Linear models for the impact of order flow on prices. II. The Mixture Transition Distribution model

DAMIAN EDUARDO TARANTO*,†, GIACOMO BORMETTI‡, JEAN-PHILIPPE BOUCHAUD§, FABRIZIO LILLO† and BENCE TÓTH§

†Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy
‡Department of Mathematics, University of Bologna, Piazza di Porta San Donato 5, 40126 Bologna, Italy
§Capital Fund Management, 23-25, Rue de l’Université, 75007 Paris, France

(Received 7 June 2016; accepted 20 October 2017; published online 16 March 2018)

Modelling the impact of the order flow on asset prices is of primary importance to understand the behaviour of financial markets. Part I of this paper reported the remarkable improvements in the description of the price dynamics which can be obtained when one incorporates the impact of past returns on the future order flow. However, impact models presented in Part I consider the order flow as an exogenous process, only characterised by its two-point correlations. This assumption seriously limits the forecasting ability of the model. Here we attempt to model directly the stream of discrete events with a so-called Mixture Transition Distribution (MTD) framework, introduced originally by Raftery [J. R. Stat. Soc. Ser. B, 1985, 528–539]. We distinguish between price-changing and non price-changing events and combine them with the order sign in order to reduce the order flow dynamics to the dynamics of a four-state discrete random variable. The MTD represents a parsimonious approximation of a full high-order Markov chain. The new approach captures with adequate realism the conditional correlation functions between signed events for both small and large tick stocks and signature plots. From a methodological point of view, constraining the MTD within the class of ergodic Markov models, and exploiting the buy–sell symmetry of the data, we propose a weak restriction on the transition matrices which solves the problem of identifiability of mixture models. In spite of the large number of parameters, this translates into a feasible and robust estimation procedure. Out-of-sample analyses demonstrate that the model does not overfit the data.

Keywords: Market microstructure; Price formation; Markov chain; Forecasting ability

JEL Classification: G1, C25

1. Introduction

This paper is the companion of Taranto et al. (2016). In the previous, part I paper we discussed the differences and similarities between two linear models describing the impact of order flow on prices, namely the Transient Impact Model (TIM) and the History Dependent Impact Model (HDIM). In these models, the sign of the order flow is considered to be an exogenous, time-correlated process that affects price dynamics either through a ‘propagator’, i.e. a linear combination of past values (TIM) or via a ‘surprise’ mechanism, i.e. the deviation between the realised order flow and its expected level (HDIM). In reality, however, order flow is not exogenous and is itself affected by the past history of price. In Taranto et al. (2016), we partly overcame this issue by enhancing the description of the order flow to account for price-changing events and non price-changing events, in the spirit of Eisler et al. (2012a), Eisler et al. (2012b). This allows one to encode the propensity of the order flow to invert its sign after a price change, an effect that is particularly important for large tick stocks. This extended model improves significantly the description of the price process, both in terms of the lag-dependent volatility (i.e. the signature plot) and in terms of the response function computed for negative lags. However this approach is still incomplete as it does not specify the data generating process for the order flow itself, which is only described through two-point correlation functions. This does not allow one to forecast the future order flow itself, for example whether a trade is likely to change the price or not.

In this paper, we attempt to model the joint dynamics of order flow and prices. This family of models has a long tradition
in market microstructure, starting from the seminal work of Hashbrouck (1988, 1991), who proposed a Vector Autoregressive (VAR) model for the joint dynamics of order flow and prices. There are two main related limitations of this approach. The first is that VAR models are adequate for variables with continuous support (e.g. Gaussian), while the order flow (signs and events) and tick by tick price changes are more naturally described by discrete variables. Second, the standard VAR approach prescribes a linear relation between the variables, while a broader definition includes the possibility of a linear relation between past variables and the probability of observing in the future the value of a given variable.

A paradigmatic example is a finite state Markov chain $X_t$. Let $m$ be the number of states, $Q$ the $m \times m$ time-invariant transition matrix, and let $x_t = (x_t(1), \ldots, x_t(m))$ be a row vector such that $x_t(i) = 1$ if $X_t = i$ and zero otherwise. The probability vector $\hat{x}_t = (\mathbb{P}(X_t = 1), \ldots, \mathbb{P}(X_t = m))$ is determined by the linear system of equations

$$\hat{x}_t = x_{t-1} Q.$$  

(1)

Therefore a natural way to describe the joint dynamics of discrete valued variables (such as the order flow sign and price changes) in a linear setting is with a Markov process of large order. In fact, we have shown in Taranto et al. (2016) that for large tick stocks the model with two propagators (TIM2) corresponding to price-changing and non price-changing trades gives constant (in time) propagators when estimated on real data (see top left panel of figure 7 of Taranto et al. (2016)). This means that the knowledge of the order flow and the information on whether a trade changes the price completely characterises the price dynamics. Thus, in the framework of linear models, it is natural to describe the system with a Markov process with $m = 4$ states, $(\epsilon_t, \pi_t) \in \{(-1, C), (-1, NC), (+1, NC), (+1, C)\}$, corresponding to buys ($\epsilon_t = +1$) and sells ($\epsilon_t = -1$) and price-changing ($\pi_t = C$) and non price-changing ($\pi_t = NC$) trades.

However, the main limitation of Markov models comes from the long memory of the order flow (Bouchaud et al. 2004, Lillo and Farmer 2004). Since the order flow sign is very persistent, a low-order Markov process cannot be suitable to describe real markets. On the other hand, Markov processes of high-order $p$ depend in general on a very large number of parameters ($O(m^p)$ and might result in inefficient estimation when a limited amount of data is available. For this reason in this paper we propose to use a parsimonious, yet versatile class of high-order Markov processes termed the Mixed Transition Distribution (MTD) model (Raftery 1985) and its generalisation (MTDg) (Berchtold 1995). Thanks to a simple structure, where each lag contributes to the prediction of the current state in a separate and additive way, the dimension of model parameter space grows only linearly with the order of the MTDg model, i.e. as $O(m^2 p)$. The model can be estimated via Maximum Likelihood or via the Generalized Method of Moments. Moreover, in the case of $m = 2$ states (such as the signs of the order flow), the version of the MTDg model proposed in this paper reduces to the Discrete Autoregressive (DAR) model (Jacobs and Lewis 1978), which has been used to model the order flow in Taranto et al. (2014) and in the companion of this paper (Taranto et al. 2016). Hence MTD and MTDg aim at providing a natural generalisation of the DAR(p) model to account for an arbitrary number of $m \geq 2$ states, while avoiding the exponential increase ($O(m^p)$) of the number of parameters of the full Markov model. Perhaps surprisingly, this class of models has not been applied to financial data and the present paper attempts to fill this gap. The main methodological innovation of our work is a weakly constrained MTDg model which can be estimated even when the number of parameters is very large, as required to account for the correlation structure of financial data. The restriction consists in constraining the MTDg model within the class of ergodic Markov model. Ergodicity allows to write all the transition matrices in terms of a first component, which depends linearly on the stationary distribution, and a second term, whose kernel contains the stationary distribution. Exploiting the buy–sell symmetry present in the data, the latter term significantly simplifies and, as a result, this translates in a feasible estimation procedure.

We consider in this paper MTD and MTDg models as promising models for the joint dynamics of order flow and price changes for large tick stocks. Compared to the models investigated in Taranto et al. (2016), we provide here an explicit model for the order flow, and in particular its response to past price dynamics. Thus we aim at reproducing with the MTD model the complex conditional correlation functions of signed events for large tick stocks (see left panel of figure 4 of Taranto et al. (2016) whose curves are reproduced also in figure 1). Moreover, our modelling approach allows one to perform out of sample analyses of the MTD’s forecasting ability of the order flow and future price changes. Still, this framework has limitations when estimated on anonymised order flow because one cannot easily disentangle order flow correlations coming from ‘herding’ and coming from ‘order splitting’. In other words, although MTDs give explicit predictions for the response of the order flow to a single event (impulse response), one has to be careful in interpreting the result, as it might not describe the true reaction of the market to an isolated, exogenous order (see Tóth et al. 2012, 2015, 2017).

The paper is organised as follows. In section 2, we review the definition, main properties and estimation methods of MTD and MTDg. In section 3 we present our parametrisation of the model and some proposed improvements for the estimation. This Section also contains our empirical results on real financial data. Section 4 describes the results of some out of sample analyses for predicting price changes and order flow and in section 5 we draw some remarks and conclusions, and discuss the limitations of our approach.

2. The Mixture Transition Distribution model

2.1. Definition

We start from a simple, but restrictive, definition of MTD models. Let $(X_t)_{t \in \mathbb{Z}}$ be a sequence of random variables taking values in the finite set $\mathcal{X} = \{1, \ldots, m\}$: This random sequence

†More recent modelling in continuous time makes use of Hawkes processes (Bacry and Muzy 2014), which bear some degree of similarity with the models considered in the present paper.

‡We remind the reader that large tick stocks have the property that the ratio between tick size and price is relatively high and as a consequence spread is almost always equal to one tick.
II. The Mixture Transition Distribution model

is said to be a $p$th order MTDg sequence if for all $t > p$ and for all $(i, i_1, \ldots, i_p) \in \mathcal{X}^{p+1}$,

$$
P(X_t = i|X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \sum_{g=1}^{p} \lambda_g q_{i,g}^g, \quad (2)$$

where the vector $\lambda = (\lambda_1, \ldots, \lambda_p)$ is subject to the constraints:

$$
\sum_{g=1}^{p} \lambda_g = 1. \quad (3)
$$

The matrices $Q^g = [q_{i,j}^g]; i, j \in \mathcal{X}; 1 \leq g \leq p$ are positive $m \times m$ stochastic matrices, i.e. they satisfy

$$
q_{i,j}^g \geq 0 \quad \text{and} \quad \sum_{j=1}^{m} q_{i,j}^g = 1, \quad \forall g \in \{1, \ldots, p\}, \forall i, j \in \mathcal{X}. \quad (5)
$$

Raftery (1985) has originally defined the model with the same transition matrix $Q^g = Q$ for each lag $g = 1, \ldots, p$ and this model is called the MTD. Later, Berchtold (1995) has introduced the more general definition of MTD models as a mixture of transitions from subsets of lagged variables $\{X_{t-1}, \ldots, X_{t-p}\}$ to the present one $X_{t}$. In other words, the order of the transition matrices $Q^g$ can be larger than one.

Berchtold and Raftery (2002) have published a complete review of the MTD model. They recall theoretical results on the limiting behaviour of the model and on its autocorrelation structure. In particular, they prove that if conditions of equations (3)–(5) are satisfied, then the model of equation (2) is a well-defined high-order Markov chain and its stationary distribution $\tilde{\eta} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_m)$ exists and it is unique. The Mixture Transition Distribution models are Markov models where each lag $X_{t-1}, X_{t-2}, \ldots$ contributes additively to the distribution of the random variable $X_{t}$. Hence, the model is linear in the sense described in the introduction.

In words, this class of models means the following: in order to determine the type of event $X_{t}$, occurring at time $t$, start choosing a reference time $t - g$ in the past, where $g$ is drawn at random with probability $\lambda_g$. If the event $X_{t-g}$ that occurred at time $t - g$ is of type $i$, then choose the event at time $t$ to be of type $i$ with probability $q_{i,g}$. This model can thus be interpreted as a probabilistic mixture of Markov processes. For this interpretation, the fact that $(\lambda_g)_{g=1,..,p}$ is a probability vector and $Q^g$ are stochastic matrices is critical. However, as already noted in the original papers Raftery (1985), Berchtold (1995), the MTDg model can also defined when these parameters are negative or larger than one, provided that the conditions

$$
0 \leq \sum_{g=1}^{p} \lambda_g q_{i,g}^g \leq 1, \quad \forall (i, i_1, \ldots, i_p) \in \mathcal{X}^{p+1}. \quad (6)
$$

are satisfied, in such a way that all transition probabilities are well defined. As we shall see below, estimated parameters do not necessarily abide to the probabilistic interpretation.

In this paper, we will consider a specific class of MTDg models where the matrices $Q^g$ share the same stationary state, i.e. the same left eigenvector $\tilde{\eta}$ corresponding to the eigenvalue 1. Under this assumption, generalising a result of Berchtold (1995), we can prove the following theorem of the existence and uniqueness of the stationary distribution.

**Theorem 2.1** Suppose that a sequence of random variables $\{X_n\}_{n \in \mathbb{N}}$ taking values in the finite set $\mathcal{X} = \{1, \ldots, m\}$ is defined by

$$
P(X_t = i|X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \sum_{g=1}^{p} \lambda_g q_{i,g}^g, \quad (7)
$$

where $Q^g = [q_{i,j}^g]; i, j \in \mathcal{X}$ are matrices with normalised rows, $\sum_{j=1}^{m} q_{i,j}^g = 1$, $\sum_{g=1}^{p} \lambda_g = 1$, and assume that $\sum_{g=1}^{p} \lambda_g q_{i,g}^g = \tilde{\eta}$, $\forall g$. If the vector $\tilde{\eta}$ is such that $\tilde{\eta}_i > 0$, $i \in \mathcal{X}$ and $\sum_{i \in \mathcal{X}} \tilde{\eta}_i = 1$, then

$$
0 < \sum_{g=1}^{p} \lambda_g q_{i,g}^g < 1, \quad \forall (i, i_1, \ldots, i_p) \in \mathcal{X}^{p+1}, \quad (8)
$$

then

$$
\lim_{t \to \infty} \mathbb{P}(X_{t+\ell} = i|X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \tilde{\eta}_i. \quad (9)
$$

The proof of the theorem is in appendix I. Notice that in this theorem we do not need to assume that the parameters $(\lambda_g, Q^g)_{1 \leq g \leq p}$ are between zero and one, but the probabilistic interpretation is guaranteed by the condition of equation (8). Finally, notice that the condition on $\tilde{\eta}$ implies that $Vg$ we can write $Q^g = Q + \tilde{\eta} \tilde{\eta}$, where $\tilde{\eta} Q = \tilde{\eta}$ and $\tilde{\eta} \tilde{\eta} = 0$.

2.2. Estimation

Despite being parsimonious with respect to full Markov models, the MTDg parameters $\theta = (\lambda_g, Q^g)_{1 \leq g \leq p}$ are difficult to estimate because they have to comply with the normalisation constraints of transition matrices. In the literature many different estimation methods have been proposed (Berchtold and Raftery, 2002), but in our paper we will focus on two specific methodologies: the maximum likelihood estimation (MLE) and the generalised method of moments (GMM). Let us introduce the details of these methods.

2.2.1. Maximum likelihood estimation. For a given data sequence with length $n$, $(X_1, X_2, \ldots, X_n)$ we define $(X_1^g = x_1^g)$ as the sequence of events $(X_1, x_1, X_2, x_2, \ldots, X_n, x_n)$ and $\mathbb{P}(X_1^g = x_1^g)$ is the joint distribution of $(X_1, x_1, \ldots, x_n)$. From the definition of MTDg models of order $p$, the likelihood function is

$$
L(\theta) = \mathbb{P}_\theta(X_1^g = x_1^g) = \mathbb{P}(X_1^g = x_1^g) \mathbb{P}_\theta(X_{p+1}^g = x_{p+1}^g|X_p^g = x_p^g) = \mathbb{P}(X_1^g = x_1^g) \prod_{i=p+1}^{n} \left( \sum_{g=1}^{p} \lambda_g q_{x_i,g}^g \right). \quad (10)
$$

To estimate the parameters of MTDg model, we have excluded $\mathbb{P}(X_1^g = x_1^g)$ from the likelihood function. Therefore, the log-likelihood function that we consider is

$$
\ell(\theta) = \log \mathbb{P}_\theta(X_{p+1}^g = x_{p+1}^g|X_p^g = x_p^g) = \sum_{i=p+1}^{n} \log \left( \sum_{g=1}^{p} \lambda_g q_{x_i,g}^g \right). \quad (11)
$$
where \( \Theta = (\lambda_g, Q^g)_{1 \leq g \leq p} \) satisfies all the constraints of equations (3)–(5) or (8). Hence, the maximum likelihood estimation of the parameters \( \hat{\Theta} = (\hat{\lambda}_g, \hat{Q}^g)_{1 \leq g \leq p} \) is the solution of the following constrained non-linear optimisation problem:

\[
\left( \hat{\lambda}_g, \hat{Q}^g \right)_{1 \leq g \leq p} = \arg \max \sum_{t=0}^{n} \log \left( \sum_{g=1}^{p} \lambda_g q_{t-g}^{g, \pi_{t-g}} \right),
\]

s.t. \( \sum_{g=1}^{p} \lambda_g = 1, \lambda_g \geq 0, \forall g \in \{1, \ldots, p\} \)

\( q_{i,j}^g \geq 0 \) and \( \sum_{j=1}^{m} q_{i,j}^g = 1 \)

\( \forall g \in \{1, \ldots, p\}, \forall i, j \in X. \quad (12) \)

Clearly, the solution of the previous optimisation problem is very hard due to the high number of constraints. Berchtold (2001) proposes an efficient iterative process with the boundary adjustment in the MLE process which leads to a modification of the Newton’s method. Under the constraints of equations (3) and (4), Lébre and Bourguignon (2008) introduce a hidden process for the coefficients of the MTDg and propose an Expectation–Maximisation approach for the parameters estimation. Chen and Lio (2009) note that all the previous constraints can be rewritten in a box-constrained form, which is easier to handle.

### 3. MTD for order flow and price impact

We consider the joint dynamics of order flow and price changes in transaction time \( t \in \mathbb{N} \). Each event is a transaction which has a positive sign \((\epsilon_t = +1)\) if it is buyer initiated or negative \((\epsilon_t = -1)\) if is seller initiated. For the price we distinguish two possibilities, namely that the trade changes the price \((\pi_t = C)\) or not \((\pi_t = NC)\). Notice that we are not considering the amplitude if the immediate price changes. For large tick stocks this is a minor problem, since price changes almost always of ±1 tick, while for small tick stocks this is not true and we lose the information on the size of price change. In this paper, we use the MTDg model to describe the sequence of signed events

\[
(\{\epsilon_t, \pi_t\})_{t \in \mathbb{N}} \rightarrow \{X_t\}_{t \in \mathbb{N}}, \quad (15)
\]

hence the number of states of the model is \( m = 4 \). The relation between the states of the model and the signed events is obtained with the arbitrary mapping

\[
\epsilon_t = -1, \pi_t = C \rightarrow X_t = 1,
\]

\[
\epsilon_t = -1, \pi_t = NC \rightarrow X_t = 2,
\]

\[
\epsilon_t = +1, \pi_t = NC \rightarrow X_t = 3,
\]

\[
\epsilon_t = +1, \pi_t = C \rightarrow X_t = 4. \quad (16)
\]

The main quantity of interest is the cross and autocorrelation functions \( C_{\pi_t, \pi_{t+\tau}}(\ell) \), already introduced in Eisler et al. (2012a, 2012b, Taranto et al. 2016). Since

\[
\hat{\eta} = P(X_t) \equiv P(\epsilon_t, \pi_t),
\]

\[
B(\ell) = P(X_t; X_{t+\ell}) \equiv P(\epsilon_t, \pi_t; \epsilon_{t+\ell}, \pi_{t+\ell}) \quad (17)
\]

these correlations

\[
C_{\pi_t, \pi_{t+\tau}}(\ell) = \frac{\mathbb{E}[\epsilon_t I(\pi_t = \pi_1) \cdot \epsilon_{t+\ell} I(\pi_{t+\ell} = \pi_2)]}{\mathbb{P}(\pi_1) \mathbb{P}(\pi_2)}
\]

\[
= \sum_{\ell_t \in \ell} \sum_{\pi_t, \pi_{t+\ell}} \epsilon_t I(\pi_t = \pi_1) \epsilon_{t+\ell} I(\pi_{t+\ell} = \pi_2) \mathbb{P}(\epsilon_t, \pi_t; \epsilon_{t+\ell}, \pi_{t+\ell}) \mathbb{P}(\pi_t) \mathbb{P}(\pi_{t+\ell})
\]

\( \mathbb{P}(NC) = \hat{\eta}_2 + \hat{\eta}_3 \),

\( C_{NC,NC}(\ell) = \frac{b_{2,3}(\ell) - b_{2,3}(\ell) + b_{1,3}(\ell)}{(\hat{\eta}_2 + \hat{\eta}_3)^2}. \quad (19) \)

In the next two subsections, we will estimate MTDg models on real financial data of the US markets. We will consider two different parametrisations and estimation methods. The first one, used as a benchmark case, is based on MLE and uses a parametrisation which preserves the probabilistic interpretation of the mixture, i.e. it assumes that the parameters \((\lambda_g, Q^g)_{1 \leq g \leq p}\) are between zero and one. Moreover, in order to be able to apply MLE, we will impose a very strong structure of \((\lambda_g, Q^g)_{1 \leq g \leq p}\), reducing the number of parameters from \(p(m^2 - m + p - 1) \sim 1, 300 \) for \( p = 100 \) to 11.

In the second case, we relax the constraint that \((\lambda_g, Q^g)_{1 \leq g \leq p}\) are between zero and one and we use GMM. We
show that a suitable parametrisation allows to reduce the estimation to the solution of a constrained linear system, which we prove to be a convex problem. This model is weakly constrained and we are able to estimate reliably 500 parameters, improving significantly the performance of the model with respect to the benchmark case.

3.1. Strongly constrained MTDg model

Estimation methods for the MTDg model proposed so far in the literature have dealt with low-order models. Unfortunately, our case requires the estimation of a high-order version of the model to capture the long-ranged dependence measured for the flow of trade signs. The log-likelihood function of equation (11) is highly non-linear and the solution of the optimisation problem could be very hard to find for large values of \( p \).

3.1.1. Parametrisation. In order to reduce the number of parameters and to avoid non-linear constraints, we impose a functional form for the parameters which automatically satisfies all the constraints. For the \( \lambda_g \), it is very natural to assume a power law scaling, \( \lambda_g = N_B g^{-\beta} \), where \( N_B^{-1} = \sum_{i=1}^{p} g^{-\beta} \). The reason behind this choice is that the values of \( \lambda_g \) influence the correlations for large lags \( \ell \), which empirically decay slowly with the lag. Another significant simplification of the problem can be achieved by assuming a buy/sell symmetry, which leads to the definition of centro-symmetric matrices \( Q_g \). This assumption leads to

\[
q_{ij}^g = q_{m-i+1,m-j+1}^g, \quad \text{for } i, j = 1, \ldots, m, \tag{20}
\]

and for the first-order stationary distribution of the process

\[
\hat{\eta}_i = \hat{\eta}_{m-i+1}, \quad \text{for } i = 1, \ldots, m. \tag{21}
\]

For instance, \( q_{12}^g = q_{34}^g \) since the influence of a sell order price-changing event at time \( t-g \) on the probability of a sell order non-price-changing event at time \( t \) is equal to the influence of a buy order price-changing event at time \( t-g \) on the probability of a buy order non-price-changing event at time \( t \).

As mentioned above (see Theorem 2.1), we consider matrices \( \hat{Q}^g \) sharing the same left eigenvector with eigenvalue one. Writing \( \hat{Q}^g = \hat{Q} + \tilde{Q}^g \), we make the following strongly parametrised ansatz:

\[
Q = \begin{pmatrix} B_1 A_1 A_1 B_1 \\ B_2 A_2 B_2 A_2 B_2 \\ B_1 A_1 B_1 A_1 B_1 \end{pmatrix},
\]

\[
\tilde{Q}^g = \begin{pmatrix} -\mu_1 e^{-\alpha_1 g} & -\nu_1 e^{-\alpha_2 g} & -\nu_1 e^{-\alpha_2 g} & \nu_1 e^{-\alpha_1 g} & \mu_1 e^{-\alpha_1 g} \\ -\mu_2 e^{-\alpha_2 g} & \nu_2 e^{-\alpha_2 g} & -\nu_2 e^{-\alpha_2 g} & -\nu_2 e^{-\alpha_2 g} & -\mu_2 e^{-\alpha_2 g} \\ \mu_1 e^{-\alpha_1 g} & \nu_1 e^{-\alpha_2 g} & -\nu_1 e^{-\alpha_2 g} & -\nu_1 e^{-\alpha_2 g} & -\mu_1 e^{-\alpha_1 g} \end{pmatrix}, \tag{22}
\]

where \( \alpha_{ij} \geq 0 \). Imposing

\[
A_1 = 1/2 - B_1, \quad A_2 = 1/2 - B_2, \quad 0 \leq B_1 \leq 1/2, \quad 0 \leq B_2 \leq 1/2, \quad -B_1 \leq \mu_1 \leq B_1, \quad -B_2 \leq \mu_2 \leq B_2, \quad B_1 - 1/2 \leq \nu_1 \leq 1/2 - B_1, \quad B_2 - 1/2 \leq \nu_2 \leq 1/2 - B_2. \tag{23}
\]

we automatically satisfy all the constraints of the model. Moreover, it is immediate to see that \( \hat{\eta} \hat{Q}^g = 0 \), as required, where

\[
\hat{\eta} = \begin{pmatrix} B_2 & 1-2B_1 & 1-2B_1 & 1-2B_1 & 1-2B_1 \\ 1-2B_1 & 2-4B_1+4B_1^2 & 2-4B_1+4B_1^2 & 2-4B_1+4B_1^2 & 2-4B_1+4B_1^2 \\ 1-2B_1 & 1-2B_1 & 2-4B_1+4B_1^2 & 2-4B_1+4B_1^2 & 2-4B_1+4B_1^2 \\ 1-2B_1 & 1-2B_1 & 1-2B_1 & 2-4B_1+4B_1^2 & 2-4B_1+4B_1^2 \\ 1-2B_1 & 1-2B_1 & 1-2B_1 & 1-2B_1 & 2-4B_1+4B_1^2 \end{pmatrix}, \tag{24}
\]

Finally, the parametrisation in equation (22) with the linear constraints of equation (23) guarantees that the matrices have the right normalisation on the rows, \( \sum_i q_{ij}^g = 1, \forall g, i \) and \( 0 < q_{ij}^g < 1, \forall g, i, j \).

The specification of the functional forms of \( \lambda_g \) and of the entries of \( \hat{Q}^g \) is motivated by the following reasonings. The parameters \( q_{ij}^g \) determine the correlations between the event \( i \) at time \( t \) and the event \( j \) at time \( t-g \). From the left panel in figure 4 in Taranto et al. (2016), reporting the empirical correlations measured for the large tick stock Microsoft, we see a quite different behaviour depending on the conditioning event. For instance, the order flow correlations among non-price-changing events is extremely persistent. It is therefore quite natural to model the decay of the pre-factor \( \lambda_g \) in terms of a hyperbolic function. On the contrary, to reproduce the faster decay of the empirical correlations which involve price-changing events, and since \( \lambda_g \) multiplies all entries of the matrices \( Q^g \), we include in the definition of the \( \hat{Q}^g \) exponential damping factors with rates \( \alpha_{ij} \) (i.e. \( j = 1, 2 \)).

The parameters of this model can be obtained via MLE. The optimisation problem is non-trivial since the likelihood problem are linear inequalities. The total number of parameters is 11, \( \theta = \{ \beta, B_i, \mu_i, \nu_i, \alpha_{ij} \} \) with \( i, j = 1, 2 \).

3.1.2. Results. In figures 1 and 2, we plot the correlation functions computed from a Monte Carlo simulation of the MTDg(100) model with parameter values obtained from MLE on Microsoft (MSFT) and Apple (AAPL) data. In table 1, we report the summary statistics for both assets during the period February 2013–April 2013 with a total of 63 trading days. The average tick size price ratios motivate why we dubb AAPL a small tick asset and MSFT a large tick asset. We compare the auto and cross-correlations \( C_{\pi_1,\pi_2}(\ell) \) for price-changing and non-price-changing events with the empirical ones. As can be noted, for the small tick stock the model can reproduce the structure of the correlations for short-time scales, but not their persistence. For the large tick stock, the persistence of the empirical correlations is not well reproduced either. The quality of the fit varies across the different correlations. The behaviour of the C,C and NC,NC correlations is recovered quite well both at short- and long-time scales (for the yellow curve, it is important to stress that the scale of the plot is logarithmic on both axes and an apparently large deviation corresponds to a small difference in the linear scale). The MTDg model describes quite well the NC,NC correlation for short lags, but the quality of the fit worsens for larger lag values. The lack in the persistence of the simulated correlations can be explained by the fact that the estimated exponent \( \beta \) is too high. Finally, the behaviour of the C,NC curve is recovered only for the very
Table 1. Summary statistics for AAPL and MSFT stocks: average traded daily volume, volatility, average spread, average tick size price ratio and trade frequency.

|          | Average traded volume (M$) | Volatility (bps) | Average spread (tick) | Average tick size price ratio | Trade frequency (trade/seconds) |
|----------|-----------------------------|------------------|-----------------------|-------------------------------|---------------------------------|
| AAPL     | 1695.13                     | 1.05             | 9.14                  | 0.22                          | 1.14                            |
| MSFT     | 451.80                      | 1.21             | 1.00                  | 3.56                          | 4.46                            |

first lags, then both the amplitude and the sign reproduced by the MTDg do not match the empirical correlations.

The robustness of the numerical results in this and in the next sections has been tested with extensive Monte Carlo simulations. The weakly and strongly constrained models have been estimated on different subsamples of the data, then simulated and re-estimated on the synthetic time series. All the experiments, whose detailed results are available under request, have shown that both parametrisations—weak and strong—and estimation procedures—MLE and GMM—are robust to the choice of the functional forms of \( \lambda_g \) and \( Q^g \). Moreover, only a minor dependence of the parameter values on the estimation periods can be reported.

We conclude that, for both small and large tick stocks, the restrictions imposed on the matrices \( Q^g \) are too binding to reproduce the different decays of the empirical correlations. Nonetheless, it is worth to comment that these modelling assumptions guarantee a fast estimation procedure, even for very high-order Markov models. This fact can motivate the use of the strongly constrained MTDg model in spite of its mild performances.

3.2. Weakly constrained MTDg model

3.2.1. Model definition. Here we introduce the main methodological innovation of this paper, namely a parametrisation of the MTDg model which can be estimated with GMM even when the number of parameters is very large. To motivate it, let us consider the DAR(p) process with

\[
\begin{align*}
X_t &= \lambda_g(\mathbf{Q}_g)^T \tilde{\eta}_g + \epsilon_t, \\
\mathbf{Q}_g &= \begin{pmatrix}
\lambda_g \\
\mathbf{A}_g
\end{pmatrix}, \\
\tilde{\eta}_g &= \begin{pmatrix}
\eta_g \\
\mathbf{B}_g
\end{pmatrix}, \\
\mathbf{A}_g &= \begin{pmatrix}
\mathbf{A}_{g,1} & \cdots & \mathbf{A}_{g,p}
\end{pmatrix}, \\
\mathbf{B}_g &= \begin{pmatrix}
\mathbf{B}_{g,1} & \cdots & \mathbf{B}_{g,p}
\end{pmatrix},
\end{align*}
\]

\( \lambda_g \) is a vector of \( p \) positive parameters, \( \mathbf{A}_g \) and \( \mathbf{B}_g \) are matrices of \( m \times m \) and \( m \times p \) respectively, and \( \eta_g \) is a \( m \times 1 \) vector. The process can be modelled as

\[
\begin{align*}
\mathbf{Q}_g &= \begin{pmatrix}
\tilde{\eta}_g & \mathbf{A}_g \\
\mathbf{B}_g & \mathbf{I}
\end{pmatrix}, \\
\tilde{\eta}_g &= \begin{pmatrix}
\tilde{\eta}_g_1, \ldots, \tilde{\eta}_g_m
\end{pmatrix},
\end{align*}
\]

with \( \tilde{\eta}_g_1, \ldots, \tilde{\eta}_g_m \) being the \( m \) left eigenvectors of \( \mathbf{Q}_g \) corresponding to the \( m \) eigenvalues \( \lambda_g \). The restriction imposed on the matrices \( \mathbf{A}_g \) and \( \mathbf{B}_g \) can be relaxed by imposing that \( \mathbf{Q}_g \) is a \( (m + p) \times (m + p) \) matrix with

\[
\begin{align*}
\mathbf{Q}_g &= \begin{pmatrix}
\tilde{\eta}_g & \mathbf{A}_g \\
\mathbf{B}_g & \mathbf{I}
\end{pmatrix}, \\
\tilde{\eta}_g &= \begin{pmatrix}
\tilde{\eta}_g_1, \ldots, \tilde{\eta}_g_m
\end{pmatrix},
\end{align*}
\]

and \( \tilde{\eta}_g \tilde{\mathbf{Q}}^g = 0 \). Moreover, normalisation of \( \mathbf{Q}^g \) imposes that each row of \( \tilde{\mathbf{Q}}^g \) sums to zero, hence these matrices will have negative elements. As in Theorem 2.1, all the \( \mathbf{Q}^g \) share the same left eigenvector \( \tilde{\eta}_g \) with eigenvalue 1. It is easy to show that the conditional probabilities of this model can be written as

\[
\mathbb{P}(X_t = i_X | X_{t-1} = i_{X-1}, \ldots, X_{t-p} = i_p) = \tilde{\eta}_g + \sum_{g=1}^{p} a_{i,g,i}^g,
\]

(28)

where \( a_{i,g,i}^g \) is the \( g \)th element of \( \mathbf{A}_g^g \). Thus the matrices \( \mathbf{A}_g^g = \lambda_g \mathbf{Q}_g^g \) describe the deviations of the \( p \thinspace -order transition probability from the stationary value given by \( \tilde{\eta}_g \). Finally, as shown in appendix 3, the system of equations of Theorem 2.1 for this model is

\[
\mathbf{B}(k) - \tilde{\eta}_g^T \tilde{\eta}_g = \sum_{g=1}^{p} \mathbf{B}(k - g) \mathbf{A}_g^g.
\]

(29)

This linear system can be used to estimate the model, i.e. the matrices \( \mathbf{A}_g^g \), from the knowledge of the stationary probabilities \( \tilde{\eta}_g \) and the bivariate distributions \( \mathbf{B}(k) \). There are, however, two technical problems:

- The estimated model might not have a probabilistic interpretation, i.e. the estimated model might generate transition probabilities larger than one or smaller than zero.
- The solution of equation (29) gives the matrix \( \mathbf{A}_g^g \), while one might need \( \lambda_g \) and \( \mathbf{Q}_g^g \) separately. Note however that the dynamics of the model is independent from this decomposition.

To fix the first problem, and to be able to use Theorem 2.1, it must also hold that

\[
0 < \tilde{\eta}_g + \sum_{g=1}^{p} a_{i,g,i}^g < 1, \quad \forall (i_1, \ldots, i_p) \in \mathcal{X}^{p+1},
\]

(30)

which correspond to \( 2mp^{p+1} \) constraints. Following Proposition 1 in Raftery and Tavaré (1994), they are equivalent to the conditions

\[
\begin{align*}
\tilde{\eta}_g + \sum_{g=1}^{p} \max_{i_g} \left( a_{i,g,i}^g \right) < 1, \quad \forall i \in \mathcal{X},
\end{align*}
\]

(31)

\[
\begin{align*}
\tilde{\eta}_g + \sum_{g=1}^{p} \min_{i_g} \left( a_{i,g,i}^g \right) > 0, \quad \forall i \in \mathcal{X},
\end{align*}
\]

(32)

which reduce to \( 2m \) inequality constraints. Under these conditions the process is well defined and possesses a unique stationary solution.
The estimation of the model can be performed solving the optimisation programme

$$\hat{q} = \arg\min_{q \in \mathbb{R}^{p(m^2 - 2m + 1)}} \left\| d - K \cdot q \right\|^2$$

s.t. \( \hat{\eta} + \sum_{g=1}^{p} \max_{t_g} (a_{t_g,i}^g) \leq 1, \quad \forall i \in \mathcal{X} \)  \hspace{1cm} (33)

where the elements of the \( p(m^2 - 2m + 1) \)-dimensional vector \( d \) correspond to left-hand side of equation (29), namely

$$d = (b_{1,1}^{p}, \ldots, b_{1,m-1}^{p}, \ldots, b_{m-1,1}^{p}, \ldots, b_{m-1,m-1}^{p}, \ldots)$$

$$\bar{b}_{1,1}^{p}, \ldots, \bar{b}_{1,m-1}^{p}, \ldots, \bar{b}_{m-1,1}^{p}, \ldots, \bar{b}_{m-1,m-1}^{p} \quad (34)$$
with
\[
\mathbf{b}^g_{i,j} = b^g_{k,j} - \hat{\eta}_i \hat{\eta}_j,
\]
the vector \( \mathbf{q} \) corresponds to the parameters of the model \( \lambda_{qg} \).
\[
\mathbf{q} = (a^1_1, \ldots, a^1_{m-1}, \ldots, a^1_{m-n}, \ldots, a^1_m, \ldots, a^p_{m-1}, \ldots, a^p_{m-n}, \ldots, a^p_m, \ldots, a^p_{m-1}, \ldots, a^p_{m-n}, \ldots)
\]
and the elements of the matrix \( \mathbf{K} \) are linear combinations of \( b^g_{i,j} \), according to equation (29) (we do not report the matrix since its form is not transparent).

The reason for the choice of the constraints in equation (33) is that we prove in appendix 4 the following proposition:

**PROPOSITION 3.1** If \( \mathbf{K} \) is not singular, the optimisation programme of equation (33) is strictly convex in \( \mathbb{R}^{(m^2-2m+1)} \).

Therefore if a local minimum exists, then it is a global minimum. The convexity property solves the issue of the high dimensionality of the problem and the model can be estimated also for large order \( p \).

### 3.2.2. Application to order flow and impact.

We now consider the application of the above-described \( \text{MTDg} \) model to the \( m = 4 \) process describing jointly the order flow and the price changes. As done in the previous section, we reduce the dimensionality of the system by exploiting the buy/sell symmetry, which leads to centrosymmetric \( \hat{\eta} \) and \( \mathbf{B}(k) \). In fact, for \( m = 4 \) we have that
\[
b^k_{i,j} = b^k_{m-i+1,m-j+1},
\]
and for the stationary distribution
\[
\hat{\eta}_i = \hat{\eta}_{m-i+1}, \quad \text{for } i = 1, \ldots, m.
\]

The buy/sell symmetry and the normalisation of matrices \( \mathbf{B}(k) \) reduces the number of independent variables in \( \mathbf{B}(k) \) to 5\( p \), 5 for each lag \( k \). Thus, we have that
\[
\mathbf{B}(k) = \begin{pmatrix}
  b^k_{1,1} & b^k_{1,2} & b^k_{1,3} \\
  b^k_{2,1} & b^k_{2,2} & b^k_{2,3} \\
  \hat{\eta}_2 - b^k_{2,1} - b^k_{2,2} & \hat{\eta}_1 - b^k_{2,1} - b^k_{2,2} & \hat{\eta}_2 - b^k_{2,1} - b^k_{2,2} - b^k_{3,2} \\
  \hat{\eta}_1 - b^k_{1,2} + b^k_{2,2} - b^k_{3,2} & \hat{\eta}_2 - b^k_{1,2} + b^k_{2,2} - b^k_{3,2} & \hat{\eta}_1 - b^k_{1,2} + b^k_{2,2} - b^k_{3,2} - b^k_{3,2}
\end{pmatrix}
\]

In order to find a solution of the problem of equation (33), we assume that the imposed centrosymmetry of \( \mathbf{B}(k) \) and \( \hat{\eta} \) does not change the rank of the matrix \( \mathbf{K} \). In this case, the solution is unique and it is easy to show that also \( \mathbf{Q}^g \) must be centrosymmetric, as
\[
\mathbf{Q}^g = \begin{pmatrix}
  \tilde{q}^g_{1,1} & \tilde{q}^g_{1,2} & \tilde{q}^g_{1,3} \\
  -\tilde{q}^g_{2,1} - \tilde{q}^g_{2,2} - \tilde{q}^g_{2,3} & -\tilde{q}^g_{1,2} - c_2(\tilde{q}^g_{2,2} + \tilde{q}^g_{2,3}) & -\tilde{q}^g_{1,3} - c_2(\tilde{q}^g_{2,2} + \tilde{q}^g_{2,3}) \\
  \tilde{q}^g_{2,1} & \tilde{q}^g_{2,2} & \tilde{q}^g_{2,3}
\end{pmatrix}
\]

where \( c_2 = \hat{\eta}_2/\hat{\eta}_1 \). With this definition the number of independent parameters in \( \mathbf{Q}^g \) is also equal to 5 for each \( g \).

We can now solve the system of equation (33) whose unknowns are the components of the matrix \( \mathbf{A}^g \). This way we obtain the value of the products \( \lambda_{qg} \), but not the value of the components \( \lambda_{qg} \) and \( \tilde{q}^g_{i,j} \) separately. For this reason, we impose that one of the five components among \( \tilde{q}^g_{1,1}, \tilde{q}^g_{1,2}, \tilde{q}^g_{2,1}, \tilde{q}^g_{2,2}, \) and \( \tilde{q}^g_{3,3} \) is independent of the lag \( g \). We arbitrarily fix \( \tilde{q}^g_{2,1} \equiv \tilde{q}^g_{1,1} \).

We are left with 4\( p \) free parameters from \( \mathbf{Q}^g \) (4 for each \( g \)), \( p - 1 \) parameters from \( \lambda_\gamma \) and \( \tilde{q}^g_{1,1} \). In total we have 5\( p \) free parameters, which is exactly the same number of independent components \( b^g_{i,j} \). The values of the products \( \lambda_{qg} \) and \( \tilde{q}^g_{i,j} \) define the MTDg model. Different choices of \( \tilde{q}^g_{i,j} = \tilde{q}^g_{j,i} \) give different factorisations, but lead to the same high-order Markov chain. The arbitrariness of the choice is an evidence of the well-known identifiability problem of all mixture models.

In the literature, there exist many algorithms which solve iteratively the constrained optimisation problem of equation (33). A widely used class belongs to the Sequential Quadratic Programming (SQP) family (Boggs and Tolle 1995). However, an issue of our optimisation is that constraints are non-smooth functions, which is a necessary condition required by the usual SQP algorithms. In a recent paper, Curtis and Overton (2012) have proposed the Sequential Quadratic Programming Gradient Sampling algorithm (SQP-GS), which can be applied to non-smooth, non-linear objective and constraint functions. We have implemented this algorithm in order to solve our optimisation problem.

### 3.2.3. Results.

We estimated the above MTDg(100) model on MSFT and AAPL. Before showing the results, it is worth to stress again an important aspect of our approach. Preliminary, we have estimated the model using equation (29) without the additional constraints of equations (31) and (32). We have found negative transition probabilities. Thus, the data reject a probabilistic description based on the unconstrained MTD model. To obtain a meaningful, even though approximated, description of the data, the MTD model parameters have to be restricted according to equations (31) and (32). Since the constraints are binding, the resulting Markov model of order
\[
\hat{\eta}_1 - b^k_{1,2} + b^k_{2,2} - b^k_{3,2} \\
\hat{\eta}_2 - b^k_{1,2} + b^k_{2,2} - b^k_{3,2}
\]

\( p \) is not ergodic, i.e. some of the estimated parameters lie on the boundary of the parameter domain. To apply the results of Theorem 2.1, which ensures the existence and uniqueness of the stationary distribution, we have restricted the model within
\[
\tilde{q}^g_{1,2} - c_2(\tilde{q}^g_{2,2} + \tilde{q}^g_{2,3}), \tilde{q}^g_{1,2} + c_2(\tilde{q}^g_{2,2} + \tilde{q}^g_{2,3}) \\
\tilde{q}^g_{2,1} - \tilde{q}^g_{2,2} - \tilde{q}^g_{2,3}, \tilde{q}^g_{2,1} + \tilde{q}^g_{2,2} + \tilde{q}^g_{2,3}
\]

the class of ergodic Markov models of order \( p \). Consistently, we have replaced the inequalities (31) and (32) with
\[
\hat{\eta}_i + \sum_{g=1}^{p} \max_{j \neq i} (a^g_{i,j}) \leq 1 - \epsilon,
\]
Another way to assess the quality of the MTDg model is to minimize the distance to all correlation function is, as presented above, only approximately well defined in $[0, 1]$. Clearly, the estimation shows that the probabilistic mixture discussed at the beginning is, perhaps meaningfully, not suitable for the present data.

Figures 5 and 6 show correlation functions $C_{\tau_1,\tau_2}(\ell)$ of signed events computed from a Monte Carlo simulation of the estimated model and compared with real data. As can be noted, for small tick stocks we have significantly improved the results of figure 2. Compared with the benchmark, the new estimation method reproduces the high persistence of the correlations of large tick stocks, whose correlations present an order signs independently from the conditioning events. In the case of the large tick stocks, whose correlations present an highly non-trivial structure, the GMM methodology greatly improves the results with respect to figure 1. In particular, the high persistence of non-price-changing events is very well reproduced. Moreover, the $C_{NC,C}(\ell)$ curve decays faster as compared to the previous method, and is thus closer to data.

### 3.3. Large tick stock signature plot

Another way to assess the quality of the MTDg model is to analyse how well it describes the volatility of prices. As noted above and in Taranto et al. (2016), the impact of a price-changing event is nearly price independent for large tick stocks (within the TIM2 model). This means that the signature plot is simply given by:

$$D_{\text{TIM2}}(\ell) \approx D_{LF} + G_C(1)^2 \mathbb{P}(C) + 2 G_C(1)^2 \mathbb{P}(C)^2 \sum_{0 < a < \ell} C_{C,C}(n),$$

which is completely determined by the correlation function $C_{C,C}(\ell)$ (once the value of $G_C(1)$ has been estimated). This correlation function is, as presented above, only approximately reproduced by the MTDg model, although it is estimated to minimise the distance to all $C_{\tau_1,\tau_2}(\ell)$. In the context of financial applications, it is therefore interesting to replot the difference between the MTDg $C_{C,C}(\ell)$ and empirical data in terms of the signature plot $D_{\text{TIM2}}(\ell)$, which involves the integral of the correlation function.

In figure 7 we show the curves corresponding to equation (41) for the strongly and weakly constrained versions of the MTDg model proposed above, where the extra fitting parameter $D_{LF}$ is optimised with OLS in order to minimise the distance between the empirical and the theoretical curves of the model. We see that in terms of the signature plot of the model, the weakly constrained and strongly constrained MTDg fare nearly equally well. We also show the predictions of the TIM2 model that uses the empirical $C_{C,C}(\ell)$; the nearly perfect fit in this case is a consequence of the fact that $G_C(\ell) \approx G_C(1)$ for large tick stocks.

Note that the TIM2 price process is strictly diffusive only if the quantity $D_{\text{TIM2}}(\ell + 1)(\ell + 1) - D_{\text{TIM2}}(\ell) \ell$ is a constant independent from $\ell$. In fact, we have that

$$D_{\text{TIM2}}(\ell + 1)(\ell + 1) - D_{\text{TIM2}}(\ell) \ell = D_{LF} + G_C(1)^2 \mathbb{P}(C) + 2 G_C(1)^2 \mathbb{P}(C)^2 \sum_{0 < a \leq \ell} C_{C,C}(n),$$

which means that the price process becomes diffusive for $\ell > \ell^*$ only if $C_{C,C}(\ell) > C_{C,C}(\ell^*) = 0$. Figures 1 and 7 suggest that this is indeed the case for $\ell^* \approx 10$.

### 4. Out-of-sample analysis

In the previous sections, we have presented two MTDg models—strongly and weakly constrained—and discussed two alternative estimation methodologies based on MLE and GMM. Since they differ both in the number of parameters and in estimation efficiency, it is important to compare their performances testing the predictive power of the models in an out-of-sample analysis. We consider as a measure of the performance the expected prediction error (EPE) defined as

$$EPE(\theta) = \mathbb{E}[L(X_t, \hat{X}_t^\theta)],$$

where $X_t$ is the observed process, $\hat{X}_t^\theta$ is the predictor of $X_t$ based on the model with parameter set $\theta$, and the $p$ past observations of the process $X_t$.

As common in the literature for categorical data, we use as loss function the log-likelihood $L(X_t, \hat{X}_t^\theta) = -2 \sum_{i=1}^m I(X_t = i) = -2 \log(\hat{x}_t) X_t$, also called cross-entropy. We remind that $\hat{x}_t$ is the $m$-probability vector describing the prediction of the model and in the previous formula we take the $X_t$th component. For the MTDg(p) model, this probability vector is

$$\hat{x}_t = \sum_{g=1}^p x_{t-g}^{\lambda_g} Q^g$$

where, as before, $x_{t-g}$ is a $m$-vector of zeros with the exception of the realised component $x_{t-g}$. This quantity can be easily computed once the model is estimated, since it depends on the transition probabilities. EPE values are in the range $[0, +\infty)$, and it is zero if all probabilities $(\hat{x}_t) X_t$ of the sample are equal to one (perfect prediction), and it is infinity if all probabilities $(\hat{x}_t) X_t$ of the sample are zero (prediction of impossible events).

We evaluate the best-performing model as the model with the lowest EPE and benchmark the MTDg with a model with the unconditional probabilities as predictors of future signed events. Table 2 reports all the EPE values for different models of the predictor estimated on MSFT, Bank of America-Citigroup (BAC), General Electric (GE), Cisco (CSCO), AAPL and Amazon (AMZN) data. The scheme of the out-of-sample analysis is the following: the model is trained on a time
period of 10 days, then we compute the loss functions in the following trading day using the parameter set provided by MLE (strongly constrained) or by GMM (weakly constrained). We repeat the procedure by shifting the estimation by one trading day ahead. Finally, we compute the global loss by averaging all measured loss functions. The financial interpretation of the EPE values is clear in the case of the large tick stocks, because a price-changing event moves the price by one tick with probability almost one and thus there exists a direct relation between the states of the MTDg model and the price return. Hence, for large tick stocks the EPE value can be employed as a proxy of the predictability of returns at high-frequency timescale.

From table 2, we see that both MTDg models outperform the benchmark. More importantly, there is a clear evidence that the weakly constrained model with the highest number of parameters (Model C) outperforms the strongly constrained MTDg, for all considered stocks. These results exclude the overfitting hypothesis, and support the claim that weakly constrained MTDg models are good candidates to capture the high-frequency dynamics of signed events.
II. The Mixture Transition Distribution model

Figure 5. *GMM calibration of the weakly constrained MTDg*. Comparison between the auto and cross-correlation functions $C_{\pi_1,\pi_2}(\ell)$ of signed events from a simulation of the MTDg(100) model estimated on MSFT data (triangles) and the empirical curves (solid lines). The error bars correspond to one standard deviation. The scale for values close to zero and bounded by horizontal solid lines is linear, whereas outside this region the scale is logarithmic.

Figure 6. *GMM calibration of the weakly constrained MTDg*. Comparison between the auto and cross-correlation functions $C_{\pi_1,\pi_2}(\ell)$ of signed events from a simulation of the MTDg(100) model estimated on AAPL data (triangles) and the empirical curves (solid lines). The error bars correspond to one standard deviation.

5. Discussion and conclusion

The companion paper (Taranto et al. 2016) has established that treating all market orders on the same basis produces erroneous predictions both for the ‘response functions’ (average lagged impact) at negative lags and the signature plot. Single-propagator models and history-dependent impact models are not designed to capture the feedback effects between past price returns and future order flow. These serious discrepancies have been significantly reduced by introducing the extended versions of the linear impact models (TIM and HDIM) which consider a richer set of signed events (see Eisler et al. 2012a, 2012b). The argument which has motivated our generalisation of the impact models is the observation that price-changing and non price-changing events have to be treated differently. This is particularly evident for large tick stocks, where price-moving events are extremely rare but very informative. This apparently minor modification has lead to an extended class of propagator models which describe with remarkable realism the intertwined high-frequency dynamics of prices and order flow. Nonetheless, the linear description of the market dynamics achieved in Part I (Taranto et al. 2016)
Figure 7. Signature plot for MSFT data: Empirical data (crosses), weakly constrained (GMM) MTDg(100) model with $D_{LF} = 0.41$ (dashed line), strongly constrained (MLE) MTDg(100) model with $D_{LF} = 0.43$ (dashed-dotted line) and the theoretical prediction of the estimated TIM2 model Taranto et al. (2016).

Table 2. EPE values and standard errors (SE) for MSFT, BAC, GE, CSCO, AAPL and AMZN data. Model A: Unconditional probabilities as predictor. Model B: Strongly constrained MTDg(100) estimated via MLE according to equation (22). Total number of parameters: 11. Model C: Weakly constrained MTDg(100) model estimated via GMM with matrices as in equation (40). Total number of parameters: 500.

|       | Model A | Model B | Model C |
|-------|---------|---------|---------|
| MSFT  | 1.928   | 1.199   | 1.181   |
| SE    | 0.003   | 0.004   | 0.004   |
| BAC   | 1.744   | 0.799   | 0.785   |
| SE    | 0.003   | 0.004   | 0.004   |
| GE    | 1.922   | 1.169   | 1.153   |
| SE    | 0.004   | 0.005   | 0.005   |
| CSCO  | 1.919   | 1.112   | 1.098   |
| SE    | 0.004   | 0.005   | 0.005   |
| AAPL  | 2.643   | 2.211   | 2.192   |
| SE    | 0.001   | 0.002   | 0.002   |
| AMZN  | 2.579   | 2.196   | 2.183   |
| SE    | 0.002   | 0.004   | 0.004   |

is still too rigid: these models are designed to describe the evolution of the market with an exogenously specified order flow. This fact seriously limits the forecasting capabilities of linear impact models.

The Mixture Transition Distribution model partly solves the above issue by introducing an explicit stochastic model for the order flow, treated as an endogenous component of the dynamics. It is specially designed for variables which are inherently discrete—a feature of great relevance for price returns of large tick stocks. In this paper, we have presented a class of so-called MTDg models as a natural extension of the Discrete Autoregressive DAR(p) in a multi-event context. Our aim was to test how well an estimated MTDg model can account for the statistics of the order flow, i.e. the string of four events: buy/sell - price-changing/non changing events. One of the most interesting aspects of our work is methodological, and concerns the nature of the restrictions to impose on large models. We have restricted the MTD within the class of ergodic Markov models of order $p$. The existence and uniqueness of the stationary solutions for the ergodic model, and the buy/sell symmetry of the order book data have motivated the introduction in section 3.2 and appendix 3 of the class of weakly constrained MTDg models. They represent a rich family of discrete models, where the number of free parameters equates the number of independent observable correlation functions. This fact allows the introduction of a numerical procedure which solves the estimation of the model parameters in a remarkably robust way. This result is rooted on the proof that the optimisation problem is convex in the parameter space. From the financial viewpoint, we have shown that—perhaps surprisingly—a weakly constrained version of the MTDg models captures the dynamics of signed events with greater realism than alternative and more parsimonious versions. Despite the large number of parameters, the out-of-sample analysis confirms that such good performances are achieved without overfitting the data.
The improvement brought by the MTDg models and the new estimation methodology is remarkable, but still some discrepancies persist when comparing the model predictions with the empirical correlation functions. Several reasons may be responsible for these deviations. The first one was already pointed out by Raftery (1985) where he has shown that there exist regions of correlations which simply cannot be reproduced by MTDg models. A second reason is that, even if the MTDg model was the correct data generating process, the estimation methodology which involves information only coming from second-order conditions, may lack efficiency with respect to the MLE approach. Finally, the MTDg model represents a parsimonious approximation of a full Markov chain of order $p$. This parsimony may come at expense of the realism of the model.

From a microstructure point of view, we can hypothesise that the string of past signed events $X_{t-1}, \ldots, X_{t-p}$ is not informative enough to predict the value $X_t$. In particular, for large tick stocks price-changing events $\pi = C$ are much rarer than non-price-changing event $\pi = NC$. Therefore, a $\pi = C$ event is by construction difficult to predict with past information based only on realised signs and trades. Hence, the behaviour that we observe may be ascribed to a problem of missing explanatory variables. A natural candidate in this respect could be the volume of orders outstanding at the opposite side of the limit order book before the execution of a trade order, i.e. the local order book imbalance. A similar reasoning suggests that an important role—different from that of the trade orders—should be played by limit orders and cancellations. Their impact on quote revision is not taken under consideration explicitly in the present version of the paper. A richer description could be considered in a future implementation of the MTD model, which, however, is beyond the scope of the present investigation. In this respect, it will be also interesting to investigate whether it is convenient to move to a description of the limit order book and price formation process in the physical or event time—the clock moves every time something changes in the order book—instead of the present description based on the trade time. We leave this extension for future research, too.

From a more fundamental point of view, we should also point out that the MTDg estimated kernel, which gives the probability that an event at time $t = 0$ will trigger similar or opposite events at time $t = g$ later, must be interpreted with care. Indeed, this kernel receives contributions both from order splitting, which increases the probability that an agent places an order of the same sign in the future, and from genuine reactions of the rest of the market to this event (Tóth et al. 2012, Tóth et al. 2015). These reactions can be herding (copy cat trades) or, on the contrary, trades in the other direction (coming e.g. from liquidity providers). The response of the order flow to a single, isolated trade is thus expected to be rather different from the impulse function obtained by calibrating an MTDg model to the full order flow since order splitting contributions will be absent in the former, but contribute to the latter. The distinction between the two effects requires trade identification to be resolved. We hope to come back to this issue in a forthcoming work (Tóth et al. 2017).

Acknowledgements

We acknowledge several interesting comments from two anonymous referees. We want to thank Z. Eisler, J. Donier and I. Mastromatteo for very useful discussions. D. E. Taranto acknowledges CFM for supporting his extended visit at CFM where part of this research was done.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Giacomo Bormetti http://orcid.org/0000-0003-3542-2505

References

Bacry, E. and Muzy, J.F., Hawkes Model for price and trades high-frequency dynamics. Quant. Finance, 2014, 14(7), 1147–1166.
Berchtold, A., Autoregressive modeling of Markov chains. Statistical Modelling: Proceedings of the 10th International Workshop on Statistical Modelling, Springer-Verlag, Innsbruck, Austria, pp. 19–26, 1995.
Berchtold, A., Estimation in the mixture transition distribution model. J. Time Ser. Anal., 2001, 22(4), 379–397.
Berchtold, A. and Raftery, A.E., The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. Stat. Sci., 2002, 328–356.
Boggs, P.T. and Tolle, J.W., Sequential quadratic programming. Acta Numer., 1995, 4, 1–51.
Bouchaud, J.-P., Gefen, Y., Potters, M. and Wyatt, M., Fluctuations and response in financial markets: The subtle nature of ‘random’ price changes. Quant. Finance, 2004, 4(2), 176–190.
Chen, D.G. and Lio, Y.L., A novel estimation approach for mixture transition distribution model in high-order Markov chains. Commun. Stat. Simulat., 2009, 38(5), 990–1003.
Curtis, F.E. and Overton, M.L., A sequential quadratic programming algorithm for nonconvex, nonsmooth constrained optimization. Siam J. Optimiz., 2012, 22(2), 474–500.
Eisler, Z., Bouchaud, J.-P. and Kockelkoren, J., The price impact of order book events: market orders, limit orders and cancellations. Quant. Finance, 2012, 12(9), 1395–1419.
Eisler, Z., Bouchaud, J.-P. and Kockelkoren, J., Models for the impact of all order book events. In Market Microstructure: Confronting Many Viewpoints, edited by F. Abergel, J.P. Bouchaud, T. Foucault, C.-A. Lehalle, and M. Rosenbaum, 2012 (John Wiley & Sons Ltd.: Oxford).
Hasbrouck, J., Trades, quotes, inventory and information. J. Financ. Econ., 1988, 22, 229–252.
Hasbrouck, J., Measuring the information content of stock trades. J. Financ., 1991, 46, 179–207.
Jacobs, P.A. and Lewis, P.A., Discrete time series generated by mixtures: I: Correlational and runs properties. J. R. Stat. Soc. Ser. B, 1978, 94–105.
Lèbre, S. and Bourguignon, P.Y., An EM algorithm for estimation in the mixture transition distribution model. J. Stat. Comput. Simul., 2008, 78(6), 713–729.
Lillo, F. and Farmer, J.D., The long memory of the efficient market. Stud. Nonlinear Dynam. Econometrics, 2004, 8(3).
Pegram, G.G.S., An autoregressive model for multilag Markov chains. J. Appl. Probab., 1980, 17, 350–362.
Appendix 1. Existence and uniqueness of the stationary distribution of the MTDg model

Theorem A.1 Suppose that a sequence of random variables \( \{X_t\}_{t \in \mathbb{N}} \) taking values in the finite set \( \mathcal{X} = \{1, \ldots, m\} \) is defined by

\[
P(X_t = i | X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \sum_{g=1}^{p} \lambda_g q_{g,i}^{'},
\]

where \( Q = \left[ q_{i,j}^{'}\right]_{i,j \in \mathcal{X}} \) are matrices with normalised row sums, \( \sum_j q_{i,j}^{'}, = 1, \) and \( \sum_{g=1}^{p} \lambda_g = 1, \) and assume that \( \lambda_q Q = \lambda_q \), \( \forall q \). If the vector \( \lambda_q \) is such that \( \lambda_q > 0, \) \( \forall i \in \mathcal{X} \), then

\[
\lim_{t \to \infty} P(X_{t+\ell} = i | X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \lambda_q.
\]

Proof. Let \( T \) be the \( m^p \times m^p \) transition matrix for the Markov chain with the \( m^p \) possible values of \( (X_{t-1}, \ldots, X_{t-p}) \) as states. The elements of \( T \) are

\[
T_{i_1,\ldots,i_p} = \sum_{g=1}^{p} \lambda_g q_{g,i_1} \cdots q_{g,i_p} \delta_{j_1 \cdots j_p}.
\]

Each column of \( T \) represents the \( p \)-vector \( (i_1, \ldots, i_p) \) of arrival states, which are ordered in such a way that \( i \) varies most slowly, \( i_1 \), second most slowly, and so on. Similarly, the rows of \( T \) represents the values of \( (j_1, \ldots, j_p) \) with \( j_p \) varies most slowly, and so on. 

The assumption of equation (A2) guarantees that all states of \( T \) intercommunicate, so \( T \) is irreducible. Amongst the diagonal elements of \( T \) are aperiodic, then, since \( T \) is irreducible, all states are aperiodic. Hence, \( T \), being finite, specifies an ergodic Markov chain and has a unique equilibrium distribution \( \lambda_q \) satisfying \( T \lambda_q = \lambda_q \) with elements

\[
\lambda_q = \lim_{t \to \infty} P(X_{t+\ell} = i | X_{t-1} = j_1, \ldots, X_{t-p} = j_p).
\]

Equating equations (A6) and (A10), we have that \( \lambda_q = \lambda_q \), \( \forall i \).

Appendix 2. System of matrix equations of the MTDg model

Proposition B.1 Suppose that a sequence of random variables \( \{X_t\}_{t \in \mathbb{N}} \) taking values in the finite set \( \mathcal{X} = \{1, \ldots, m\} \) is defined by equation (2) and is stationary. Let \( B(k) \) be a \( m \times m \) matrix with elements

\[
b_{i,j} = P(X_t = i, X_{t+k} = j), \quad i, j \in \mathcal{X}; k \in \mathbb{Z}
\]

and \( B(0) = \text{diag}(\lambda_1, \ldots, \lambda_m) \). Then

\[
B(k) = \sum_{g=1}^{p} \lambda_g B(k-g) Q^g.
\]

Proof. First consider the case where \( k = 1, \ldots, p \). Let

\[
y_k^l = \{X_{t+k-g} : g = 1, \ldots, p; g \neq k\},
\]

then

\[
b_{i,j} = \sum_{g=1}^{p} \lambda_g P(X_t = i, X_{t+k} = j) = \sum_{g=1}^{p} \lambda_g P(X_t = i, X_{t+k} = j | y_k^l) P(y_k^l).
\]

We write the same matrix for the model (28)

\[
R = \sum_{g=1}^{p} \lambda_g U_g \quad \text{and} \quad A_{g,k} = \begin{cases} \frac{Q^g}{1} & \text{if } g = k \\ Q^g & \text{if } g \neq k \end{cases}
\]

where \( U_g = A_{g,1} \otimes \cdots \otimes A_{g,p} \)

and \( \otimes \) is the Kronecker product and \( 1^T \) is a \( m \times 1 \) vector of ones. We now calculate \( \xi R \) in another way. The \( k \)-th column of \( U_g \) is

\[
\sum_{i_1, \ldots, i_p} q_{g,i_1}^{'}, \xi_{i_1, \ldots, i_p} = \sum_{i_1, \ldots, i_p} q_{g,i_1}^{'}, \xi_{i_1, \ldots, i_p} = \sum_{i_1} q_{g,i_1}^{'}, \omega_{i_1 \neq g}
\]

which is also the \( k \)-th column of \( \lambda_q Q^g \). Thus

\[
\xi R = \sum_{g=1}^{p} \lambda_g \omega Q^g.
\]
Appendix 3. A general class of MTDg models

Let $B(k)$ be an $m \times m$ matrix whose elements are
\[ b_{i,j}^k = \mathbb{P}(X_t = i, X_{t+k} = j), \quad i, j = 1, \ldots, m, k \in \mathbb{Z}, \] (C1)
where $B(0) = \text{diag}(\hat{\eta}_1, \ldots, \hat{\eta}_m)$. The matrices $B(k)$ represent the bivariate distributions of the random variable $X_t$. Then, we have that $B(k)$
\[ B(k) = \begin{pmatrix} b_{1,1}^k & \cdots & b_{1,m-1}^k \\ \vdots & \ddots & \vdots \\ b_{m-1,1}^k & \cdots & b_{m-1,m-1}^k \\ \hat{\eta}_1 - \sum_{i=1}^{m-1} b_{i,1}^k & \cdots & \hat{\eta}_{m-1} - \sum_{i=1}^{m-1} b_{i,m-1}^k \\ -\sum_{i=1}^{m-1} c_i b_{i,1}^k & \cdots & -\sum_{i=1}^{m-1} c_i b_{i,m-1}^k \\
\end{pmatrix}, \] (C2)

where the total number of independent elements is $m^2 - 2m + 1$ for each $k$. The parameters of the MTD model of order $p$ consists in the vector $\lambda = (\lambda_1, \ldots, \lambda_p)$ and the matrices $Q^g$, such that
\[ Q^g = 1^T \hat{\eta} + \tilde{Q}^g, \]
\[ \tilde{Q}^g = \begin{pmatrix} \tilde{q}_{1,1}^g & \cdots & \tilde{q}_{1,m-1}^g \\ \vdots & \ddots & \vdots \\ \tilde{q}_{m-1,1}^g & \cdots & \tilde{q}_{m-1,m-1}^g \\ -\sum_{i=1}^{m-1} \tilde{q}_{i,1}^g & \cdots & -\sum_{i=1}^{m-1} \tilde{q}_{i,m-1}^g \\ -\sum_{i=1}^{m-1} c_i \tilde{q}_{i,1}^g & \cdots & -\sum_{i=1}^{m-1} c_i \tilde{q}_{i,m-1}^g \\
\end{pmatrix}, \] (C3)

where $\hat{\eta} \tilde{Q}^g = 0$, $\forall g$ and $c_i = \hat{\eta}_i/\hat{\eta}_m$. Consistently with these definitions, the conditional probabilities of the $p$-order Markov chain read
\[ \mathbb{P}(X_t = i | X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \hat{\eta}_i + \sum_{g=1}^{p} \hat{a}_{i,g,i}^g, \] (C4)

where $\hat{a}_{i,g,i}^g \equiv \lambda_g \hat{\eta}_{i,g,i}^g$.

Within this framework, the bivariate distributions and the matrices $\tilde{Q}^g$ satisfy the following system of matrix equations
\[ B(k) - \hat{\eta}^T \hat{\eta} + \sum_{g=1}^{p} B(k - g) A^g, \] (C5)

where $A^g \equiv \lambda_g \tilde{Q}^g$. Employing the empirical bivariate distributions, above linear system can be inverted in order to find the parameters of the model. Resulting parameters have to satisfy the following conditions in order to characterise a well-defined $p$-order Markov model
\[ \hat{\eta}_i + \sum_{g=1}^{p} \max_{i \in \mathcal{X}} \left( a_{i,g,i}^g \right) < 1, \quad \forall i \in \mathcal{X}; \]
\[ \hat{\eta}_i + \sum_{g=1}^{p} \min_{i \in \mathcal{X}} \left( a_{i,g,i}^g \right) > 0, \quad \forall i \in \mathcal{X}. \] (C6)

Appendix 4. Convexity of the optimisation problem

PROPOSITION D.1 If $K$ is not singular, the following constrained optimisation problem
\[ \hat{q} = \arg\min_{q \in \mathbb{R}^{m^2-2m+1}} \|d - K \cdot q\|^2 \]
\[ \text{s.t.} \quad \hat{\eta}_i + \sum_{g=1}^{p} \max_{i \in \mathcal{X}} \left( a_{i,g,i}^g \right) < 1, \quad \forall i \in \mathcal{X} \]
\[ \hat{\eta}_i + \sum_{g=1}^{p} \min_{i \in \mathcal{X}} \left( a_{i,g,i}^g \right) > 0, \quad \forall i \in \mathcal{X} \] (D1)
is convex in $\mathbb{R}^{p(m^2-2m+1)}$.

Proof. This is true if the objective function and all the constraints are convex functions. First of all, it is straightforward to show that the Hessian of the objective function $2K K^T$ is a positive semi-definite matrix. The constraints are convex in $q$, if they are convex in the parameters $a_{i,j}^g$ because they are affine functions of the components of $q$. Let $a$ be the vector of parameters $(a_{i,j}^g)_{i,j \in \mathcal{X}; 1 \leq g \leq p}$, we need to prove that the function
\[ f(a) = \sum_{g=1}^{p} \max_{i,j \in \mathcal{X}} \left( a_{i,j}^g \right), \quad \forall i \in \mathcal{X} \] (D2)
is convex in $\mathbb{R}^{p(m^2-2m+1)}$.

If we prove it for a fixed $i$, then it is true for all $i \in \mathcal{X}$ and also for the constraints with the minimum function. The function $f(a)$ satisfies, for $0 \leq \theta \leq 1$, different vectors of parameters $a, b \in \mathbb{R}^{m^2p}$, and fixed $i$
\[ f(\theta a + (1 - \theta)b) = \sum_{g=1}^{p} \max_{i \in \mathcal{X}} \left( \theta a_{i,g,i}^g + (1 - \theta)b_{i,g,i}^g \right) \]
\[ \leq \theta \sum_{g=1}^{p} \max_{i \in \mathcal{X}} \left( a_{i,g,i}^g \right) + (1 - \theta) \sum_{g=1}^{p} \max_{i \in \mathcal{X}} \left( b_{i,g,i}^g \right) \]
\[ = \theta f(a) + (1 - \theta) f(b). \] (D3)

Therefore, we conclude that the function $f(a)$ is convex in $\mathbb{R}^{p(m^2-2m+1)}$. \qed