

Predicting short-term Bitcoin price fluctuations from buy and sell orders

Tian Guo
ETH Zurich, COSS
Zurich, Switzerland
tian.guo@ethz.ch

Nino Antulov-Fantulin
ETH Zurich, COSS
Zurich, Switzerland
anino@ethz.ch

ABSTRACT

Bitcoin is the first decentralized digital cryptocurrency, which has showed significant market capitalization growth in last few years. It is important to understand what drives the fluctuations of the Bitcoin exchange price and to what extent they are predictable. In this paper, we study the ability to make short-term prediction of the exchange price fluctuations (measured with volatility) towards the United States dollar. We use the data of buy and sell orders collected from one of the largest Bitcoin digital trading offices in 2016 and 2017. We construct a generative temporal mixture model of the volatility and trade order book data, which is able to outperform the current state-of-the-art machine learning and time-series statistical models. With the gate weighting function of our generative temporal mixture model, we are able to detect regimes when the features of buy and sell orders significantly affects the future high volatility periods. Furthermore, we provide insights into dynamical importance of specific features from order book such as market spread, depth, volume and ask/bid slope to explain future short-term price fluctuations.

KEYWORDS

Cryptocurrency, Mixture models, Time series

1 INTRODUCTION

Bitcoin (BTC) [37] is a novel digital currency system which functions without central governing authority. Instead, payments are processed by a peer-to-peer network of users connected through the Internet. Bitcoin users announce new transactions on this network, which are verified by network nodes and recorded in a public distributed ledger called the blockchain. Bitcoin is the largest of its kind in terms of total market capitalization value. They are created as a reward in a competition in which users offer their computing power to verify and record transactions into the blockchain. Bitcoins can also be exchanged for other currencies, products, and services. The exchange of the Bitcoins with other currencies is done on the exchange office, where "buy" or "sell" orders are stored on the order book. "Buy" or "bid" offers represent an intention to buy certain amount of Bitcoins at some price while "sell" or "ask" offers represent an intention to sell certain amount of Bitcoins at some price. The exchange is done by matching orders by price from order book into a valid trade transaction between buyer and seller.

Volatility [2, 18] as a measure of price fluctuations has a significant impact on trade strategies and investment decisions [12] as well as on option pricing [3, 10] and measures of systemic risk [8, 21, 32, 39]. The order book data can give us the additional information to predict future volatility [38] by providing insights into the liquidity and trading intentions [24, 36]. In this paper, we focus

on the short-term prediction of volatility from order book data from machine learning perspective, with no intention to produce another financial model of volatility [2, 18] or order book separately.

Bitcoin, as a pioneer in the blockchain financial renaissance [7, 37] plays a dominant role in a whole cryptocurrency market capitalization ecosystem. Therefore, it is of great interest of data mining and machine learning community to be able to: (i) predict Bitcoin price fluctuations and (ii) give insights to understand what drives the Bitcoin volatility and better estimate associated risks in cryptocurrency domain.

The main contributions are: (1) We formulate the problem of predicting short-term Bitcoin price fluctuations from order book as learning predictive models over both volatility series and order book features; (2) We propose interpretable temporal mixture models to capture the dynamical effect of order book on the volatility evolution; (3) The comprehensive experimental evaluation demonstrates the superior performance of the mixture model in comparison to numerous time series models [18, 19, 42] and ensemble methods [5, 6, 13, 23, 31, 33]; (4) By analyzing the components of the mixture model, we detect regimes when order book dynamics affect future short-term volatility; (5) In addition, we adopt rolling and incremental learning and evaluation schemes to study the robustness of models with respect to the look-back horizon of historical data.

The remainder of the paper is structured as follows. In section 2, we give overview of the related work. Section 3, gives the general description of the order book data as well as the feature extraction step. Section 4 and 5, explain the proposed generative temporal mixture model that we use to learn volatility and order book data and associated learning and evaluation methodology. Finally, in section 6, we explain the experimental procedures along with interpretations and discussions.

2 RELATED WORK

Different studies have tried to explain various aspects of the Bitcoin such as its price formation, price fluctuations, systems dynamics and economic value.

From general complex systems perspective, different studies quantify either the evolution of transaction network [28] or the evolution of whole cryptocurrency ecosystem [11]. From economy perspective, the main studies [4, 20, 29] are focused around the fundamental and speculative value of Bitcoin.

Prediction of price was done with the following studies: (i) Garcia et. al. used autoregression techniques and identified two positive feedback loops (word of mouth, and new Bitcoin adopters) that lead to price bubbles [16], (ii) Amjad et. al. used the historical time series price data for price prediction and trading strategy [1] and (iii)

Garcia et. al. also showed that the increases in opinion polarization and exchange volume precede rising of Bitcoin prices [15].

The price fluctuations were studied from different data sources: (a) Kondor et. al. used the Principal Component Analysis of the blockchain transaction networks data to find correlations between principal variables and changes in the exchange price, (b) Kim et. al. used cryptocurrency web communities data to extract sentiment and predict price fluctuations [26] and (c) Donier et. al. used order book data and found that the lack of buyers stoked the panic [9] before April 10th 2013 Bitcoin price crash.

Separately from cryptocurrency markets, a huge amount of volatility models [2, 18, 25, 46] and order book models [24, 36, 38] for standard financial markets exist. However, in this paper, we focus on the short-term prediction of cryptocurrency volatility from machine learning perspective, with no intention to produce another financial model of volatility or order book separately, but to construct a generative temporal mixture model of volatility and order book.

3 DATA AND PREPROCESSING

In this paper we use the actual historical hourly volatility data of Bitcoin market, which refers to the standard deviation of minute returns over one hour time range [18]. The return is defined as the relative change in consecutive prices of BTC. Our dataset contains time series of hourly volatility spanning more than one year, which outlines the fluctuation of Bitcoin market over time.

In addition, we have the order book data from the OKCoin, which is a digital asset trading platform providing trading services between fiat currencies and cryptocurrencies. In the last 2 years, trading volume of BTC at the OKCoin exchange office was approximately 39 % of the total traded BTC volume, which implies that our data source (OkCoin) can be used as a good proxy for BTC trading. Order book was collected through the exchange API with the granularity of one minute, with negligible missing values due to the API downtime or communication errors. Each order contains two attributes, price and amount. For instance, in the middle panel of Fig. 1 the green and red areas respectively show the accumulated amount of ask and bid orders w.r.t. the prices in one minute. Such a figure is commonly used to interpret the market intention and potential movements.

Volatility series and order book data are of different types and time granularities. The volatility series carries the long-term contextual information. Order book provides fine-grained selling and buying information characterizing the local behavior of the market. Therefore, in forecasting the volatility it is highly desirable to develop a systematical way to model the complementary dependencies in volatility and order book data. Meanwhile, the proposed model should be interpretable, in the sense that it enables to observe how such two types of data interact to drive the evolution of volatility.

Intuitively, our idea is to first transform order book data into features over time, referred as feature series and then to develop probabilistic models to consume volatility and feature series simultaneously. Concretely, from each minute snapshot of order book we extract the features related to: price spread, weighted spread,

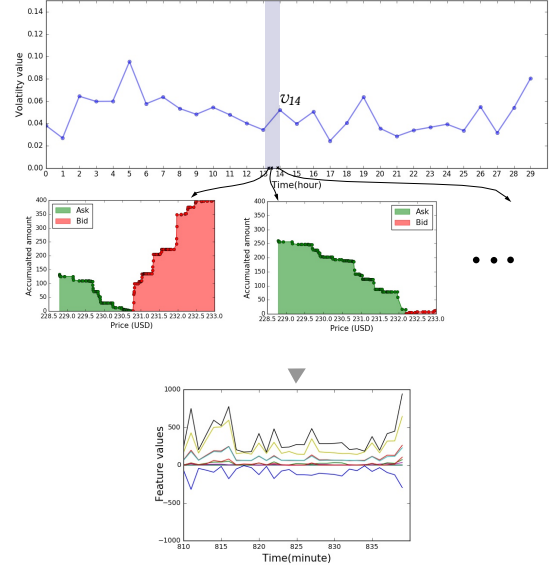


Figure 1: Volatility data with feature series from order book. Top panel: volatility series on time offset in hours w.r.t. to 00:00AM, September 1st, 2016. Middle panel: order book snapshots on minutes within hour 13. Bottom panel: order book snapshots are transformed to feature series.

ask/bid volume and their difference, ask/bid depth and their difference and ask/bid slope. The 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). Depth is the number of orders on the bid or ask side. Volume is the number of BTCs on the bid or ask side. Weighted spread is the difference between cumulative price over 10 % of bid depth and the cumulative price over 10 % of ask depth. Slope is estimated as the volume until δ price offset from the current traded price, where δ is estimated by the bid price at the order that has at least 10 % of orders with the higher bid price. For instance, Fig. 1 illustrates how a volatility observation and the associated order book feature series are organized. The shaded area in the top panel demonstrates time period corresponding to volatility v_{13} . The bottom panel shows that the order book data at each minute instant within the shaded area are transformed into feature series, which will be used to learn models for volatility forecasting.

We formulate the problem to resolve in this paper as follows. Given an hourly series of volatility $\{v_0, \dots, v_H\}$, the h -th observation is denoted by $v_h \in \mathbb{R}^+$. The features of order book at each minute are denoted by a vector $\mathbf{x}_m \in \mathbb{R}^n$, where n is the dimension of the feature vector. We define an index mapping function $i(\cdot)$ to map an hour index to the starting minute index of that hour (e.g. $i(0)=1$, $i(1)=61$), such that order book features associated with volatility observation v_h is denoted by a matrix $\mathbf{X}_{[i(h), -l_b]} = (\mathbf{x}_{i(h)-1}, \dots, \mathbf{x}_{i(h)-l_b}) \in \mathbb{R}^{n \times l_b}$, where l_b is the look-back time horizon. Likewise, a set of historical volatility observations w.r.t. v_h is denoted by $\mathbf{v}_{[h, -l_v]} = (v_{h-1}, \dots, v_{h-l_v}) \in \mathbb{R}^{l_v}$. Given historical volatility observations $\mathbf{v}_{[h, -l_v]}$ and order book

features $\mathbf{X}_{[i(h), -l_b]}$, we aim to predict the one-step ahead volatility v_h . In addition, the proposed model should be able to uncover how such two types of data contribute to the prediction.

4 MODELS

In this section, we first briefly describe the conventional idea of modeling time series with external features. Then, we present the temporal mixture model to adaptively learn volatility and order book data.

4.1 Times series model with external features

Time series model is the natural choice for volatility series. As for the order book data, it can be incorporated as exogenous variable. This idea gives rise to ARIMAX model, which adds in external features on the right hand side of ARIMA model [23]:

$$\phi(L)(1-L)^d v_h = \sum_{j=1}^{l_b} \beta_j^\top \mathbf{x}_{i(h)-j} + \theta(L)\epsilon_h$$

where d is the d -th difference operator, L is a lag operator, $\phi(L)$ is the autoregressive polynomial $(1 - \phi_1 L - \dots - \phi_{l_v} L^{l_v})$, $\theta(L)$ is the moving average polynomial $(1 + \theta_1 L + \dots + \theta_q L^q)$, $\beta_j \in \mathbb{R}^n$ and $\beta_j^\top \mathbf{x}_{i(h)-j}$ is the regression term on order book features. ϵ_h is a white noise process (i.e., zero mean and i.i.d).

This model suggests that the order book data constantly affects the evolution of volatility via the regression terms. However, considering the time-varying behaviors of order book snapshots we have observed, it is desirable to have a model, which is capable of capturing the time-varying behaviors of order book as well as providing interpretable results about the effect on volatility [14].

4.2 Temporal mixture model

In this part, we present the temporal mixture model with the aim to adaptively using volatility series and order book features in forecasting.

Mixture model is popular in data mining and machine learning areas and has been used in numerous regression, classification, clustering, and fusion applications in healthcare, finance, and pattern recognition [17, 47]. The model is a weighted sum of component models. It allows for great flexibility in the specification of each component models and the structure of the mixture based on some knowledge of the problem domain. Individual component models can specialize on different part of the data. The weights dependent on input data enable the model to adapt to non-stationary data. Moreover, it provides interpretable results [35, 47].

Specifically, we start with building a joint probabilistic density function of volatility observations conditional on order book features as follows:

$$\begin{aligned} p(v_1, \dots, v_H | \{\mathbf{x}_m\}; \Theta) &= \prod_h p(v_h | \mathbf{v}_{[h, -l_v]}, \mathbf{X}_{[i(h), -l_b]}) \\ &= \prod_h \left[p(v_h, z_h = 0 | \mathbf{v}_{[h, -l_v]}, \mathbf{X}_{[i(h), -l_b]}) \right. \\ &\quad \left. + p(v_h, z_h = 1 | \mathbf{v}_{[h, -l_v]}, \mathbf{X}_{[i(h), -l_b]}) \right]. \end{aligned} \quad (1)$$

The Eq. 1 is obtained with iteratively applying chain rule. It models the distribution of volatility v_h conditional on historical volatility and feature series with two components by introducing a binary latent variable $z_h \in \{0, 1\}$. Variable $z_h = 0$ corresponds

to the case when volatility is dependent on the historical volatility data itself, while $z_h = 1$ stands for the case of dependence on order book.

Then, by defining a conditional density term g_h for latent variable z_h , Eq. 1 can be rewritten as:

$$\begin{aligned} p(v_1, \dots, v_H | \{\mathbf{x}_m\}; \Theta) &= \prod_h \left[p(v_h | \mathbf{v}_{[h, -l_v]}, z_h = 0) \cdot g_h \right. \\ &\quad \left. + p(v_h | \mathbf{X}_{[i(h), -l_b]}, z_h = 1) \cdot (1 - g_h) \right], \end{aligned} \quad (2)$$

where $g_h := P(z_h = 0 | \mathbf{v}_{[h, -l_v]}, \mathbf{X}_{[i(h), -l_b]})$. The mixture components $p(v_h | \mathbf{v}_{[h, -l_v]}, z_h = 0)$ and $p(v_h | \mathbf{X}_{[i(h), -l_b]}, z_h = 1)$ define the data generating process by modeling the distribution of v_h conditional on volatility itself and order book feature series.

Gate function g_h represents the weight for mixture components, i.e. the probability of volatility v_h is driven solely by its history, given $\mathbf{v}_{[h, -l_v]}$ and $\mathbf{X}_{[i(h), -l_b]}$. Consequently, $1 - g_h$ indicates the weight for the order book component. Depending on volatility and order book features, g_h dynamically adjusts the contribution from volatility and order book to forecasting.

In the inference phase, two component models compute means conditioned on their respective input data. The weighted combination of their means by g_h is taken as the prediction of the mixture model. Moreover, this temporal mixture model is interpretable in the sense that by observing the gate weights and predictions of component models, we can understand when and to what extent order book contributes to the evolution of volatility. We demonstrate this in the experiment section.

In the following, we describe two realizations of the mixture model respectively based on Gaussian and log normal distributions.

4.2.1 Gaussian temporal mixture model.

In this part, we choose Gaussian distribution to model the conditional density of v_h under different states of latent variable z_h . Specifically, they are represented as:

$$\begin{aligned} v_h | \{\mathbf{v}_{[h, -l_v]}, z_h = 0\} &\sim \mathcal{N}(\mu_{h,0}, \sigma_{h,0}^2) \\ v_h | \{\mathbf{X}_{[i(h), -l_b]}, z_h = 1\} &\sim \mathcal{N}(\mu_{h,1}, \sigma_{h,1}^2), \end{aligned} \quad (3)$$

where $\mu_{h,\cdot}$ and $\sigma_{h,\cdot}^2$ are the mean and variance of individual component models.

$$\begin{aligned} \mu_{h,0} &= \sum \phi_i v_{h-i} \\ \mu_{h,1} &= U^\top \mathbf{X}_{[i(h), -l_b]} V, \end{aligned} \quad (4)$$

where $\phi_i \in \mathbb{R}$, $U \in \mathbb{R}^n$ and $V \in \mathbb{R}^{l_b}$ are the parameters to learn. The set of parameters $\{\phi_i, U, V\}$ is denoted by Θ_v .

In Eq. 4, we use autoregressive model to capture the dependence of v_h on historical volatility. Order book features are organized as a matrix with temporal and feature dimensions. Therefore we make use of bilinear regression, where parameters U and V respectively capture the temporal and feature dependence. As a result, the importance of each feature can be interpreted with ease, which is illustrated in the experiment section. The variance term $\sigma_{h,\cdot}^2$ in each component is obtained by simply performing linear regression on the input of that component.

Then, the gate g_h is defined by the softmax function

$$g_h := \frac{\exp(\sum \theta_i v_{h-i})}{\exp(\sum \theta_i v_{h-i}) + \exp(A^\top X_{[i(h), -l_b]} B)}, \quad (5)$$

where $\theta_i \in \mathbb{R}$, $A \in \mathbb{R}^n$ and $B \in \mathbb{R}^{l_b}$ are the parameters to learn. Denoted by Θ_z is the set of parameters in the gate function. Likewise, we utilize autoregression and bilinear regression, thereby facilitating the understanding of the feature importance in determining the contribution of volatility history and order book features.

During the inference, the conditional mean of the mixture distribution is taken as the predicted value \hat{v}_h :

$$\begin{aligned} \hat{v}_h &= \mathbb{E}(v_h | v_{[h, -l_v]}, X_{[i(h), -l_b]}) \\ &= g_h \cdot \mu_{h,0} + (1 - g_h) \cdot \mu_{h,1}. \end{aligned} \quad (6)$$

We define $\Theta = \{\phi_i, U, V, \theta_i, A, B\}$ as the entire set of parameters in the mixture model. We present the objective functions for learning Θ and in the next section we describe the detailed learning algorithm.

In learning the parameters in Gaussian temporal mixture model, the loss function to minimize is defined as:

$$\begin{aligned} O(\Theta) &= - \underbrace{\sum_h \log \left[g_h \mathcal{N}(v_h | \mu_{h,0}, \sigma_{h,0}^2) + (1 - g_h) \mathcal{N}(v_h | \mu_{h,1}, \sigma_{h,1}^2) \right]}_{\text{negative log likelihood}} \\ &\quad + \underbrace{\lambda \|\Theta\|_2^2 + \alpha \sum_h [\max(0, \delta - \mu_{h,0}) + \max(0, \delta - \mu_{h,1})]}_{\text{Non-negative mean regularization}}. \end{aligned} \quad (7)$$

In addition to the L2 Regularization over parameter set Θ for preventing over-fitting, we introduce two hinge terms to regularize the predictive mean of each component model, i.e. $\mu_{h,0}$ and $\mu_{h,1}$. Since the value of volatility lies in the non-negative domain of real values, we impose hinge loss on the mean of each component model to penalize negative values. The parameter δ is the margin parameter, which is set to zero in experiments.

4.2.2 Log-normal temporal mixture model.

Instead of enforcing non-negative mean of component model by regularization, in this part we present the temporal mixture model using log-normal distribution, which naturally fits non-negative values.

Specifically, for a random non-negative variable of log-normal distribution, the logarithm of this variable is normally distributed [44]. Thus, by assuming v_h is log-normally distributed, we represent component models of the temporal mixture model as:

$$\begin{aligned} \log(v_h) | \{v_{[h, -l_v]}, z_h = 0\} &\sim \mathcal{N}(\mu_{h,0}, \sigma_{h,0}^2) \\ \log(v_h) | \{X_{[i(h), -l_b]}, z_h = 1\} &\sim \mathcal{N}(\mu_{h,1}, \sigma_{h,1}^2). \end{aligned} \quad (8)$$

The conditional mean of v_h in one component model becomes $\mathbb{E}(v_h | \cdot) = \exp(\mu + 0.5\sigma)$ [44]. Regarding the gate function g_h , we can use the same form as in Eq. 5. In prediction, the mean of the log-normal temporal mixture model is a gate weighted sum of component means analogous to Eq. 6. In the objective function, we can safely get rid of the non-negative regularization, due to the

non-negative nature of $\mathbb{E}(v_h | \cdot)$ and obtain the loss function as:

$$\begin{aligned} O(\Theta) &= - \sum_h \log \left[g_h \cdot p(v_h | \mu_{h,0}, \sigma_{h,0}^2) + (1 - g_h) \cdot p(v_h | \mu_{h,1}, \sigma_{h,1}^2) \right] \\ &\quad + \lambda \|\Theta\|^2, \end{aligned} \quad (9)$$

where $p(v_h | \cdot)$ is the density function of log-normal. Due to the limitation of pages, we skip the details.

5 MODEL LEARNING AND EVALUATION

The standard approach of learning and evaluating time series models is to split the entire time series at a certain time step, where the front part is taken as training and validation data while the rest is used as testing data [22, 30].

However, such a method could overlook the time-varying behavior of volatility series and order book [1]. Consequentially, the derived model has to compromise the non-stationarity in data. Therefore, we adopt a rolling strategy to learn and evaluate models [1], such that it enables to study the performance of models on different time periods of the data as well as the effect of the look-back time horizon for model learning.

In the following of this section, we first describe the learning methods for aforementioned models. Then we present the rolling learning and evaluation scheme.

5.1 Learning methods

In this part, we mainly present the learning algorithms of the temporal mixture model. For the ARIMAX style models mentioned in Sec. 4.1 we use standard maximum likelihood estimation to learn the parameters [27].

Our mixture model involves both latent states and coupled parameters and thus we iteratively minimize the objective function defined in Eq. 7 and 9 [40, 47]. We use the Gaussian temporal mixture model to illustrate the algorithm below, but the same methodology is applied to the log-normal model as well.

Specifically, the learning algorithm consists of two main steps. First, fix all component model parameters and **update the parameters of gate function** g_h , by using a gradient descent method [35, 40]. Due to the page limitation, We present the gradients of certain parameters w.r.t. the objective function below. The rest can be derived analogously.

$$\begin{aligned} \frac{\partial O}{\partial \theta_i} &= - \sum_h \left[\frac{1}{a(v_h | \Theta)} g_h v_{h-i} (1 - g_h) \mathcal{N}(v_h | v_{[h, -l_v]}, z_h = 0) \right. \\ &\quad \left. - g_h v_{h-i} (1 - g_h) \mathcal{N}(v_h | X_{[i(h), -l_b]}, z_h = 1) \right] + 2\lambda \theta_i, \end{aligned} \quad (10)$$

where $a(v_h | \Theta)$ denotes $g_h \mathcal{N}(v_h | \mu_{h,0}, \sigma_{h,0}^2) + (1 - g_h) \mathcal{N}(v_h | \mu_{h,1}, \sigma_{h,1}^2)$.

For the coupled parameters of bilinear regression, it can be broken into two convex tasks, where we individually learn A as follows [43]:

$$\begin{aligned} \frac{\partial O}{\partial A} = & - \sum_h \frac{X_{[i(h), -l_b]}^B}{a(v_h | \Theta)} \left[\frac{(g_h - 1)g_h}{\exp(\sum \theta_i v_{h-i})} \mathcal{N}(v_h | v_{[h, -l_v]}, z_h = 0) \right. \\ & \left. - \frac{(g_h - 1)g_h}{\exp(\sum \theta_i v_{h-i})} \mathcal{N}(v_h | X_{[i(h), -l_b]}, z_h = 1) \right] + 2\lambda A. \end{aligned} \quad (11)$$

Second, fix the gate function and **update the parameters in component models**: $P(v_h | v_{[h, -l_v]}, z_h = 0)$ and $P(v_h | X_{[i(h), -l_b]}, z_h = 1)$.

$$\begin{aligned} \frac{\partial O}{\partial \phi_i} = & - \sum_h \frac{2g_h \cdot \mathcal{N}(v_h | v_{[h, -l_v]}, z_h = 0)}{a(v_h | \Theta) \sigma_{h,0}^2} \left(\sum_j \phi_j v_{h-j} - v_h \right) v_{h-i} \\ & + 2\lambda \phi_i + \alpha \sum_h \mathbb{1}_{>0} \{ \max(0, \delta - \mu_{h,0}) \} (-v_{h-i}). \end{aligned} \quad (12)$$

$$\begin{aligned} \frac{\partial O}{\partial V} = & - \sum_h \frac{2(1 - g_h) \cdot \mathcal{N}(v_h | X_{[i(h), -l_b]}, z_h = 1)}{a(v_h | \Theta) \sigma_{h,1}^2} \left(U^T X_{[i(h), -l_b]} V - v_h \right) \\ & \cdot X_{[i(h), -l_b]}^T U + 2\lambda V + \alpha \sum_h \mathbb{1}_{>0} \{ \max(0, \delta - \mu_{h,1}) \} (-X_{[i(h), -l_b]}^T U). \end{aligned} \quad (13)$$

5.2 Evaluation scheme

The idea of **rolling learning and evaluation** is as follows. We divide the whole time range of data into non-overlapping intervals, for instance, each interval corresponds to one month (see Figure 2). We perform the following two steps repeatedly over each interval: (i) we take data within one interval as the testing set and the data in the previous N intervals as the training and validation set, (ii) each time that testing data is built from a new interval, the model is retrained and evaluated on the current training and testing data. Eventually, we obtain the model prediction result on each testing interval. Such a scheme provides a natural evaluation for non-stationary models. Fig. 2 provides a toy example. In the top panel, testing set A and B are sequentially selected as the testing set. Given that $N = 2$, the shadow areas show the temporal range of their corresponding training sets.

For comparison, we also use an **incremental evaluation** method. This approach amounts to always use all the data preceding to the current testing set for model training. In particular, the testing set is iteratively selected the same way as the rolling method, while the training data for the corresponding testing data is incrementally increased by adding the expired test set to the current training set. Then, the model is retrained with the enlarged training data and used for predicting the current test set. The bottom panel in Fig. 2 demonstrates the process of incremental evaluation.

6 EXPERIMENTS

In this section we present a comprehensive evaluation and reasoning behind the results of the approaches. We first introduce the dataset, baselines and set-up details. Then, prediction performance as well as insights drawn from the model interpretation are reported.

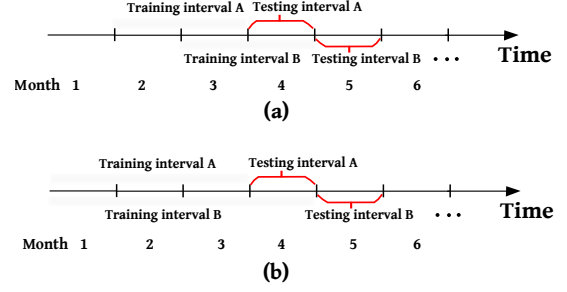


Figure 2: Top panel: rolling scheme. Bottom panel: incremental scheme.

6.1 Dataset

We have collected volatility and order book data ranging from September 2015 to April 2017. It consists of 13730 hourly volatility observations and 701892 order book snapshots. Each order book snapshot contains several hundreds of ask and bid orders. The maximum number of ask and bid orders at each minute are 1021 and 965. For the volatility time series, Augmented Dickey-Fuller (ADF) test is rejected at 1% significance level and therefore there is no need for differencing [27]. Seasonal patterns of volatility series is examined via periodogram and the result shows no existence of strong seasonality [45].

6.2 Baselines

The **first category of statistics baselines** are only trained either on volatility or return time series. The included methods are as follows:

EWMA represents the exponential weighted moving average approach, which simply predicts volatility by performing moving average over historical ones [23].

GARCH refers to generalized autoregressive conditional heteroskedasticity model. It is a widely used approach to estimate volatility of returns and prices [18]. The basic idea of such methods is to model the variance of the error term of a time series as a function of the the previous error terms.

BEGARCH represents the Beta-t-EGARCH model [19]. It extends upon GARCH models by letting conditional log-transformed volatility dependent on past values of a t-distribution score.

STR is the structural time series model [41, 42]. It is formulated in terms of unobserved components via the state space method and used to capture local trend variation in time series.

Under this category, we also have plain **ARIMA** model.

The **second category of machine learning baselines** learns volatility and order book data simultaneously.

RF refers to random forests. Random forests are an ensemble learning method consisting several decision trees for classification, regression and other tasks [31, 34].

GBT is the gradient boosted tree, which is the application of boosting methods to regression trees [13]. GBT trains a sequence of simple regression trees and then adds together the prediction of individual trees to provide final prediction.

XGT refers to the extreme gradient boosting [6]. It was developed to improve the training efficiency and model performance of

GBT by efficiently making use of computing resources and a more regularized model formalization.

ENET represents elastic-net, which is a regularized regression method combining both L1 and L2 penalties of the lasso and ridge methods [33].

GP stands for the Gaussian process based regression [5, 46], which has been successfully applied to time series. It is a supervised learning method which provides a Bayesian nonparametric approach to smoothing and interpolation.

STRX is the **STR** method augmented by adding regression terms on external features, similar to the way of **ARIMAX**.

For RF, GBT, XGT, ENET, and GP methods, input features are built by concatenating historical volatility and order book features. The Gaussian and log-normal temporal mixture will be respectively denoted by **TM-G** and **TM-LOG**¹.

As it is still nontrivial to decipher variable importance of multi-variable series from deep models, we do not take into account deep models like recurrent neural networks in the present paper. The aim of the current work is the first step towards fundamentally understanding the data using models with good interpretability.

6.3 Evaluation set-up

In GARCH and BEGARCH, the orders of autoregressive and moving average terms for the variance are both set to one [18]. Smoothing parameter in EWMA is chosen from $\{0.01, 0.1, 0.2, \dots, 0.9\}$. In ARIMA and ARIMAX, the orders of auto-regression and moving-average terms are set via the correlogram and partial autocorrelation. For decision tree based approaches including RF, GBT, XGT, hyper-parameter tree depth and the number of iterations are chosen from range $[3, 10]$ and $[3, 200]$ via grid search. For XGT, L2 regularization is added by searching within $\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$. As for ENET, the coefficients for L2 and L1 penalty terms are selected from the set of values $\{0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$. In GP, we adopt the radial-basis function (RBF) kernel, white noise, and periodic kernels[5]. The hyper-parameters in GP are optimized via maximum likelihood estimation. In TM-G and TM-LOG, the regularization coefficient is chosen from the set of values $\{1, 0.1, 0.01, 0.001, 0.0001\}$. The volatility autoregressive order l_v is set as in ARIMA, while order book look-back horizon l_b is empirically set to 30. We will study the effect of l_b in the results below.

We use root mean squared error (RMSE) and mean absolute error (MAE) as evaluation metrics [18], which are defined as follows: $RMSE = 1/n^2 \sqrt{\sum_i (v_i - \hat{v}_i)^2}$ and $MAE = 1/n \sum_i |v_i - \hat{v}_i|$. Furthermore, we also report the significance between the error distributions from different models by the two-sample Kolmogorov-Smirnov test [35].

Regarding the evaluation process, volatility and order book data are divided into a set of monthly data. Totally, we have 12 testing months, namely we will retrain models for each testing month and evaluate the performance in the corresponding testing month. In the rolling evaluation scheme, the period of training and validation data prior to a certain testing month is set to 2 months.

6.4 Prediction performance

In this part, we report on the prediction performance over time intervals of each approach respectively trained in the rolling and incremental ways. The time horizon of order book features, i.e. parameter l_b , is set to 30, that is, order book features within the 30 minutes prior to the prediction hour are fed into models. Then, the sensitivity of prediction performance w.r.t. l_b is also presented.

Tab. 4 shows prediction errors over each testing interval obtained by the rolling learning. The results are reported in three groups of approaches according to the category described in the baseline subsection. In general, mixture models, i.e. TM-LOG and TM-G constantly outperform others. Basically, approaches using both volatility and order book features perform better than those using only volatility or return series, however the simple EWMA can beat all others except for mixture models in some intervals. Particularly, in the middle group of Tab. 4, ARIMAX and STRX using both volatility and order book data fail to outperform their counterparts, i.e. ARIMA and STR in the top group. It suggests that simply adding features from order book does not necessarily improve the performance. Ensemble and regularized regression perform better, e.g., XGT and ENET perform the best in most of cases, within this group.

Gaussian temporal mixture, i.e. TM-G model outperforms other approaches in most of cases. Specifically, it can achieve 50% less errors at most. Although the volatility lies only on non-negative range of values, surprisingly the TM-LOG is still inferior to TM-G. One possible explanation is that on the short time scales the random variables $v_h | \{v_{[i(h), -l_v]}, z_h = 0\}$ and $v_h | \{X_{[i(h), -l_b]}, z_h = 1\}$ are better explained with the Gaussian since the variations in short time scale can not lead to the heavy-tail of the log-normal distribution.

Tab. 5 shows the prediction errors over each testing interval obtained by incremental learning. Over the testing intervals, incremental learning uses all the historical data that has been seen to learn the models. Due to the page limitation, we list the results from interval 7 to 11. By comparing the errors in Tab. 5 and Tab. 4 w.r.t. a certain approach, we aim to study the effect of learning the model with increasingly accumulated historical data. It is observed that in the top group of Tab. 5, such methods do not ingest order book features, and therefore the errors of EWMA, GARCH, BEGARCH, and STR are relatively similar to the corresponding ones in Tab. 4, though for ARIMA, the errors in the last few intervals, e.g. 10 and 11, present little ascending pattern. However, for the approaches in the middle group, ARIMAX and ENET exhibit decreasing and then increasing error pattern, compared with their counterparts in Tab. 4. This suggests that increasing the amount of data benefits the model for prediction. Nevertheless, because of the time-varying behaviors in order book, too old data could deteriorate the prediction performance of the models instead. Due to the adaptive weighting mechanism for order book features, TM-G is robust to increasing amount of data for learning, though in comparison with the ones in Tab. 4 the errors decline in some intervals. Above observations in Tab. 4 and Tab. 5 apply to MAE results as well. Please find MAE results in the appendix section.

Tab. 3 demonstrates the effect of time horizon (i.e. l_b) of order book features on prediction performance of models using order

¹The code and data can be provided upon request.

Table 1: Test errors over each time interval of rolling learning (RMSE)

Model \ Interval	1	2	3	4	5	6	7	8	9	10	11	12
EWMA	0.082	0.126*	0.265*	0.182	0.096*	0.027*	0.034*	0.030*	0.056*	0.054*	0.064*	0.096*
GARCH	0.136**	0.225**	0.464**	0.286**	0.135**	0.038**	0.046**	0.047**	0.097**	0.104**	0.086**	0.127**
BEGARCH	0.134**	0.223**	0.462**	0.283**	0.133**	0.037**	0.045**	0.045**	0.096**	0.101**	0.083**	0.123**
STR	0.111**	0.207**	0.408**	0.252**	0.105**	0.042**	0.058**	0.031**	0.082**	0.232**	0.065*	0.111**
ARIMA	0.106**	0.194**	0.409**	0.247**	0.101**	0.117**	0.037**	0.031**	0.084**	0.211**	0.066**	0.096**
ARIMAX	0.126**	0.282**	0.443**	0.271**	0.134**	0.156**	0.072**	0.041**	0.094**	0.247**	0.076**	0.107**
STRX	0.159**	0.249**	0.414**	0.242**	0.176**	0.125**	0.044**	0.045**	0.104**	0.255**	0.078**	0.138**
GBT	0.083**	0.157**	0.284**	0.214**	0.096**	0.070**	0.060**	0.031**	0.060**	0.054**	0.062*	0.085*
RF	0.082**	0.151**	0.296**	0.212**	0.096**	0.076**	0.060**	0.029*	0.066**	0.051**	0.061*	0.085*
XGT	0.076*	0.144**	0.264**	0.194**	0.090	0.096**	0.046**	0.031**	0.061**	0.050*	0.061**	0.088*
ENET	0.080**	0.138**	0.270**	0.184**	0.102**	0.045**	0.035**	0.028	0.064**	0.051**	0.059*	0.084
GP	0.082**	0.137**	0.373**	0.196*	0.108**	0.067**	0.042**	0.028	0.061**	0.053**	0.063**	0.085*
TM-LOG	0.083	0.186	0.339	0.172	0.104	0.033	0.039	0.034	0.098	0.067	0.381	0.150
TM-G	0.075	0.118	0.259	0.189	0.089	0.025	0.031	0.027	0.051	0.047	0.058	0.083

Symbols * and ** respectively indicate that the error of TM-G in the table is significantly different from the corresponding one at level 5% and 1% (two sample KS test on error distributions).

Table 2: Test errors over each time interval of incremental learning (RMSE)

Model	7	8	9	10	11
EWMA	0.034*	0.030*	0.056**	0.054**	0.064**
GARCH	0.046**	0.046**	0.097**	0.105**	0.086**
BEGARCH	0.044**	0.044**	0.095**	0.102**	0.084**
STR	0.041**	0.037**	0.083**	0.188**	0.068**
ARIMA	0.039**	0.031*	0.083**	0.222**	0.067**
ARIMAX	0.058**	0.050**	0.154**	0.302**	0.105**
STRX	0.043**	0.048**	0.138**	0.263**	0.095**
GBT	0.038**	0.039**	0.124**	0.059**	0.070**
RF	0.044**	0.039**	0.059**	0.057**	0.068**
XGT	0.035*	0.029*	0.053*	0.050*	0.060*
ENET	0.038**	0.036**	0.076**	0.058**	0.070**
GP	0.043**	0.041**	0.097**	0.061**	0.069**
TM-LOG	0.038	0.033	0.098	0.067	0.381
TM-G	0.030	0.027	0.048	0.047	0.058

book features. The results are obtained by evaluating each model on interval 1 with increasing size of l_b . It exhibits that short term order book features is sufficient for most of the models and further more data, e.g. 40 and 50 minutes of order book features, lead to no improvement of the performance. In particular, models like ARIMAX and STRX are prone to overfit by redundant data of long horizon, while our mixture models, ensemble method XGT, and ENET are relatively robust to the horizon.

6.5 Model interpretation

In this part, we provide insights into the data by analyzing the components of the model. Specifically, in Fig. 3 each column of figures corresponds to a sample period from the testing month. The top panel shows the model prediction and true values of the period. The second panel demonstrates the distribution of gate values in

Table 3: Test error sensitivity to time horizon of order book features (RMSE)

Model	10	20	30	40	50
ARIMAX	0.112**	0.125**	0.126**	0.126**	0.135**
STRX	0.138**	0.142**	0.159**	0.157**	0.161**
GBT	0.081*	0.081*	0.083**	0.084**	0.085
RF	0.081**	0.082**	0.082**	0.083**	0.083**
XGT	0.077*	0.077*	0.076*	0.077**	0.076
ENET	0.080*	0.080**	0.080**	0.080**	0.081**
GP	0.095**	0.081	0.082**	0.084**	0.085
TM-LOG	0.082	0.082	0.083	0.083	0.084
TM-G	0.075	0.075	0.075	0.076	0.076

the mixture model corresponding to the top panel. The dark area corresponds to the gate values $1 - g_h$ of the component model w.r.t. the order book at each time step. The sum of dark and light areas at each time step is equal to one and therefore the lower the dark area reaches, the higher gate value is assigned to the component model of order book. The time evolution of the inferred gate values g_h and $1 - g_h$ in our mixture model explains the dynamical importance and interplay of the order book features for high volatility regimes. Thus, it implies that the order book features can encode the future short-term price fluctuations from the trade orders.

The bottom two panels exhibit two order book features with high coefficients in the mixture model. It demonstrates the correspondence between feature and gate values over time. Recall that each hourly volatility observation has the associated order book features over a time horizon l_b and therefore the value range within l_b of each feature at each hour is shown in the figures. If we look at the column (a) the large price fluctuations at the offset of 20 hours from 2016 June, 20th, 00:00 am are mostly driven by the negative market depth feature (panel four) which implies larger buying demand for the Bitcoins coupled together with the larger spread between bid

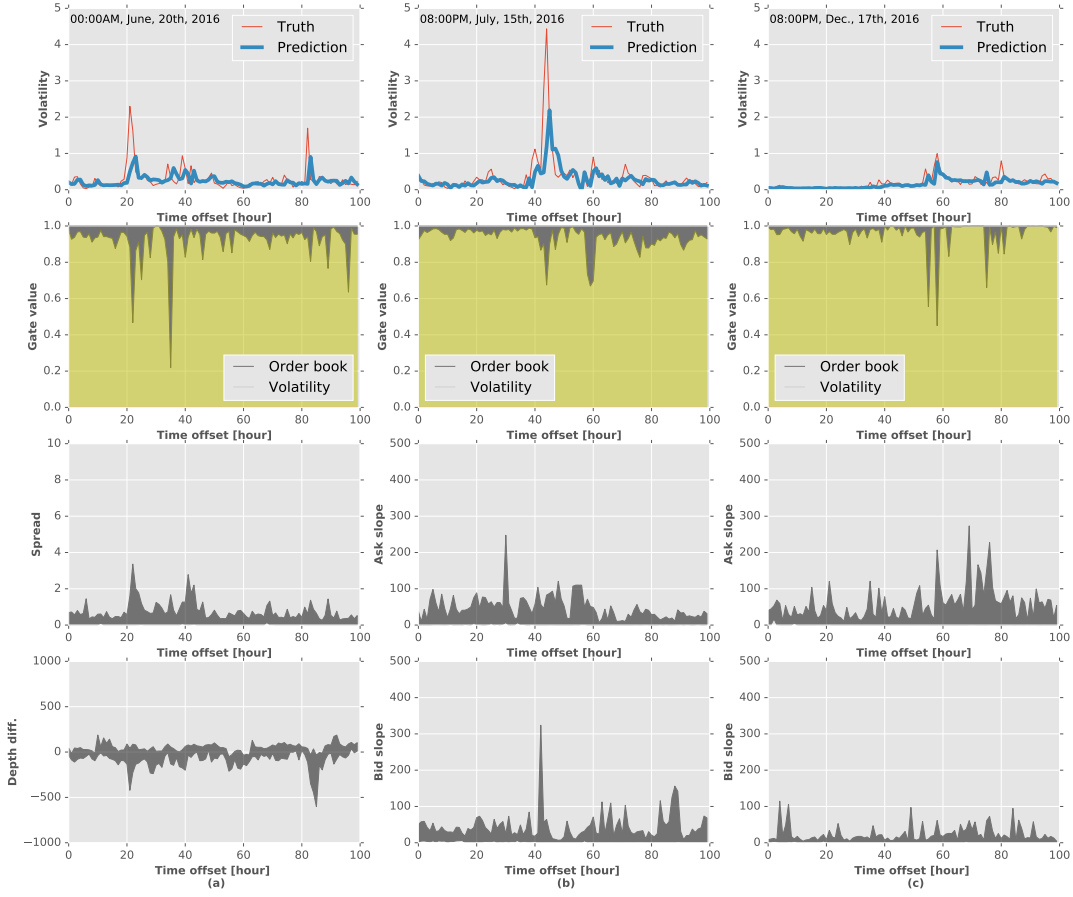


Figure 3: Mixture model visualization. Each column of figures represents the results from the mixture model on a sample testing period, where the time offset is measure in hours w.r.t. to (a) 00:00AM, June, 20th, 2016, (b) 08:00PM, July, 15th, 2016 and (c) 08:00PM, December, 17th, 2016. Top panel: prediction and true values of three sample periods. Second panel: mixture gate value g_h (in yellow color for the weight of past volatility) and $1 - g_h$ (in gray color for the weight of past volatility and order book) over time. Bottom two panels: order book feature values over time. (Best viewed in color.)

and ask price (panel three). Similarly, the large volatility around the offset of 40 hours from 2016 July, 15th, 08:00 pm in panel (b) are driven by the large bid slope i.e. buying demand near the current traded price w.r.t. much smaller ask slope i.e. selling offer near the current traded price. In contrast, in panel (c) at the offset of around 60 hours from 2016 December, 17th, 08:00 pm the medium price fluctuations are driven by larger selling offer near the current traded price. These three cases show the interpretability of the dynamical effects in our mixture model to learn the future short-term price fluctuations from order book.

7 CONCLUSION

In this paper, we study the short-term fluctuation of Bitcoin market by using real data from a major Bitcoin exchange office. The dataset comprises realized volatility observations and order book snapshots of different time scales and data types. By reformatting the order book data into feature series, we formulate the volatility

prediction problem as learning predictive models over both volatility and order book feature series. The conventional way to tackle such type of a problem is to train time series models with external features. However, the limitation of simply combining volatility and order book features lies in a fact that it could overlook the time-varying contribution from order book to volatility and it is not straightforward to interpret the obtained model as well.

Therefore, we propose temporal mixture models to capture the dynamical effect of order book features on the volatility evolution and to provide interpretable results. The comprehensive experimental evaluation compares numerous approaches including time series models, ensemble methods and the temporal mixture model. The results demonstrate that conventional time series models with external features lead to degraded prediction performance. The mixture model outperforms other approaches in most of cases. Meanwhile, by visualizing the mixture gate values and associated order book features over time, it is convenient to understand the effect of order book on the market.

In addition, we adopt rolling and incremental learning and evaluation schemes to study the robustness of the models with respect to the look-back horizon of historical data. The results show that in the time-varying environment of Bitcoin market, redundant historical data degenerates the volatility prediction performance of most of regression and ensemble models. Both our mixture models and XGT methods are more robust.

Acknowledgement

The work of T.G. and N.A.-F. has been funded by the EU Horizon 2020 SoBigData project under grant agreement No. 654024.

REFERENCES

- [1] Muhammad Amjad and Devavrat Shah. 2017. Trading Bitcoin and Online Time Series Prediction. In *NIPS 2016 Time Series Workshop*. 1–15.
- [2] Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Paul Labys. 2003. Modeling and Forecasting Realized Volatility. *Econometrica* 71, 2 (2003), 579–625.
- [3] Fischer Black and Myron Scholes. 1973. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81, 3 (1973), 637–654.
- [4] Wilko Bolt. 2016. On the Value of Virtual Currencies. *SSRN Electronic Journal* (2016).
- [5] Sofiane Brahimi-Belhouari and Amine Bermak. 2004. Gaussian process for non-stationary time series prediction. *Computational Statistics & Data Analysis* 47, 4 (2004), 705–712.
- [6] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *SIGKDD*. ACM, 785–794.
- [7] D.L.K. Chuen. 2015. *Handbook of Digital Currency: Bitcoin, Innovation, Financial Instruments, and Big Data*. Academic Press.
- [8] Jamil Civitarese. 2016. Volatility and correlation-based systemic risk measures in the US market. *Physica A: Statistical Mechanics and its Applications* 459 (2016), 55–67.
- [9] Jonathan Donier and Jean-Philippe Bouchaud. 2015. Why Do Markets Crash? Bitcoin Data Offers Unprecedented Insights. *PLOS ONE* 10, 10 (2015), 1–11.
- [10] Huu Nhan Duong, Petko S. Kalev, and Chandrasekhar Krishnamurti. 2009. Order aggressiveness of institutional and individual investors. *Pacific-Basin Finance Journal* 17, 5 (2009), 533–546.
- [11] Abeer ElBahrawy, Laura Alessandretti, Anne Kandler, Romualdo Pastor-Satorras, and Andrea Baronchelli. 2017. Evolutionary dynamics of the cryptocurrency market. *Royal Society Open Science* 4, 11 (2017), 170623.
- [12] Jeff Fleming, Chris Kirby, and Barbara Ostdiek. 2003. The economic value of volatility timing using “realized” volatility. *Journal of Financial Economics* 67, 3 (2003), 473–509.
- [13] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.
- [14] João Gama, Indrè Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)* 46, 4 (2014), 44.
- [15] David Garcia and Frank Schweitzer. 2015. Social signals and algorithmic trading of Bitcoin. *Royal Society Open Science* 2, 9 (2015), 150288.
- [16] D. Garcia, C. J. Tessone, P. Mavrodiev, and N. Perony. 2014. The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of The Royal Society Interface* 11, 99 (2014), 20140623–20140623.
- [17] David Hallac, Sagar Vare, Stephen Boyd, and Jure Leskovec. 2017. Toeplitz inverse covariance-based clustering of multivariate time series data. In *SIGKDD*. ACM, 215–223.
- [18] Peter R. Hansen and Asger Lunde. 2005. A forecast comparison of volatility models: does anything beat a GARCH(1, 1)? *Journal of Applied Econometrics* 20, 7 (2005), 873–889.
- [19] Andrew C Harvey and Tirthankar Chakravarty. 2008. Beta-t-(e) garch. (2008).
- [20] Adam Hayes. 2015. Cryptocurrency Value Formation: An Empirical Analysis Leading to a Cost of Production Model for Valuing Bitcoin. *SSRN Electronic Journal* (2015).
- [21] Dirk Helbing. 2013. Globally networked risks and how to respond. *Nature* 497, 7447 (2013), 51–59.
- [22] Bryan Hooi, Shenghua Liu, Asim Smailagic, and Christos Faloutsos. 2017. BeatLex: Summarizing and Forecasting Time Series with Patterns. In *ECML/PKDD*. Springer, 3–19.
- [23] Rob J Hyndman and George Athanasopoulos. 2014. *Forecasting: principles and practice*. OTexts.
- [24] Pankaj K. Jain, Pawan Jain, and Thomas H. McInish. 2011. The Predictive Power of Limit Order Book for Future Volatility, Trade Price, and Speed of Trading. *SSRN Electronic Journal* (2011).
- [25] Paraskevi Katsiampa. 2017. Volatility estimation for Bitcoin: A comparison of GARCH models. *Economics Letters* 158 (2017), 3–6.
- [26] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeon Kim, Shin Jin Kang, and Chang Hun Kim. 2016. Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies. *PLOS ONE* 11, 8 (2016), e0161197.
- [27] Gebhard Kirchgässner and Jürgen Wolters. 2007. *Introduction to modern time series analysis*. Springer Science & Business Media.
- [28] Daniel Kondor, Marton Posfai, Istvan Csabai, and Gabor Vattay. 2014. Do the Rich Get Richer? An Empirical Analysis of the Bitcoin Transaction Network. *PLOS ONE* 9, 2 (2014), 1–10.
- [29] Ladislav Kristoufek. 2015. What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis. *PLOS ONE* 10, 4 (2015), e0123923.
- [30] Nikolay Laptev, Jason Yosinski, Li Erran Li, and Slawek Smyl. 2017. Time-series extreme event forecasting with neural networks at Uber. In *ICML*.
- [31] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.
- [32] Jianxu Liu, Songsak Sriboonchitta, Panisara Phochanachan, and Jiechen Tang. 2015. Volatility and Dependence for Systemic Risk Measurement of the International Financial System. In *Lecture Notes in Computer Science*. Springer International Publishing, 403–414.
- [33] Yan Liu, Alexandru Niculescu-Mizil, Aurelie C Lozano, and Yong Lu. 2010. Learning temporal causal graphs for relational time-series analysis. In *ICML*. 687–694.
- [34] Christopher Meek, David Maxwell Chickering, and David Heckerman. 2002. Autoregressive tree models for time-series analysis. In *SDM*. SIAM, 229–244.
- [35] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.
- [36] Randi Næs and Johannes A. Skjeltorp. 2006. Order book characteristics and the volume–volatility relation: Empirical evidence from a limit order market. *Journal of Financial Markets* 9, 4 (2006), 408–432.
- [37] Satoshi Nakamoto. 2008. Bitcoin: A peer-to-peer electronic cash system. (2008). <http://bitcoin.org/bitcoin.pdf>
- [38] Roberto Pascual and David Veredas. 2008. Does the Open Limit Order Book Matter in Explaining Informational Volatility? *SSRN Electronic Journal* (2008).
- [39] Matija Piskorec, Nino Antulov-Fantulin, Petra Kralj Novak, Igor Mozetič, Miha Grčar, Irena Vodenska, and Tomislav Šmuc. 2014. Cohesiveness in Financial News and its Relation to Market Volatility. *Scientific Reports* 4, 1 (2014).
- [40] Jiezhong Qiu, Jie Tang, Tracy Xiao Liu, Jie Gong, Chenhui Zhang, Qian Zhang, and Yufei Xue. 2016. Modeling and predicting learning behavior in MOOCs. In *WSDM*. ACM, 93–102.
- [41] Kira Radinsky, Krysta Svore, Susan Dumais, Jaime Teevan, Alex Bocharov, and Eric Horvitz. 2012. Modeling and predicting behavioral dynamics on the web. In *WWW*. ACM, 599–608.
- [42] Steven L Scott and Hal R Varian. 2014. Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation* 5, 1-2 (2014), 4–23.
- [43] Jianing V Shi, Yangyang Xu, and Richard G Baraniuk. 2014. Sparse bilinear logistic regression. *arXiv preprint arXiv:1404.4104* (2014).
- [44] Brandon K Vaughn. 2008. Data analysis using regression and multi-level/hierarchical models, by Gelman, A., & Hill, J. 45, 1 (2008), 94–97.
- [45] Xiaozhe Wang, Kate Smith, and Rob Hyndman. 2006. Characteristic-based clustering for time series data. *Data mining and knowledge Discovery* 13, 3 (2006), 335–364.
- [46] Yue Wu, José Miguel Hernández-Lobato, and Zoubin Ghahramani. 2014. Gaussian process volatility model. In *NIPS*. 1044–1052.
- [47] Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. 2012. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems* 23, 8 (2012), 1177–1193.

Table 4: Test errors over each time interval of rolling learning (MAE)

Interval Model	1	2	3	4	5	6	7	8	9	10	11	12
EWMA	0.047	0.063	0.158*	0.075*	0.050	0.018	0.018*	0.019*	0.026*	0.030*	0.032*	0.040*
GARCH	0.99**	0.133**	0.333**	0.147**	0.091**	0.019**	0.029**	0.035**	0.054**	0.075**	0.057**	0.084**
BEGARCH	0.96**	0.131**	0.330**	0.142**	0.087**	0.016**	0.027**	0.033**	0.052**	0.071**	0.053**	0.079**
STR	0.065**	0.105**	0.254**	0.141**	0.051	0.028**	0.040**	0.022**	0.036**	0.220**	0.034*	0.065**
ARIMA	0.059**	0.091**	0.255**	0.123**	0.053*	0.115**	0.020**	0.022**	0.037**	0.204**	0.032**	0.051**
ARIMAX	0.083**	0.172**	0.295**	0.140**	0.092**	0.050**	0.027**	0.030**	0.045**	0.236**	0.042**	0.062**
STRX	4.03**	1.520**	7.219**	11.963**	15.510**	5.068**	2.207**	1.953**	3.514**	1.759**	2.402**	3.81**
GBT	0.052**	0.082**	0.162**	0.094**	0.061**	0.062**	0.040**	0.020**	0.030**	0.034**	0.036*	0.041*
RF	0.052**	0.089**	0.174**	0.095**	0.060**	0.071**	0.031**	0.018	0.031**	0.031**	0.037**	0.040*
XGT	0.047*	0.077**	0.159**	0.080**	0.057**	0.051**	0.031**	0.020**	0.029**	0.031**	0.034**	0.041*
ENET	0.052**	0.074**	0.166**	0.093**	0.064**	0.036**	0.024**	0.017	0.028**	0.030**	0.032**	0.035
GP	0.050**	0.070**	0.216**	0.084*	0.069**	0.055**	0.025**	0.018	0.028**	0.034**	0.037**	0.035*
TM-LOG	0.053	0.078	0.176	0.70	0.061	0.018	0.024	0.041	0.037	0.037	0.103	0.053
TM-G	0.044	0.063	0.153	0.079	0.050	0.014	0.017	0.017	0.024	0.026	0.029	0.032

Symbols * and ** respectively indicate that the error of TM-G in the table is significantly different from the corresponding one at level 5% and 1% .

Table 5: Test errors over each time interval of incremental learning (MAE)

Model	7	8	9	10	11	12
EWMA	0.018	0.019*	0.026*	0.030*	0.032*	0.040**
GARCH	0.028**	0.035**	0.054**	0.075**	0.057*	0.083**
BEGARCH	0.025**	0.032**	0.051**	0.071**	0.053*	0.080**
STR	0.022*	0.030**	0.037**	0.179**	0.032*	0.054**
ARIMA	0.020*	0.021*	0.037**	0.215**	0.032*	0.073**
ARIMAX	0.037**	0.037**	0.078**	0.291**	0.064**	0.089**
STRX	1.843**	5.088**	1.164**	3.05**	3.770**	2.110*
GBT	0.025**	0.027**	0.053**	0.040**	0.039*	0.040**
RF	0.034**	0.031**	0.039**	0.041**	0.048**	0.048**
XGT	0.023**	0.021*	0.032**	0.030*	0.032*	0.037*
ENET	0.026**	0.025**	0.041**	0.038**	0.052**	0.052**
GP	0.028**	0.028**	0.050**	0.039**	0.040*	0.042**
TM-LOG	0.039	0.034	0.098	0.067	0.381	0.150
TM-G	0.017	0.017	0.024	0.027	0.030	0.032