an estimation window of 736 observations, and then with a larger window of 1500 observations. In the first case, we selected a bandwidth h of one month (as in Section 3.2), while in the second case of two months. In both cases, the kernel K has a Gaussian shape.

In Tables 3.6 and 3.7, we report the sum of the ABS and MSE loss functions for each market and for the sum of the loss functions of the seven indices (denoted as "Total"). The symbols *,**,*** indicate when the accuracy of the considered model is significantly worse than the corresponding benchmark model, that is, the AL-vMEM for the shortest window ($T = 736$) and the AL-SPvMEM for the longest window ($T = 1500$), according to the one-sided Diebold-Mariano test. These models are used as benchmarks since they deliver the best out-of-sample forecasts in most cases.

With the shortest window ($T = 736$), the univariate MEM performs well, while full multivariate models (vMEM, Full SPvMEM) are disadvantaged by the presence of a higher number of parameters. When the larger window is considered ($T = 1500$), the vMEM improves its performance when compared to the simpler univariate specification. In general, the Adaptive Lasso method brings interesting improvements in volatility prediction over other specifications considered, especially with the shorter estimation window. Also the GtS approach deliver good forecasts, often comparable to the AL.

3.5 Conclusions

In this chapter, we have taken the issue of market volatility interdependence within selected European stock markets represented by realized volatilities on major indices. We have applied the Adaptive Lasso method for vector Multiplicative Error Models to study the dynamic interdependence of the volatility series, reconstructing the networks of volatility spillovers which crucially relies on the isolation of zeros in the correspond-

model	loss f.	DAX	CAC40	AEX	SMI	IBEX	FTSEMIB	FTSE100	Total
MEM	ABS	23.167	23.205	20.313	13.281	28.950	22.588	11.585	143.088
	MSE	0.859	0.898	0.732	0.489	1.600	0.741	0.214	5.534
vMEM	ABS	23.350	23.375	20.482	13.347	29.152	22.794	11.730	144.229
	MSE	0.883	0.915	0.741	0.487	1.622	0.784	0.224	5.656
GtS-vMEM	ABS	23.280^{***}	23.302^{**}	20.418^{*}	13.316^{**}	28.990	22.946^{**}	11.676^{*}	143.928^{***}
	MSE	0.860^{*}	0.903^{**}	0.736^{*}	0.492	1.599	0.746	0.217^{**}	5.553
AL-vMEM	ABS	23.084	23.137	20.310	13.248	28.807	22.751	11.621	142.958
	MSE	0.853	0.894	0.731	0.491	1.586	0.741	0.214	5.509
SPvMEM	ABS	23.227	23.421	20.375	13.449	28.836	22.871	11.723	143.902
	MSE	0.860	0.908	0.738	0.492	1.596	0.750	0.219	5.563
Full SPvMEM	ABS	23.353	23.577^{*}	20.540	13.554^{**}	29.431^{*}	23.431^{***}	11.825	145.711^{**}
	MSE	0.877	0.920	0.743	0.499	1.649	0.799^{**}	0.227^{**}	5.715^{*}
GtS-SPvMEM	ABS	22.974	23.145	20.238	13.400	29.064	23.201^{***}	11.616	143.638
	MSE	0.848	0.900	0.730	0.497^{*}	1.612	0.761^{**}	0.218	5.567
AL-SPvMEM	ABS	23.072	23.158	20.268	13.326	28.936	23.237^{***}	11.646	143.645
	MSE	0.857	0.904	0.736	0.497^{*}	1.616^{*}	0.767^{***}	0.219	5.595

Table 3.6: The sum of the ABS and MSE loss functions with a rolling window of 736 observations. The symbols ***,**,* signal that the accuracy of the considered model is significantly worse than the AL-vMEM at the 1%, 5%, 10% significance level according to the one sided Diebold-Mariano test. The column "Total" indicates the sum of the loss functions of the 7 indexes. The value of v is equal to 10.

model	loss f.	DAX	CAC40	AEX	SMI	IBEX	FTSEMIB	FTSE100	Total
MEM	ABS	23.354*	23.433	20.430	13.453	29.053	22.748	11.705	144.175
	MSE	0.866*	0.904	0.735	0.494	1.614	0.741	0.215	5.569
vMEM	ABS	23.216	23.211	20.309	13.318	28.747	22.679	11.743	143.223
	MSE	0.867	0.896	0.728	0.488	1.582	0.754	0.223	5.538
GtS-vMEM	ABS	23.312	23.288	20.409	13.362	28.914	23.069	11.756	144.110
	MSE	0.887	0.921	0.749	0.494	1.617	0.786	0.229	5.683
AL-vMEM	ABS	23.179	23.222	20.302	13.263	28.812	22.845	11.725	143.346
	MSE	0.854	0.891	0.726	0.487	1.591	0.748	0.218	5.514
SPvMEM	ABS	23.098	23.182	20.379	13.490**	28.539	22.450	11.890**	143.027
	MSE	0.853	0.894	0.729	0.492	1.602	0.736	0.219	5.526
Full SPvMEM	ABS	23.108	23.190	20.297	13.356*	28.717	22.841	11.742	143.251
	MSE	0.855	0.893	0.725	0.491	1.587	0.755	0.221	5.528
GtS-SPvMEM	ABS	22.955	23.141	20.243	13.318**	28.727	22.999	11.641	143.022
	MSE	0.842	0.886	0.719	0.488	1.596	0.752	0.214	5.497
AL-SPvMEM	ABS	23.017	23.132	20.243	13.231	28.748	22.915	11.646	142.932
	MSE	0.844	0.887	0.723	0.489	1.597	0.749	0.215	5.504

Table 3.7: The sum of the ABS and MSE loss functions with a rolling window of 1500 observations. The symbols ***,**,* signal that the accuracy of the considered model is significantly worse than the AL-SPvMEM at the 1%, 5%, 10% significance level according to the one sided Diebold-Mariano test. The column "Total" indicates the sum of the loss functions of the 7 indexes. The value of v is equal to 20.

ing adjacency matrix. This have allowed us to identify which dynamic links from one market to another are insignificant.

We have focused on the parameter stability during and after the European debt crisis. Our results show a time-varying network of interactions during and after the crisis period: as one would expect, the network is dense during the crisis with less connectedness in the following years. We have also addressed the issue of commonality versus interdependence, with the idea that a common low frequency component extracted as in the multivariate SeminonParametric vector MEM (SPvMEM) of Barigozzi et al. [2014] may change the dynamic profile of short run dynamic interactions. The empirical evidence shows that this is not the case: the interdependence profile is robust to such an extraction.

Since the parametric assumptions we choose for our model (log-normal distribution, and linear specification of the conditional mean) are essential for assuring the oracle property of the Adaptive Lasso for vMEM (Theorem 2), we calculated a number of diagnostics, both in terms of informational content left in the residuals (autocorrelation tests) and of test about the distributional assumptions. These results show that the log-normal distribution is performing well in capturing the marginal distribution of the innovations, while it has troubles in fully capturing the contemporaneous dependence among the error terms. In the light of the simulation results of Section 2.5, this seems to be of lesser importance for the overall performance of the procedure.

Finally, we have identified eight models, four without and four with this low-frequency component, leading them to a forecasting comparison using the Adaptive Lasso approaches as a benchmark, since they delivers the best out-of-sample performance for most series. Formal Diebold Mariano tests confirm the different behavior across subsamples and show that there is indeed a superior performance by our approach.

Conclusions

The contribution of the thesis is twofold.

First, we have proposed a discrete-time forecasting model for integer-valued time series with seasonality and pseudo-long-memory patterns. The model, called SHARP, has been formulated for a generic integer-valued random variable, but in our empirical applications, we have put the focus on the bid-ask spreads of financial equity stocks, given their prominent role in optimal execution. Although the model, formally, does not generate long-memory patterns, we have proved, with a panel of bid-ask spreads of ten NYSE equity stocks, that its forecasting accuracy outperforms that of the Long Memory ACP model of Groß-Klußmann and Hautsch [2013]. As a consequence, in order to obtain reliable forecasts of the bid-ask spread, it is not necessary to adopt genuine long-memory processes.

We have also proposed an extension of the model which is thought to better exploit the filtration generated by the bid-ask spread, as in the MIDAS approach of Ghysels et al. [2004]. The new model, called MIDAS-SHARP, achieves undoubtedly the highest forecasting accuracy, keeping a fast estimation procedure. Finally, we have shown how bid-ask spread forecasts obtained with the SHARP provide a significant reduction of the total transaction costs compared to other benchmark strategies.

Second, we have introduced the use of Adaptive Lasso techniques for variable selection in vector Multiplicative Error Models (vMEM), proving that they provide the oracle property, that is, asymptotic consistency in variable selection and the same efficiency as if the set of true predictors were known in advance. With a Monte Carlo exercise we have

demonstrated the good performances of this estimator and, with an empirical application, we have detailed the effectiveness of this approach for the study of the network of volatility spillovers among European financial indices, during and after the sovereign debt crisis. Then we have shown the robustness of the networks of volatility interactions to the introduction of a common secular trend among the series, as an extension of the SPvMEM of Barigozzi et al. [2014]. Finally, we have proved that the adoption of the Adaptive Lasso method provides superior volatility forecasts compared to several MEM specifications.

List of Figures

List of Tables

Bibliography

Aurélien Alfonsi, Antje Fruth, and Alexander Schied. Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, 10(2):143–157, 2010.

Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001.

Robert F Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied mathematical finance*, 10(1):1–18, 2003.

Torben G Andersen and Tim Bollerslev. Intraday periodicity and volatility persistence in financial markets. *Journal of empirical finance*, 4(2):115–158, 1997.

Theodore W Anderson and Donald A Darling. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, pages 193–212, 1952.

Donald W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, 1991.

Francesco Audrino and Lorenzo Camponovo. Oracle properties and finite sample inference of the adaptive Lasso for time series regression models. *arXiv:1312.1473*, 2013.

Francesco Audrino and Simon D Knaus. Lassoing the HAR model: A model selection perspective on realized volatility dynamics. *Econometric Reviews*, 35(8-10):1485–1521, 2016.

Richard T Baillie. Long memory processes and fractional integration in econometrics. *Journal of econometrics*, 73(1):5–59, 1996.

Matteo Barigozzi, Christian Brownlees, Giampiero M Gallo, and David Veredas. Disentangling systematic and idiosyncratic dynamics in panels of volatility measures. *Journal of econometrics*, 182(2):364–384, 2014.

Ole E Barndorff-Nielsen, Peter Reinhard Hansen, Asger Lunde, and Neil Shephard. Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536, 2008.

Luc Bauwens and Pierre Giot. The logarithmic ACD model: an application to the bid-ask quote process of three NYSE stocks. *Annales d'Economie et de Statistique*, (60):117–149, 2000.

Luc Bauwens, Sébastien Laurent, and Jeroen V. K. Rombouts. Multivariate garch models: a survey. *Journal of Applied Econometrics*, 21(1):79–109, 2006. doi: 10.1002/jae.842.

Christian Bayer, Peter Friz, and Jim Gatheral. Pricing under rough volatility. *Quantitative Finance*, 16(6):887–904, 2016.

Monica Billio and Loriana Pelizzon. Volatility and shocks spillover before and after EMU in European stock markets. *Journal of Multinational Financial Management*, 13(4): 323–340, 2003.

Monica Billio, Mila Getmansky, Andrew W Lo, and Loriana Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics*, 104(3):535–559, 2012.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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

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

Mehmet Caner. Lasso-type GMM estimator. *Econometric Theory*, 25(01):270–290, 2009.

Luca Cattivelli and Giampiero M. Gallo. Adaptive Lasso for Vector Multiplicative Error Models. *Available at SSRN: https://ssrn.com/abstract=3220432*, 2018.

Luca Cattivelli and Davide Pirino. A SHARP model of bid-ask spread forecasts. *Available at SSRN: https://ssrn.com/abstract=2899105*, to appear in the International Journal of Forecasting, 2019.

Ray Yeu-Tien Chou. Forecasting financial volatilities with extreme values: the Conditional AutoRegressive Range (CARR) model. *Journal of Money, Credit, and Banking*, 37 (3):561–582, 2005.

F Cipollini, RF Engle, and GM Gallo. A model for multivariate positive valued processes in financial econometrics. Technical report, Working Paper 2007/16, Universita di Firenze, Dipartimento di Statistica, 2007.

Fabrizio Cipollini and Giampiero M Gallo. Automated variable selection in vector Multiplicative Error Models. *Computational Statistics & Data Analysis*, 54(11):2470–2486, 2010.

Fabrizio Cipollini, Robert F. Engle, and Giampiero M. Gallo. Vector multiplicative error models: Representation and inference. Technical Report 12690, National Bureau of Economic Research, 2006.

Fabrizio Cipollini, Robert F Engle, and Giampiero M Gallo. Semiparametric vector MEM. *Journal of Applied Econometrics*, 28(7):1067–1086, 2013.

Fabrizio Cipollini, Robert F Engle, and Giampiero M Gallo. Copula–based vMEM specifications versus alternatives: The case of trading activity. *Econometrics*, 5(2):16, 2017.

Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

Fulvio Corsi, Stefan Mittnik, Christian Pigorsch, and Uta Pigorsch. The volatility of realized volatility. *Econometric Reviews*, 27(1-3):46–78, 2008.

Russell Davidson and James G MacKinnon. *Econometric theory and methods*, volume 5. Oxford University Press New York, 2004.

Richard A Davis, Tina Hviid Rydberg, Neil Shephard, and Sarah B Streett. The CBin model for counts: testing for common features in the speed of treading, quote changes, limit and market order arrivals. *Discussion paper, Nuffield College, Oxford.*, 2001.

Khalil Dayri and Mathieu Rosenbaum. Large tick assets: implicit spread and optimal tick size. *Market Microstructure and Liquidity*, 1(01):1550003, 2015.

Francis X Diebold and Roberto S Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Zhuanxin Ding and Robert F Engle. Large scale conditional covariance matrix modeling, estimation and testing. *NYU Working Paper No. FIN-01-029. Available at SSRN: https://ssrn.com/abstract=1294569*, 2001.

Alfonso Dufour, Robert F Engle, et al. The acd model: predictability of the time between consecutive trades. *University of Reading and University of California at San Diego*, 35(3): 463–497, 2000.

Bradley Efron. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81(395):709–721, 1986.

Zoltan Eisler, Jean-Philippe Bouchaud, and Julien Kockelkoren. The price impact of order book events: market orders, limit orders and cancellations. *Quantitative Finance*, 12(9):1395–1419, 2012.

Robert Engle. New frontiers for ARCH models. *Journal of Applied Econometrics*, 17(5): 425–446, 2002.

Robert F. Engle and Giampiero M. Gallo. A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, 131(1–2):3 – 27, 2006.

Robert F. Engle and Joe Mezrich. GARCH for groups. *Risk*, 9:36–40, 1996.

Robert F Engle and Joseph Mezrich. Grappling with GARCH. *Risk*, 8(9):112–117, 1995.

Robert F Engle and Jeffrey R Russell. Autoregressive Conditional Duration: a new model for irregularly spaced transaction data. *Econometrica*, pages 1127–1162, 1998.

Robert F Engle, Giampiero M Gallo, and Margherita Velucchi. Volatility spillovers in East Asian financial markets: a MEM-based approach. *Review of Economics and Statistics*, 94(1):222–223, 2012.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan, Heng Peng, et al. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

René Ferland, Alain Latour, and Driss Oraichi. Integer-valued GARCH process. *Journal of Time Series Analysis*, 27(6):923–942, 2006.

Konstantinos Fokianos, Anders Rahbek, and Dag Tjøstheim. Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439, 2009.

Thierry Foucault, Ohad Kadan, and Eugene Kandel. Limit order book as a market for liquidity. *Review of Financial Studies*, 18(4):1171–1217, 2005.

Christian Francq and Jean-Michel Zakoïan. Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli*, 10:605–637, 2004.

R Keith Freeland and Brendan PM McCabe. Forecasting discrete valued low count time series. *International Journal of Forecasting*, 20(3):427–434, 2004.

Giampiero M Gallo and Margherita Velucchi. Market interdependence and financial volatility transmission in East Asia. *International Journal of Finance & Economics*, 14(1): 24–44, 2009.

Jim Gatheral and Alexander Schied. Optimal trade execution under geometric Brownian motion in the Almgren and Chriss framework. *International Journal of Theoretical and Applied Finance*, 14(03):353–368, 2011.

Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov. The MIDAS touch: Mixed Data Sampling Regression Models. CIRANO Working Papers 2004s-20, 2004. URL `https://ideas.repec.org/p/cir/cirwor/2004s-20.html`.

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

Axel Groß-Klußmann and Nikolaus Hautsch. Predicting bid–ask spreads using long-memory autoregressive conditional Poisson models. *Journal of Forecasting*, 32(8):724–742, 2013.

James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

Andrew C Harvey and Clara Fernandes. Time series models for count or qualitative observations. *Journal of Business & Economic Statistics*, 7(4):407–417, 1989.

N. Hautsch. Capturing common components in high-frequency financial time series: A multivariate stochastic multiplicative error model. *Journal of Economic Dynamics and Control*, 32:3978 – 4015, 2008.

Nikolaus Hautsch. *Econometrics of financial high-frequency data*. Springer Science & Business Media, 2011.

Nikolaus Hautsch, Peter Malec, and Melanie Schienle. Capturing the zero: a new class of zero-augmented distributions and multiplicative error processes. *Journal of financial econometrics*, 12(1):89–121, 2013.

Gerd Heber, Asger Lunde, Neil Shephard, and Kevin Sheppard. Oxford-Man Institute's realized library, version 0.2, 2009.

Andréas Heinen. Modelling time series count data: an autoregressive conditional Poisson model. *Available at SSRN: http://dx.doi.org/10.2139/ssrn.1117187*, 2003.

N Henze and B Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods*, 19(10):3595–3617, 1990.

Yongmiao Hong and Yoon-Jin Lee. Detecting misspecifications in autoregressive conditional duration models and non-negative time-series processes. *Journal of Time Series Analysis*, 32(1):1–32, 2011. doi: 10.1111/j.1467-9892.2010.00681.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.2010.00681.x`.

Nan-Jung Hsu, Hung-Lin Hung, and Ya-Mei Chang. Subset selection for vector autoregressive processes using Lasso. *Computational Statistics & Data Analysis*, 52(7): 3645–3657, 2008.

Roger D Huang and Ronald W Masulis. Fx spreads and dealer competition across the 24-hour trading day. *Review of Financial Studies*, 12(1):61–93, 1999.

Joann Jasiak. Persistence in intertrade durations. *Available at SSRN: http://dx.doi.org/10.2139/ssrn.162008*, 1999.

Igor L. Kheifets. Specification tests for nonlinear dynamic models. *The Econometrics Journal*, 18(1):67–94, 2015. doi: 10.1111/ectj.12040. URL `http://dx.doi.org/10.1111/ectj.12040`.

A. B. Kock and L. A. Callot. Oracle efficient estimation and forecasting with the adaptive lasso and the adaptive group lasso in vector autoregressions. In M. Meitz N. Haldrup

and P. Saikkonen, editors, *Essays in Nonlinear Time Series Econometrics*, chapter 10. Oxford University Press, 2014.

Anders Bredahl Kock and Laurent Callot. Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344, 2015.

Sang-Won Lee and Bruce E. Hansen. Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory*, 10(1):29–52, 1994. doi: 10.1017/S0266466600008215.

Erich Leo Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Hubert W Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967.

Iain L MacDonald and Walter Zucchini. *Hidden Markov and other models for discrete-valued time series*, volume 110. CRC Press, 1997.

Simone Manganelli. Duration, volume and volatility impact of trades. *Journal of Financial markets*, 8(4):377–399, 2005.

Kanti V Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

Eddie McKenzie. Ch. 16. discrete variate time series. *Handbook of statistics*, 21:573–606, 2003.

Daniel B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59:347–370, 1991.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Whitney K Newey and Kenneth D West. A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix, 1986.

Indeewara Perera and Mervyn J. Silvapulle. Specification tests for multiplicative error models. *Econometric Theory*, 33(2):413–438, 2017. doi: 10.1017/S026646661500047X.

Peter CB Phillips. Automated inference and the future of econometrics: A colloquium for et's 20th anniversary. *Econometric Theory*, 21(1):1–2, 2005.

Silviu Predoiu, Gennady Shaikhet, and Steven Shreve. Optimal execution in a general one-sided limit-order book. *SIAM Journal on Financial Mathematics*, 2(1):183–212, 2011.

Eduardo Rossi and Dean Fantazzini. Long memory and periodicity in intraday volatility. *Journal of Financial Econometrics*, 13(4):922–961, 2015.

TH Rydberg and N Shephard. Bin models for trade-by-trade data. modelling the number of trades in fixed interval of time. *Technical report 0740, Econometric Society.*, 740: 28, 2000.

Mark Schmidt, Glenn Fung, and Rmer Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *European Conference on Machine Learning*, pages 286–297. Springer, 2007.

Daniel Straumann and Thomas Mikosch. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics*, 34(5):2449–2495, 10 2006. doi: 10.1214/009053606000000803. URL `https://doi.org/10.1214/009053606000000803`.

N. Taylor and Y. Xu. The logarithmic vector multiplicative error model: an application to high frequency NYSE stock data. Cardiff Economics working papers e2013/7, Cardiff University, Cardiff Business School, Economics Section, 2013.

N. Taylor and Y. Xu. The logarithmic vector multiplicative error model: an application to high frequency NYSE stock data. *Quantitative Finance*, 17(7):1021–1035, 2017. doi: 10.1080/14697688.2016.1260756. URL `https://doi.org/10.1080/14697688.2016.1260756`.

Nicholas Taylor. The economic and statistical significance of spread forecasts: Evidence from the London Stock Exchange. *Journal of banking & finance*, 26(4):795–818, 2002.

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

Gregor NF Weiß and Hendrik Supper. Forecasting liquidity-adjusted intraday value-at-risk with vine copulas. *Journal of Banking & Finance*, 37(9):3334–3350, 2013.

Scott L Zeger. A regression model for time series of counts. *Biometrika*, 75(4):621–629, 1988.

Hui Zou. The Adaptive Lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.