Decisions in Economics and Finance (2021) 44:905–940
https://doi.org/10.1007/s10203-021-00344-9

Temporal mixture ensemble models for probabilistic forecasting of intraday cryptocurrency volume

Nino Antulov-Fantulin[1,2] · **Tian Guo**[3] · **Fabrizio Lillo**[4]

Received: 30 March 2020 / Accepted: 30 June 2021 / Published online: 10 August 2021

Abstract
We study the problem of the intraday short-term volume forecasting in cryptocurrency multi-markets. The predictions are built by using transaction and order book data from different markets where the exchange takes place. Methodologically, we propose a temporal mixture ensemble, capable of adaptively exploiting, for the forecasting, different sources of data and providing a volume point estimate, as well as its uncertainty. We provide evidence of the clear outperformance of our model with respect to econometric models. Moreover our model performs slightly better than Gradient Boosting Machine while having a much clearer interpretability of the results. Finally, we show that the above results are robust also when restricting the prediction analysis to each volume quartile.

Keywords Econometrics · Machine learning · Cryptocurrency markets · Temporal mixture ensemble

JEL Classifications C53 · C58 · G12

1 Introduction

Cryptocurrencies recently attracted massive attention from public and researcher community in several disciplines such as finance and economics (Urquhart 2016; Bolt

N. Antulov-Fantulin, T. Guo: Shared first authorship.
T. Guo: Work done when at ETH Zürich.

✉ Fabrizio Lillo
fabrizio.lillo@unibo.it

1 ETH Zürich, Zürich, Switzerland

2 Aisot Technologies AG Zurich, Zürich, Switzerland

3 RAM Active Investments, Geneva, Switzerland

4 Department of Mathematics, University of Bologna and Scuola Normale Superiore, Pisa, Italy

2016; Cheah and Fry 2015; Chu et al. 2015; Donier and Bouchaud 2015; Ciaian et al. 2016), computer science (Ron and Shamir 2013; Jang and Lee 2018; Amjad and Shah 2016; Alessandretti et al 2018; Guo et al 2018; Beck et al 2019) or complex systems (Garcia and Schweitzer 2015; Wheatley et al. 2019; Gerlach et al. 2019; Antulov-Fantulin et al. 2018; Kondor et al. 2014; ElBahrawy et al. 2017). It originated from a decentralized peer-to-peer payment network (Nakamoto 2008), relying on cryptographic methods (Bos et al. 2014; Mayer 2016) like elliptical curve cryptography and the SHA-256 hash function. When new transactions are announced on this network, they have to be verified by network nodes and recorded in a public distributed ledger called the blockchain (Nakamoto 2008). Cryptocurrencies are created as a reward in the verification competition (see Proof of work Jakobsson and Juels 1999), in which users offer their computing power to verify and record transactions into the blockchain. Bitcoin is one of the most prominent decentralized digital cryptocurrencies and it is the focus of this paper, although the model developed below can be adapted to other cryptocurrencies with ease, as well as to other "ordinary" assets (equities, futures, FX rates, etc.).

The exchange of Bitcoins with other fiat or cryptocurrencies takes place on exchange markets, which share some similarities with the foreign exchange markets (Baumöhl 2019). These markets typically work through a continuous double auction, which is implemented with a limit order book mechanism, where no designated dealer or market maker is present and limit and market orders to buy and sell arrive continuously. Moreover, as observed for traditional assets, the market is *fragmented*, i.e. there are several exchanges where the trading of the same asset, in our case the exchange of a cryptocurrency with a fiat currency, can simultaneously take place.

The automation of the (cryptocurrency) exchanges lead to the increase of the use of automated trading (Chaboud et al. 2014; Hendershott et al. 2011) via different trading algorithms. An important input for these algorithms is the prediction of future trading volume. This is important for several reasons. First, trading volume is a proxy for liquidity which in turn is important to quantify transaction costs. Trading algorithms aim at minimizing these costs by splitting orders in order to find a better execution price (Frei and Westray 2015; Barzykin and Lillo 2019) and the crucial part is the decision of when to execute the orders in such a way to minimize market impact or to achieve certain trading benchmarks (e.g. VWAP) (Brownlees et al. 2010; Satish et al. 2014; Chen et al. 2016; Bialkowski et al. 2008; Calvori et al. 2013; Kawakatsu 2018). Second, when different market venues are available, the algorithm must decide where to post the order and the choice is likely the market where more volume is predicted to be available. Third, volume is also used to model the time-varying price volatility process, whose relation is also known as "Mixture of Distribution Hypothesis" (Andersen 1996).

In this paper, we study the problem of intraday short-term volume prediction on multi-market of cryptocurrency, as is shown in Fig. 1, intending to obtain not only point estimate but also the uncertainty on the point prediction (Chu et al. 2015; Urquhart 2016; Katsiampa 2017; Balcilar et al. 2017). Moreover, conventional volume predictions focuses on using data or features from the same market. Since cryptocurrency markets are traded on several markets simultaneously, it is reasonable to use cross-market data not only to enhance the predictive power, but also to help understanding the interaction between markets. In particular, we investigate the exchange rate of Bitcoin

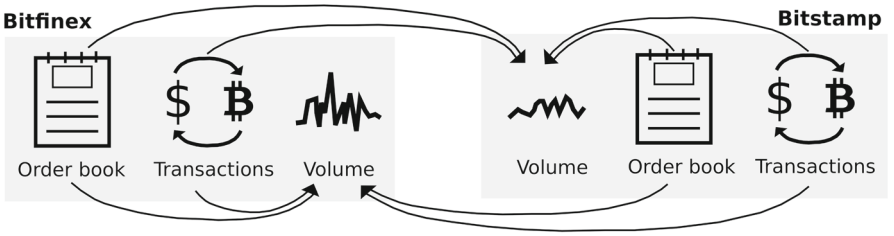

Fig. 1 Illustration of the probabilistic volume predicting in the multi-source data setting of Bitfinex and Bitstamp markets. The left and right panel respectively depict the order book and transaction data of each market. The arrows represent the data used to model the volume of each market. Note that the volume information is implicitly contained in the transaction data of one market and thus there is no arrow linking the volumes from the two markets

(BTC) with a fiat currency (USD) on two liquid markets: Bitfinex and Bitstamp. The first market is more liquid than the second, since its traded volume in the investigated period from June 2018 to November 2018 is 2.5 times larger.[1] Thus one expects an asymmetric role of the past volume (or other market variables) of one market on the prediction of volume in the other market.

Specifically, the contribution of this paper can be summarized as follows:

- We formulate the cross-market volume prediction as a supervised multi-source learning problem. We use multi-source data, i.e. transactions and limit order books from different markets, to predict the volume of the target market.
- We propose the **T**emporal **M**ixture **E**nsemble (TME), which models individual source's relation to the target and adaptively adjusts the contribution of the individual source to the target prediction.
- By equipping with modern ensemble techniques, the proposed model can further quantify the predictive uncertainty since the conditional distribution is a mixture of log-normal distributions and confidence intervals can be computed (see below).
- As main benchmarks for volume dynamics, we use different time-series and machine learning models (clearly with the same regressors/features used in our model). We observe that our dynamic mixture ensemble is often having superior out-of-sample performance on conventional prediction error metrics e.g. root mean square error (RMSE) and mean absolute error (MAE). More importantly, it presents much better calibrated results, evaluated by metrics taking into account predictive uncertainty, i.e. normalized negative log-likelihood (NNLL), uncertainty interval width (IW).
- We discuss the prediction performance conditional to volume. Since our choice of modeling log-volume is tantamount to considering a multiplicative noise model for volumes, when using relative RMSE and MAE machine learning methods outperforms econometric models in providing more accurate forecasts.

[1] Recently, there have been few reports that are showing fake reported volume for certain Bitcoin exchange markets. In this paper, we investigate Bitcoin exchange markets that have regulatory status (Hougan et al. 2019) either with the Money Services Business (MSB) license or BitLicense from the New York State Department of Financial Services, and have been independently verified to report true values.

The paper is organized as follows: in Sect. 2 we present the investigated markets, the data, and the variables used in the modeling. In Sect. 3 we present our benchmark models. In Sect. 4 we present our empirical investigations on the cryptocurrency markets for the prediction of intraday market volume. Finally, Sect. 5 presents some conclusions and outlook for future work. Most of the technical description of models and algorithms, as well as some additional empirical results, are presented in an appendix.

2 Multiple market cryptocurrency data

Our empirical analyses are performed on a sample of data over the period from May 31, 2018 9:55 pm (UTC) until September 30 2018 9:59 pm (UTC) from two exchange markets, Bitfinex[2] and Bitstamp,[3] where Bitcoins can be exchanged with US dollars. These markets work through a limit order book, as many conventional exchanges. For each of the two markets we consider two types of data: transaction data and limit order book data.

From **transaction data** we extract the following features on each 1-min interval:

- *Buy volume*–number of BTCs traded in buyer initiated transactions
- *Sell volume*–number of BTCs traded in seller initiated transactions
- *Volume imbalance*–absolute difference between buy and sell volume
- *Buy transactions*–number of executed transactions on buy side
- *Sell transactions*–number of executed transactions on sell side
- *Transaction imbalance*–absolute difference between buy and sell number of transactions

We remind that a buyer (seller) initiated transaction in a limit order book market is a trade where the initiator is a buy (sell) market order or a buy (sell) limit order crossing the spread.

From **limit order book data** we extract the following features each minute (Gould et al. 2013; Rambaldi et al. 2016):

- *Spread* is the difference between the highest price that a buyer is willing to pay for a BTC (bid) and the lowest price that a seller is willing to accept (ask).
- *Ask volume* is the number of BTCs on the ask side of order book.
- *Bid volume* is the number of BTCs on the bid side of order book.
- *Imbalance* is the absolute difference between ask and bid volume.
- *Ask/bid Slope* is estimated as the volume until δ price offset from the best ask/bid price. δ is estimated by the bid price at the order that has at least 1%, 5% and 10 % of orders with the highest bid price.
- *Slope imbalance* is the absolute difference between ask and bid slope at different values of price associated to δ. δ is estimated by the bid price at the order that has at least 1%, 5% and 10 % of orders with the highest bid price.

The target variable that we aim at forecasting is the trading volume of a given target market including both buy and sell volume. In the proposed modeling approaches

[2] https://www.bitfinex.com.

[3] https://www.bitstamp.net.

(described in Sect. 3) we consider different data sources at each time can affect the probability distribution of trading volume in the next time interval in a given market. As is illustrated in Fig. 1, given the setting presented above, there are four sources, namely one for transaction data and one for limit order book data for the two markets.

Before going into the details of models, in order to choose the appropriated variable to investigate, we visualize some characteristics of the data in Fig. 2. In the 1st and 3rd panel, we show the quantiles of intraday 1-min trading volume of BTC/USD rates, for the two different markets. We also show as vertical lines the opening times of four major stock exchanges. Although we do not observe abrupt changes in volume distribution (possibly because cryptocurrency exchanges are weakly related with stock exchanges), some small but significant intraday pattern (Andersen and Bollerslev 1997) is observed.

For this reason, we pre-process the raw volume data to remove intraday patterns as follows. Let us denote v_t the volume traded at the time t, in units of Bitcoins. $I(t)$ is a mapping function of the time t, which returns the intraday time interval index of t. We use $a_{I(t)}$ to represent the average of volumes at the same intraday index $I(t)$ across days. Next, to remove intraday patterns, we process the raw volume by the following operation:

$$y_t \triangleq \frac{v_t}{a_{I(t)}} \tag{1}$$

In practice, in order to avoid leaking future information to the training phase, $a_{I(t)}$ is calculated only based on the training data, and shared to both validating and testing data.

The histogram of $\ln y_t$ for the two markets[4] is shown in the 2nd and 4th panel of Fig. 2 along with first four cumulants of empirical distribution of log-volumes. We observe that the distribution is approximately normal, even if a small negative skew is present. For this reason, our modeling choice is to consider y_t as log-normally distributed.

3 Models

Econometric modeling of intra-daily trading volume relies on a set of empirical regularities (Brownlees et al. 2010; Satish et al. 2014; Chen et al. 2016) of volume dynamics. These include fat tails, strong persistence and an intra-daily clustering around the "U"-shaped periodic component. Brownlees et al. (2010) proposed Component Multiplicative Error Model (CMEM), which is the extension of Multiplicative Error Model (MEM) (Engle 2002). The CMEM volume model has a connection to the component-GARCH (Engle and Sokalska 2012) and the periodic P-GARCH (Bollerslev and Ghysels 1996). Satish et al. (2014), proposed four-component volume forecast model composed of: (i) rolling historical volume average, (ii) daily ARMA for serial correlation across daily volumes, (iii) deseasonalized intra-day ARMA volume model and (iv) a dynamic weighted combination of previous models. Chen et al. (2016),

[4] The frequency of time intervals with zero volume and how we handle them is detailed in Sect. 2.

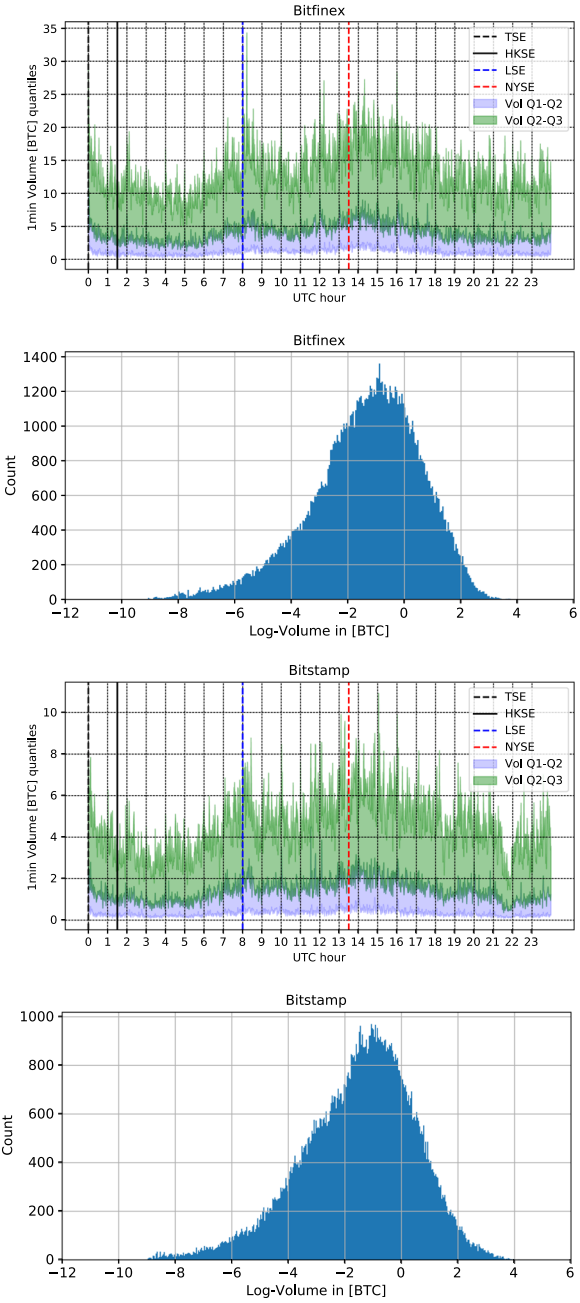

Fig. 2 **1st and 3rd panel**: interquartile range $Q_1 - Q_2$ and $Q_2 - Q3$ of intraday 1-min transaction volume of BTC/USD rate on Bitfinex (1st panel) and Bitstamp (3rd panel) market, along with openings of major exchanges (TSE, HKSE, LSE, NYSE) that are denoted with vertical lines. **2nd and 4th panel**: histogram of deseasonalized log volume at 1-min resolution. The statistics of these variable are: Bitfinex: mean = -1.3627, variance = 3.7658, skewness = -0.6336, and kurtosis = 0.6886; Bitstamp: mean = -2.4224, variance = 3.9894, skewness = -0.5307, kurtosis = 0.4114

simplify the multiplicative volume model (Brownlees et al. 2010) into an additive one by modeling the logarithm of intraday volume with the Kalman filter.

3.1 Problem setting

In this paper, we focus on the probabilistic forecasting of the intraday volume in the multi-source data setting. We propose the use of TME, presented below, and we benchmark its performance against two econometric baseline models (ARMA-GARCH and ARMAX-GARCH) and one Machine Learning baseline model (Gradient Boosting Machine).

As mentioned above, we assume y_t follows the log-normal distribution and thus all the models used in the experiments will be developed to learn the log-normal distribution of y_t. However, our proposed TME is flexible to be equipped with different distributions for the target, and the study in this direction will be the future work.

When evaluating the performance of the forecasting procedure with real data experiments, we choose to use evaluation metrics defined on the original volume v_t, since in real world application the interest is in forecasting volume rather than log-volume. Finally, for understanding the performance of TME and Machine Learning and econometric baselines in different setups, we will evaluate them using three different time intervals of volumes, namely 1 min, 5 min and 10 min.

Regarding the multi-source data setting, on one hand, it includes the features from the target market. This data is believed to be directly correlated with the target variable. On the other hand, there is an alternative market, which could interact with the target market. Together with the target market, the features from this alternative market constitute the multi-source data.

In this paper, we mainly focus on Bitfinex and Bitstamp markets. For each market, we have the features from both transaction and order book data, thereby leading to $S = 4$ data sources. In particular, we indicate with $\mathbf{x}_{s,t} \in \mathbb{R}^{d_s}$ the features from source s at time step t, and d_s the dimensionality of source data s. Given the list of features presented in Sect. 2, we have $d_s = 6$ when the source is transaction data in any market, while $d_s = 13$ for order book data. Then, these multi-source data will be used to model the volume of each market, as is shown in Fig. 1.

3.2 Overview of TME

In this paper, we construct a Temporal Mixture Ensemble (TME), belonging to the class of of mixture models (Waterhouse et al. 1996; Yuksel et al. 2012; Wei et al. 2007; Bazzani et al. 2016; Guo et al. 2019), which takes previous transactions and limit order book data (Gould et al. 2013; Rambaldi et al. 2016) from multiple markets simultaneously into account. Though mixture models have been widely used in machine learning and deep learning (Guo et al 2018; Schwab et al. 2019; Kurle et al. 2019), they have been hardly explored for prediction tasks in cryptocurrency markets. Moreover, our proposed ensemble of temporal mixtures can provide predictive uncertainty of the target volume by the use of the Stochastic Gradient Descent (SGD) based ensemble technique (Lakshminarayanan et al. 2017; Maddox et al. 2019; Snoek

et al. 2019). Predictive uncertainty reflects the confidence of the model over the prediction. It is valuable extra information for model interpretability and reliability. The model developed below is flexible to consume multi-source data of arbitrary number of sources and dimensionalities of individual source data.

In principle, TME exploits latent variables to capture the contributions of different sources of data to the future evolution of the target variable. The source contributing at a certain time depends on the history of all the sources.

For simplicity, we will use one data sample of the target y_t to present the proposed model. In reality, the training, validation, and testing data contain the samples collected in a time period. More quantitatively, the generative process of the target variable y_t conditional on multi-source data $\{\mathbf{x}_{1,t}, \ldots, \mathbf{x}_{S,t}\}$ is formulated as the following probabilistic mixture process:

$$\begin{aligned} &p(y_t|\{\mathbf{x}_{1,<t}, \ldots, \mathbf{x}_{S,<t}\}, \Theta) \\ &= \sum_{s=1}^{S} p_{\theta_s}(y_t|z_t = s, \mathbf{x}_{s,<t}) \cdot \mathbb{P}_\omega(z_t = s\,|\mathbf{x}_{1,<t}, \ldots, \mathbf{x}_{S,<t}). \end{aligned} \tag{2}$$

The latent variable z_t is a discrete random variable defined on the set of values $\{1, \ldots, S\}$, each of which represents the corresponding data source. The quantity $p_{\theta_s}(y_t|z_t = s, \mathbf{x}_{s,<t})$ models the predictive probabilistic density of the target based on the historical data $\mathbf{x}_{s,<t}$ from a certain source s. The quantity[5] $\mathbb{P}_\omega(z_t = s\,|\mathbf{x}_{1,<t}, \ldots, \mathbf{x}_{S,<t})$ characterizes a time-varying categorical distribution dependent on multi-source data. It adaptively adjusts the contribution of the data source specific density $p_{\theta_s}(y_t|z_t = s, \mathbf{x}_{s,<t})$ at each time step. Clearly, it holds $\sum_{s=1}^{S} \mathbb{P}_\omega(z_t = s\,|\mathbf{x}_{1,<t}, \ldots, \mathbf{x}_{S,<t}) = 1$. Finally, $\Theta \triangleq \{\theta_1, \ldots, \theta_S, \omega\}$ represents the parameters in data source specific components and the latent variable's probability function, and it will be learned in the training phase discussed below.

3.3 Model specification

We now specify in detail the mathematical formulation of each component in the temporal mixtures. Without loss of generality, we present the following model specification for the volume in cryptocurrency exchange of this paper's interest.

To specify the model, we need to define the predictive density function of individual sources, i.e. $p_{\theta_s}(y_t|z_t = s, \mathbf{x}_{s,<t})$ and the probability function of latent variable, i.e. $\mathbb{P}_\omega(z_t = s\,|\mathbf{x}_{1,<t}, \ldots, \mathbf{x}_{S,<t})$. We make a general assumption for both these functions that data from different sources are taken within the same time window w.r.t. the target time step. We denote by h the window length, i.e. the number of past time steps which enter in the conditional probabilities. We assume that this value is the same for each source. Equation 2 is thus simplified as:

$$\sum_{s=1}^{S} p_{\theta_s}(y_t|z_t = s, \mathbf{x}_{s,(-h,t)}) \cdot \mathbb{P}_\omega(z_t = s|\mathbf{x}_{1,(-h,t)}, \ldots, \mathbf{x}_{S,(-h,t)}), \tag{3}$$

[5] We indicate with p_θ probability densities and $\mathbb{P}_\omega$ probability mass functions.

where $\mathbf{x}_{s,(-h,t)}$ represents the data from source s within the time window from $t-h$ to $t-1$ and $\mathbf{x}_{s,(-h,t)} \in \mathbb{R}^{d_s \times h}$.

As for $p_{\theta_s}(y_t|z_t = s, \mathbf{x}_{s,(-h,t)})$, due to the non-negative nature of the volume in cryptocurrency exchange and its statistical properties (Bauwens et al. 2008), we choose the log-normal distribution to model y_t as:

$$\ln y_t \,|z_t = s, \mathbf{x}_{s,(-h,t)} \sim \mathcal{N}\big(\mu_{t,s}, \sigma^2_{t,s}\big) \tag{4}$$

As a result, the probability density function is expressed as:

$$p_{\theta_s}(y_t|z_t = s, \mathbf{x}_{s,(-h,t)}) = \frac{1}{\sqrt{2\pi} \cdot y_t \sigma_{t,s}} \exp\left(-\frac{(\ln y_t - \mu_{t,s})^2}{2\sigma^2_{t,s}}\right) \tag{5}$$

Given $\mathbf{x}_{s,(-h,t)} \in \mathbb{R}^{d_s \times h}$, we choose bi-linear regression to parameterize the mean and variance of the log transformed volume as follows:

$$\mu_{t,s} \triangleq L^\top_{\mu,s} \cdot \mathbf{x}_{s,(-h,t)} \cdot R_{\mu,s} + b_{\mu,s} \tag{6}$$

$$\sigma^2_{t,s} \triangleq \exp(L^\top_{\sigma,s} \cdot \mathbf{x}_{s,(-h,t)} \cdot R_{\sigma,s} + b_{\sigma,s}), \tag{7}$$

where $L_{\mu,s}, L_{\sigma,s} \in \mathbb{R}^{d_s}$ and $R_{\mu,s}, R_{\sigma,s} \in \mathbb{R}^h$. $b_{\mu,s}$, while $b_{\sigma,s} \in \mathbb{R}$ are bias terms. Note that above parameters are data source specific and then the trainable set of parameters is denoted by $\theta_s \triangleq \{L_{\mu,s}, L_{\sigma,s}, R_{\mu,s}, R_{\sigma,s}, b_{\mu,s}, b_{\sigma,s}\}$.

Based on the properties of log-normal distribution (MacKay and Mac Kay 2003; Cohen and Whitten 1980), the mean and variance of the intraday-pattern-free volume modeled by individual data sources can be derived from the mean and variance of log transformed volume as:

$$\mathbb{E}[y_t \mid z_t = s, \mathbf{x}_{s,(-h,t)}, \theta_s] = \exp\left\{\mu_{t,s} + \frac{1}{2}\sigma^2_{t,s}\right\} \tag{8}$$

and

$$\mathbb{V}[y_t \mid z_t = s, \mathbf{x}_{s,(-h,t)}, \theta_s] = \exp\{\sigma^2_{t,s} - 1\} \cdot \exp\{2\mu_{t,s} + \sigma^2_{t,s}\} \tag{9}$$

Then, we define the probability distribution of the latent variable z_t using a softmax function as follows:

$$\mathbb{P}_\omega(z_t = s \mid \{\mathbf{x}_{k,(-h,t)}\}_{k=1}^S) \triangleq \frac{\exp(\, f_s(\mathbf{x}_{s,(-h,t)})\,)}{\sum_{k=1}^S \exp(\, f_k(\mathbf{x}_{k,(-h,t)})\,)}, \tag{10}$$

where

$$f_s(\mathbf{x}_{s,(-h,t)}) \triangleq L^\top_{z,s} \cdot \mathbf{x}_{s,(-h,t)} \cdot R_{z,s} + b_{z,s} \tag{11}$$

$\omega \triangleq \{L_{z,s}, R_{z,s}, b_{z,s}\}_{s=1}^{S}$ denotes the set of trainable parameters regarding the latent variable z_t, i.e. $L_{z,s}, R_{z,s} \in \mathbb{R}^{d_s}$ and $b_{z,s} \in \mathbb{R}$ is a bias term.

In the following, we will present how the distributions modeled by individual data sources and the latent variable will be used to learn the parameters in the training phase and to perform probabilistic forecasting of the target volume in the predicting phase.

3.4 Learning

The learning process of TME is based on SGD optimization (Ruder 2016; Kingma and Ba 2015). It is able to give rise to a set of parameter realizations for building the ensemble which has been proven to be an effective technique for enhancing the prediction accuracy as well as enabling uncertainty estimation in previous works (Lakshminarayanan et al. 2017; Maddox et al. 2019).

We first briefly describe the training process of SGD optimization. Denote the set of the parameters by $\Theta \triangleq \{\theta_1, \ldots, \theta_S, \omega\}$. The whole training dataset denoted by $\mathcal{D}$ is consisted of data instances, each of which is a pair of y_t and $\{\mathbf{x}_{s,(-h,t)}\}_1^S$. t is a time instant in the period $\{1, \ldots, T\}$.

Starting from a random initialized value of Θ, in each iteration SGD samples a batch of training instances to update the model parameters as follows:

$$\Theta(i) = \Theta(i-1) - \eta \nabla \mathcal{L}(\Theta(i-1); \mathcal{D}_i), \tag{12}$$

where i is the iteration step, $\Theta(i)$ represents the values of Θ at step i, i.e. a snapshot of Θ. η is the learning rate, a tunable hyperparameter to control the magnitude of gradient update. $\nabla \mathcal{L}(\Theta(i-1); \mathcal{D}_i)$ is the gradient of the loss function w.r.t. the model parameters given data batch $\mathcal{D}_i$ at iteration i. The iteration stops when the loss converges with negligible variation. The model parameter snapshot at the last step or with the best validation performance will be taken as one realization of the model parameters.

The learning process of TME is to minimize the loss function defined by the negative log likelihood of the target volume as:

$$\mathcal{L}(\Theta; \mathcal{D}) \triangleq -\sum_{t=1}^{T} \ln l(\Theta\,;\, y_t, \{\mathbf{x}_{s,(-h,t)}\}_{s=1}^{S}) + \lambda \left\|\Theta\right\|_2^2, \tag{13}$$

where $\|\Theta\|_2^2$ is the L2 regularisation with the hyper-parameter λ. The likelihood function is denoted by $l(\Theta\,;\, y_t, \{\mathbf{x}_{s,(-h,t)}\}_{s=1}^{S})$. Based on the model specification in Sect. 3.3, it is expressed as:

$$\begin{aligned} &l(\Theta\,;\, y_t, \{\mathbf{x}_{s,(-h,t)}\}_{s=1}^{S}) \\ &= \sum_{s=1}^{S} \frac{1}{\sqrt{2\pi} \cdot y_t \sigma_{t,s}} \exp\Big(-\frac{(\ln y_t - \mu_{t,s})^2}{2\sigma_{t,s}^2} \Big) \cdot \frac{\exp(\, f_s(\mathbf{x}_{s,(-h,t)})\,)}{\sum_{k=1}^{S} \exp(\, f_k(\mathbf{x}_{k,(-h,t)})\,)}, \end{aligned} \tag{14}$$

In the SGD optimization process, different initialization of Θ leads to distinct parameter iterate trajectories. Recent studies show that the ensemble of independently initialized and trained model parameters empirically often provide comparable performance on uncertainty quantification w.r.t. sampling and variational inference based methods (Lakshminarayanan et al. 2017; Snoek et al. 2019; Maddox et al. 2019). Our ensemble construction follows this idea by taking a set of model parameter realizations from different training trajectories, and each parameter realization is denoted by $\Theta_m = \{\theta_{m,1}, \ldots, \theta_{m,S}, \omega_m\}$. For more algorithmic details about the collecting procedure of these parameter realizations, please refer to the appendix. The code for the TME model is available at our Git repository.[6]

3.5 Prediction

In this part, we present the probabilistic predicting process given model parameter realizations $\{\Theta_m\}_1^M$. Specifically, the predictive mean of intraday-pattern-free volume y_t is expressed as:

$$\mathbb{E}[y_t|\{\mathbf{x}_{s,(-h,t)}\}_{s=1}^S, \mathcal{D}] \approx \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}[y_t \mid \{\mathbf{x}_{s,(-h,t)}\}_{s=1}^S, \Theta_m], \tag{15}$$

where $\mathbb{E}[y_t|\{\mathbf{x}_{s,(-h,t)}\}_{s=1}^S, \Theta_m]$ is the conditional mean given one realization Θ_m. In TME, it is a weighted sum of the predictions by individual data sources as:

$$\begin{aligned}\mathbb{E}[y_t|\{\mathbf{x}_{s,(-h,t)}\}_{s=1}^S, \Theta_m] = &\sum_{s=1}^{S}\mathbb{P}_{\omega_m}(z_t = s|\{\mathbf{x}_{s,(-h,t)}\}_{s=1}^S) \\ &\cdot\mathbb{E}[y_t|z_t = s, \mathbf{x}_{s,(-h,t)}, \theta_{m,s}]\end{aligned} \tag{16}$$

Note that y_t is the volume deseasonalized from the intraday patterns. Since our interest is to obtain the volume prediction in the original scale, the predicted volume is derived as:

$$\hat{v}_t = a_{I(t)} \cdot \mathbb{E}[y_t|\{\mathbf{x}_{s,(-h,t)}\}_{s=1}^S, \mathcal{D}] \tag{17}$$

Besides the point prediction $\hat{v}_t$, TME enables to characterize the distributional information of the prediction. In particular, the predictive probability density of the volume in the original scale is:

$$p(v_t \,|\{\mathbf{x}_{s,(-h,t)}\}_{s=1}^S, \{\Theta_m\}_1^M) = \frac{1}{M \cdot a_{I(t)}}\sum_{m=1}^{M} p\Big(\frac{v_t}{a_{I(t)}} \,\Big|\, \{\mathbf{x}_{s,(-h,t)}\}_{s=1}^S, \Theta_m\Big), \tag{18}$$

where $p\Big(\frac{v_t}{a_{I(t)}} \,\Big|\, \{\mathbf{x}_{s,(-h,t)}\}_{s=1}^S, \Theta_m\Big)$ corresponds to the density function of y_t in Eq. 2.

[6] https://github.com/weilai0980/Probabilistic_Mixture_Ensemble.

Furthermore, the cumulative distribution function of v_t is:

$$P(v_t \,|\{\mathbf{x}_{s,(-h,t)}\}_{s=1}^{S}, \{\Theta_m\}_{m=1}^{M}) = \frac{1}{M}\sum_{m=1}^{M}\sum_{s=1}^{S} P_{\theta_{m,s}}(y_t|z_t = s, \mathbf{x}_{s,(-h,t)}) \cdot \mathbb{P}_{\omega_m}(z_t = s|\{\mathbf{x}_{k,(-h,t)}\}_{k=1}^{S}), \quad (19)$$

where $y_t = \frac{v_t}{a_{I(t)}}$ and $P_{\theta_{m,s}}(\cdot)$ represents the cumulative distribution function generated by the component model parameterized by $\theta_{m,s}$. In particular, for log-normal distribution, it is expressed as:

$$P_{\theta_{m,s}}(y_t \,|z_t = s, \mathbf{x}_{s,(-h,t)}) = \Phi\left(\frac{\ln y_t - \mu_{t,s}}{\sigma_{t,s}}\right), \quad (20)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. With Eqs. 17–20, we are able to assess the real volume predictions via both conventional error metrics and probabilistic performance metrics, which will be shown in the experiment section. Specifically, by using the distribution function we find numerically the quantiles needed to compute at each time the confidence intervals.

4 Experiments

In this section, we report the overall experimental evaluation. More detailed results are in the appendix section.

4.1 Data and metrics

Data: We collected the limit order book and transaction data respectively from two exchanges, Bitfinex and Bitstamp and extracted features defined in Sect. 2 from the order book and transactions of each exchange for the period from May 31, 2018 9:55 pm (UTC) until September 30, 2018 9:59 pm (UTC). Then, for each exchange, we build three datasets of different prediction horizons, i.e. 1 min, 5 min, 10 min, for training and evaluating models. Depending on the prediction horizon, each instance in the dataset contains a target volume and the time-lagged features from order book and transactions of two exchanges. In particular, for Bitfinex, the sizes of datasets for 1 min, 5 min, 10 min are respectively 171727, 34346 and 17171. For Bitstamp, the sizes of datasets are respectively 168743, 33749 and 16873. In all the experiments, data instances are time ordered and we use the first 70% of points for training, the next 10% for validation, and the last 20% of points for out-of-sample testing. Note that all the metrics are evaluated on the out of sample testing data.

For modeling the log-normal distribution, the baseline models and TME need to perform the log-transformation on y_t and thus in the data pre-processing step, we have filtered out the data instances with zero trading volume. Empirically, we found out

that these zero-volume data instances account for less than 2.25% and 3.98% of the entire dataset for Bitfinex and Bitstamp markets respectively.[7]

For the baseline methods not differentiating the data source of features, each target volume has a feature vector built by concatenating all features from order book and transactions of two markets into one vector. For our TME, each target volume has four groups of features respectively corresponding to the order book and transaction data of two markets.

Metrics: Note that baseline models and TME are trained to learn the log-normal distribution of the deseasonalized volume, however, the following metrics are evaluated on the original scale of the volume, because our aim is to quantify the performance on the real volume scale, that is of more interest for practical purposes.

In the following definitions, $\hat{v}_t$ corresponds to the prediction of the raw volume. $\hat{v}_t$ is derived from the predictive mean of y_t multiplied by $a_{I(t)}$, according to the definition of y_t in Sect. 3. For baseline models, the predictive mean of y_t is derived from the mean of logarithmic transformed y_t via Eq. 8 used by individual sources in TME. $\overline{T}$ is the number of data instances in the testing dataset.

The considered performance metrics are:

RMSE: is the root mean square error as $\text{RMSE} = \sqrt{\frac{1}{\overline{T}} \sum_{t=1}^{\overline{T}} (v_t - \hat{v}_t)^2}$.

MAE: is the mean absolute error as $\text{MAE} = \frac{1}{\overline{T}} \sum_{t=1}^{\overline{T}} |v_t - \hat{v}_t|$.

NNLL: is the predictive Negative Log-Likelihood of testing instances normalized by the total number of testing points. For TME, the likelihood is calculated based on Eq. 14. For baseline models, the likelihood of the real v_t is that of the corresponding intraday-pattern free y_t scaled by $\frac{1}{a_{I(t)}}$, according to the definition in Sect. 3. The likelihood of y_t can be straightforwardly derived from the predictive mean and variance of $\ln y_t$ (MacKay and Mac Kay 2003; Cohen and Whitten 1980).

IW: is the averaged width of the prediction intervals of testing instances corresponding to a certain probability. It is meant to evaluate the uncertainty in probabilistic forecasting. The ideal model is expected to provide tight intervals, which imply the model is confident in the prediction even if this metric must always be used jointly with NNLL. Notice that the IW is the width of the confidence interval.

Specifically, in this paper the prediction interval is derived by the quantiles of the two boundary points defining the probability. We will report the prediction interval corresponding to 68% probability of the predictive distribution, i.e. the quantiles of 16% and 84%. For TME producing multi-modal distributions, the cumulative distribution function in Eq. 19 enables to numerically obtain the interval. As the baselines mainly generate unimodal predictive distributions, we can use the quantile function of the log-normal distribution to obtain the quantiles of intraday pattern-free volume and consequentially the prediction interval of the real volume.

[7] The choice of removing zero volume observations is clearly non optimal and more sophisticated methods might be adopted. In our case, the fraction of zero volume observations is small enough that the impact of the removal should be negligible.

Table 1 Statistics of GARCH(1,1) parameters (ω, α, β) on log-volume residuals, trained with ARMA(p, q)-GARCH(1,1) model for different markets and prediction horizons

BITFINEX	ω	Std.Err(ω)	α	Std.Err(α)	β	Std.Err(β)
1 min	0.0177*	0.0011	0.0259*	0.0008	0.9677*	0.0000
5 min	0.0119*	0.0015	0.0218*	0.0017	0.9663*	0.0000
10 min	0.0062*	0.0011	0.0152*	0.0018	0.9762*	0.0000
BITSTAMP	ω	Std.Err(ω)	α	Std.Err(α)	β	Std.Err(β)
1 min	0.0112*	0.0007	0.0203*	0.0006	0.9759*	0.0000
5 min	0.0175*	0.0023	0.0277*	0.0022	0.9561*	0.0000
10 min	0.0262*	0.0056	0.0291*	0.0042	0.9387*	0.0000

(*) indicate p-values $< 10^{-5}$ for estimated parameters. Parameters p, q, were selected by minimizing AIC and are reported in Tables 2, 3 and 4

4.2 Baseline models and TME setup

As mentioned above, we benchmark the performance of TME against two econometric and one Machine Learning models. We present them below.

ARMA-GARCH is the autoregressive moving average model (ARMA) plus generalized autoregressive conditional heteroskedasticity (GARCH) model (Brownlees et al. 2010; Satish et al. 2014; Chen et al. 2016). It is aimed to respectively capture the conditional mean and conditional variance of the logarithmic volume. Then, the predictive mean and variance of the logarithmic volume are transformed to the original scale of the volume for evaluation.

The number of autoregressive and moving average lags in ARMA are selected in the range from 1 to 10, by minimization of Akaike Information Criterion, while the orders of lag variances and lag residual errors in GARCH are found to affect the performance negligibly and thus are both fixed to one. The residual diagnostics for ARMA-GARCH models is given in the Appendix Figs. 5 and 6.

In Table 1 we report the estimated GARCH parameters in the ARMA-GARCH model on the log-volume, together with the standard errors and the p-value. All parameters are statistically different from zero at all time scales, indicating the significant existence of heteroskedasticity.

ARMAX-GARCH is the variant of ARMA-GARCH by adding external feature terms. In our scenario, external features are obtained by concatenating all features from order book and transaction data of two exchanges into one feature vector. The hyper-parameters in ARMAX-GARCH are selected in the same way as for ARMA-GARCH.

GBM is the gradient boosting machine (Friedman 2001). It has been empirically proven to be highly effective in predictive tasks across different machine learning challenges (Gulin et al. 2011; Taieb and Hyndman 2014) and more recently in finance (Zhou et al. 2015; Sun et al. 2018). The feature vector fed into GBM is also the concatenation of features from order book and transaction data of two markets.

The hyper-parameters (Pedregosa et al. 2011) of GBM are selected by random search in the ranges: number of trees in {100, 200, . . . , 1000}, max number of features

used by individual trees in $\{0.1, 0.2, \ldots, 0.9, 1.0\}$, minimum number of samples of the leaf nodes in $\{2, 3, \ldots, 9\}$, maximum depth of individual trees in $\{4, 5, \ldots, 9\}$, and the learning rate in $\{0.005, 0.01, 0.025, 0.05\}$.

TME is implemented by TensorFlow.[8] The hyper-parameters tuned by the random search process are mainly the learning rate from the continues range $[0.0001, 0.001]$, batch size in the continues range from 10 to 300, and the l2 regularization term λ in the continues range from 0.1 to 5.0. The number of model parameter realizations for building the ensemble is set to 20. Beyond this number, we found no significant performance improvement. The window length is set to 9, when TME shows desirable performance in all experiments.

4.3 Results

In this section, we present the results on the predictions of volume in both markets at different time scales. Tables 2, 3, and 4 show the error metrics on the testing data for the two markets and the four models. We observe that in all cases the smallest RMSE is achieved by TME while the smallest MAE is achieved by GBM. Concerning NNLL, in Bitfinex TME outperforms the other models for 1-min and 5-min cases, while for Bitstamp markets econometric model have (slightly) lower values. Finally, the smallest IW is always achieved by TME.

The fact that GBM has superior performance on MAE but not RMSE might be related to the fact that GBM has been trained to minimize point prediction, while ARMA-GARCH, ARMAX-GARCH and TME were trained with a maximum likelihood objective, that handles better the larger deviations.

By comparing RMSE and MAE in both markets, we observe that for ARMA-GARCH model in Bitfinex, external features and extra information from Bitstamp are lowering MAE and RMSE errors (except on 1 min interval on Bitfinex market). In Bitstamp market, for ARMAX-GARCH model external features only help on 1-min interval prediction. This phenomenon could indicate the one-directional information relevance across two markets. However, due to the data source specific components and temporal adaptive weighting schema, our TME is able to yield more accurate prediction consistently, compared to ARMA-GARCH.

As for NNLL, similar pattern is observed in ARMA-GARCH family, i.e. additional features impair the NNLL performance instead, while TME retains comparable performance. More importantly, TME has much lower IW. Together with the lower RMSE, it implies that when TME predicts the mean closer to the observation, the predictive uncertainty is also lower. Remind that GBM does not provide probabilistic predictions. Overall we find that TME outperforms quite often the baseline benchmarks, by providing smaller RMSE and tighter intervals.

In order to disentangle the different contributions to TME predictions and, at the same time, understanding why choosing different predicting features at each time step is important, we present in Figs. 3 and 4 the predictive volume and interval of 5-min volume prediction in a sample time period of the testing data for the two markets. The prediction interval is obtained via the same method as IW, i.e. the quantiles

[8] https://github.com/tensorflow/tensorflow.

Table 2 Results of 1-min volume prediction

BITFINEX MARKET	RMSE ↓	MAE ↓	NNLL ↓	IW ↓
ARMA(5,5)-GARCH(1,1)	24.547	14.227	2.660	19.123
ARMAX(5,5)-GARCH(1,1)	24.629	14.189	2.664	19.051
GBM	21.026	**7.978**	NA	NA
TME	**20.142**	10.204	**2.654**	**17.436**
BITSTAMP MARKET	RMSE ↓	MAE ↓	NNLL ↓	IW ↓
ARMA(3,3)-GARCH(1,1)	14.587	7.688	**1.719**	8.637
ARMAX(3,3)-GARCH(1,1)	14.292	7.487	1.719	8.413
GBM	11.740	**3.515**	NA	NA
TME	**11.378**	4.299	1.720	**6.462**

The arrow symbols indicate the direction of the metrics for better models. GBM does not provide probabilistic output and thus NNLL and IW results are not available (NA). Results in bold indicate the best performance among models

Table 3 Results of 5-min volume prediction

BITFINEX MARKET	RMSE ↓	MAE ↓	NNLL ↓	IW ↓
ARMA(3,3)-GARCH(1,1)	64.999	39.909	4.642	80.014
ARMAX(3,3)-GARCH(1,1)	64.456	39.150	4.641	77.999
GBM	64.964	**32.888**	NA	NA
TME	**63.855**	39.527	**4.636**	**77.738**
BITSTAMP MARKET	RMSE ↓	MAE ↓	NNLL ↓	IW ↓
ARMA(5,4)-GARCH(1,1)	38.300	17.606	**3.732**	32.478
ARMAX(5,4)-GARCH(1,1)	40.273	18.887	3.766	34.821
GBM	39.196	**14.714**	NA	NA
TME	**38.223**	17.287	3.765	**31.087**

The arrow symbols indicate the direction of the metrics for better models. GBM does not provide probabilistic output and thus NNLL and IW results are not available (NA). Results in bold indicate the best performance among models

corresponding to 5% and 95% probability of the predictive distribution. For individual data sources, the prediction interval is solely based on the predictive distribution by the corresponding component model of the mixture.

Panel (a) shows how well the TME (pointwise and interval) predictions follows the actual data. The TME is able to quantify at each time step the contribution of each source to the target forecasting. In Panel(b) we show the dynamical contribution scores of the four sources. These scores are simply the average of the probability of the latent variable value corresponding to each data source, i.e. $z_t = s$. We notice that the relative contributions varies with time and we observe that the external order book source from the less liquid market (Bitstamp) does not contribute much to predictions. On the contrary, in Panel (b) of Fig. 4, where the data for Bitstamp are shown, external order

Table 4 Results of 10-min volume prediction

BITFINEX MARKET	RMSE ↓	MAE ↓	NNLL ↓	IW ↓
ARMA(3,2)-GARCH(1,1)	112.872	72.505	5.409	148.807
ARMAX(3,2)-GARCH(1,1)	111.987	71.149	**5.373**	**145.024**
GBM	110.197	**61.506**	NA	NA
TME	**109.878**	68.382	5.386	151.797
BITSTAMP MARKET	RMSE ↓	MAE ↓	NNLL ↓	IW ↓
ARMA(3,2)-GARCH(1,1)	66.486	31.942	**4.452**	60.285
ARMAX(3,2)-GARCH(1,1)	68.067	32.795	4.457	62.324
GBM	67.128	**27.719**	NA	NA
TME	**66.234**	31.460	4.507	**60.193**

The arrow symbols indicate the direction of the metrics for better models. GBM does not provide probabilistic output and thus NNLL and IW results are not available (NA). Results in bold indicate the best performance among models

book and external transaction features from the more liquid market (Bitfinex) play a more dominant role. Then, when we further look at the predictions by individual data sources in TME, see Panel(c)–(f),[9] what we observe is in line with the pattern of contribution scores in Panel (b). For instance, the component model in TME responsible for the data source from more liquid Bitfinex captures more uncertainty, thereby being given high contribution scores in the volatile period. The addition of the features from the other markets allows to shrink the uncertainty intervals. We have also repeated these experiments for 1-min and 10-min volume prediction. The results are collected in the figures and tables in the appendix section.

Finally, we discuss how well the different predictors perform conditionally to the volume size. To this end, we compute for each model, market, and time interval the RMSE and MAE conditional to the volume quartile. However, finding the suitable metrics to compare forecasting performance of a model across different quartiles is a subtle issue. Take for example a linear model for the log-volume. When considering it as a model for (linear) volume, it is clear that the additive noise in the log-volume becomes multiplicative in linear volume. Thus the RMSE conditional to volume becomes proportional to (or increasing as a power-law of) the volume and therefore one expects to see RMSE for high quartiles to be larger than the one for bottom quartiles. This is exactly what we observe in Table 5 for 5 min horizons (for 1 and 10 min, see the appendix) when looking at RMSE and MAE.

In order to take into account this statistical effect, essentially caused by the choice of modeling log-volume and presenting error metrics for linear volume, in Table 5 we show also *relative* error metrics, namely relative root mean squared error (RelRMSE) and mean absolute percentage error (MAPE) conditioned to volume quartile. (RelRMSE is defined as $\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\frac{v_t-\hat{v}_t}{v_t})^2}$), and MAPE is defined as $\frac{1}{T}\sum_{t=1}^{T}|\frac{v_t-\hat{v}_t}{v_t}|$.) In

[9] The log-normal distribution is asymmetric and it is easy to show that the mean value could fall out of the prediction interval corresponding to the probability range from 5% to 95%. A similar behavior is observed in the figures for TME whose conditional distribution is a mixture of log-normal distributions.

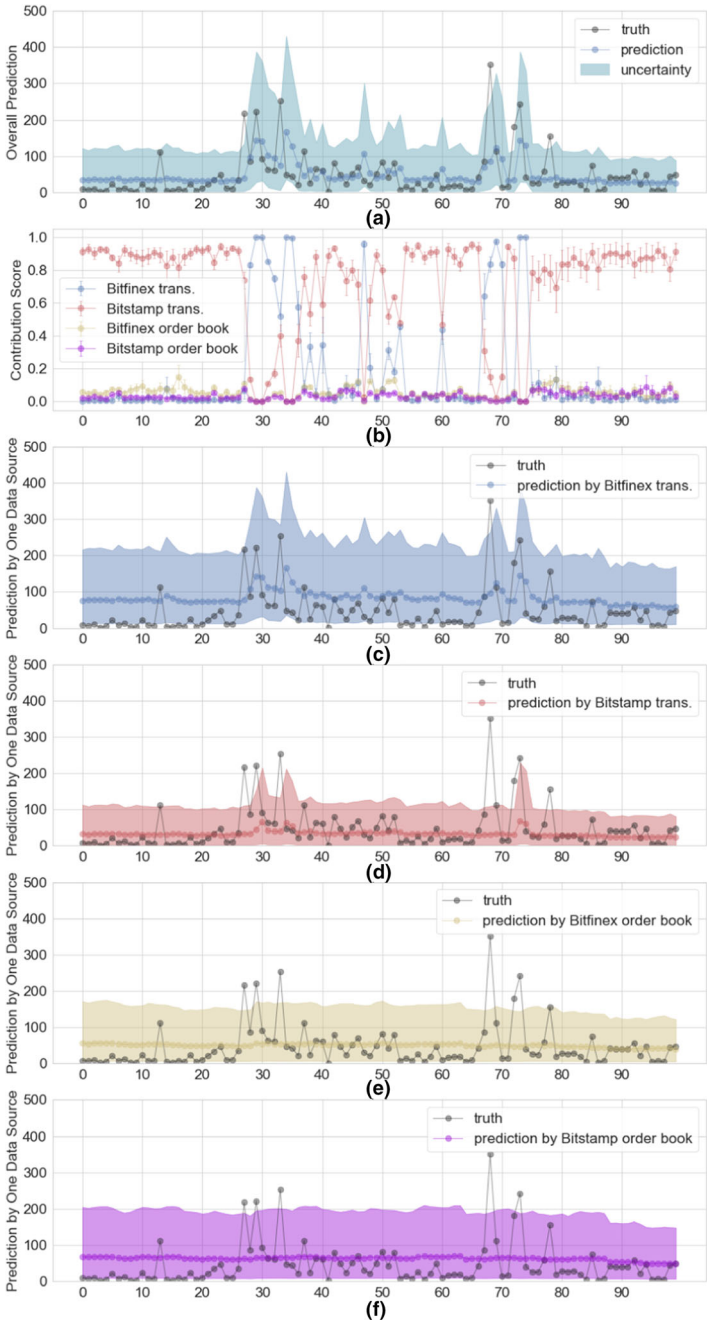

Fig. 3 Visualization of TME in a sample period of Bitfinex for 5-min volume predicting. Panel (**a**): Predictive mean and interval w.r.t. $5\% - 95\%$ probability. Panel (**b**): Data source contribution scores (i.e. average of latent variable probabilities) over time. Panel (**c**)–(**f**): Each data source's predictive mean and interval w.r.t. $5\% - 95\%$ probability. The color of each source's plot corresponds to that of the contribution score in Panel(b). (best viewed in colors)

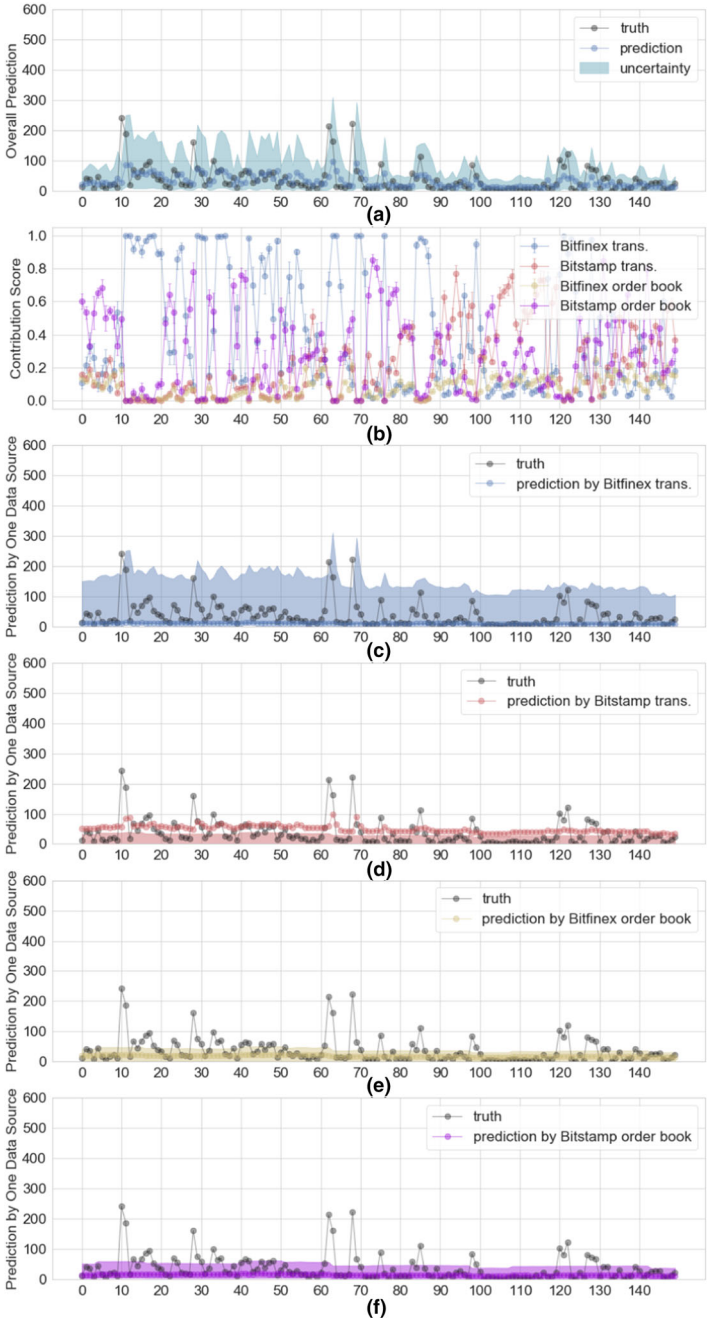

Fig. 4 Visualization of TME in a sample period of Bitstamp for 5-min volume predicting. Panel (**a**): Predictive mean and interval w.r.t. $5\% - 95\%$ probability. Panel (**b**): Data source contribution scores (i.e. average of latent variable probabilities) over time. Panel (**c**)–(**f**): Each data source's predictive mean and interval w.r.t. $5\% - 95\%$ probability. The color of each source's plot corresponds to that of the contribution score in Panel (b). (best viewed in colors)

Table 5 5-min volume prediction errors conditional on the quartile of the true volume values

BITFINEX MARKET	RMSE Q1	RMSE Q2	RMSE Q3	RMSE Q4
ARMA-GARCH	32.862	38.594	43.073	**111.660**
ARMAX-GARCH	31.939	36.887	41.005	112.020
GBM	**17.632**	**17.332**	**22.056**	125.604
TME	32.200	32.976	35.928	112.280
	RelRMSE Q1	RelRMSE Q2	RelRMSE Q3	RelRMSE Q4
ARMA-GARCH	19.417	2.769	1.299	0.609
ARMAX-GARCH	18.845	2.657	1.243	0.594
GBM	**12.672**	**1.268**	**0.615**	0.623
TME	26.235	2.378	1.098	**0.578**
	MAE Q1	MAE Q2	MAE Q3	MAE Q4
ARMA-GARCH	26.781	28.635	29.622	74.576
ARMAX-GARCH	26.145	27.625	28.190	74.612
GBM	**14.719**	**11.985**	**16.674**	88.122
TME	26.317	28.784	23.338	**74.484**
	MAPE Q1	MAPE Q2	MAPE Q3	MAPE Q4
ARMA-GARCH	9.506	2.003	0.871	0.497
ARMAX-GARCH	9.272	1.935	0.831	**0.493**
GBM	**5.717**	**0.846**	**0.460**	0.570
TME	11.424	1.823	0.690	0.503
BITSTAMP MARKET	RMSE Q1	RMSE Q2	RMSE Q3	RMSE Q4
ARMA-GARCH	13.465	15.304	16.199	**72.046**
ARMAX-GARCH	14.2	16.835	20.0009	74.8532
GBM	**6.432**	**6.679**	**8.732**	77.317
TME	13.617	14.492	14.583	73.778
	RelRMSE Q1	RelRMSE Q2	RelRMSE Q3	RelRMSE Q4
ARMA-GARCH	22.171	2.563	1.175	0.671
ARMAX-GARCH	24.259	2.784	1.429	0.889
GBM	**12.578**	**1.154**	**0.596**	0.658
TME	29.289	2.429	1.084	**0.586**
	MAE Q1	MAE Q2	MAE Q3	MAE Q4
ARMA-GARCH	10.668	11.648	11.020	37.094
ARMAX-GARCH	11.0917	12.275	12.1831	39.9903
GBM	**5.037**	**4.575**	**6.705**	42.506
TME	11.778	11.342	10.138	**36.827**
	MAPE Q1	MAPE Q2	MAPE Q3	MAPE Q4
ARMA-GARCH	11.049	1.910	0.792	0.523
ARMA-GARCH	11.656	2.007	0.873	0.603
GBM	**5.838**	**0.755**	**0.456**	0.599
TME	14.117	1.871	0.720	**0.493**

Measures: *RMSE Qx* root mean squared error conditioned on *x*th quantile, *RelRMSE Qx* relative root mean squared error conditioned on *x*th quantile, *MAE Qx* mean average error conditioned on *x*th quantile, *MAPE Qx* mean absolute percentage error conditioned on *x*th quantile

this case the scenario changes completely. First, across all models the relative error is smaller for top volume quartiles and larger for bottom quartiles. Especially relative errors for the lowest quartile are large, likely because small volumes at the denominator create large fluctuations. Second, and more important, TME outperforms mostly in Q4, while GBM is superior in Q1, Q2, and Q3 frequently. The difference between models in Q4 is somewhat smaller, while in Q1-Q3 the out-performance of machine learning methods is up to a factor 2 with respect to econometric methods. Thus also when considering large volumes and considering relative errors, TME and GBM provide more accurate predictions.

5 Conclusion and discussion

In this paper, we analyzed the problem of predicting trading volume and its uncertainty in cryptocurrency exchange markets. The main innovations proposed in this paper are (i) the use of transaction and order book data from different markets and (ii) the use of TME, a class of models able to identify at each time step the set of data locally more useful in predictions.

By investigating data from BTC/USD exchange markets, we found that time series models of the ARMA-GARCH family do provide fair basic predictions for volume and its uncertainty, but when external data (e.g. from order book and/or from other markets) are added, the prediction performance does not improve significantly. Our analysis suggests that this might be due to the fact that the contribution of this data to the prediction could be not constant over time, but depending on the "market state". The temporal mixture ensemble model is designed precisely to account for such a variability. Indeed we find that this method outperforms time series models both in point and in interval predictions of trading volume. Moreover, especially when compared to other machine learning methods, the temporal mixture approach is significantly more interpretable, allowing the inference of the dynamical contributions from different data sources as a core part of the learning procedure. This has important potential implications for decision making in economics and finance.

Also when conditioning to volume quartile, TME and GBM outperform econometric methods especially in the first three quartiles. For large volumes, likely due to the presence of unexpected bursts of volume which are very challenging to forecast, the performances of the methods are more comparable. However by using relative RMSE and MAPE the forecasting errors for large volumes are small.

Finally, although the method has been proposed and tested for cryptocurrency volume in two specific exchanges, we argue that it can be successfully applied (in future work) to other cryptocurrencies and to more traditional financial assets.

Acknowledgements This work has been funded by the European Program scheme 'INFRAIA-01-2018-2019: Research and Innovation action', grant agreement #871042 'SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics'. We would like also to thank two anonymous referees for their very detailed reports. Their suggestions contributed significantly to the improvement of the paper.

Funding Open access funding provided by Scuola Normale Superiore within the CRUI-CARE Agreement.

Appendix

ARMAX-GARCH

As mentioned, our benchmarks belong to the ARMAX-GARCH class with external regressors. More specifically, the volume process is modelled with the following:

$$\Phi(L)(\ln(y_t) - \mu_t) = \Theta(L)\epsilon_t, \tag{21}$$

where $\Phi(L)$, $\Theta(L)$ denote polynomials of the lag operator L. The time varying mean μ_t is modeled as

$$\mu_t = \mu + \sum_{s=1}^{S} \sum_{j=1}^{d_s} \psi_{s,j} x_{s,t-1}(j) \tag{22}$$

where $x_{s,t-1}(j)$ denotes the jth feature from external feature vector $\mathbf{x}_{s,t-1}$ at time $t-1$ from source s. The total number of sources $S = 4$, which includes transactions and limit order book data of the two markets. The parameters for ARMAX-GARCH, were inferred jointly (Ghalanos et al. 2019). Since the variance of volume might exhibit time clustering, we assume that the residuals ϵ_t are modelled by a GARCH process (Bollerslev 1986; Brownlees et al. 2010; Satish et al. 2014; Chen et al. 2016):

$$\epsilon_t = \sigma_t e_t \qquad e_t \sim \mathcal{N}(0, 1) \tag{23}$$

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{24}$$

where ω is a constant term.

Gradient boosting

In the following, we summarize how GBM is the used in the context of volume predicting. For more details of GBM, we suggest referring to Friedman (2001).

At time t, the target label is $u_t = \ln y_{t+1}$, is the logarithm of deseasonalized volume at next time segment. Gradient boosting approximates the target variable u_t with a function $F(\mathbf{x}_t)$ that has the following additive expansion (similar to other functional approximation methods like radial basis functions, neural networks, wavelets, etc.):

$$\hat{u_t} = F(\mathbf{x}_t) = \sum_{m=0}^{M} \beta_m h(\mathbf{x}_t; \mathbf{a}_m), \tag{25}$$

where $\mathbf{x}_t$ denotes the feature vector, that is constructed as a concatenation from different sources[10] $\mathbf{x}_t = (\mathbf{x}_{s=1,(-h,t)}, \mathbf{x}_{s=2,(-h,t)}, \mathbf{x}_{s=3,(-h,t)}, \mathbf{x}_{s=4,(-h,t)})$.

For a given training sample $\{u_t, \mathbf{x}_t\}_{t=1}^{T}$, our goal is to find a function $F^*(\mathbf{x})$ such that the expected value of loss function $\frac{1}{2}(u - F(\mathbf{x}))^2$ (squared loss) is minimized over the joint distribution of $\{u, \mathbf{x}\}$

$$F^*(\mathbf{x}) = \arg\min_{F(x)} \mathbb{E}_{u,\mathbf{x}}(u - F(\mathbf{x}))^2. \tag{26}$$

Under the additive expansion $F(\mathbf{x}) = \sum_{m=0}^{M} \beta_m h(\mathbf{x}; \mathbf{a}_m)$ with parameterized functions $h(\mathbf{x}; \mathbf{a}_m)$, we proceed by making the initial guess $F_0(\mathbf{x}) = \arg\min_c \sum_{t=1}^{T}(u_t - c)^2$ and then parameters are jointly fit in a forward incremental way $m = 1, \ldots, M$:

$$(\beta_m, \mathbf{a}_m) = \arg\min_{\beta,\mathbf{a}} \sum_{t=1}^{T}(u_t - (F_{m-1}(\mathbf{x}_t) + \beta h(\mathbf{x}_t; \mathbf{a})))^2 \tag{27}$$

and

$$F_m(\mathbf{x}_t) = F_{m-1}(\mathbf{x}_t) + \beta_m h(\mathbf{x}_t; \mathbf{a}_m). \tag{28}$$

First, the function $h(\mathbf{x}_t; \mathbf{a})$ is fit by least-squares to the pseudo-residuals $\widetilde{u}_{t,m}$

$$\mathbf{a}_m = \arg\min_{\mathbf{a},\rho} \sum_{t=1}^{T}[\widetilde{u}_{t,m} - \rho h(\mathbf{x}_t; \mathbf{a})]^2, \tag{29}$$

which at stage m is a residual $\widetilde{u}_{t,m} = (u_t - F_{m-1}(\mathbf{x}_t))$. Pseudo-residual at arbitrary stage is defined as

$$\widetilde{u}_{t,m} = -\left[\frac{\partial \frac{1}{2}(u_t - F(\mathbf{x}_t))^2}{\partial F(\mathbf{x}_t)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}. \tag{30}$$

The parameter ρ acts as the optimal learning rate in the steepest-descent step, for more

[10] Note, that we have omitted the transpose operators in the next line, as the concatenation is simple operation and to avoid confusion with index of time.

details check (Friedman 2001). Now, we just find the coefficient β_m for the expansion as

$$\beta_m = \arg\min_\beta \sum_{t=1}^{T} \frac{1}{2}(u_t - (F_{m-1}(\mathbf{x}_t) + \beta h(\mathbf{x}_t; \mathbf{a}_m)))^2. \tag{31}$$

Each base learner $h(\mathbf{x}_t; \mathbf{a}_m)$, parameterized with $\mathbf{a}_m$ partitions the feature space $\mathbf{x}_t \in \mathbf{X}$ into L_m-disjoint regions $\{R_{l,m}\}_1^{L_m}$ and predicts a separate constant value in each:

$$h(\mathbf{x}_t; \{R_{l,m}\}_1^{L_m}) = \sum_{l=1}^{L_m} \bar{u}_{l,m} \mathbf{1}(\mathbf{x}_t \in R_{l,m}), \tag{32}$$

where $\bar{u}_{l,m}$ is the mean value of pseudo-residual (Eq. 30) in each region $R_{l,m}$

$$\bar{u}_{l,m} = \frac{\sum_{t=1}^{T} \widetilde{u}_{t,m} \mathbf{1}[\mathbf{x}_t \in R_{l,m}]}{\sum_{t=1}^{T} \mathbf{1}[\mathbf{x}_t \in R_{l,m}]}. \tag{33}$$

We have used the GBM implementation from Scikit-learn library (Pedregosa et al. 2011) for all our experiments. Furthermore, note that different variants of tree boosting have been empirically proven to be state-of-the-art methods in predictive tasks across different machine learning challenges (Bentéjac et al. 2020; Chen and Guestrin 2016; Lu and Mazumder 2020; Taieb and Hyndman 2014; Gulin et al. 2011) and more recently in finance (Zhou et al. 2015; Sun et al. 2018). Note, that although by default GBM does not provide confidence intervals, we are not claiming that is not possible to construct confidence intervals for GBM. However, this adaptation falls out-of-scope of our current work.

SGD-based model ensemble

In stochastic gradient descent (SGD) based optimization, stochasticity comes from two places:

- SGD trajectory. The iterates $\{\Theta(0), \ldots, \Theta(i)\}$ forms a exploratory trajectory, as $\Theta(i)$ is updated by randomly data sample $\mathcal{D}_i$. Recent works (Mandt et al. 2017; Gur-Ari et al. 2018) studied the connection of trajectory iterates to an approximate Markov chain Monte Carlo sampler by analyzing the dynamics of SGD.
- Model initialization. Different initialization of model parameters, i.e. $\Theta(0)$, leads to distinct trajectories. It has been shown that ensembles of independently initialized and trained models empirically often provide comparable performance in prediction and uncertainty quantification w.r.t. sampling and variational inference

based methods, even though it does not apply conventional Bayesian grounding (Lakshminarayanan et al. 2017; Snoek et al. 2019).

In this paper, we make a hybrid approach, that uses both sources of stochasticity to obtain parameter realizations $\{\Theta_m\}$ as follows:

$$\{\Theta_m\} \triangleq \bigcup_j \{\Theta^j(i), \ldots, \Theta^j(I)\} \tag{34}$$

Equation 34 indicates that from each independently trained SGD trajectory (indexed by j), we skip the beginning few epochs as a "burn-in" step. We choose the remaining as samples from this trajectory. Then, we further take the union of samples from independent trajectories as the samples used by the inference in Sect. 3.5.

In our experiments, we use Adam optimization, a variant of SGD, which has been widely used in machine learning (Kingma and Ba 2015). We found that 5 to 25 independent training processes can give rise to decently accurate and calibrated forecasting. Moreover, by parallel computing on GPU, we perform each training process in parallel without loss of efficiency.

Residual diagnostics of ARMA-GARCH models

In Figs. 5 and 6, we show auto-correlation function of the residuals for Bitfinex and Bitstamp market, respectively. In all cases of ARMA(p, q)-GARCH(1,1) models, parameters p, q were fitted to Akaike Information Criterion. For 1 min, 5 min and 10 min target the majority of residuals ACF are within significance area and only a small number falls very close to the significance area, which suggests that the residuals are almost uncorrelated and models well specified.

Disentangling TME contributions on 1- and 10-min intervals

We report in Figs. 7 and 8 the contributions to TME prediction in the two markets when the sampling interval is 1 min, while in Figs. 9 and 10 the same figures for 10 min intervals.

Volume predictions conditional to volume quartile

Tables 6 and 7 show the performance of the forecast conditioned to volume quartile for the four models and the two markets. As in Table 5 in the main text, we present both absolute metrics (RMSE and MAE) and relative ones (relative MSE and MAPE).

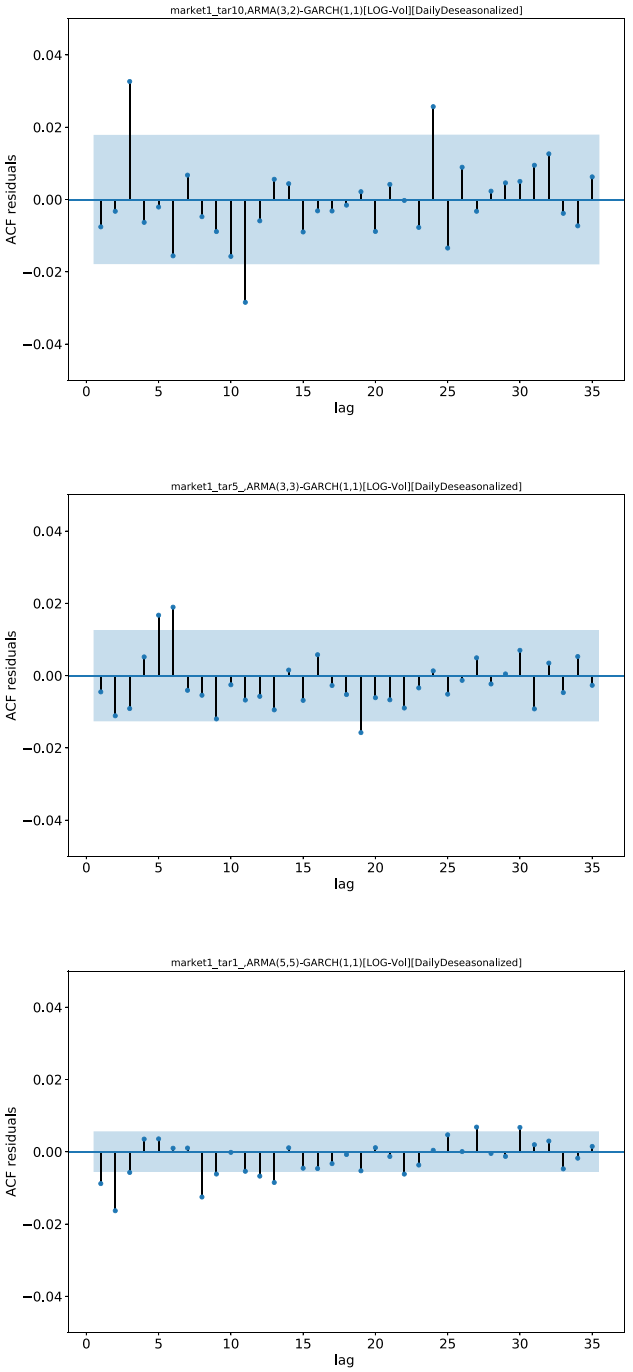

Fig. 5 **Residual diagnostics** for Bitfinex market: **First row** ACF of ARMA(3,2)-GARCH(1,1) model residuals on log-volume 10 min. **Second row** ACF of ARMA(3,3)-GARCH(1,1) model residuals on log-volume 5 min. **Third row** ACF of ARMA(5,5)-GARCH(1,1) model residuals on log-volume 1 min

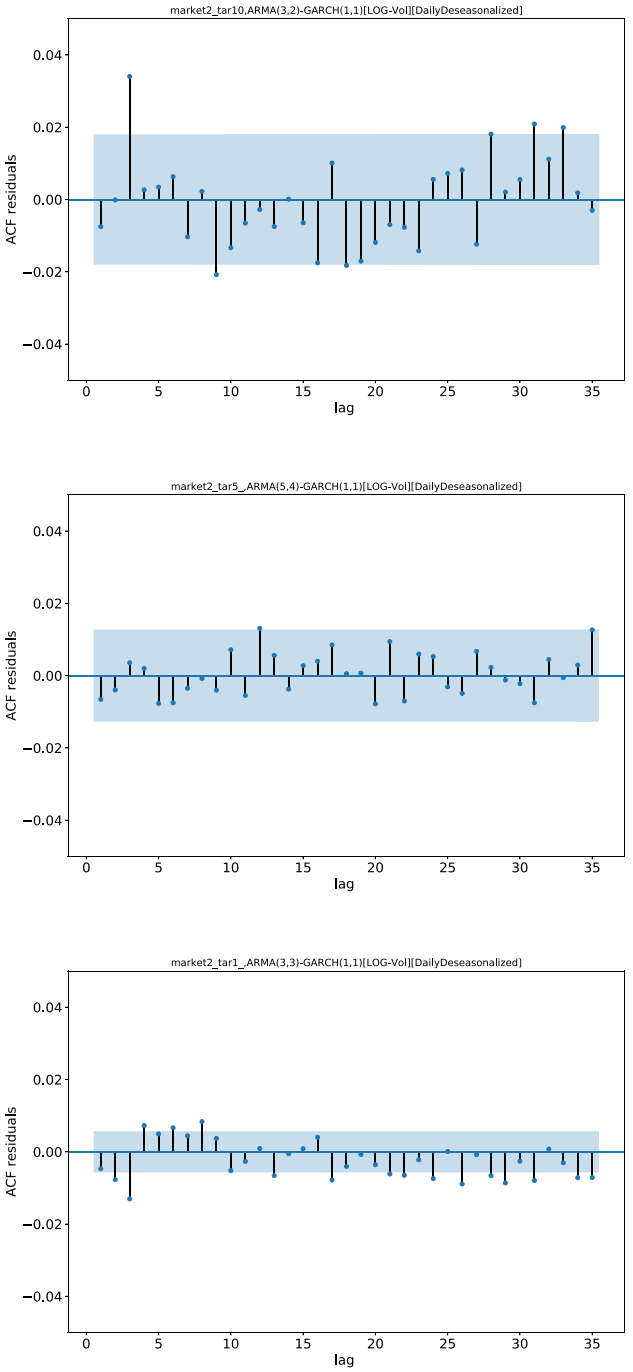

Fig. 6 **Residual diagnostics** for Bitstamp market: **First row** ACF of ARMA(3,2)-GARCH(1,1) model residuals on log-volume 10 min. **Second row** ACF of ARMA(5,4)-GARCH(1,1) model residuals on log-volume 5 min. **Third row** ACF of ARMA(3,3)-GARCH(1,1) model residuals on log-volume 1 min

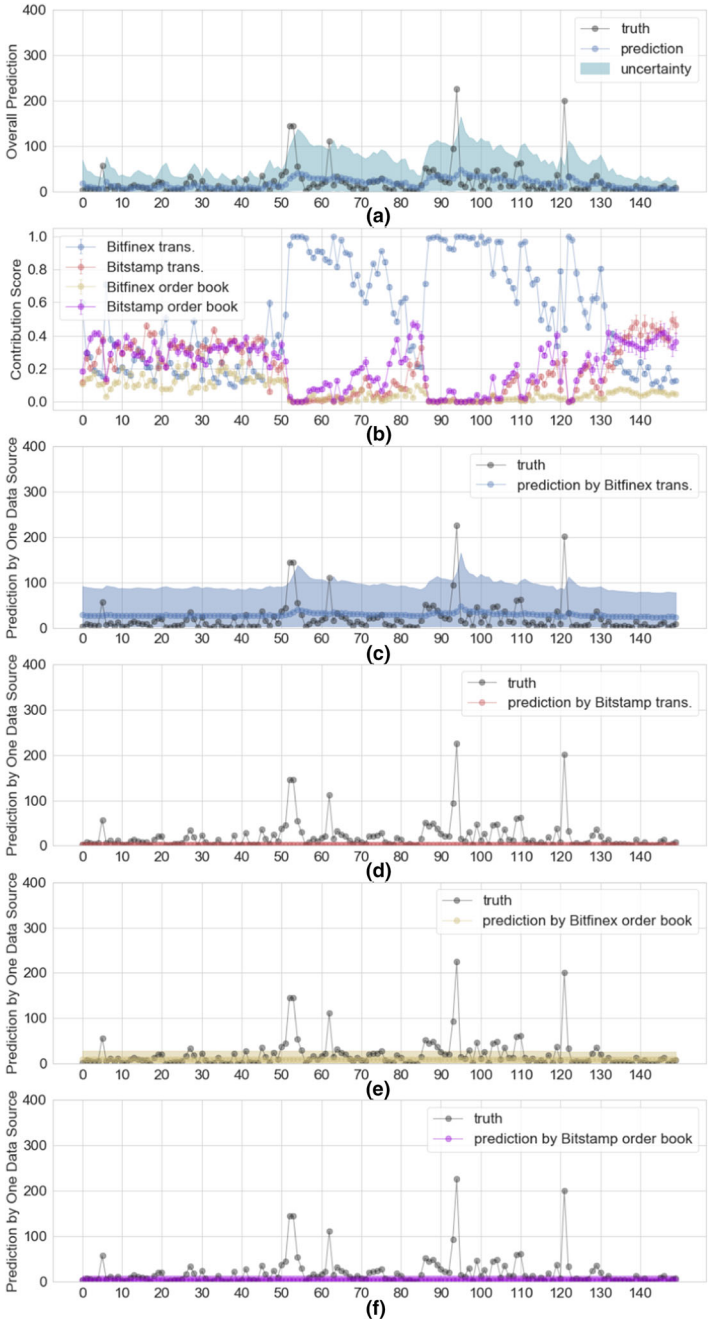

Fig. 7 Visualization of TME in a sample period of Bitfinex for 1-min volume predicting. Panel (**a**): Predictive mean and interval w.r.t. 5% − −95% probability. Panel (**b**): Data source contribution scores (i.e. average of latent variable probabilities) over time. Panel (**c**)–(**f**): Each data source's predictive mean and interval w.r.t. 5%–95% probability. The color of each source's plot corresponds to that of the contribution score in Panel (**b**). (best viewed in colors)

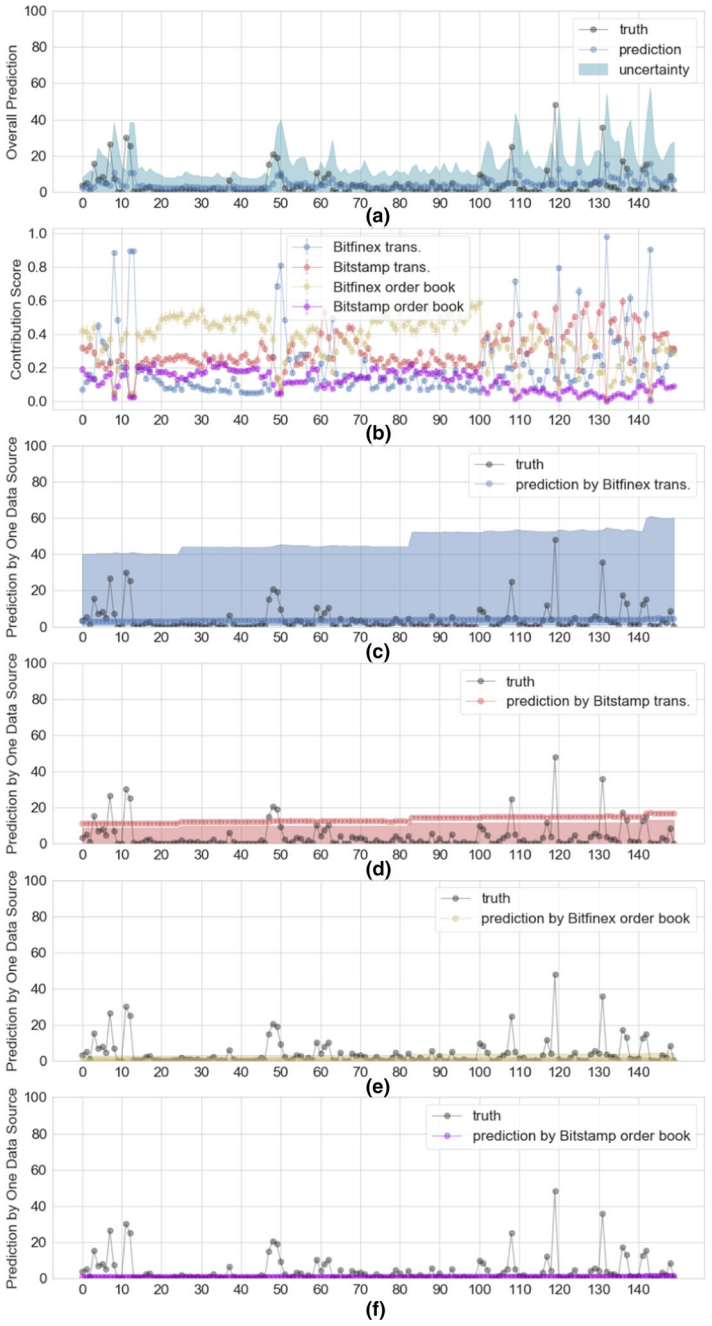

Fig. 8 Visualization of TME in a sample period of Bitstamp for 1-min volume predicting. Panel (**a**): Predictive mean and interval w.r.t. 5%–95% probability. Panel (**b**): Data source contribution scores (i.e. average of latent variable probabilities) over time. Panel (**c**)–(**f**): Each data source's predictive mean and interval w.r.t. 5%–95% probability. The color of each source's plot corresponds to that of the contribution score in Panel(b). (best viewed in colors)

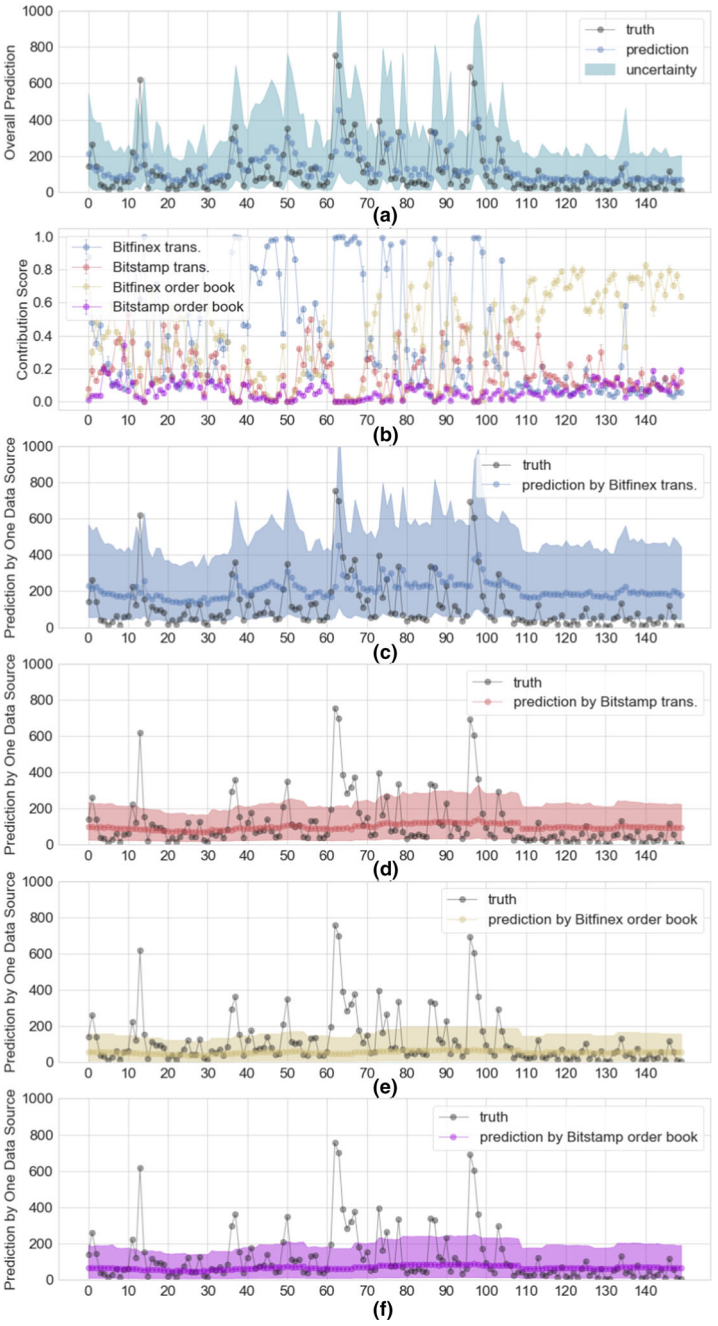

Fig. 9 Visualization of TME in a sample period of Bitfinex for 10-min volume predicting. Panel (**a**): Predictive mean and interval w.r.t. 5%–95% probability. Panel (**b**): Data source contribution scores (i.e. average of latent variable probabilities) over time. Panel (**c**)–(**f**): Each data source's predictive mean and interval w.r.t. 5%–95% probability. The color of each source's plot corresponds to that of the contribution score in Panel (**b**). (best viewed in colors)

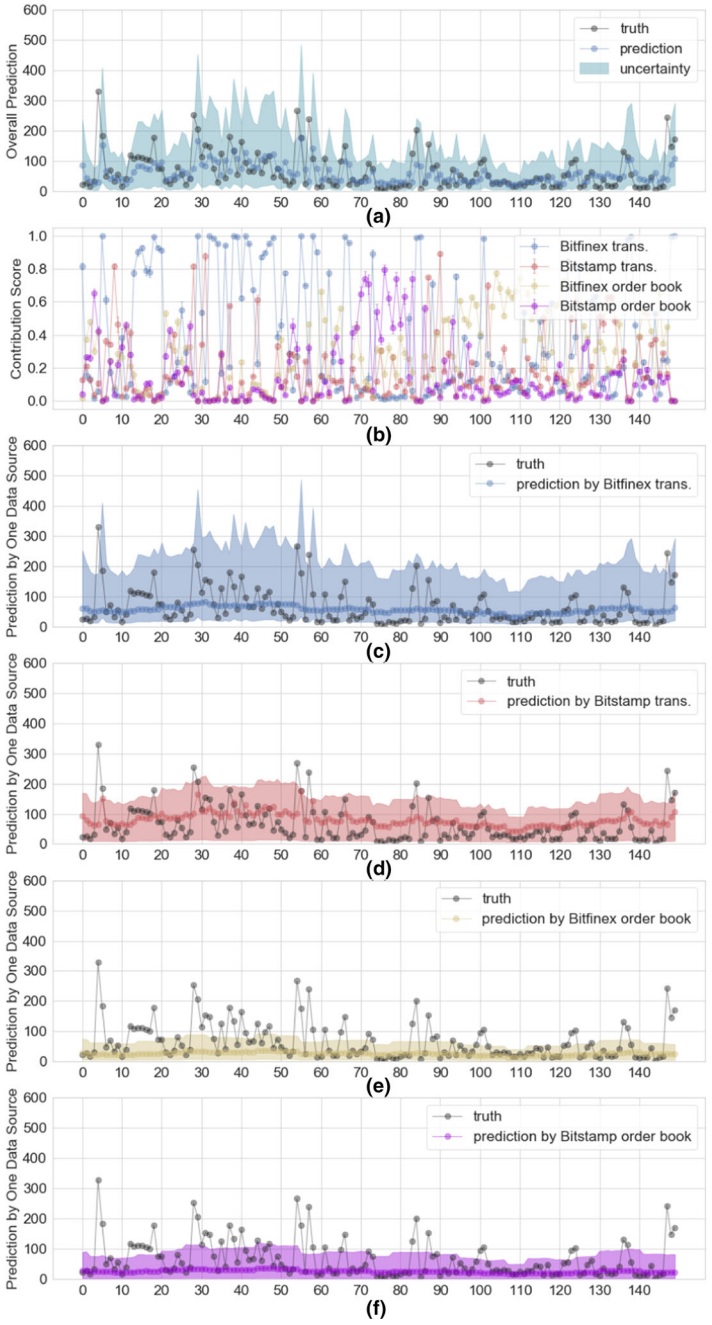

Fig. 10 Visualization of TME in a sample period of Bitstamp for 10-min volume predicting. Panel (**a**): Predictive mean and interval w.r.t. 5%–95% probability. Panel (**b**): Data source contribution scores (i.e. average of latent variable probabilities) over time. Panel (**c**)–(**f**): Each data source's predictive mean and interval w.r.t. 5%–95% probability. The color of each source's plot corresponds to that of the contribution score in Panel (**b**). (best viewed in colors)

Table 6 1-min volume prediction errors conditional on the quantile of the true volume values

BITFINEX MARKET	RMSE Q1	RMSE Q2	RMSE Q3	RMSE Q4
ARMA-GARCH	13.618	17.314	20.287	38.903
ARMAX-GARCH	13.601	17.350	20.206	39.141
GBM	**2.159**	**2.907**	**4.323**	41.671
TME	8.700	10.032	10.154	**36.916**
	RelRMSE Q1	RelRMSE Q2	RelRMSE Q3	RelRMSE Q4
ARMA-GARCH	93998.678	16.755	4.992	1.427
ARMAX-GARCH	96451.132	16.508	4.973	1.425
GBM	**22735.646**	**2.764**	**0.968**	0.754
TME	26968.219	9.269	2.539	**0.737**
	MAE Q1	MAE Q2	MAE Q3	MAE Q4
ARMA-GARCH	9.173	11.300	12.659	23.777
ARMAX-GARCH	9.158	11.249	12.556	23.791
GBM	**1.540**	**1.665**	**3.054**	25.652
TME	8.225	8.620	7.252	**20.629**
	MAPE Q1	MAPE Q2	MAPE Q3	MAPE Q4
ARMA-GARCH	1319.606	10.151	2.989	0.858
ARMAX-GARCH	1344.985	10.097	2.966	0.855
GBM	**296.446**	**1.502**	**0.646**	0.707
TME	318.682	7.372	1.692	**0.568**
BITSTAMP MARKET	RMSE Q1	RMSE Q2	RMSE Q3	RMSE Q4
ARMA-GARCH	8.2704	8.9885	11.0034	24.1002
ARMAX-GARCH	7.9412	8.6546	10.5348	23.8335
GBM	**0.929**	**1.085**	**1.721**	23.373
TME	3.047	3.426	3.540	**22.005**
	RelRMSE Q1	RelRMSE Q2	RelRMSE Q3	RelRMSE Q4
ARMA-GARCH	16842.969	27.709	6.423	1.974
ARMAX-GARCH	16501.630	26.679	6.135	1.923
GBM	**2082.545**	**3.190**	**0.845**	0.786
TME	7279.497	10.951	2.107	**0.696**
	MAE Q1	MAE Q2	MAE Q3	MAE Q4
ARMA-GARCH	5.400	6.125	7.003	12.223
ARMAX-GARCH	5.239	5.944	6.778	11.985
GBM	**0.599**	**0.605**	**1.316**	11.540
TME	2.664	2.780	2.259	**9.615**
	MAPE Q1	MAPE Q2	MAPE Q3	MAPE Q4
ARMA-GARCH	1043.922	17.211	3.891	1.097
ARMAX-GARCH	1018.872	16.694	3.767	1.066
GBM	**118.256**	**1.720**	**0.632**	0.741
TME	500.100	8.035	1.301	**0.571**

Measures: *RMSE Qx* root mean squared error conditioned on *x*th quantile, *RelRMSE Qx* relative root mean squared error conditioned on *x*th quantile, *MAE Qx* mean average error conditioned on *x*th quantile, *MAPE Qx* mean absolute percentage error conditioned on *x*th quantile

Table 7 10-min volume prediction errors conditional on the quartile of the true volume values

BITFINEX MARKET	RMSE Q1	RMSE Q2	RMSE Q3	RMSE Q4
ARMA-GARCH	61.494	70.540	74.188	191.332
ARMAX-GARCH	59.438	67.474	71.237	192.155
GBM	**39.380**	**39.576**	**46.774**	207.957
TME	57.252	64.705	67.354	**185.136**
	RelRMSE Q1	RelRMSE Q2	RelRMSE Q3	RelRMSE Q4
ARMA-GARCH	9.262	1.998	1.037	0.544
ARMAX-GARCH	8.822	1.911	0.990	**0.538**
GBM	**6.154**	**1.136**	**0.587**	0.571
TME	11.434	1.878	0.945	0.543
	MAE Q1	MAE Q2	MAE Q3	MAE Q4
ARMA-GARCH	50.542	54.148	53.169	131.891
ARMAX-GARCH	48.985	52.116	51.257	132.081
GBM	**33.788**	**28.760**	**34.892**	148.550
TME	46.680	48.797	43.890	**128.578**
	MAPE Q1	MAPE Q2	MAPE Q3	MAPE Q4
ARMA-GARCH	5.109	1.503	0.705	0.462
ARMAX-GARCH	4.938	1.448	0.678	**0.461**
GBM	**3.561**	**0.804**	**0.434**	0.518
TME	6.739	1.499	0.590	0.465
BITSTAMP MARKET	RMSE Q1	RMSE Q2	RMSE Q3	RMSE Q4
ARMA-GARCH	25.000	27.672	27.541	124.527
ARMAX-GARCH	24.963	29.748	29.545	126.985
GBM	**14.355**	**14.700**	**16.749**	131.497
TME	25.469	27.221	24.905	**123.804**
	RelRMSE Q1	RelRMSE Q2	RelRMSE Q3	RelRMSE Q4
ARMA-GARCH	7.210	1.860	0.949	0.619
ARMAX-GARCH	6.926	1.974	0.984	0.739
GBM	**5.638**	**0.993**	**0.541**	0.625
TME	12.308	1.838	0.858	**0.574**
	MAE Q1	MAE Q2	MAE Q3	MAE Q4
ARMA-GARCH	20.014	21.343	19.398	66.899
ARMAX-GARCH	19.633	21.685	19.718	70.021
GBM	**11.622**	**10.490**	**12.965**	75.674
TME	21.233	20.034	17.787	**64.918**
	MAPE Q1	MAPE Q2	MAPE Q3	MAPE Q4
ARMA-GARCH	4.866	1.410	0.646	0.487
ARMAX-GARCH	4.716	1.429	0.653	0.528
GBM	**3.341**	**0.695**	**0.414**	0.555
TME	7.205	1.457	0.550	**0.478**

Measures: *RMSE Qx* root mean squared error conditioned on *x*th quantile, *RelRMSE Qx* relative root mean squared error conditioned on *x*th quantile, *MAE Qx* mean average error conditioned on *x*th quantile, *MAPE Qx* mean absolute percentage error conditioned on *x*th quantile

References

Alessandretti, L., ElBahrawy, A., Aiello, L.M., and Baronchelli, A.: "Machine learning the cryptocurrency market," arXiv preprint arXiv:1805.08550, (2018)

Amjad, M., Shah, D.: Trading bitcoin and online time series prediction, in NIPS. Time Series Workshop **2017**, 1–15 (2016)

Andersen, T.G.: Return volatility and trading volume: An information flow interpretation of stochastic volatility. J. Finance **51**(1), 169–204 (1996)

Andersen, T.G., Bollerslev, T.: Intraday periodicity and volatility persistence in financial markets. J. Empir. Finance **4**(2–3), 115–158 (1997)

Antulov-Fantulin, N., Tolic, D., Piskorec, M., Ce, Z., Vodenska, I.: "Inferring short-term volatility indicators from the bitcoin blockchain," In: International Conference on Complex Networks and their Applications. Springer, pp. 508–520 (2018)

Balcilar, M., Bouri, E., Gupta, R., Roubaud, D.: Can volume predict bitcoin returns and volatility? a quantiles-based approach. Econ. Model. **64**, 74–81 (2017)

Barzykin, A., Lillo, F.: "Optimal vwap execution under transient price impact," arXiv preprint arXiv:1901.02327, (2019)

Baumöhl, E.: Are cryptocurrencies connected to forex? a quantile cross-spectral approach. Finance Res. Lett. **29**, 363–372 (2019)

Bauwens, L., Galli, F., Giot, P.: Moments of the log-acd model. Quant. Qualitative Anal. Soc. Sci. **2**, 1–28 (2008)

Bazzani, L., Larochelle, H., Torresani, L.: "Recurrent mixture density network for spatiotemporal visual attention," arXiv preprint arXiv:1603.08199 (2016)

Beck, J., Huang, R., Lindner, D., Guo, T., Ce, Z., Helbing, D., and Antulov-Fantulin, N.: "Sensing social media signals for cryptocurrency news," in Companion Proceedings of The 2019 World Wide Web Conference, pp. 1051–1054 (2019)

Bentéjac, C., Csörgő, A., Martinez-Munoz, G.: A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev. **54** 1–31 (2020)

Bialkowski, J., Darolles, S., Le Fol, G.: Improving vwap strategies: A dynamic volume approach. J. Banking Finance **32**(9), 1709–1722 (2008)

Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. J. Econom. **31**(3), 307–327 (1986)

Bollerslev, T., Ghysels, E.: Periodic autoregressive conditional heteroscedasticity. J. Business Econom. Stat. **14**(2), 139–151 (1996)

Bolt, W., Van Oordt, M.R.C.: On the value of virtual currencies. J Money Credit Bank **52**(4), 835–862 (2020)

Bos, J.W., Halderman, J.A., Heninger, N., Moore, J., Naehrig, M., Wustrow, E.: "Elliptic curve cryptography in practice," In: International Conference on Financial Cryptography and Data Security. Springer, pp. 157–175 (2014)

Brownlees, C.T., Cipollini, F., Gallo, G.M.: Intra-daily volume modeling and prediction for algorithmic trading. J. Financial Econom. **9**(3), 489–518 (2010)

Calvori, F., Cipollini, F., Gallo, G.M.: "Go with the flow: A gas model for predicting intra-daily volume shares," Available at SSRN 2363483 (2013)

Chaboud, A.P., Chiquoine, B., Hjalmarsson, E., Vega, C.: Rise of the machines: Algorithmic trading in the foreign exchange market. J. Finance **69**(5), 2045–2084 (2014)

Cheah, E.-T., Fry, J.: Speculative bubbles in bitcoin markets? an empirical investigation into the fundamental value of bitcoin. Econom. Lett. **130**, 32–36 (2015)

Chen, R., Feng, Y., Palomar, D.: "Forecasting intraday trading volume: A kalman filter approach," Available at SSRN 3101695, (2016)

Chen, T., Guestrin, C.:"Xgboost: A scalable tree boosting system," In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

Chu, J., Nadarajah, S., Chan, S.: Statistical analysis of the exchange rate of bitcoin. PLoS ONE **10**(7), 1–27 (2015)

Ciaian, P., Rajcaniova, M., Kancs, D.: The economics of bitcoin price formation. Appl. Econom. **48**, 1799–1815 (2016)

Cohen, A.C., Whitten, B.J.: Estimation in the three-parameter lognormal distribution. J. Am. Stat. Assoc. **75**(370), 399–404 (1980)

Donier, J., Bouchaud, J.-P.: Why do markets crash? bitcoin data offers unprecedented insights. PLoS ONE **10**, 1–11 (2015)
ElBahrawy, A., Alessandretti, L., Kandler, A., Pastor-Satorras, R., Baronchelli, A.: Evolutionary dynamics of the cryptocurrency market. Royal Soc. Open Sci. **4**(11), 170623 (2017)
Engle, R.: New frontiers for arch models. J. Appl. Econom. **17**(5), 425–446 (2002)
Engle, R.F., Sokalska, M.E.: Forecasting intraday volatility in the us equity market multiplicative component garch. J. Financial Econom. **10**(1), 54–83 (2012)
Frei, C., Westray, N.: Optimal execution of a vwap order: a stochastic control approach. Math. Finance **25**(3), 612–639 (2015)
Friedman, J.H.: "Greedy function approximation: a gradient boosting machine," Ann. Stat. **29**, 1189–1232 (2001)
Garcia, D., Schweitzer, F.: Social signals and algorithmic trading of bitcoin. Royal Society Open Science **2**(9), 150288 (2015)
Gerlach, J.-C., Demos, G., Sornette, D.: Dissection of bitcoin's multiscale bubble history from January 2012 to February 2018. Royal Soc. Open Sci. **6**(7), 180643 (2019)
Ghalanos, A., Ghalanos, M.A., Rcpp, L.: "Package 'rugarch'," (2019)
Gould, M.D., Porter, M.A., Williams, S., McDonald, M., Fenn, D.J., Howison, S.D.: Limit order books. Quant. Finance **13**(11), 1709–1742 (2013)
Gulin, A., Kuralenok, I., Pavlov, D.: "Winning the transfer learning track of yahoo!'s learning to rank challenge with yetirank," In: Proceedings of the Learning to Rank Challenge, pp. 63–76 (2011)
Guo, T., Bifet, A., and Antulov-Fantulin, N.:"Bitcoin volatility forecasting with a glimpse into buy and sell orders," In: 2018 IEEE International Conference on Data Mining (ICDM). IEEE, pp. 989–994 (2018)
Guo, T., Lin, T., Antulov-Fantulin, N.: "Exploring interpretable lstm neural networks over multi-variable data," In: International Conference on Machine Learning, pp. 2494–2504, (2019)
Gur-Ari, G., Roberts, D. A., Dyer, E.: "Gradient descent happens in a tiny subspace," arXiv preprint arXiv:1812.04754 (2018)
Hendershott, T., Jones, C., Menkveld, A.: Does algorithmic trading improve liquidity? J. Finance **66**(1), 1–33 (2011)
Hougan, M., Kim, H., Lerner, M., Management, B.A.: "Economic and non-economic trading in bitcoin: Exploring the real spot market for the world's first digital commodity," Bitwise Asset Management, (2019)
Jakobsson, M., Juels, A.: "Proofs of work and bread pudding protocols," In: Preneel, B. (ed.) Secure Information Networks. Springer, pp. 258–272 (1999)
Jang, H., Lee, J.: An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. IEEE Access **6**, 5427–5437 (2018)
Katsiampa, P.: Volatility estimation for bitcoin: A comparison of GARCH models. Econom. Lett. **158**, 3–6 (2017)
Kawakatsu, H.: Direct multiperiod forecasting for algorithmic trading. J. Forecasting **37**(1), 83–101 (2018)
Kingma, D.P., Ba, J.: "Adam: A method for stochastic optimization," In: International Conference on Learning Representations (2015)
Kondor, D., Csabai, I., Szule, J., Posfai, M., Vattay, G.: Inferring the interplay between network structure and market effects in bitcoin. New J. Phys. **16**, 125003 (2014)
Kurle, R., Günnemann, S., van der Smagt, P.: "Multi-source neural variational inference," In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 4114–4121 (2019)
Lakshminarayanan, B., Pritzel, A., Blundell, C.: "Simple and scalable predictive uncertainty estimation using deep ensembles," In: Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, pp. 6402–6413 (2017)
Lu, H., Mazumder, R.: Randomized gradient boosting machine. SIAM J. Optim. **30**(4), 2780–2808 (2020)
MacKay, D.J., Mac Kay, D.J.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge (2003)
Maddox, W.J., Izmailov, P., Garipov, T., Vetrov, D. P., Wilson, A. G.: "A simple baseline for bayesian uncertainty in deep learning," In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, pp. 13 132–13 143 (2019)
Mandt, S., Hoffman, M.D., Blei, D.M.: Stochastic gradient descent as approximate bayesian inference. J. Mach. Learn. Res. **18**(1), 4873–4907 (2017)

Mayer, H.: "Ecdsa security in bitcoin and ethereum: a research survey," CoinFaabrik, June, vol. 28, p. 126, (2016)
Nakamoto, S.: "Bitcoin: A peer-to-peer electronic cash system," [Online]. Available: http://bitcoin.org/bitcoin.pdf (2008)
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
Rambaldi, M., Bacry, E., Lillo, F.: The role of volume in order book dynamics: a multivariate hawkes process analysis. Quant. Finance **17**(7), 999–1020 (2016)
Ron, D., Shamir, A.: "Quantitative analysis of the full bitcoin transaction graph," In: International Conference on Financial Cryptography and Data Security. Springer, pp. 6–24 (2013)
Ruder, S.: "An overview of gradient descent optimization algorithms," arXiv preprint arXiv:1609.04747, 2016
Satish, V., Saxena, A., Palmer, M.: Predicting intraday trading volumeand volume percentages. J. Trading **9**(3), 15–25 (2014)
Schwab, P., Miladinovic, D., Karlen, W.: "Granger-causal attentive mixtures of experts: Learning important features with neural networks," In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 4846–4853 (2019)
Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J., Ren, J., Nado, Z.: "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," In: Wallach, H., Larochelle, H., Beygelzimer A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, pp. 13 969–13 980 (2019)
Sun, X., Liu, M., Sima, Z.: A novel cryptocurrency price trend forecasting model based on lightgbm. Finance Res. Lett. (2018)
Taieb, S.B., Hyndman, R.J.: A gradient boosting approach to the kaggle load forecasting competition. Int. J. Forecast. **30**(2), 382–394 (2014)
Urquhart, A.: The inefficiency of bitcoin. Econom. Lett. **148**, 80–82 (2016)
Waterhouse, S., MacKay, D., Robinson, T.: Bayesian methods for mixtures of experts. In: Touretzky, D., Mozer, M.C., Hasselmo, M. (eds.) Proceedings of the 8th International Conference on Neural Information Processing Systems (NIPS'95), pp. 351–357. MIT Press, Cambridge, MA, USA (1995)
Wei, X., Sun, J., Wang, X.: "Dynamic mixture models for multiple time-series." In: IJCAI, vol. 7, (2007), pp. 2909–2914
Wheatley, S., Sornette, D., Huber, T., Reppen, M., Gantner, R.N.: Are bitcoin bubbles predictable? combining a generalized metcalfe's law and the log-periodic power law singularity model. Royal Soc. Open Sci. **6**(6), 180538 (2019)
Yuksel, S.E., Wilson, J.N., Gader, P.D.: Twenty years of mixture of experts. IEEE Trans. Neural Netw. Learning Syst. **23**, 1177–1193 (2012)
Zhou, N., Cheng, W., Qin, Y., Yin, Z.: Evolution of high-frequency systematic trading: a performance-driven gradient boosting model. Quant. Finance **15**(8), 1387–1403 (2015)