Endogenous and Exogenous Volatility in the Foreign Exchange Market

LEONARDO BARGIGLI AND GIULIO CIFARELLI

WORKING PAPER N. 17/2020

DISEI, Università degli Studi di Firenze
Via delle Pandette 9, 50127 Firenze (Italia) www.disei.unifi.it

The findings, interpretations, and conclusions expressed in the working paper series are those of the authors alone. They do not represent the view of Dipartimento di Scienze per l’Economia e l’Impresa.
Endogenous and Exogenous Volatility in the Foreign Exchange Market

L. Bargigli\textsuperscript{1} and G. Cifarelli\textsuperscript{2}

\textsuperscript{1,2} Università di Firenze

September 25, 2020

Abstract

We identify two sources of heteroskedasticity in high-frequency financial data. The first source is the endogenous changing participation of heterogeneous speculators to the market, coupled with the time varying behavior of the market maker. The second source is the exogenous flow of market relevant information. We model the first one by means of a Markov switching (MS) SVAR process, and the second one by means of a GARCH process for the MS-SVAR structural errors. Using transaction data of the EUR/USD market in 2016, we detect three regimes characterized by different levels of endogenous volatility. The impact of structural shocks on the market depends on both sources, but the exogenous information is channeled to the market mostly through price. This suggests that the market maker is better informed than the speculators, who act as momentum traders. The latter are able to profit from trade because, unlike noise traders, they respond immediately to price shocks.

Keywords: heteroskedasticity, asset pricing model, heterogeneous beliefs, market making, foreign exchange market, Markov switching, GARCH, SVAR, high frequency data.

JEL codes: G12, D84, F31, C32, C55
1 Introduction

The efficiency of financial markets is a subject of great controversy among economists. Usually the claim that markets are efficient is associated to the supposed “perfect” rationality of market participants and to the free availability of information. Thus, if some agents hold private information or if some others are unable to properly elaborate the available public information, markets are doomed to be inefficient and market excess volatility and eventually bubbles ensue.

A growing stream of literature assumes that market participants might deviate from the paradigm of full rationality. In particular, models with heterogeneous agents (HAM) assume that the deviations of the price of assets from their fundamental value arise because the heterogeneous beliefs of market participants are myopic and evolve endogenously (Brock and Hommes, 1998). In parallel, the market microstructure approach provides powerful insights into the working of financial markets under condition of asymmetry of information (Vitale, 2007; King et al., 2013). In this paper we incorporate both approaches in a model which explains how the heteroskedasticity of price variations and demand might arise as a consequence of endogenous factors such as market making and market participation of bounded rational speculators. At the same time we argue that excess volatility also depends on exogenous sources, since the market must adjust to real world events which are highly unpredictable and unevenly distributed over time.

In order to quantify these two sources of volatility, we develop an articulated econometric framework which is applied to high-frequency transaction data on the EUR/USD market. The main result of our analysis is that endogenous factors play a role in determining the volatility of the market, although the most part of the latter is explained by the exogenous flow of information. The two sources of volatility tend to correlate so that, when exogenous information arrives, the speculators react more strongly to shocks and this amplifies volatility further. In these situations we observe that demand reacts positively to a positive price shock, while price reacts negatively to a positive demand shock. This result suggests that market makers are better informed than speculators and that the price they charge reflects the relevant information. Accordingly we claim that the EUR/USD market is efficient, notwithstanding bounded rationality and asymmetric information. Indeed, in our framework price reflects all the available information exactly because market makers are better informed than speculators (otherwise this role would belong to demand). Speculators are nevertheless able to gain because, acting as momentum traders, they react quickly to price shocks. At the opposite, many market participants trade with a loss, since we observe
a negative correlation of the structural errors. We will argue that this result supports the established view that noise traders act as market stabilizers on the foreign exchange market (Black 1986; Evans and Lyons 2002; King et al. 2013).

The remaining of the paper is organized as follows. In Sec. 2 we present our theoretical model. In Sec. 3 we describe the dataset we employ for our empirical analysis. In Sec. 4 we describe our estimation strategy. In Sec. 5 we present the main results of the analysis. Finally, Sec. 6 concludes.

2 The model

Our model incorporates two basic assumptions. The first one, coming from the market microstructure literature, is the dependency of price variations on the contemporaneous net order flow (Evans and Lyons 2002). The second one, coming from the literature on models with heterogeneous agents (HAM), is the dependency of current demand on price variations (Brock and Hommes 1998). The motivation for the first assumption comes from the behavior of a market maker who might find optimal to increase / decrease her bid and/or ask price following a positive / negative net order flow (Madhavan and Smidt 1993). The motivation for the second assumption comes from the myopic nature of speculators who form their expectation of tomorrow’s returns based on the returns observed to date. This general mechanism of expectation formation encompasses different typologies of speculators which are popular in the literature such as chartists, fundamentalists, contrarians etc. One frequent shortcoming of HAM models is that they assume that agents react to price signals with a delay. Although bounded rationality represents a good approximation for the actual behavior of market participants, it is not very realistic to assume that the latter do not react as quickly as possible to price variations. Actually, we should expect the contrary. An instantaneous reaction, which is the assumption we adopt in this paper, appears to be much more realistic.

In our model the market maker is a profit maximizing monopolist who trades a zero yielding asset with a large number of different types of speculators whose participation to the market evolves endogenously. We assume that the market maker knows the optimal demand of each type of speculator and that she employs this information when she solves her optimization problem. Then we suppose that the market is liquid enough to allow the market maker to adjust in advance her inventory at the current price, in order to match the projected orders of speculators. After this adjustment she announces the optimal price, taking into account a quadratic cost of inventory maintenance.
Once the new price is revealed to speculators, the latter trade according to their optimal demand in such a way that, at the end of the period, the net variation of the inventory position of the market maker is zero. Finally, the market price is updated to the optimal price of the market maker.

Figure 1: Timeline of events occurring within a single period of the model.

The timeline of events in each period of the model is pictured in Fig. 1. We remark that the assumption we make on market liquidity is analogous to the one made by [Evans and Lyons (2002)], who suppose that FX dealers trade on the wholesale market before the price adjustment on the retail market. It is also in line with evidence from the FX market and in particular with the practice of “hot potato” trading which allows FX dealers to profit from retail trading [King et al. (2013)].

The current profit of the market maker is specified as follows:

$$\Pi_{d,t} = (p_t - p_{t-1}) z_t - \frac{\omega_t}{2} z_t^2$$

where $z_t$ is the net market demand and $\frac{\omega_t}{2} > 0$ measures the time varying impact of the quadratic inventory cost $z_t^2$ on $\Pi_{d,t}$. We remark that, according to our hypotheses, the inventory cost is linked to $z_t$ since at the end of each period the inventory is brought down to zero. In practice, it is the cost of holding an amount of risky asset equal to $z_t > 0$ until it is resold to the speculators.\footnote{In the case in which $z_t < 0$, we might think that the risky asset must be exchanged against some other asset and that holding this other asset is costly too.}

The market maker maximizes $\Pi_{d,t}$ with respect to $p_t$. Taking into account the effect of the optimization variable on $z_t$ the FOC reads:

$$z_t + (p_t - p_{t-1}) \frac{\partial}{\partial p_t} z_t - \omega_t z_t \frac{\partial}{\partial p_t} z_t = 0$$

Setting $\gamma_{0,t} \equiv \frac{\partial}{\partial p_t} z_t$, we obtain

$$\Delta p_t = \left( \frac{\omega_t}{\gamma_{0,t}} - \frac{1}{\gamma_{0,t}} \right) z_t$$
We follow the literature on asset pricing with heterogeneous speculators mentioned in Sec. 1 and exclude, for simplicity, a more complex forward looking setting, e.g. like the one of Madhavan and Smidt (1993). In the HAM setting speculators are myopic mean-variance maximizers, since their future wealth is uncertain. In practice, speculators of each type $i$ maximize their expected risk adjusted profit in the next period:

$$\max_{z_{i,t}} \left\{ E_{it} [\Pi_{i,t+1}] - \frac{1}{2D} V_{it} [\Pi_{i,t+1}] \right\}$$  \hspace{1cm} (4)

where $D^{-1}$ is a risk aversion parameter linked to the variance of future profit $V_{it} [\Pi_{i,t+1}]$. Indeed the profit of speculators at $t + 1$ is determined by their net demand at $t$ and they ignore the future market price when taking their decision. The current profit of speculators of type $i$ is written as follows:

$$\Pi_{i,t} = (p_t - p_{t-1}) z_{i,t-1}$$  \hspace{1cm} (5)

where the right hand side of the equation above represents the profit obtained buying the amount $z_{i,t-1}$ at the price $p_{t-1}$ and reselling the same amount at the price $p_t$.

Then we have:

$$E_{it} [\Pi_{i,t+1}] = (E_{it} [p_{t+1}] - p_t) z_{i,t}$$  \hspace{1cm} (6)

$$V_{it} [\Pi_{i,t+1}] = V_{it} [p_{t+1}] z_{it}^2$$  \hspace{1cm} (7)

Taking into account Eqs. (6) and (7), supposing for simplicity that $V_{it} [p_{t+1}] = \sigma^2$ is constant across time and investor’s types and finally letting this term be absorbed by $D^{-1}$, we may rewrite the objective as follows:

$$\max_{z_{i,t}} \left\{ E_{i,t} [\Delta p_{t+1}] z_{i,t} - \frac{z_{i,t}^2}{2D} \right\}$$  \hspace{1cm} (8)

where $E_{i,t} [\Delta p_{t+1}]$ stands for the type $i$’s expectation of $\Delta p_{t+1}$.

We solve the FOC for $z_{i,t}$ to obtain the optimal demand of a generic speculator of type $i$:

$$z_{i,t} = DE_{i,t} [\Delta p_{t+1}]$$  \hspace{1cm} (9)

Then the total market demand is

$$z_t = D \sum_{i=1}^{S} E_{i,t} [\Delta p_{t+1}] N_{i,t}$$  \hspace{1cm} (10)
where \( N_{i,t} \) is the number of speculators of type \( i \) at \( t \) and \( S \) is the number of different types of speculators on the market. We introduce the following specification for \( E_{i,t}[\Delta p_{t+1}] \):

\[
E_{i,t}[\Delta p_{t+1}] = \sum_{k=0}^{K} g_{i,k} \Delta p_{t-k}
\]  

(11)

where the \( g_{i,k} \) are type specific fixed coefficients which measure the impact of the price increment at lag \( k \) on the expectation of speculators of type \( i \). Substituting (11) in (10) we obtain

\[
z_t = D \sum_{i=1}^{S} N_{i,t} \sum_{k=0}^{K} g_{i,k} \Delta p_{t-k}
\]

(12)

Changing the order of summation we obtain

\[
z_t = \sum_{k=0}^{K} \Delta p_{t-k} \sum_{i=1}^{S} D N_{i,t} g_{i,k}
\]

(13)

and defining

\[
\gamma_{k,t} \equiv \sum_{i=1}^{S} D N_{i,t} g_{i,k}
\]

(14)

we obtain that market demand is a time varying function of current and past prices:

\[
z_t = \sum_{k=0}^{K} \gamma_{k,t} \Delta p_{t-k}
\]

(15)

We see that, according to (15), \( \frac{\partial z_t}{\partial p_t} = \gamma_{0,t} \) which is consistent with our previous definition in Eq. (3). Thus we end up with a system of 2 simultaneous equations in \((\Delta p_t, z_t)\) plus lagged values of price increments:

\[
\begin{cases}
\Delta p_t = \left( \omega_t - \frac{1}{\gamma_{0,t}} \right) z_t + \epsilon_{0,t} \\
z_t = \sum_{k=0}^{K} \gamma_{k,t} \Delta p_{t-k} + \epsilon_{1,t}
\end{cases}
\]

(16)

The bivariate random process \( \epsilon_t = (\epsilon_{0,t}, \epsilon_{1,t}) \) is added to take into account all the exogenous factors which impact price and demand. We expect that this process has more structure than a simple i.i.d. white noise since it reflects the uneven flow of information to the market. In particular we allow...
for time varying conditional correlation and heteroskedasticity by assuming the following:

$$\epsilon_t = H_t^{\frac{1}{2}} v_t$$

(17)

where $v_t$ is a bivariate i.i.d. vector with scaled first and second moments and $H_t^{\frac{1}{2}}$ is a positive definite matrix such that $H_t$ is the conditional covariance matrix of $\epsilon_t$.

The system (16) is equivalent to a restricted SVAR model with time varying coefficients:

$$\begin{bmatrix} \Delta p_t & z_t \end{bmatrix} \begin{bmatrix} 1 & -\gamma_{0,t} \\ \omega_t - \frac{1}{\gamma_{0,t}} & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} \Delta p_{t-1} & \Delta p_{t-2} & \cdots & \Delta p_{t-K} \end{bmatrix} \begin{bmatrix} 0 & \gamma_{1,t} \\ 0 & \gamma_{2,t} \\ \cdots & \cdots \\ 0 & \gamma_{K,t} \end{bmatrix} + \begin{bmatrix} \epsilon_{0,t} \\ \epsilon_{1,t} \end{bmatrix}$$

(18)

In the empirical estimation we prefer to lift all the restrictions on the r.h.s. since there is evidence that microstructural effects on the market lead to autocorrelation of both price variations and demand. In particular, we assume that $z_t$ is autocorrelated because of microstructural effects such as order-splitting or back-stop orders. Although the lack of autocorrelation of price increments has been widely documented at the longer time scales and is viewed as a proof that the market is efficient, autocorrelation of $\Delta p_t$ cannot be ruled out at short time scales, which are those the market needs to react to new information. In particular, there is evidence of negative first-order autocorrelation of prices in high-frequency FX markets (Zhou, 1996; Cont, 2001). So, setting $y_t = (\Delta p_t, z_t)$, we end up with the following formulation:

$$A_{0,t} y_t = A_{1,t} y_{t-1} + \cdots + A_{q,t} y_{t-K} + H_t^{\frac{1}{2}} v_t$$

(19)

where

$$A_{0,t} = \begin{bmatrix} 1 & -\left(\omega - \frac{1}{\gamma_{0,t}}\right) \\ -\gamma_{0,t} & 1 \end{bmatrix}$$

(20)

and all the coefficients in $A_{1,t}, \ldots, A_{q,t}$ are unrestricted. Regarding the coefficients in $A_{0,t}$, the only ex ante hypothesis we can make is $\omega_t - \frac{1}{\gamma_{0,t}} > 0$, since we expect that market makers raise (lower) prices when there is a positive (negative) excess demand. We choose not to impose this sign restriction ex ante as a constraint and instead to verify it from the data.
3 Data description

We analyze tick-by-tick transaction data, recorded on the EBS FX Spot trading platform and provided by NEX data. For the purpose of our analysis, the data are sampled with a 5 minutes frequency. More precisely, we investigate the following time series: the (last) Euro/Dollar midpoint between bid and offer prices \( p_t \) expressed in USD cents; the total values in period \( t \) of the bid \( BSZ_t \) and offer \( OSZ_t \) order flows expressed in Millions of Euros. The sample spans the time interval from 3 January 2016, 17:55 (5.55 pm.), to 30 December 2016, 21.55 (9.55 pm.). It includes relevant events related to the final spasms of the EMU crisis, which interact with the vagaries of the British Brexit referendum campaign. A painstaking synchronization of the time series has reduced the sample length to 36,207 observations per continuous time series. Weekends, holidays and late-evening / night periods are excluded.

Fig. 2 exhibits the EUR/USD exchange rate in levels \( p_t \) and in 5 minutes first differences expressed \( \Delta p_t \), together with the 5 minutes difference between the total values of the offer and bid order flows, which is our measure of net demand: \( z_t = OSZ_t - BSZ_t \). The exchange rate \( p_t \) is clearly non-stationary, as corroborated by a unit root analysis (data not shown), and is affected by the outcome of the Brexit referendum (June 24) and by the election of Donald Trump as president of the US (November 11), which are marked with a thick vertical line in the graph. The appreciation of the dollar in the wake of the two events clearly stands out, although the price and demand movements which are associated with these events are by no means the largest in magnitude. Indeed, the largest price swings followed a controversial announcement by the ECB of a further expansion of Quantitative Easing on March 10\(^2\) and the announcement of weak US jobs data on June 3\(^3\), both of which lead to a stark appreciation of the Euro. The largest demand swing instead was triggered by the Euro falling below the critical 1.04 $ benchmark for the first time in 14 years on December 15\(^4\).

It should be noted that large demand swings do not necessarily correspond to large price swings and vice versa. The explanation is twofold: on the one hand, most public information is incorporated directly into price and, on the other, the EUR/USD market is liquid enough to absorb large demand imbalances without large price effects. This goes without denying that the two variables influence each other: indeed the Pearson correlation coefficient between \( \Delta p_t \) and \( z_t \) is 0.305. The source of this correlation is twofold.

\(^2\)https://www.ft.com/content/02ec97ea-e6d9-11e5-bc31-138df2ae9ee6
\(^3\)https://www.ft.com/content/eb77d7d6-2937-11e6-8ba3-cdd781d02d89
\(^4\)https://www.ft.com/content/25fba186-fe9f-3bf2-a020-82efe29f1f7b
On the one hand, we have the behavioral interaction of market makers and speculators as depicted by the model of Sec. 2. On the other we have a common factor affecting both price and demand, namely the exogenous flow of information which acts simultaneously on both market makers and speculators. The first source is captured by the coefficients $A_0,t, A_1,t, \ldots, A_q,t$ of the model (19), the second one is captured by the coefficients of a parametric representation of the process which drives $H_t$. The purpose of the estimation described the next section is to distinguish and quantify each of the two sources. At the same time we must be aware of a potential additional factor at work. This is represented by the fact that we are dealing with a wholesale market whose participants do not necessarily trade at their own initiative. Indeed, according to the practice of “hot potato” trading mentioned in the previous section, dealers replicate on the wholesale market the trades of their customers in order to offload their inventory (King et al., 2013). Thus the trades in our dataset do not necessarily represent the goals and beliefs of those who submit them. Instead they might represent goals and beliefs of different classes of traders, among which liquidity / noise traders might pre-
4 Estimation Strategy

Previous attempts to estimate a model such as (19)-(20) are apparently not available in the econometric literature. Primiceri (2005) has proposed a Bayesian approach to estimate a time-varying SVAR, where both the coefficients of the SVAR and the covariance matrix of the shocks are allowed to vary over time. But this comes at the cost of imposing the usual a priori identification restrictions on $A_{0,t}$, which are unfortunately unwelcome in our case. Indeed feedback trading (which implies $\gamma_{0,t} \neq 0$) is ubiquitous in foreign exchange market since it is associated with price-contingent trading strategies such as stop-loss and profit-taking orders (Osler 2003, 2005; Danielsson and Love 2006). Neither we can suppose that $\omega_t = 0$, since the latter restriction would imply that market makers are not adverse to inventory risk, contrary to a widespread evidence (King et al. 2013). Finally, we cannot even assume that $\omega_t - \frac{1}{\gamma_{0,t}} = 0$, since this would contradict our basic intuition about the behavior of market maker as discussed at the end of Sec. 2.

A well established option, when a priori restrictions cannot be made, is to find good instruments. This is impossible to achieve in our case given that market efficiency rules out by principle that good predictors of one-period-ahead returns are available. An increasingly popular alternative is based on the observation that, if the variances of the structural shocks change over time, the coefficients of interest can be identified from the different reduced-form covariances (Rigobon, 2003). Under this approach, in order to obtain identification we need to impose some parametric restrictions on the structure of the covariance matrix, which might take the form of variance regimes (Rigobon, 2003; Lanne et al. 2010), restricted or unrestricted multivariate ARCH or GARCH processes (Rigobon 2002; Sentana and Fiorentini 2001), constant or dynamic conditional correlation processes (Weber 2010), stochastic volatility (Lewis 2018).

In our study we envisage that a GARCH specification for volatility is the best option in order to exploit the information contained in our dataset,
provided the preliminary detection of ARCH effects in the data (not shown).

In addition, according to simulation studies, a GARCH based Gaussian ML approach provides the best results when the underlying d.g.p. is conditionally heteroskedastic even if the distribution of the structural errors is misspecified \cite{Herwartz2019}, so that this approach lends itself naturally to quasi-ML estimation. The resulting complete model is:

\begin{align}
A_{0,t} y_t &= A_{1,t} y_{t-1} + \cdots + A_{K,t} y_{t-K} + \epsilon_t \\
A_{0,t} &= \begin{bmatrix} 1 & -\left(\omega_t - \frac{1}{\gamma_0,t}\right) \\
-\gamma_{0,t} & 1 \end{bmatrix} \\
\epsilon_t &= H_t^\frac{1}{2} v_t \\
v_t &\sim \mathcal{N}(0, I) \\
H_t &= \text{diag}(h_t) \\
h_t &= \bar{h} + \text{diag}(\alpha) \epsilon_{t-1} \odot \epsilon_{t-1} + \text{diag}(\beta) h_{t-1}
\end{align}

where \( \bar{h} \), \( \alpha \) and \( \beta \) are vectors of parameters and \( \odot \) stands for the Hadamard product.

We estimate the model (19)-(24) under the assumption that the participation of the heterogeneous speculators to the market follows a Markov switching (MS) autoregressive model \cite{Hamilton1989}. For this purpose it is convenient to introduce the regime vector \( \xi_t \):

\begin{equation}
\xi_t = \begin{bmatrix} 1 \{ s_t = 1 \} \\
\vdots \\
1 \{ s_t = M \} \end{bmatrix}
\end{equation}

where \( 1 \{ s_t = i \} \) represents the indicator function that the system is in state \( i \) at \( t \). Thus, for a generic number of states \( M \), we specify the time varying structural coefficients in (19) and (20) as follows:

\begin{equation}
A_{j,t} = \sum_{i=1}^{M} \xi_{i,t} A_{i,j}^t \quad (j = 0, \ldots, K)
\end{equation}

where the matrices \( A_{i,j}^t \) contain the structural parameter values at lag \( j \) when the prevailing regime is \( i \).

\footnote{In particular \cite{Herwartz2019} employ skewed and leptokurtic distributions which reflect the characteristics of financial time series."
The MS approach allows us to test if a linear model, i.e. a model with no heterogeneity, is outperformed by a non linear model, i.e. a model with heterogeneity according to our framework. The estimation proceeds through the following steps. Firstly, the reduced form parameters of the following MS-VAR model are estimated:

\[
y_t = B_{1,t} y_{t-1} + \cdots + B_{K,t} y_{t-K} + u_t
\]

\[
u_t = A_{0,t}^{-1} \epsilon_t
\]

\[
B_{j,t} = A_{0,t}^{-1} A_{j,t} = \sum_{i=1}^{M} \xi_{i,t} B_{i,j}^j \quad (j = 1, \ldots K)
\]

\[
B_{i,j}^j = A_{i,0}^{-1} A_{i,j}^j
\]

In the second step, according to the assumptions of the MS model, the following log-likelihood is maximized:

\[
L(\theta) = \sum_{t=1}^{T} \ln f_t(\theta)
\]

where

\[
f_t(\theta) = \sum_{i=1}^{M} P(s_t = i|I_t) f_{i,t}(\theta)
\]

\[
f_{i,t}(\theta) = 2\pi^{-\frac{n}{2}}|H_t^i|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} u_t^i A_{i}^i H_t^{i-1} A_{i}^i u_t \right\}
\]

We obtain the filtered conditional probability \( P(s_t = i|I_t) \) on the information set \( I_t \) from the MS-VAR model at the first step. The vector \( \theta \) contains the regime dependent parameters \( \omega = (\omega_1, \ldots, \omega_M) \) and \( \gamma_0 = (\gamma_{0,1}, \ldots, \gamma_{0,M}) \) plus the regime independent parameters of the GARCH equation (24) which takes the following regime dependent form:

\[
h_t = \sum_{i=1}^{M} \xi_{i,t} h_t^i
\]

\[
h_t^i = \bar{h} + \text{diag}(\alpha)(A_0^i u_{t-1}) \odot (A_0^i u_{t-1}) + \text{diag}(\beta) h_{t-1}^i \quad (i = 1, \ldots, M)
\]

We choose not to make the GARCH parameters depend on regimes because we wish to distinguish between the endogenous and exogenous components of volatility, where the former is linked to the variation of regimes and the latter is linked to the GARCH process.\(^8\)

\(^8\)For an identification approach based regime dependent volatility see [Lanne et al. (2010)] or [Herwartz and Lütkepohl (2014)].
In Appendix A we provide the details of the estimation procedure for the MS-SVAR-GARCH model described by eqs. (19)-(24), (26), (34)-(35), together with a discussion of the identification conditions of the structural parameter vectors $\gamma_0$ and $\omega$. Here we limit ourselves to the following considerations. The identification condition of Rigobon (2003) for a linear model is based on the existence of two distinct non proportional covariance matrices. This condition is reinforced by the GARCH setting of this paper, since introducing additional non proportional covariance matrices makes the structural coefficients of the linear model overidentified. Indeed, the non proportionality of the time varying covariance matrices is a.s. assured if there are ARCH effects in the errors. Following this line of reasoning, Milunovich and Yang (2013) have provided a set of sufficient identification conditions for linear models with ARCH type errors, which state that identification in a GARCH setting is achieved if no structural shock is degenerate (i.e. $h_i > 0$ in eq. (24)) and at most one structural shock $i$ is homoskedastic (i.e. $\alpha_i = 0$ for at most one shock in (24)). These conditions can be verified directly from the GARCH estimation. That they are sufficient but not necessary is obvious because, according to the argument of Rigobon, a SVAR-GARCH model (i.e. a model where the coefficients of the conditional mean equations do not vary over time) is a.s. overidentified if the GARCH process satisfies the conditions stated by Milunovic and Yang. This leaves us with enough degrees of freedom to estimate the value of the structural parameters for a limited number of different regimes as we do.

As a final step, we extend further our analysis by relaxing the assumption of orthogonality of $H_t$ and specify the covariance matrix with a Dynamic Conditional Correlation (DCC) model (Engle 2002):

$$H_t = D_t R_t D_t$$  \hspace{1cm} (36)

$$D_t = \text{diag}(\sqrt{h_t})$$  \hspace{1cm} (37)

$$R_t = \text{diag}(Q_t)^{-\frac{1}{2}} Q_t \text{diag}(Q_t)^{-\frac{1}{2}}$$  \hspace{1cm} (38)

$$Q_t = \sum_{i=1}^{M} \xi_{i,t-1} Q^i_t$$  \hspace{1cm} (39)

$$Q^j_t = (1 - a - b)\overline{Q} + a \left( D_{t-1}^{-1} A_0^t u_{t-1} \right) \left( D_{t-1}^{-1} A_0^t u_{t-1} \right) + b Q_{t-1}$$  \hspace{1cm} (40)

where $a, b$ are scalars and $h_t = (h_{0,t}, h_{1,t})$ is still given by eqs. (34)-(35). This extension relaxes the identification restrictions of the approaches based on conditional heteroskedasticity, which require that the structural errors are uncorrelated (Rigobon 2002, 2003, Ehrmann et al. 2011). This is accomplished thanks to a two step estimation procedure similar to the
standard DCC model (Engle 2002). Indeed, given eq. (36), the regime dependent likelihood (33) can be written as follows:

\[
f_{i,t}(\theta_2|\theta_1) = 2\pi^{-\frac{1}{2}}|D_t|^{-\frac{1}{2}}|R_t|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} u_t^t A_0^t D_t^{-1} R_t^{-1} D_t^{-1} A_0^t u_t \right\}
\] (41)

The parameter vector \(\theta_1\) contains the same parameters of the bivariate GARCH model. The parameter vector \(\theta_2\) contains the unknown parameters of eq. (40), \(a, b\) and \(Q\), which are estimated conditionally on the value of the parameters in \(\theta_1\).

Thanks to the proposed MS-SVAR-GARCH-DCC framework, we are able to decompose the conditional heteroskedasticity of data in two components. The first component is endogenous and relates to the time variations of \(A_0\) as a consequence of the changing participation to the market of the heterogeneous speculators and of the adjusted behavior of the market maker. The second component is exogenous and relates to the GARCH process which reflects the uneven flow of information which impacts both on heterogeneous speculators and the market maker. We believe that the possibility of discriminating between these two components of volatility is the main advantage of the proposed approach.

5 Results

We start by estimating a three alternative reduced-form models: a VAR model with \(K = 24\) lags along with a 2-state and a 3-state Markov Switching models with the same number of lags as the VAR. Additionally, in the 2-state setting we restrict the coefficients respectively of the price and demand equation to be fixed across regimes in order to evaluate if time is relevant for both or just one of the two equations.

In our theoretical setting, the VAR model should fit better to the data only in the unlikely case that heterogeneity is absent. Indeed we see from Tab. 1 that the VAR model is strongly penalized according to all information criteria. Both restrictions on the 2-states MS-VAR model are rejected by the LR test, showing that non linearity matters for both price and demand. The 3-state model is selected against the 2-state model according to all information criteria.

The expected duration of regimes is on the scale of a few periods (Tab. 2). In real time this means that the highest expected duration is of 50
Table 1: Comparison of reduced-form models

<table>
<thead>
<tr>
<th>Model</th>
<th>$\mathcal{L}$</th>
<th>AIC</th>
<th>BIC</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAR</td>
<td>-120,964</td>
<td>242,123</td>
<td>242,956</td>
<td>-</td>
</tr>
<tr>
<td>2-states MS-VAR (price restr.)</td>
<td>-104,935</td>
<td>210,181</td>
<td>211,605</td>
<td>286.94***</td>
</tr>
<tr>
<td>2-states MS-VAR (demand restr.)</td>
<td>-104,981</td>
<td>210,271</td>
<td>211,696</td>
<td>377.86***</td>
</tr>
<tr>
<td>2-states MS-VAR</td>
<td>-104,792</td>
<td>209,992</td>
<td>211,867</td>
<td>-</td>
</tr>
<tr>
<td>3-states MS-VAR</td>
<td>-102,100</td>
<td>204,818</td>
<td>207,658</td>
<td>-</td>
</tr>
</tbody>
</table>

*** significant at the 1% level. Note: $\mathcal{L}$ = likelihood; AIC = $-2 \times \mathcal{L} + 2 \times n$ of free parameters; BIC = $-2 \times \mathcal{L} + T \times n$ of free parameters; LR = loglikelihood ratio test statistics.

minutes and lowest of a bit less than 15 minutes. With low frequency data it is customary to explain regime shifts in terms of exogenous events which trigger a structural change in the system under study. In our case, considered the low duration of regimes, it seems unlikely that these might be correlated with exogenous “once-in-a-year” events. Instead the regime shifts reflect the adjustment of the market to the high-frequency flow of ordinary news, which trigger seamless actions and reactions from market participants.

If this hypothesis is true, the regime shifts should reflect themselves in variations of volatility, since the arrival of news is likely to coincide with spikes of volatility. Indeed from Tab. 3 we see that the regimes are drastically separated in terms of volatility, whose values change by one order of magnitude. In particular, we see that the high and low volatility regimes of the 2-state model split further, in the 3-state model, into an extremely high volatility regime and an intermediate one. By looking at the transition probabilities of this model (Tab. 4), we see that the market is not likely to calm down immediately, after a spike of volatility, since the intermediate volatility state 3 is most likely to succeed to the high volatility state 2. On the overall, the market jumps between the low and intermediate volatility states most of the time. Entering into a high volatility state is less likely, but once this regimes takes hold, it is likely to persist for some time.

Since the MS-VAR is a non-linear model, impulse response analysis must be adapted accordingly. Ehrmann et al. (2003) have proposed to calculate the impulse response functions (IRFs) under the assumption that a particular regime is prevailing over the entire time span covered by the IRFs. In practice, they propose to employ the IRFs of the VAR submodels of the MS-VAR model. The problem with this approach is that the probability that these IRFs are representative of the dynamics of the MS-VAR model is rapidly converging to zero along with the probability of staying in the initial regime. We follow instead the approach proposed by Krolzig (2006) who, in
<table>
<thead>
<tr>
<th>State 1</th>
<th>State 2</th>
<th>State 3</th>
<th>State 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-State MS-VAR</td>
<td>3-State MS-VAR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>State 1</td>
<td>9.84</td>
<td>6.05</td>
<td></td>
</tr>
<tr>
<td>State 2</td>
<td>2.94</td>
<td>2.98</td>
<td></td>
</tr>
<tr>
<td>State 3</td>
<td>4.28</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Expected duration periods of regimes for the 2-state and 3-state MS-VAR models

<table>
<thead>
<tr>
<th>Σ1</th>
<th>Σ2</th>
<th>Σ3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0008 0.1862</td>
<td>0.0006 0.1091</td>
<td>0.0020 0.5185</td>
</tr>
<tr>
<td>323.13</td>
<td>160.18</td>
<td>1205.87</td>
</tr>
</tbody>
</table>

Table 3: Estimated state covariance matrices of the MS-VAR models

<table>
<thead>
<tr>
<th>State 1</th>
<th>State 2</th>
<th>State 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>State 1</td>
<td>0.83</td>
<td>0.00</td>
</tr>
<tr>
<td>State 2</td>
<td>0.01</td>
<td>0.50</td>
</tr>
<tr>
<td>State 3</td>
<td>0.16</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Table 4: Transition matrix of the 3-state MS-VAR model

the spirit of Koop et al. (1996), proposes the following definition:

$$\text{IRF}_u(i, h) = \frac{\partial E[y_{t+h}|u_t, \xi_t]}{\partial u_i,t}$$

(42)

where $\xi_t$ is the regime vector (25). This means that the IRFs are conditioned on the prevailing regime at $t$ but not bounded to stay in a specific regime for all the subsequent periods.\footnote{For the detailed computation of the IRFs introduced in this Section see Appendix B.}
From the cumulative IRFs computed according to definition (42) (Fig. 3), we see that in all regimes demand responds positively and price negatively to their own shocks, as expected (see Sec. 2). The lagged cross-effects between price and demand are small, as we could expect from market efficiency considerations, and negative. In particular, considering that price variations are expressed in USD cents and the unit of demand is one million Euros, a positive demand shock worth approximately one standard deviation (38 million Euros) causes a (negative) price variation worth as little as -0.0023 USD cents across regimes in the subsequent 2 hours (i.e. $h = 24$ periods), equivalent to the 4.88% of the sample volatility of price. Instead, a positive EUR/USD price shock worth one standard deviation (0.0466 cents) causes a subsequent drop of demand of up to 6.15 millions Euros across regimes in the subsequent 2 hours, equivalent to the 16% of the sample volatility of demand. A small price impact of past demand shocks is what we expect in an efficient and extremely liquid market, like the one we consider. Instead, a more pronounced impact of past price shocks on demand is what we expect in a market which is populated by speculators who try to react as fast as possible to price shocks but may do so with some delay and in addition are forced by the microstructure of the market to split their trades over different periods.

The residuals of the MS-VAR estimation strongly reject the null hypothesis of homoskedasticity for a number of tests. In particular, both series display ARCH effects as required by the identification conditions discussed in Sec. 4 (data not shown). Then we are justified in going to the second step of the estimation as described in the same section. In order to proceed, we suppose that the variances of the structural errors follow a GARCH(1,1) process. The results, which are obtained through the maximization of the likelihood function defined by eqs. (31)-(33), are summarized in Tab. 5. All the GARCH coefficients are highly significant, as we expected from the analysis of the MS-VAR residuals. The standardized residuals and squared residuals obtained from the GARCH estimation show no signs of serial correlation (data not shown). The signs of the parameters in $\omega$ and $\gamma_0$ are consistent with the basic hypotheses of the model in Sec. 2 ($\omega > 0$), the basic intuitions regarding market making ($\omega - 1/\gamma_0 > 0$) and the prevalence of momentum traders among speculators across regimes ($\gamma_0 > 0$). We also tested all the combinations of equality restrictions among the parameters in $\omega$ and $\gamma_0$ by means of LR tests, which rejected all of them (data not shown). This results confirms that all the parameters are identified by the data and thus that three distinct regimes exist in the foreign exchange market.

The economic implication of our estimates of $\omega$ and $\gamma_0$ is interesting, since
Figure 3: Cumulative IRFs to a unit shock for the MS-VAR model (see Eq. 42)
Table 5: MS-SVAR-GARCH parameters’ estimates (eqs. (19)-(24) and (34)-(35))

<table>
<thead>
<tr>
<th></th>
<th>( \Delta p_t )</th>
<th>( z_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{h} )</td>
<td>0.000095</td>
<td>26.88</td>
</tr>
<tr>
<td></td>
<td>(0.0000)</td>
<td>(0.0000)</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>0.195222</td>
<td>0.184590</td>
</tr>
<tr>
<td></td>
<td>(0.0071)</td>
<td>(0.0045)</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.783096</td>
<td>0.760720</td>
</tr>
<tr>
<td></td>
<td>(0.0061)</td>
<td>(0.0036)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State 1</th>
<th>State 2</th>
<th>State 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \omega )</td>
<td>0.005364</td>
<td>0.000627</td>
</tr>
<tr>
<td></td>
<td>(0.000013)</td>
<td>(0.000015)</td>
</tr>
<tr>
<td>( \gamma_0 )</td>
<td>206.10</td>
<td>2088.79</td>
</tr>
<tr>
<td></td>
<td>(0.0061)</td>
<td>(0.0036)</td>
</tr>
<tr>
<td>( \omega - \frac{1}{\gamma_0} )</td>
<td>0.000511</td>
<td>0.000149</td>
</tr>
<tr>
<td></td>
<td>(0.000013)</td>
<td>(0.000016)</td>
</tr>
</tbody>
</table>

it shows that the exchange rate is stabilized by the existence of momentum traders. Indeed, from Eq. (3) we see that the simultaneous price impact of demand is equal to \( \omega_t - \frac{1}{\gamma_0} \). From the table we see that this quantity is reduced by up to 10 times thanks to the contribution of the positive values of \( \gamma_0 \). The economic rationale of this result is the following: thanks to the fact that an increase of the price quoted by market makers coincides with a simultaneous positive net order flow, the profits of market makers are maximized with a smaller price increase than if this positive feedback was absent.

The high volatility state 2 stands out since, in this regime, market makers are much less adverse to inventory build up and speculators are much more reactive to price increases. It is plausible that the market makers adjust their behavior to the behavior of the speculators since the overall price impact (bottom line of the table) is much closer to the other regimes than the respective values of the \( \omega \)’s. This result suggests that the our basic model should be extended in order to endogenize \( \omega_t \). It might appear contradictory that market makers become less risk adverse when volatility is high, but it all depends on who owns valuable information, either the market maker her-
self or the speculators. Only if the latter are better informed, it is rational
for the former to become more risk adverse, since in this case the valuable
information owned by speculators is enshrined in demand shocks. Other-
wise, information is mostly reflected in price shocks and speculators will find
themselves in the necessity to react adjusting demand as quickly as possible
to the latter.

We can get some indications on this issue from the CIRFs obtained from
our estimates, which are depicted in Fig. 4. These are computed by adapting
the definition \( \text{IRF}(i, h) = \frac{\partial E[y_{t+h} | \xi_t, A_{t-1}]}{\partial \epsilon_t} \) as follows:

\[
\text{IRF}(i, h) = \frac{\partial E[y_{t+h} | \xi_t, A_{t-1}]}{\partial \epsilon_t} \tag{43}
\]

We see that now the cross effects between price and demand are both
positive thanks to the simultaneous positive impact of both shocks. The
initial positive effects fade away over time except that in the high volatility
state 2, where a positive price shock is still causing demand to increase after
2 hours. In this regime a positive price shock worth one sample standard
deviation causes an increase in Euro demand worth 39 millions after 2 hours,
equal to 104\% of the sample standard deviation of demand itself. Instead, a
positive demand shock worth one sample standard deviation causes a price
deviation of 0.0054 cents after two hours, equal to 11.6\% of the sample stan-
dard deviation of price variations. Given the arguments above, these results
suggest that market makers have a leading role on the market when volatility
is high, i.e. when valuable public information arrives.

The last step is the estimation of the conditional correlation of the struc-
tural shocks according to Eqs. (40)-(41). The results are summarized in the
following specification of eq. (40):

\[
Q_i = \begin{pmatrix}
    1 - 0.015282 - 0.893275 \\ 0.001756 (0.0101756) \\
    (0.013087)
\end{pmatrix}
\begin{pmatrix}
    1,00519 & -0.29799 \\
    -0.29799 & 1.63862
\end{pmatrix}
\begin{pmatrix}
    0.015282 \\
    (0.001756) \\
    (0.0101756)
\end{pmatrix}
+ 0.015282 \ (D_{t-1}A_0u_{t-1}) \ (D_{t-1}A_0u_{t-1})' + 0.893275 \ Q_{t-1} \tag{44}
\]

The time evolution of the conditional variances \( h_{0,t} \) and \( h_{1,t} \) and the con-
ditional correlation \( \rho_t \) are depicted in Fig. 5. We signal with a vertical line
the following critical days (see Sec. 3): March 10 (QE expansion); June 3
(US Jobs announcement); June 24 (Brexit referendum); November 9 (Trump
election); December 15 (Euro falls below $1.04). We see that price volatility
is more clearly related to exogenous events (March 10, June 3, June 24, Nov
9), while demand volatility is related to endogenous events like the crossing
of a critical support price (December 15). This result is consistent with the
Figure 4: Cumulative IRFs to a unit shock for the MS-SVAR-GARCH model (see Eq. (43))
idea that the external information is channeled to the market mostly through price variations, as suggested by the analysis of the IRFs.

The volatility spikes of the two variables show a tendency to move in the same direction (with a Pearson correlation of 0.47). At the same time the unconditional correlation of the structural shocks is negative (-0.26), and the conditional correlation is mostly negative with only some rare exceptions. These results are robust to an alternative estimation performed assuming that the errors follow a Student distribution. Indeed the exogenous events do not appear to have large effects on $\rho_t$, with the big exception of June 3, when $\rho_t$ becomes strongly positive. Moreover, the movements of $\rho_t$ result to be negatively correlated to those of $h_{0,t}$ (-0.10) and $h_{1,t}$ (-0.09).

A candidate explanation for the negative correlation of the structural errors is the action of liquidity / noise traders on the foreign exchange market. The latter are generally assumed to trade with a loss, a feature which might explain why a positive (negative) price shock occurs along with a negative (positive) demand shock. Moreover, the effect of the negative correlation is to balance the market. This stabilizing role is exactly the one assigned to liquidity / noise traders from the market microstructure literature (Evans and Lyons, 2002; King et al., 2013). In particular, as explained above (see Sec. 3), we must take into consideration that many of the transactions we observe on the EBS platform occur because dealers wish to clear the trades of their customers, many of which are likely to be noise traders. These transactions are likely to appear as exogenous shocks in our estimations, as they don’t result from the speculative behavior we assume in our model. Under this perspective, it is consistent that the magnitude of the negative correlation between the structural errors becomes stronger when $h_0$ and $h_1$ are higher (i.e. when external information arrives), because it is exactly in those moments that noise traders are likely to lose more.

In order to account for the covariance of the structural errors, we adapt our definition of impulse response function as follows:

$$\text{IRF}_v(i,h) = \frac{\partial E \left[ y_{t+h} | A_{0,t}^{-1} H^2 \xi_t, \xi_t \right]}{\partial v_{i,t}}$$ (45)

where $H$ is the (regime independent) unconditional covariance matrix of the structural errors. We can also define a time varying counterpart which is dependent on the conditional covariance of the structural errors:

$$\text{IRF}^t_v(i,h) = \frac{\partial E \left[ y_{t+h} | A_{0,t}^{-1} H^2 \xi_t, \xi_t \right]}{\partial v_{i,t}}$$ (46)
The CIRFs obtained from Eq. (45) are depicted in Fig. 6. We see that the cumulative impact of a demand shock on price deviation turns negative under all regimes. This is consistent with the idea that the structural shocks to demand are driven in large part by noise traders, who are uninformed by definition. Instead the impact of a price shock on demand is still positive under all regimes, and largest by magnitude in the high volatility regime 2. These results confirm the leading informational role of market makers.

We represent the time-varying CIRFs obtained using eq. (46) for $h = 24$ by means of the boxplots in Fig. 7. The cumulative impact of price and demand shocks shifts dramatically its magnitude and direction over time, mostly as a consequence of the shifting conditional volatility and correlation of the structural errors. A look at the graph suggests that the actual dynamics of the market is driven by exogenous volatility much more than by endogenous volatility, since the impact of shocks appears visually to change more within states than across the states. The only exception regards the effect of price shocks on demand, which under regime 2 turns much more towards the positive. This is consistent with the previous results that suggest that price shocks are highly informative under this regime. The visual impression is not confirmed by a 2-sample KS test which rejects the null of equal distribution for all pairwise combinations across regimes. Thus we
Figure 6: Cumulative IRFs for the MS-SVAR-GARCH-DCC model (see Eq. 45)
conclude that endogenous volatility, although playing a quantitatively minor role, contributes significantly to the overall volatility of the market.

The last issue which is worth mentioning is the dynamic relationship between exogenous and endogenous volatility. We attempt to represent it by means of a simple Pearson correlation between $h_{0,t}$, $h_{0,1}$, $\rho_t$ on the one side, and the filtered probabilities of the three Markov regimes on the other. The results of Tab. 6 are unequivocal: when exogenous volatility is higher, both speculators and the market maker are more likely to react more strongly to exogenous shocks. The fact that the evolution of $\rho_t$ is unconnected to the endogenous regimes confirms that the conditional correlation does not relate to the behavior of speculators or the market maker but instead to noise trading.

6 Conclusions

Since the MS-VAR model outperforms the VAR model, we vindicate the claim of HAMs that demand on financial markets is affected by the shifting expectations of heterogeneous speculators and their changing participation to the market. At the same time, the reduced form MS-VAR coefficients explain only a small part of the volatility of price and demand. This is by no means a surprise because the EUR/USD currency pair is exchanged on a highly liquid market, where we expect that lagged effects are small. Indeed, if most of the action on the market occurs at the very short time scale, it is very likely that this action pops up in the form of simultaneous effects which contribute to the heteroskedasticity of the reduced form errors of the MS-VAR model.

We have called this component of conditional volatility endogenous since, according to the model of Sec. 2 it depends on the actions of the market maker and of the speculators. We have addressed the task of quantifying endogenous volatility by means of the MS-SVAR-GARCH-DCC model, which shows that in the foreign exchange market the simultaneous impacts of price variation on demand and viceversa are positive across regimes. This result is consistent with the basic intuitions regarding market making, which require that $\omega - \frac{1}{\gamma_0} > 0$, and with the prevalence of momentum traders among speculators across regimes, which requires $\gamma_0 > 0$.

After taking into account the unconditional covariance matrix of the structural errors, the cumulative effect of a demand shock on price turns out to be negative in all regimes, while the opposite holds for the cumulative effect of a price shock on demand (Fig. 4). We relate this asymmetry to the leading informational role of the market maker. Indeed any piece of valuable
information that is likely to raise / lower the price charged by the latter is also likely to raise /lower the demand of momentum traders and viceversa. Thus the fact that demand increases after a positive price shock means that valuable information is reflected first in the price, i.e in the information set of the market maker. Conversely, the fact that the price drops after a positive demand shock means that speculators are less informed than the market maker. In fact, if the speculators were better informed than her, the reaction of a rational market maker would be to raise and not to lower prices. Instead, lowering (raising) the price is rational for the market maker, if the buyer (seller) is considered to be uninformed, in order to prevent liquidity based speculation to catch on [Vitale 2000, Jeanne and Rose 2002]. Moreover, when volatility rears up, the market maker becomes less risk adverse ($\omega$ falls). This result can only be rationalized if momentum traders are no more informed than the market maker herself.

We synthesize our results by claiming that the foreign exchange market is efficient notwithstanding the existence of momentum traders and asymmetry of information. In our framework, the price charged by the market
Table 6: Pearson correlation between $h_{0,t}$, $h_{1,t}$, $\rho_t$ and filtered Markov probabilities.

|          | $P(s_t = 1 | I_t)$ | $P(s_t = 2 | I_t)$ | $P(s_t = 3 | I_t)$ |
|----------|-------------------|-------------------|-------------------|
| $h_{0,t}$| -0.39             | 0.40              | 0.21              |
| $h_{1,t}$| -0.42             | 0.32              | 0.29              |
| $\rho_t$ | -0.02             | 0.01              | 0.01              |

maker reflects all the available information because she is able on the one hand to make profits out of the trades of the most reactive among them, and on the other hand to avoid to be outplayed by the other market participants. Momentum traders behave rationally too since they are able to gain from their quick reaction to price shocks. On the other hand, this prize is reserved only to a few winners. The negative correlation of the structural errors suggests in fact that many market participants trade with a loss. We argue that the losers are noise traders. Indeed, for the reasons explained a few lines above, it is not possible to justify that the structural price and demand shocks are negatively correlated if we assume that momentum traders are the only source of demand. Since the effect of a negative conditional correlation of price and demand shocks is to help balancing the market, our interpretation is consistent with the role assigned to liquidity / noise traders from the market microstructure literature [Evans and Lyons, 2002; King et al., 2013]. Moreover, the temporal evolution of the conditional correlation of the structural errors is unconnected to the evolution of endogenous volatility, as measured by the filtered probabilities of Markov regimes, which suggests that $\rho_t$ is unrelated to the behavior of the market maker and of momentum traders. At the opposite, the conditional volatilities $h_{0,t}$ and $h_{1,t}$ are strongly correlated to the evolution of endogenous volatility. This suggests that they are linked to the information sets of the market maker and of momentum traders. And indeed $\rho_t$ is only weakly correlated with them as well.

We thank the participants to the virtual World Finance Conference held online on September 4-6 2020 for their useful comments. All the usual disclaimers apply.
A Identification conditions and estimation procedure

It is known that, if the identification conditions based on conditional heteroskedasticity hold, the values of the estimated parameters are unique up to a reordering, change of sign and renormalization of the columns of $A_0$ (Rigobon 2003; Ehrmann et al. 2011; Lewis 2018). This means that we need additional information to identify the structural coefficients of the model.

In order to see why this is necessary, let’s start from a generic bivariate simultaneous system:

$$x_t = a y_t + \epsilon_{0,t} \quad (A.1)$$
$$y_t = b x_t + \epsilon_{1,t} \quad (A.2)$$

where $x_t$ and $y_t$ are observed variables and $\epsilon_{0,t}$ and $\epsilon_{1,t}$ unobserved errors. After exchanging the order of the equations, the system can be rewritten equivalently as follows:

$$x_t = \frac{1}{b} y_t + \epsilon_{0,t} \quad (A.3)$$
$$y_t = \frac{1}{a} x_t + \epsilon_{1,t} \quad (A.4)$$

Indeed, since we don’t observe the errors, we cannot distinguish between $\epsilon_{0,t}$ and $\epsilon_{1,t}$, between $\epsilon_{0,t}$ and $\epsilon_{0,t}' = \frac{\epsilon_{0,t}}{a}$ or between $\epsilon_{1,t}$ and $\epsilon_{1,t}' = \frac{\epsilon_{1,t}}{b}$. Thus from the two equivalent formulations we obtain the same representation:

$$x_t = \theta_0 y_t + \epsilon_{0,t} \quad (A.5)$$
$$y_t = \theta_1 x_t + \epsilon_{1,t} \quad (A.6)$$

In order to simplify the argument we follow Rigobon (2003) and we assume that there are two different regimes $s = 0, 1$ for the variance of $\epsilon_{0,t}$ and $\epsilon_{1,t}$. In this case the system (A.1)-(A.2) satisfies exactly the order condition under the assumption that $\epsilon_{0,t}$ and $\epsilon_{1,t}$ are uncorrelated under each regime. In particular, $a$ and $b$ satisfy the following couple of equations:

$$a = \frac{w_{0,1,s} - w_{1,1,s} b}{w_{0,0,s} - w_{0,1,s} b} \quad s = 0, 1 \quad (A.7)$$

where, under each regime, $w_{0,0,s}, w_{1,1,s}, w_{0,1,s}$ are the variances and covariance of $x_t$ and $y_t$, which can be estimated from the data. In order to obtain
(A.7), these estimates are equated to their theoretical counterparts which are on their part obtained from the following reduced form solutions:

\[ x_t = \frac{\epsilon_{0,t} + \epsilon_{1,t}a}{1 - ab} \]  
\[ y_t = \frac{\epsilon_{0,t}b + \epsilon_{1,t}}{1 - ab} \]  

(A.8)

(A.9)

From the ordering (A.3)-(A.4) we obtain a different reduced form solution:

\[ x_t = \frac{(\epsilon_{0,t}b + \epsilon_{1,t}) a}{1 - ab} \]  
\[ y_t = \frac{(\epsilon_{0,t} + \epsilon_{1,t}a) b}{1 - ab} \]  

(A.10)

(A.11)

Nevertheless, subjecting (A.10)-(A.11) to the same algebraic manipulations of (A.8)-(A.9), it’s possible to show that \( a \) and \( b \) still satisfy the equations (A.7).

Equating the r.h.s. of the system (A.7) we obtain a quadratic equation which yields two distinct real and finite solutions for \( b \) when the following holds (Rigobon, 2003):

\[ w_{0,0,0}w_{0,1,1} - w_{0,0,1}w_{0,1,0} \neq 0 \]  

(A.12)

If we solve the equations (A.7) for \( b \) we obtain instead the following system:

\[ b = \frac{w_{0,0,s}a - w_{0,1,s}}{w_{0,1,s}a - w_{1,1,s}} \]  
\[ s = 0, 1 \]  

(A.13)

Equating the r.h.s of these equations it’s easy to check that, if \( b^* \) is a solution for (A.7), then \( a^* = \frac{1}{b^*} \) is a solution for (A.13). This shows that the two solutions obtained from either (A.7) or (A.13) correspond to the values of the reduced form parameters \( \theta_0 \) and \( \theta_1 \) of eqs. (A.5)-(A.6), so that only the latter are identified while \( a \) and \( b \) are not.

Most of the existing literature addresses this problem by imposing inequality restrictions on the coefficients of the model (Ehrmann et al., 2011; Herwartz and Lütkepohl, 2014; Lütkepohl and Netsunajev, 2014; Lanne and Luoto, 2020). In our case, we can show that the constraint imposed by the theoretical model of Sec. 2 on the reduced form coefficients allows the unequivocal identification of one of the two structural coefficients. For the sake of expositional simplicity we assume that \( A_0 \) is not time varying, since the argument can be replicated for each regime dependent matrix \( A_i \) and the corresponding structural coefficients.
In order to illustrate the main point, let’s write in vector form the two ways in which the reduced form errors can be written:

\[
\begin{align*}
\{ u_t &= A_0 \epsilon_t \\
A_0 &= \begin{bmatrix} 1 & \frac{1}{\gamma_0} - \omega \\
-\frac{1}{\gamma_0} & 1 \end{bmatrix} \tag{A.14} \\
\{ u_t^* &= A_0^* \epsilon_t \\
A_0^* &= \begin{bmatrix} 1 & -\frac{1}{\gamma_0} \\
\frac{1}{\gamma_0} & 1 \end{bmatrix} \tag{A.15} 
\end{align*}
\]

Now let’s suppose that \( \theta_0 \) and \( \theta_1 \) are the reduced form coefficients of either \( A_0 \) or \( A_0^* \). This yields the following two systems:

\[
\begin{align*}
\{ \theta_0 &= \omega - \frac{1}{\gamma_0} \\
\theta_1 &= \frac{1}{\gamma_0} \tag{A.16} \\
\{ \theta_0 &= \frac{1}{\gamma_0} \\
\theta_1 &= \frac{\gamma_0}{\gamma_0 \omega - 1} \tag{A.17} 
\end{align*}
\]

Let’s solve the two systems for the structural parameters. In the first case, we obtain:

\[
\begin{align*}
\omega &= \theta_0 + \frac{1}{\theta_1} \\
\gamma_0 &= \theta_1 \tag{A.18}
\end{align*}
\]

In the second case instead we obtain:

\[
\begin{align*}
\omega &= \theta_0 + \frac{1}{\theta_1} \\
\gamma_0^* &= \frac{1}{\theta_0} \tag{A.19}
\end{align*}
\]

Thus we see that \( \omega \) is identified independently from the ordering of the system while, except in the unlikely case that \( \theta_1 = \frac{1}{\theta_0} \), \( \gamma_0 \) is not identified.

Since \( \gamma_0 \) and \( \gamma_0^* \) yield by construction the same value of the likelihood, we estimate, for each regime, \( \theta_0^* \) and \( \theta_1^* \) firstly and afterwards \( \omega_i, \gamma_0, \gamma_0 \) according to the two orderings (A.14) and (A.15). In detail, we take the following steps:

1. Estimate the MS-VAR model \([27]-[30]\)
2. Estimate equation-by-equation a bivariate GARCH model on the reduced form errors of the MS-VAR model, i.e. set $\theta_0 = \theta_1 = 0$

3. Use the GARCH parameters obtained with the previous step as starting values for an equation by equation maximization of the likelihood (31)-(33)

4. Use the GARCH parameters obtained with the previous step as starting point for a bivariate maximization of the likelihood (31)-(33) in order to obtain the estimates $\hat{\theta}_0^i, \hat{\theta}_1^i$ for each regime $i = 1, 2, 3$

5. Compute the estimates $\hat{\omega}$ and $\hat{\gamma}_0$ according to the ordering (A.14): these are the ones reported in the main text

6. Compute the estimates $\hat{\omega}$ and $\hat{\gamma}_0^*$, according to the ordering (A.15)

The results obtained under step 6 are as follows:

<table>
<thead>
<tr>
<th></th>
<th>State 1</th>
<th>State 2</th>
<th>State 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega$</td>
<td>0.005369</td>
<td>0.000628</td>
<td>0.006481</td>
</tr>
<tr>
<td></td>
<td>(0.000121)</td>
<td>(0.000007)</td>
<td>(0.000169)</td>
</tr>
<tr>
<td>$\gamma_0^*$</td>
<td>1953.68</td>
<td>6697.65</td>
<td>2223.84</td>
</tr>
<tr>
<td></td>
<td>(0.0000)</td>
<td>(0.0000)</td>
<td>(0.0000)</td>
</tr>
</tbody>
</table>

Table 7: MS-SVAR-GARCH parameters’ estimates according to the ordering (A.15)

From the comparison of Tab. 7 with Tab. 5 we see that the values $\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3$ are indistinguishable, as we expected from the previous arguments, while the value of $\hat{\gamma}_0^*$ is much larger than the value of $\hat{\gamma}_0$ reported in the main text.

As explained above, the literature tries to overcome this source of indeterminacy relying mainly on restrictions imposed on $A_0$. For instance, if we deal with a standard demand-supply system, we might impose sign restrictions on the structural coefficients which translate into sign restrictions of the reduced form coefficients, making the former identifiable. In our model, the only theoretical restriction we might think of is the positivity of the adjustment coefficient in the price equation in each regime, i.e.

$$\theta_0^i = \omega_i - \frac{1}{\gamma_{0,i}} > 0 \quad i = 1, \ldots, M$$

(A.20)
which is strongly supported by our knowledge of how market makers adjust prices. This constraint could have been informative if one of the two reduced form coefficients was negative and the other positive in at least one regime. But unfortunately this is not the case with our estimation.

Alternatively, researchers have proposed to identify equations by supposing that the structural shock which is more correlated to a given variable is the shock of interest for that variable (Lewis, 2018). This approach runs the risk of being circular, because it is almost sure to end up identifying the residuals of the equation where, say, price is the dependent variable as the structural shock of the price equation in (16). At the same time, an ordering of equations respecting this principle seems in some sense more “natural” than the alternatives.

The common denominator of the different approaches in the literature is that they do not provide more than plausible arguments to identify the “right” ordering of equations.\footnote{For a discussion see Lewis (2018).} Indeed, if the issue is one of forecasting the response of dependent variables to structural shocks, this identification is unimportant since the IRFs depend only on the coefficients \( \theta_0 \) and \( \theta_1 \). It only becomes important in those contexts where the response to a policy shock is of interest, and thus we must know which is the “right” structural error corresponding to the policy shock. In our case, we wish to distinguish price and demand shocks because, as explained in the main text, we wish to argue about which type of agents (either speculators or market maker) is more informed on the market. Thus we venture to advance two plausible arguments in favour of the estimate we report in the main text.

The first argument relies on the magnitude of the coefficients. Taking into consideration the regime dependent unconditional standard deviations of Tab. 3 and considering the two alternative estimates \( \gamma_0 \) and \( \gamma_0^* \), we obtain the values for the instantaneous response of \( z_t \) to a price shock equal to \( \sigma(\Delta p_t) \) which are reported in Tab. 8. The values on the first row range between 75% and 87% of \( \sigma(z_t) \), while those in the second row range between 279% and 719% of \( \sigma(z_t) \). We consider the latter values to be the less plausible since they entail that the average instantaneous response of \( z_t \) to an average price shock is much larger than the average volatility of \( z_t \) itself.

The second argument relies on the correlation between the structural errors obtained from our model and those obtained from a restricted counterpart which uniquely identifies the ordering of the equations. To begin with, we observe that the only possible zero restriction on \( A_0 \) is \( \theta_0 = \omega - \frac{1}{\gamma_0} = 0 \), while no zero restrictions can be made on \( A_0^* \). Thus a restricted SVAR-
\[
\begin{array}{c|c|c|c}
\text{State 1} & \text{State 2} & \text{State 3} \\
\hline
\gamma_0 \times \frac{\sigma(\Delta p_t)}{\sigma(z_t)} & 0.75 & 0.87 & 0.22 \\
\gamma_0^* \times \frac{\sigma(\Delta p_t)}{\sigma(z_t)} & 7.19 & 2.79 & 2.98 \\
\end{array}
\]

Table 8: Alternative expected simultaneous impact of \(\Delta p_t\) on \(z_t\) for a shock equal to the unconditional regime dependent s.d. of \(\Delta p_t\), relative to the unconditional regime dependent s. d. of \(z_t\).

GARCH model estimated adopting the restriction \(\theta_0 = 0\) uniquely identifies the structural errors of the equation where \(\Delta p_t\) occurs as dependent variable as those of the price equation, i.e. the first equation in (16). In fact the equivalent reordering of the equations is not possible in this restricted model. Let’s denote the structural residuals of the price equation in the restricted SVAR-GARCH estimation just described as \(\epsilon_{\Delta p_t}\). In the ordering (A.14) the price equation is the first one, while in the ordering (A.15) it is the second one. Let’s denote the structural residuals of the first equation in (A.14) as \(\epsilon'_{\Delta p_t}\) and those of the second equation in (A.15) as \(\epsilon''_{\Delta p_t}\). Correspondingly, let’s denote the structural residuals of the second equation in (A.14) as \(\epsilon'_{z_t}\) and those of the first equation in (A.15) as \(\epsilon''_{z_t}\). Since the structural residuals are the same for both orderings, of course we have that \(\epsilon'_{\Delta p_t} = \epsilon'_{z_t}\). Then we can identify the “correct” equation ordering in the unrestricted model as the one which yields the highest correlation of the assumed structural residuals of price equation (i.e. either \(\epsilon'_{\Delta p_t}\) or \(\epsilon''_{\Delta p_t} = \epsilon'_{z_t}\)) with the structural residuals which are for sure those of the price equation in the restricted model (\(\epsilon_{\Delta p_t}\)).

The values of the Pearson coefficients \(\rho(\epsilon_{\Delta p_t}, \epsilon'_{\Delta p_t})\) and \(\rho(\epsilon_{\Delta p_t}, \epsilon''_{\Delta p_t})\) across regimes are reported in the following Tab. (9). The almost perfect correlation between \(\epsilon_{\Delta p_t}\) and \(\epsilon'_{\Delta p_t}\) suggests that the estimate of \(\gamma_0\) to privilege is the one corresponding to the ordering (A.14).

B Impulse Response Functions for Markov Switching Models

In this appendix we detail the computation of the IRFs given by eqs. (42)-(43) and (45)-(46). For this purpose we adjust the approach of Krolzig (2006) which employs a linear state space representation of MS models. In the first place we write the system (19) in a stacked form:
Table 9: Correlation between the regime dependent structural errors of the price equation in a restricted SVAR model \((\epsilon_{\Delta p_t})\) and those of the price equation according to the two orderings A.14 and A.15 (respectively \(\epsilon'_{\Delta p_t}\) and \(\epsilon''_{\Delta p_t}\) defined in the text).

<table>
<thead>
<tr>
<th>State 1</th>
<th>State 2</th>
<th>State 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\rho(\epsilon_{\Delta p_t}, \epsilon'_{\Delta p_t}))</td>
<td>0.91</td>
<td>0.99</td>
</tr>
<tr>
<td>(\rho(\epsilon_{\Delta p_t}, \epsilon''_{\Delta p_t}))</td>
<td>0.06</td>
<td>-0.92</td>
</tr>
</tbody>
</table>

\[
\begin{bmatrix}
A_{0,t} & 0 & \ldots & 0 \\
0 & I_n & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \vdots & I_n
\end{bmatrix}
\begin{bmatrix}
y_t \\
y_{t-1} \\
\vdots \\
y_{t-K+1}
\end{bmatrix}
=
\begin{bmatrix}
C_t \\
A_{1,t} & A_{2,t} & \ldots & A_{K-1,t} & A_{K,t} \\
I_n & 0 & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & I_n & 0
\end{bmatrix}
\begin{bmatrix}
y_{t-1} \\
y_{t-2} \\
\vdots \\
y_{t-K}
\end{bmatrix}
+
\begin{bmatrix}
\epsilon_t \\
0 \\
\vdots \\
0
\end{bmatrix}
\]  

(B.1)

where \(n\) is the dimension of \(y_t\). Eq. (B.1) can be written as

\[
A_{0,t}y_t = C_t + A_{1,t}y_{t-1} + \epsilon_t
\]

(B.2)

It is convenient to rewrite the system (B.2) in reduced form:

\[
y_t = D_t + B_{1,t}y_{t-1} + u_t
\]

(B.3)

We can decompose the vectors \(y_t\) of Eq. (B.3) and \(\xi_t\) of Eq. (25) as follows:

\[
y_t = E[y_t|\xi_{t-1}] + u_t^* \\
\xi_t = E[\xi_t|\xi_{t-1}] + \eta_t
\]

(B.4)  

(B.5)

where \(u_t^*\) and \(\eta_t\) are martingale difference sequences. Then, introducing
\( \mathbf{1}_n \) as a column vector of ones of size \( n \), we can write

\[
\begin{bmatrix}
\xi_{1,t} y_t \\
\vdots \\
\xi_{M,t} y_t
\end{bmatrix}
= \begin{bmatrix}
p_{11} D_1 & \ldots & p_{M1} D_1 \\
\vdots & \ddots & \vdots \\
p_{1M} D_M & \ldots & p_{MM} D_M
\end{bmatrix}
\begin{bmatrix}
\xi_{1,t-1} \mathbf{1}_n \\
\vdots \\
\xi_{M,t-1} \mathbf{1}_n
\end{bmatrix}
+ \begin{bmatrix}
p_{11} B_1 & \ldots & p_{M1} B_1 \\
\vdots & \ddots & \vdots \\
p_{1M} B_M & \ldots & p_{MM} B_M
\end{bmatrix}
\begin{bmatrix}
\xi_{1,t-1} y_{t-1} \\
\vdots \\
\xi_{M,t-1} y_{t-1}
\end{bmatrix}
+ \begin{bmatrix}
\xi_{1,t-1} \mathbf{u}_t^* \\
\vdots \\
\xi_{M,t-1} \mathbf{u}_t^*
\end{bmatrix}
\]  

or in short:

\[
\psi_{y,t} = \mathbf{D} \psi_{\xi,t-1} + \mathbf{B} \psi_{y,t-1} + \varepsilon_{y,t} \]  

(B.6)

The same arrangement can be made for the regime vector:

\[
\begin{bmatrix}
\xi_{1,t} \mathbf{1}_n \\
\vdots \\
\xi_{M,t} \mathbf{1}_n
\end{bmatrix}
= \begin{bmatrix}
p_{11} \mathbf{1}_n & \ldots & p_{M1} \mathbf{1}_n \\
\vdots & \ddots & \vdots \\
p_{1M} \mathbf{1}_n & \ldots & p_{MM} \mathbf{1}_n
\end{bmatrix}
\begin{bmatrix}
\xi_{1,t-1} \mathbf{1}_n \\
\vdots \\
\xi_{M,t-1} \mathbf{1}_n
\end{bmatrix}
+ \begin{bmatrix}
p_{11} \mathbf{1}_n & \ldots & p_{M1} \mathbf{1}_n \\
\vdots & \ddots & \vdots \\
p_{1M} \mathbf{1}_n & \ldots & p_{MM} \mathbf{1}_n
\end{bmatrix}
\begin{bmatrix}
\xi_{1,t-1} \mathbf{u}_t^* \\
\vdots \\
\xi_{M,t-1} \mathbf{u}_t^*
\end{bmatrix}
\]  

or in short:

\[
\psi_{\xi,t} = \mathbf{F} \psi_{\xi,t-1} + \varepsilon_{\xi,t} \]  

(B.8)

We see that the system (B.7)-(B.9) can be written as follows:

\[
\psi_t = \mathbf{B}^* \psi_{t-1} + \varepsilon_t \]  

(B.10)

where

\[
\mathbf{B}^* = \begin{bmatrix}
\mathbf{D} & \mathbf{B} \\
\mathbf{F} & \mathbf{0}
\end{bmatrix}
\]  

(B.11)

Hence the moving average representation of \( \psi_{t+h} \) given \( \psi_t \) is

\[
\psi_{t+h} = \mathbf{B}^h \psi_t + \sum_{j=1}^{h} \mathbf{B}^{*j} \varepsilon_{t+j} \]  

(B.12)

We can write the conditional expectation of \( y_{t+h} \) given \( y_t, \xi_t \) as a function of the conditional expectation of \( \psi_{t+h} \) given \( \psi_t \):

\[
E\left[y_{t+h} | y_t, \xi_t\right] = \sum_{i=1}^{M} E\left[\xi_{i,t+h} y_{t+h} | \xi_{i,t} y_t\right] =
\]

\[
= \begin{bmatrix}
\mathbf{0}_n & \ldots & \mathbf{0}_n \\
\mathbf{I}_n & \ldots & \mathbf{I}_n
\end{bmatrix}
E\left[\psi_{t+h} | \psi_t\right] = \mathbf{H} E\left[\psi_{t+h} | \psi_t\right] \]  

(B.13)
Using Eq. (B.12) we obtain:

\[ E[\psi_{t+h}\psi_t] = B^{*h}\psi_t \]  

(B.14)

and finally we compute the IRFs:

\begin{align*}
\text{IRF}_u(i, h) = & \frac{\partial E[y_{t+h}u_{t,\xi}]}{\partial u_{i,t}} = H B^{*h} \begin{bmatrix} 0_{nm \times 1} \\ \xi_{1,t} e_t \\ \vdots \\ \xi_{M,t} e_t \end{bmatrix} \quad \text{(B.15)} \\
\text{IRF}_\epsilon(i, h) = & \frac{\partial E[y_{t+h}A_0^{-1}e_{t,\xi}]}{\partial \epsilon_{i,t}} = H B^{*h} \begin{bmatrix} 0_{nm \times 1} \\ \xi_{1,t} A_0^{-1} e_t \\ \vdots \\ \xi_{M,t} A_0^{-1} e_t \end{bmatrix} \quad \text{(B.16)} \\
\text{IRF}_v(i, h) = & \frac{\partial E[y_{t+h}A_0^{-1}H_{t}^{\frac{1}{2}}v_{t,\xi}]}{\partial v_{i,t}} = H B^{*h} \begin{bmatrix} 0_{nm \times 1} \\ \xi_{1,t} A_0^{-1} H_{t}^{\frac{1}{2}} e_t \\ \vdots \\ \xi_{M,t} A_0^{-1} H_{t}^{\frac{1}{2}} e_t \end{bmatrix} \quad \text{(B.17)} \\
\text{IRF}'_v(i, h) = & \frac{\partial E[y_{t+h}A_0^{-1}H_{t}^{\frac{3}{2}}v_{t,\xi}]}{\partial v_{i,t}} = H B^{*h} \begin{bmatrix} 0_{nm \times 1} \\ \xi_{1,t} A_0^{-1} H_{t}^{\frac{3}{2}} e_t \\ \vdots \\ \xi_{M,t} A_0^{-1} H_{t}^{\frac{3}{2}} e_t \end{bmatrix} \quad \text{(B.18)}
\end{align*}

where we recall that \( A_{0,t} = \sum_{i=1}^{M} \xi_{i,t} A_i^{0} \).
References


