Personal and Ubiquitous Computing (2023) 27:1585–1593
https://doi.org/10.1007/s00779-021-01651-z

ORIGINAL ARTICLE

The cross-interval price impact model and its empirical analysis on cryptocurrency order book

Bin Teng[1,2] · Sicong Wang[1,2] · Qinghua Ren[1,2] · Qi Hao[1,2] · Yufeng Shi[1,2]

Received: 31 May 2021 / Accepted: 31 August 2021 / Published online: 22 September 2021

Abstract
The demand for high-frequency algorithmic trading in the cryptocurrency markets is driving the research of price impact mechanisms. We propose the cross-interval price impact model (CIPIM) to explore the advanced or delayed price impact of order book events. The results of the empirical analysis show that neural network structures such as long short-term memory (LSTM) as a specific implementation of CIPIM obtain better concurrent interpretation on price impact than order flow imbalance (OFI) in Cont et al. (*J Financ Economet* 12(1):47–88, 2014). Meanwhile, the classification version of CIPIM that predicts the direction of Bitcoin price changes tends to work to some extent.

Keywords Cross-interval price impact model (CIPIM) · Limit order book (LOB) · Cryptocurrency · Deep learning · Long short-term memory (LSTM)

1 Introduction

It is known that digital currencies could be regarded as an alternative to cash but they have the advantage of real-time transaction and quick ownership transfer, as they exist in digital or electronic form. Unlike the physical currency, digital currencies have lower transaction costs, higher liquidity, and higher security without disclosure of personal information. Especially, cryptocurrency is a virtual currency which uses encryption techniques based on decentralized network. It is notable that Bitcoin, Ethereum and the central bank digital currency are exactly cryptocurrencies. In this paper, we choose one of the most popular cryptocurrencies, Bitcoin, to study its price impact of order book events, which would be likely benefit the fields of algorithm trading in cryptocurrency market.

A lot of studies focused on modeling the dependence of the price fluctuations on traded volumes (Bouchaud [3], Hiemstra and Jones [10], Karpoff [13]). Remarkably, it was found that signed volume (sell/buy) could not be used to predict the future or option prices (Schlag and Stoll [19]). Potters and Bouchaud [17] explained the logarithmic relationship between price response and volumes.

Furthermore, to derive deeper implication for the price movement, many scholars have been trying to explore financial market activities from a microstructure perspective. The issue of how prices response to different types of order book events has been the focus of a stream of papers. For example, Biais et al. [1], Farmer et al. [6] and Mu et al. [16] presented some economic and statistical properties of order flow dynamics and also provided their insights on price information and market liquidity. Weber and Rosenow [25] have suggested that the information involved in the order book could explain the price fluctuations only partially, between which there exists a strong anticorrelation.

Most of the previous literature mainly focus on the analysis of the market order (Rosenow [18] and Hasbrouck [9]). Eisler et al. [7] also considered the impact of other several order book events (e.g., limit order and cancellation) on price changes and showed their cross-correlation empirically. Chordia and Subrahmanyam [4] derived implications for the relation between imbalances and price movements by developing an intertemporal model of how prices react to imbalances. Furthermore, a simplified model describing the linear price impact of order book events was proposed by Cont et al. [5], who introduced the order flow imbalance (OFI). Based on the OFI indicator, Wang et al. [24] proposed a stationarized log-OFI indicator by observing the characteristics of high-frequency data.

✉ Yufeng Shi
yfshi@sdu.edu.cn

1 Institute for Financial Studies and School of Mathematics, Shandong University, Jinan 250100, China

2 Shandong Big Data Research Association, Jinan 250100, China

In recent years, in order to capture more complex feature of limit order books (LOB), data-driven models have been applied for order books via deep learning. Sirignano and Cont [21], Tashiro et al. [23] and Zhang et al. [26] use the state of the order book as the feature of the neural network and apply different types of network architectures (CNN, LSTM, etc.) to forecast the next price movement. Especially, it has been shown that this nonparametric approach exactly proves the existence of a nonlinear relationship between order flow and price changes (see Sirignano and Cont [21]).

With digital currency attracted considerable attention in recent years and its distinctive characterism, there are more and more researches related to virtual currency in microstructure market (McIntyre and Harjes [15], Bianchi and Dickerson [2]). Based on Cont et al. [5], when OFI was applied to the cryptocurrency market, Silantyev [20] found that trade flow imbalance outperformed order flow imbalance at explaining contemporaneous price changes. Fang et al. [8] trained the LSTM networks to predict the mid-price move in cryptocurrency market.

In this paper, we try to study the impact of order book events on price changes in different nearby intervals. This can be described as an advance or delay mechanism for price changes. Combining the ideas of order book event inscription in Cont et al. [5], we propose the cross-interval price impact model (CIPIM). To find the optimal features of order book events and the optimal price impact structure, we design the feature extractor for order book events in the form of long short-term memory (LSTM) layer, and price response model in the form of multi-layer perceptron (MLP) neural network, with a data-driven optimization search. With the help of high-frequency data from the Bitcoin market, we empirically demonstrate the usefulness of CIPIM.

The rest of the paper is structured as follows. In Section 2 we propose the cross-interval price impact model (CIPIM) and its deep learning-based implements. The data source and model parameter settings are provided in Section 3. Section 4 presents our main empirical results on BTC order book. Section 5 concludes.

2 The model

Suppose at a given time t, M-level limit order book (LOB) can be observed. We denote the level-m ask price, level-m bid price as $p_t^{m,s}$, $p_t^{m,b}$, and the corresponding size as $q_t^{m,s}$, $q_t^{m,b}$, $m=1,2,\ldots,M$. For a given time interval $[t_{k-1},t_k]$, we have a vector of state variables of order flows (price and volume), denoted as $\mathcal{X}^M_{[t_{k-1},t_k]}=\{(p_t^{1,s},q_t^{1,s},p_t^{1,b},q_t^{1,b},\ldots,p_t^{M,s},q_t^{M,s},p_t^{M,b},q_t^{M,b})^\top\}_{t\in[t_{k-1},t_k]}$. Consider another given time interval $[t_{j-1},t_j]$, notably it is not the same as the interval mentioned above, the mid-price changes can be determined by $\Delta p_{[t_{j-1},t_j]}=\frac{1}{2}(p_{t_j}^{1,s}-p_{t_{j-1}}^{1,s}+p_{t_j}^{1,b}-p_{t_{j-1}}^{1,b})$. Let $\mathbf{1}_{\{\Delta p_{[t_{j-1},t_j]}\geq 0\}}$ denote the direction of the price move, where $\mathbf{1}$ presents the 0-1 indicator function.

2.1 The cross-interval price impact model

In order to explore the impact mechanism of partial information involved in order flows $\mathcal{X}^M_{[t_{k-1},t_k]}$ on price dynamics $\Delta p_{[t_{j-1},t_j]}$, we propose the *cross-interval price impact model* (CIPIM) as follows:

$$\Delta p_{[t_{j-1},t_j]}=f(g(\mathcal{X}^M_{[t_{k-1},t_k]}),\tau_b,\tau_f)+\epsilon, \tag{1}$$

where $\tau_b=t_-t_{j-1}$, $\tau_f=t_j-t_k$, $\tau_b+\tau_f>0$, ϵ is a random error term. Function g should be specially designed to describe the features of order book events at interval $[t_{k-1},t_k]$, and the mapping f further depicts the price impact over $[t_{j-1},t_j]$ from order book events g. Remarkably, there is neither constraint $[t_{j-1},t_j]=[t_{k-1},t_k]$ nor $[t_{j-1},t_j]\cap[t_{k-1},t_k]=\emptyset$, which is the reason why we call it cross-interval price impact. In other words, we are trying to figure out the implicit advanced or delayed price response, not just the contemporaneous (or instantaneous) effects. When selecting $t_{j-1}>t_{k-1}$ (i.e., $\tau_b<t_k-t_{k-1}$), the order book events described by the model may be a leading indicator of price changes; Conversely, when selecting $t_{j-1}<t_{k-1}$ (i.e., $\tau_b>t_k-t_{k-1}$), the corresponding order book events tend to follow after the price changes. Specifically, when $t_{j-1}\geq t_k$ (i.e., $\tau_b\leq 0$), the model (1) represents the future prediction of price change $\Delta p_{[t_{j-1},t_j]}$ based on the order book event $g(\mathcal{X}^M_{[t_{k-1},t_k]})$.

Function g is assumed to extract the features of a series of consecutive order book events. In the model form (1), the event features extracted by g are independent of τ_b and τ_f. This is very useful when studying the relationship between certain specific events and price movements in different time intervals. If the model (1) is modified to the form $\Delta p_{[t_{j-1},t_j]}=f(\mathcal{X}^M_{[t_{k-1},t_k]},\tau_b,\tau_f)+\epsilon$, it is difficult to explicitly identify an order book event that is independent of the relative positions of the cross-intervals (τ_b and τ_f).

In the following, we will first provide two specific examples to illustrate the functionality of g and f. In Section 2.2 we will present the deep learning-based structure.

Example 1 A self-explanatory model structure

Consider $\forall\tau_b\in[0,t_k-t_{k-1}]$, $\tau_f\in[t_{k-1}-t_k,0]$ which satisfy $\tau_b+\tau_f>0$, and define

$$g(\mathcal{X}^M_{[t_{k-1},t_k]})=\{(p_t^{1,s},p_t^{1,b})^\top\}_{t\in[t_{k-1},t_k]},$$
$$f(g(\mathcal{X}^M_{[t_{k-1},t_k]}),\tau_b,\tau_f)=\frac{1}{2}(p^{1,s}_{t_k+\tau_f}-p^{1,s}_{t_k-\tau_b}+p^{1,b}_{t_k+\tau_f}-p^{1,b}_{t_k-\tau_b}).$$

In this case $\Delta p_{[t_{j-1},t_j]}\equiv f(g(\mathcal{X}^M_{[t_{k-1},t_k]}),\tau_b,\tau_f)$, for any target interval $[t_{j-1},t_j]\subset[t_{k-1},t_k]$. This is due to the fact that the defined event $g(\mathcal{X}^M_{[t_{k-1},t_k]})$ contains all the information to inscribe the price movement in the target interval $[t_{j-1},t_j]$.

Example 2 Order flow imbalance (OFI)

Order flow imbalance (OFI) and its price impact linear regression, proposed by Cont et al. [5], can be seen as an implementation of CIPIM, albeit under the condition that $[t_{j-1}, t_j] = [t_{k-1}, t_k]$. Based on the supply/demand contributions of the order events, OFIs are calculated according to the following formula:

$$\begin{aligned} g_{\mathrm{OFI}}(\mathcal{X}^M_{[t_{k-1},t_k]}) = \sum_{t\in(t_{k-1},t_k]} \Big\{ & q_t^{1,b}\mathbf{1}_{\{p_t^{1,b}\geq p_{t-\delta_t}^{1,b}\}} \\ & -q_{t-\delta_t}^{1,b}\mathbf{1}_{\{p_t^{1,b}\leq p_{t-\delta_t}^{1,b}\}} - q_t^{1,s}\mathbf{1}_{\{p_t^{1,s}\leq p_{t-\delta_t}^{1,s}\}} \\ & +q_{t-\delta_t}^{1,s}\mathbf{1}_{\{p_t^{1,s}\geq p_{t-\delta_t}^{1,s}\}} \Big\}, \end{aligned}$$

where $t-\delta_t$ denotes the previous moment adjacent to t in the observed data. Cont et al. [5] established a linear regression between OFIs and mid-price changes, which is equivalent to:

$$f(g_{\mathrm{OFI}}(\mathcal{X}^M_{[t_{k-1},t_k]}), t_k - t_{k-1}, 0) = \alpha + \beta g_{\mathrm{OFI}}(\mathcal{X}^M_{[t_{k-1},t_k]}).$$

By observing the formula of g_{OFI}, it can be found that although OFIs do not contain all the information of price changes as in Example 1, the indicator functions still characterize the direction of price changes accurately. This is the main reason why the concurrent linear function f is able to obtain a high goodness of fit.

Cont et al. [5] conducted the high-frequency indicator OFI calculated by a series of order prices and sizes as an order book event function, and analyzed the linear relationship between OFIs and price changes. However, it is natural to consider whether it is possible to find another event function that utilizes the same market information (or even less information) as OFI, while obtaining a more powerful explanation of price dynamics. Obviously, OFI might not be the best one. Thus we do think determining the shape of the event generating function g is a necessary phase in the price adjustments. In order to find the optimal event function g and price impact function f, represented by LOB observation states (order prices and sizes), we will make use of end-to-end neural networks and their gradient methods, as mentioned in the algorithm in the next section.

In essence, we desire to figure out the relationship between order book events and price changes, that is to say, the explanatory powers of events occurring at certain timescale to the price changes need to be discussed. In addition, how the model behaves if the range of price changing is not placed restrictions on, adjusting τ_f and τ_b randomly, is also our concerns. Based on the above consideration, we design this price impact mechanism, extending the two special cases mentioned in Cont et al. [5], to study the impacts of order book events on price changes in different nearby intervals. The empirical results are shown in the following, where we can see the varying impacts over different intervals clearly.

2.2 Model specification based on neural network

We consider a nonparametric approach to specify g and f in the CIPIM model (1) using the universal approximation capability of neural networks, and give the steps for model parameter estimation.

Let the matrix form $\mathbf{X}_{[t_{k-1},t_k]} \in \mathbb{R}^n \times \mathbb{R}^d$ consist of the subset of LOB observation states $\mathcal{X}^M_{[t_{k-1},t_k]}$ and the corresponding indicators (e.g., $\{\mathbf{1}_{\{p_t^{m,s}\leq p_{t-\delta_t}^{m,s}\}}\}_{t\in[t_{k-1},t_k]}$), where n is the number of order book events in the interval $[t_{k-1}, t_k]$, d represents dimension of states variables at each quote.

We use the recurrent neural network long short-term memory (LSTM) to implement the function g. To avoid overfitting, dropout layers are set before the input and after the output of the LSTM layer.[1] Thus, the function g can be structured as:

$$\mathbf{H}_X = \mathrm{dropout}(\mathbf{X}_{[t_{k-1},t_k]}), \tag{2}$$

$$\mathbf{h}_{\mathrm{LSTM}} = \mathrm{LSTM}(\mathbf{H}_X; \boldsymbol{\Theta}), \tag{3}$$

$$\mathbf{g} = \mathrm{dropout}(\mathbf{h}_{\mathrm{LSTM}}), \tag{4}$$

i.e., $\mathbf{g} = g(\mathbf{X}_{[t_{k-1},t_k]}; \boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ represents all the parameters in the LSTM network.[2]

Next, function f can be represented by a multi-layer perceptron (MLP) as follows:

$$\mathbf{g}_\tau = [\mathbf{g}^\top, \tau_b, \tau_f]^\top, \tag{5}$$

$$\mathbf{h}_g = \mathrm{ReLU}(\mathbf{W}_g \cdot \mathbf{g}_\tau + \mathbf{b}_g), \tag{6}$$

$$\widehat{\Delta p} = \mathbf{w}_f^\top \cdot \mathbf{h}_g + b_f, \tag{7}$$

Notice that the input layer $\mathbf{g}_\tau$ consists of $\mathbf{g}$, τ_b, τ_f. We could get the final output $\widehat{\Delta p} = f(\mathbf{g}, \tau_b, \tau_f; \boldsymbol{\Lambda})$, where the set of parameters $\boldsymbol{\Lambda} = \{\mathbf{W}_g, \mathbf{b}_g, \mathbf{w}_f, b_f\}$.

To find the optimal deep learning-based event function g and price impact function f based on the observed LOB data, we provide the following model training procedure.

[1] See more details about random dropout in Hinton et al. [11] and Srivastava et al. [22].

[2] LSTM can be formally formulated as:

$$\begin{aligned} \mathbf{z}_t &= \sigma(\mathbf{W}_z \cdot [\mathbf{h}_{t-\delta_t}^\top, \mathbf{x}_t^\top]^\top + \mathbf{b}_z), \\ \mathbf{i}_t &= \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-\delta_t}^\top, \mathbf{x}_t^\top]^\top + \mathbf{b}_i), \\ \mathbf{o}_t &= \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-\delta_t}^\top, \mathbf{x}_t^\top]^\top + \mathbf{b}_o), \\ \tilde{\mathbf{c}}_t &= \tanh(\mathbf{W}_c \cdot [\mathbf{h}_{t-\delta_t}^\top, \mathbf{x}_t^\top]^\top + \mathbf{b}_c), \\ \mathbf{c}_t &= \mathbf{z}_t * \mathbf{c}_{t-\delta_t} + \mathbf{i}_t * \tilde{\mathbf{c}}_t, \\ \mathbf{h}_t &= \mathbf{o}_t * \tanh(\mathbf{c}_t). \end{aligned}$$

See Hochreiter and Schmidhuber [12] for more details. Combining the equation (3), $\{\mathbf{x}_t\}_{t\in[t_{k-1},t_k]}$ forms the input variable $\mathbf{H}_X$. t and $t-\delta_t$ denote the adjacent moments in $[t_{k-1}, t_k]$. The output of formula (3) $\mathbf{h}_{\mathrm{LSTM}} = \mathbf{h}_{t_k}$. The parameters of LSTM can be rewritten as $\boldsymbol{\Theta} = \{\mathbf{W}_z, \mathbf{b}_z, \mathbf{W}_i, \mathbf{b}_i, \mathbf{W}_o, \mathbf{b}_o, \mathbf{W}_c, \mathbf{b}_c\}$.

See Algorithm 1 for the pseudo-code of the training procedure.

Algorithm 1 Training procedure of regressive CIPIM.

Data: the training matrix $\mathbf{X} \in \mathbb{R}^N \times \mathbb{R}^d$, which consists of a series of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, and the corresponding mid-price series $\{p_1, p_2, \ldots, p_N\}$, where N represents the total number of quotes in the training set;

Input: number of quotes $n_{\Delta t}$ in each time interval, mini-batch size K, upper & lower limit $T^b_{\min}, T^b_{\max}, T^f_{\min}, T^f_{\max}$, maximum iteration limit n_T, and learning rate η;

Result: optimal network parameters $\{\boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*\}$;

Initialize network parameters $\{\boldsymbol{\Theta}_0, \boldsymbol{\Lambda}_0\}$;

for $i = 0, 1, 2, \ldots, n_T - 1$ **do**

  Randomly select a consecutive segment of length $n_{\Delta t} \times K$ from $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, denoted as $\{\mathbf{x}^{(i)}_1, \mathbf{x}^{(i)}_2, \ldots, \mathbf{x}^{(i)}_{n_{\Delta t} \times K}\}$;

  Randomly select $\tau_b \in [T^b_{\min}, T^b_{\max}]$, $\tau_f \in [T^f_{\min}, T^f_{\max}]$;

  for $k = 1, 2, \ldots, K$ **do**

    Build the k-th matrix $\mathbf{X}_k = \{\mathbf{x}^{(i)}_{n_{\Delta t} \times (k-1)+1}, \ldots, \mathbf{x}^{(i)}_{n_{\Delta t} \times k-1}, \mathbf{x}^{(i)}_{n_{\Delta t} \times k}\}$;

    Calculate the k-th cross-interval price change $\Delta p_k \leftarrow p_{n_{\Delta t} \times k - \tau_b} - p_{n_{\Delta t} \times k + \tau_f}$;

    Get the features of order book events $\mathbf{g}_k \leftarrow g(\mathbf{X}_k; \boldsymbol{\Theta}_i)$;

    Get the estimates of price impacts $\widehat{\Delta p}_k \leftarrow f(\mathbf{g}_k, \tau_b, \tau_f; \boldsymbol{\Lambda}_i)$;

  $\ell_{\text{MSE}} \leftarrow \frac{1}{K} \sum_{k=1}^{K} (\Delta p_k - \widehat{\Delta p}_k)^2$;

  $\boldsymbol{\Theta}_{i+1} \leftarrow \boldsymbol{\Theta}_i - \eta \nabla_{\boldsymbol{\Theta}} \ell_{\text{MSE}}$;

  $\boldsymbol{\Lambda}_{i+1} \leftarrow \boldsymbol{\Lambda}_i - \eta \nabla_{\boldsymbol{\Lambda}} \ell_{\text{MSE}}$;

Get the optimal parameters $\{\boldsymbol{\Theta}^*, \boldsymbol{\Lambda}^*\} \leftarrow \{\boldsymbol{\Theta}_{n_T}, \boldsymbol{\Lambda}_{n_T}\}$;

a. Randomly select K consecutive time interval $[t_{k-1}, t_k]$ from the data set, requiring that there contains the same number of observation data in any of the time intervals.[3]
b. Randomly select $\tau_b \in [T^b_{\min}, T^b_{\max}]$, $\tau_f \in [T^f_{\min}, T^f_{\max}]$, thus determining K corresponding time intervals $[t_{j-1}, t_j]$, and calculate the price changes within these intervals, denoted as $\{\Delta p_k\}_{k=1,\ldots,K}$.
c. Based on formula (2–7), calculate $\mathbf{g}_k, \widehat{\Delta p}_k$, $\forall k = 1, \ldots, K$.
d. The mean square error loss is defined according to the following formula:

$$\ell_{\text{MSE}} = \frac{1}{K} \sum_{k=1}^{K} (\Delta p_k - \widehat{\Delta p}_k)^2. \tag{8}$$

The model parameters $\{\boldsymbol{\Theta}, \boldsymbol{\Lambda}\}$ are trained by gradient-based method.
e. Repeat the steps above.

[3] The subscript t can be considered as the index of the quotes data. In this case, $\delta_t = 1$, and $n_k = t_k - t_{k-1} + 1$ counts the number of quote events in $[t_{k-1}, t_k]$. We will use this notation in the following.

Remark 1 If we consider using f to explain the directions of Δp, then a classification form of CIPIM can be established. Once the target is replaced by $\Delta p_{[t_{j-1}, t_j]}$ with $\mathbf{1}_{\{\Delta p_{[t_{j-1}, t_j]} \geq 0\}}$, the output layer (7) is correspondingly modified to

$$\tilde{f} = \sigma(\mathbf{w}_f^\top \cdot \mathbf{h}_g + b_f), \tag{9}$$

where $\sigma(x) = \frac{1}{1+e^{-x}} \in (0, 1)$. At this point $\tilde{f}$ fits the probability of $\Delta p_{[t_{j-1}, t_j]} \geq 0$.

Accordingly, the loss function (8) should be replaced by the average cross-entropy (ACE) loss as follows:

$$\ell_{\text{ACE}} = \frac{1}{K} \sum_{k=1}^{K} \left\{ \mathbf{1}_{\{\Delta p_k \geq 0\}} \log \tilde{f}_k + \mathbf{1}_{\{\Delta p_k < 0\}} \log(1 - \tilde{f}_k) \right\}. \tag{10}$$

3 Data and model settings

The data in this paper is collected from bitcoin (BTC) in Binance, the bid/ask tick size is 0.01 dollars, the smallest order size is 0.001 btc. BTC is traded 7 × 24 h, so there is no overnight jump in price and no special data processing is needed.

Throughout the empirical analysis, 9 h BTC high-frequency data in 2020-03-01 09:00–18:00 is selected, totally $N = 57,305$ snapshot quotes data. According to statistics, this dataset contains an average of 18 quotes snapshots per 10 s, so we set $n_{\Delta t} = t_k - t_{k-1} = 17$ (≈ 10 s), and set $K = 180$ which means about 30 min as the batch size. To validate the effectiveness of deep learning CIPIM, the first two-thirds of the data (9:00–15:00) are the training set and the last one-third of the data (15:00–18:00) are used as the validation set, and the results are presented in Section 4.

The parameter setting of deep learning model is explained in detail below. We will focus on the deep learning CIPIM model given two types of information, denoted as LSTM-Size and LSTM-Price+Size, respectively:

a. (M-level LSTM-Size) $\mathbf{X}_{[t_{k-1}, t_k]}$ consists of $\{(q_t^{m,s}, q_t^{m,b})^\top\}_{t \in [t_{k-1}, t_k]}^{m=1,\ldots,M}$.

b. (M-level LSTM-Price+Size) $\mathbf{X}_{[t_{k-1},t_k]}$ consists of $\{(q_t^{m,s}, q_t^{m,b})^\top\}_{t\in[t_{k-1},t_k]}^{m=1,\ldots,M}$ and $\{(\mathbf{1}_{\{p_t^{m,s}>p_{t-\delta_t}^{m,s}\}} - \mathbf{1}_{\{p_t^{m,s}<p_{t-\delta_t}^{m,s}\}}, \mathbf{1}_{\{p_t^{m,b}>p_{t-\delta_t}^{m,b}\}} - \mathbf{1}_{\{p_t^{m,b}<p_{t-\delta_t}^{m,b}\}})^\top\}_{t\in[t_{k-1},t_k]}^{m=1,\ldots,M}$.

Obviously, the information contained in LSTM-Price+Size is exactly consistent with that of OFI, while the LSTM-Size model tends to explain the price changes through the order size. In this paper, we make empirical analysis in the case of $M = 1$.

We choose $\tau_b = -n_{\Delta t}, -n_{\Delta t}+1, \ldots, 2n_{\Delta t}$ (i.e., $T_{\min}^b = -n_{\Delta t}$, $T_{\max}^b = 2n_{\Delta t}$), $\tau_f = -2n_{\Delta t}, -2n_{\Delta t}+1, \ldots, n_{\Delta t}$ (i.e., $T_{\min}^f = -2n_{\Delta t}$, $T_{\max}^f = n_{\Delta t}$), and $\tau_b + \tau_f > 0$, as the x/y-axis shown in Fig. 1. The hidden units size of both LSTM layer (3) and fully connected layer (6) in the neural networks are set to 5, small enough to avoid overfitting. The dropout rate is 0.2. Select Adam optimizer for stochastic optimization of neural network f and g, and the learning rate $\eta = 4 \times 10^{-3}$. The maximum iteration limit $n_T = 120,000$. See more details of Adam in Kingma and Ba [14].

4 Empirical results

4.1 Deep learning-based CIPIM versus OFI

Recall that OFI proposed by Cont et al. [5] is a concrete implementation of CIPIM. To demonstrate that the CIPIM model could work as a price impact model just as the model based on OFI, validity of this extended model is proved in this subsection. Our goal is to state the difference of the price impact between two models, driven by the order book information over the interval $[t_{k-}, t_k]$. The explanatory power of the two models is measured by the average R^2 defined as:

$$\text{Avg-}R^2 = \frac{1}{n_T}\sum_{i=1}^{n_T}\left\{1 - \frac{\sum_{k=1}^{K}(\Delta p_{i,k} - \widehat{\Delta p}_{i,k})^2}{\sum_{k=1}^{K}(\Delta p_{i,k} - \overline{\Delta p}_i)^2}\right\}, \quad (11)$$

where $\overline{\Delta p}$ denotes the average price change, and the subscript $i = 1, \ldots, n_T$ represents the index of samplings in either the training and validation set. From the definition of OFI and function $g(\mathcal{X}^M_{[t_{k-1},t_k]})$, it is obvious that two aspects of information over order flow are involved in both two models, i.e., the order size and the direction of the price movement.

Figure 1 presents four contour plots of the average R^2 of OFI-based model and LSTM-Price+Size in the training and validation sets, respectively. Recall that τ_b(x-axis), τ_f(y-axis) in Section 2 represents all possible price change interval we choose, i.e., the upper triangle part in Fig. 1. Especially, Cont et al. [5] only consider two extreme cases: one is that the price impact over the contemporaneous OFI, i.e., the OFI interval coincides exactly with the price change interval ($[t_{k-1}, t_k] = [t_{j-1}, t_j]$); following which, when $t_{j-1} > t_k$, the price changes in $[t_{j-1}, t_j]$ are predicted by the OFI in $[t_{k-1}, t_k]$. These two cases exactly correspond with two parts of Fig. 1, i.e., $\tau_b = 17$, $\tau_f = 0$ and $\tau_b = 0$, $\tau_f > 0$.

The contour plots of average R^2 in the training and validation sets show the similar distribution, so only the results on the validation set are analyzed in detail. The results in the validation set show that both the OFI-based model and LSTM-Price+Size have stronger explanatory power to the contemporaneous price impact, while behaving worse on the forecasting. The goodness of fit of the price dynamics on the contemporaneous OFI is 0.36, while R^2 of LSTM-Price+Size can reach at 0.6. When predicting the price changes over the interval $[t_{j-1}, t_j]$ ($t_{j-1} > t_k$), both two models have a poor performance and obtain a negative R^2 in the validation set. We speculate that there are two reasons for such results: First, price changes can not be predicted based on the information of ask/bid price and size; Second, perhaps on higher frequency data, our model might demonstrate the better performance of predicting price change. From the overall viewpoint, whether the interval of price dynamics is contemporaneous or not, the explanatory power of CIPIM is basically comparable to OFI, so it is reasonable to assume that CIPIM is valid.

4.2 Analysis of cross-interval price impacts

In Section 4.1, we consider the price impact on the order book events when the information contained in $g(\mathcal{X}^M_{[t_{k-1},t_k]})$ is the same as that of $g_{\text{OFI}}(\mathcal{X}^M_{[t_{k-1},t_k]})$. It is notable that taking the price information into account might cause the higher goodness of fit, which is natural as unknown price information is involved in implication variables. Hence, we remove the price information and consider the price impact on order size only. We make empirical analysis on two main issues, the impact of order size on price changes and the direction of the price move, which are so-called regression and classification model.

Consider the regression model, Fig. 2 displays the explanatory power of order book events on price changes over the all available intervals. From Fig. 2, obviously, given that $\tau_b = 17$ ($t_{j-1} = t_{k-1}$), with the time interval $[t_{j-1}, t_j]$ becomes longer, the average R^2 of Regression LSTM-Size increases firstly and then decreases, reaches maximum value around $\tau_f = 0$ ($[t_{k-1}, t_k] = [t_{j-1}, t_j]$), which implies that the price changes remain strong dependence on the contemporaneous order size. In addition, start with $\tau_f = 0$, when $\tau_f < 0$, contour lines of the average R^2 are found more dense, which means the average R^2 for the regression model declines faster with τ_f increases negatively. However, when $\tau_f < 0$, it shows that the average R^2

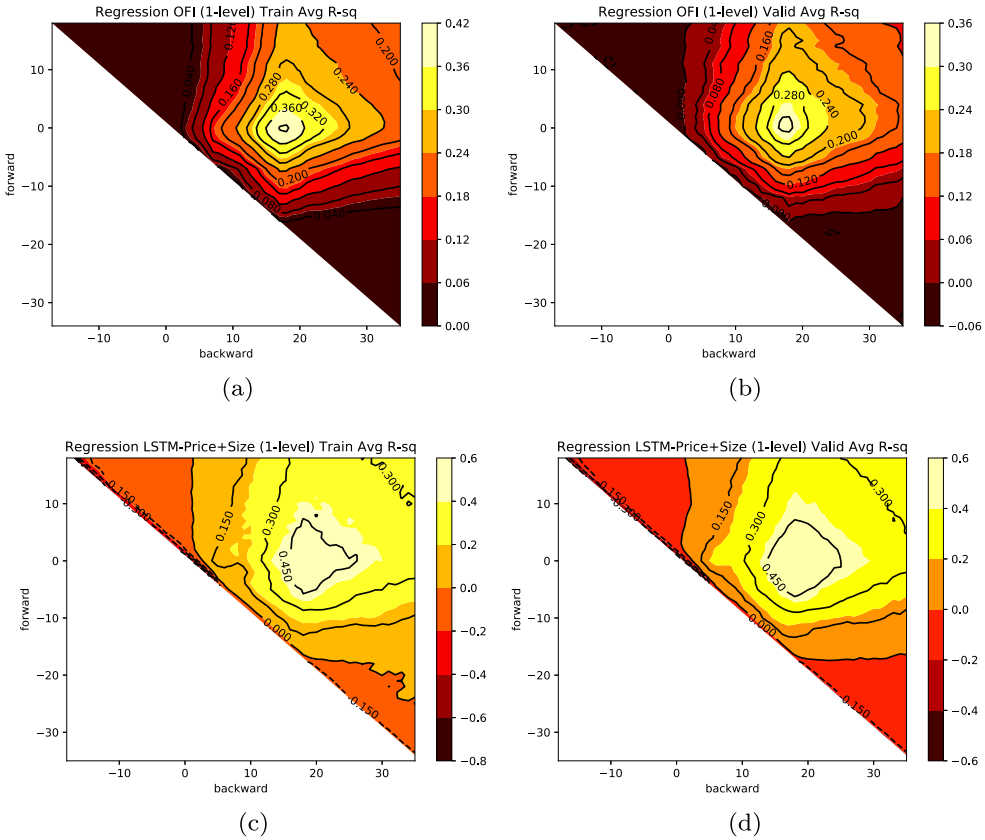

Fig. 1 Average R^2 contour plots of OFI and LSTM-Price+Size in the training and validation sets. (a) and (b) show the R^2 of the linear regression model of OFI and price change, and (c) and (d) show the goodness of fit of CIPIM using size and price information. The horizontal axis backward represents the left endpoint of the price change interval τ_b, and the vertical axis forward represents the right endpoint of the price change interval τ_f, each point in the figure represents a price change interval, and the degree of explanation of different price change intervals by order book events is shown by contour plots

declines relatively slower with τ_f increases positively, as contour lines seem more sparse. Similarly, given that $\tau_f = 0$ ($t_j = t_k$), with the time interval $[t_{j-1}, t_j]$ becomes longer, the average R^2 of Regression LSTM-Size also increases first and then decreases, reaches maximum value 0.1 around $\tau_b = 17$.

From the statistic of mean value in Table 1, we observe that the average R^2 of the Regression LSTM-Size and LSTM-Price+Size model at interval $t_{j-1} < t_{k-1} < t_j < t_k$ are both greater than the corresponding R^2 at interval $t_{k-1} < t_{j-1} < t_k < t_j$. Compared with the LSTM-Price+Size model, LSTM-Size model has significantly weak explanatory power in the contemporaneous case, but has greater R^2 in predicting the price changes, although they are both sightly less than zero.

For the classification problem, Fig. 3 demonstrates the prediction accuracy of the order book events for the direction of price change in all available intervals. The highest prediction accuracy of LSTM-Size model reaches 0.65 around $\tau_b = 17$, $\tau_f = 0$. The variation pattern of accuracy for different price intervals is similar to that of the regression problem. From Table 2, the average prediction accuracy for the interval $t_{j-1} < t_{k-1} < t_j < t_k$ is greater than that for the interval $t_{k-1} < t_{j-1} < t_k < t_j$. When $t_{j-1} \geq t_k$ (the event interval does not intersect with the price interval), the average prediction accuracy of LSTM-Size model is 0.523

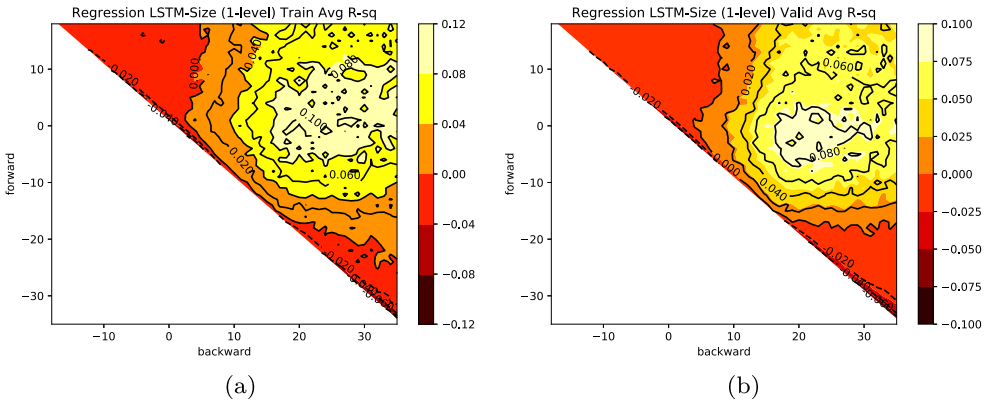

Fig. 2 (Regression) The average R^2 of CIPIM using size information to explain price changes

Table 1 (Regression) The statistical characteristics of CIPIM's Avg-R^2

		$t_{j-1} < t_{k-1} < t_j < t_k$ $(\tau_b > 17, -17 < \tau_f < 0)$		$t_{j-1} = t_{k-1}, t_j = t_k$ $(\tau_b = 17, \tau_f = 0)$		$t_{k-1} < t_{j-1} < t_k < t_j$ $(0 < \tau_b < 17, \tau_f > 0)$		$t_{j-1} \geq t_k$ $(\tau_b \leq 0, \tau_f > 0)$	
		Size	Price+Size	Size	Price+Size	Size	Price+Size	Size	Price+Size
1-level	mean	0.054	0.234	0.088	0.572	0.010	0.186	-0.012	-0.087
	max	0.098	0.566			0.073	0.509	-0.004	-0.014
	min	0.002	0.003			-0.018	-0.084	-0.045	-0.504
	std	0.025	0.146			0.022	0.129	0.006	0.114

Notes: Consider the four relationships between event intervals and price intervals: $t_{j-1} < t_{k-1} < t_j < t_k$ indicates that the price interval lags the event interval and there is an intersection between the two; $t_{j-1} = t_{k-1}, t_j = t_k$ indicates that the two intervals coincide exactly; $t_{k-1} < t_{j-1} < t_k < t_j$ indicates that the price interval exceeds the event interval and the two intersect. $t_{j-1} \geq t_k$ indicates that the price interval is ahead of the event interval and the two do not intersect

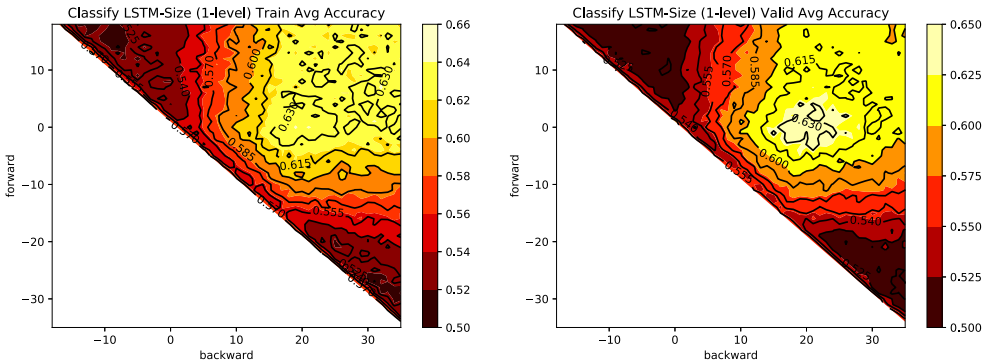

Fig. 3 (Classification) The average accuracy of CIPIM using size information to explain direction of price changes

Table 2 (Classification) Statistical characteristics of CIPIM's accuracies

		$t_{j-1} < t_{k-1} < t_j < t_k$ $(\tau_b > 17, -17 < \tau_f < 0)$		$t_{j-1} = t_{k-1}, t_j = t_k$ $(\tau_b = 17, \tau_f = 0)$		$t_{k-1} < t_{j-1} < t_k < t_j$ $(0 < \tau_b < 17, \tau_f > 0)$		$t_{j-1} \geq t_k$ $(\tau_b \leq 0, \tau_f > 0)$	
		Size	Price+Size	Size	Price+Size	Size	Price+Size	Size	Price+Size
1-level	mean	0.588	0.670	0.634	0.797	0.568	0.641	0.523	0.520
	max	0.644	0.811			0.623	0.774	0.582	0.581
	min	0.539	0.556			0.512	0.530	0.504	0.480
	std	0.025	0.060			0.027	0.053	0.016	0.019

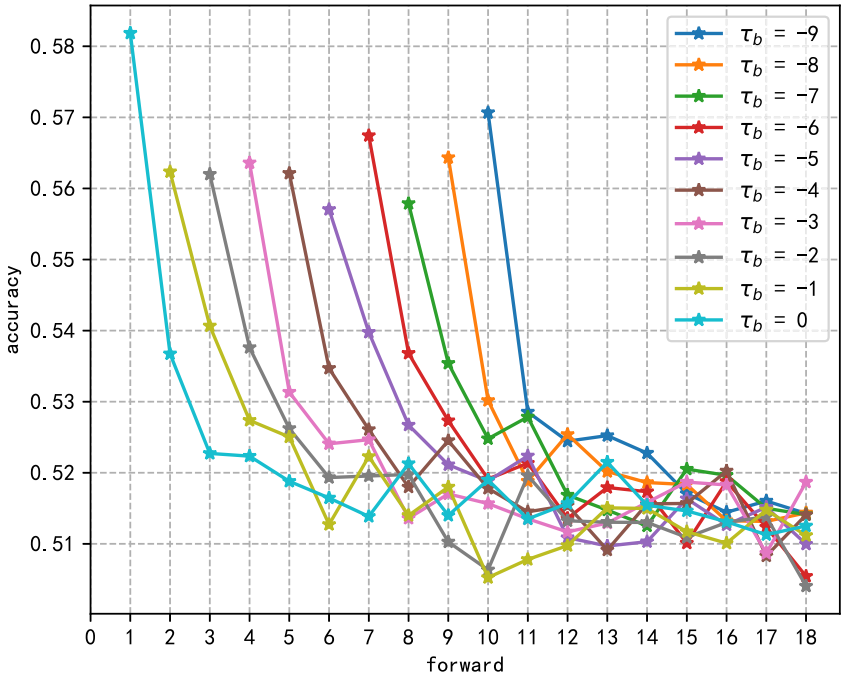

Fig. 4 (Classification) Results of the out-of-sample forecasting accuracy of CIPIM over different intervals. The horizontal axis is τ_f and the vertical axis is the prediction accuracy in the direction of price changes. We separately plot the relationship between the accuracy and τ_f when τ_b is fixed

(> 0.5) . This indicates that LSTM-Size model could be used to predict the direction of the price movement.

The prediction performance of the classification model on the validation set is shown in Fig. 4, which plots the trend of the prediction accuracy in the direction of price changes as the prediction interval becomes longer, under condition $\tau_b \leq 0$. When $\tau_b = 0$, $[t_{j-1}, t_j]$ does not intersect with $[t_{k-1}, t_k]$ and the highest prediction accuracy (≥ 0.58) is reached when $\tau_f = 1$. As the prediction interval becomes longer, the accuracy of the prediction decreases. Overall, a prediction accuracy greater than 0.5 can be obtained on all intervals. In addition, the first point of each curve roughly represents the price change in ≤ 1 s, but the prediction accuracy is highest in the 1 s closest to the present. This also means that the prediction of price direction within 1 s is a very interesting topic.

5 Conclusions

In this paper, we propose the cross-interval price impact model (CIPIM), and use the LSTM and MLP structure to explore price change mechanism of Bitcoin. We find that CIPIM with a combination of quoted size and price obtains a much higher concurrent R^2 range than OFI. For events defined only by order sizes, while concurrent R^2 is significantly lower, they have a higher predict R^2 than size+price events, albeit they are still slightly less than 0. The classification version of CIPIM is generally effective in predicting the direction of Bitcoin price movements. It is also necessary to point out that deep learning is only one implementation of CIPIM, which also has weaknesses such as unrobustness and poor interpretation. A more concise and effective implementation of CIPIM may be the direction of future works.

Acknowledgements The authors thank the editor and the three anonymous reviewers for reviewing this article for providing valuable suggestions.

Funding This work is supported by the National Key R&D Program of China (Grant No. 2018YFA0703900), and the National Natural Science Foundation of China (Grant Nos. 11871309, 11371226).

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Biais B, Hillion P, Spatt C (1995) An empirical analysis of the limit order book and the order flow in the Paris Bourse. J Finance 50(5):1655–1689
2. Bianchi D, Dickerson A (2018) Trading volume in cryptocurrency markets. Available at SSRN 3239670
3. Bouchaud JP (2010) Price impact. In Encyclopedia of Quantitative Finance, Cont R (Ed.)
4. Chordia T, Subrahmanyam A (2004) Order imbalance and individual stock returns: Theory and evidence. J Financ Econ 72(3):485–518
5. Cont R, Kukanov A, Stoikov S (2014) The price impact of order book events. J Financ Econometr 12(1):47–88
6. Farmer JD, Gillemot L, Lillo F, Mike S, Sen A (2004) What really causes large price changes? Quant Finance 4(4):383–397

7. Eisler Z, Bouchaud JP, Kockelkoren J (2012) The price impact of order book events: market orders, limit orders and cancellations. Quant Finance 12(9):1395–1419
8. Fang F, Chung W, Ventre C, Basios M, Kanthan L, Li L, Wu F (2021) Ascertaining price formation in cryptocurrency markets with machine learning. Eur J Financ (online)
9. Hasbrouck J (1991) Measuring the information content of stock trades. J Finance 46(1):179–207
10. Hiemstra C, Jones JD (1994) Testing for linear and nonlinear Granger causality in the stock price-volume relation. J Finance 49(5):1639–1664
11. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580
12. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
13. Karpoff JM (1987) The relation between price changes and trading volume: A survey. J Financ Quant Anal 22(1):109–126
14. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
15. McIntyre KH, Harjes K (2016) Order flow and the bitcoin spot rate. Appl Econ Finance 3(3):136–147
16. Mu GH, Zhou WX, Chen W, Kertész J (2010) Order flow dynamics around extreme price changes on an emerging stock market. New J Phys 12(7):075037
17. Potters M, Bouchaud JP (2003) More statistical properties of order books and price impact. Phys A 324(1-2):133–140
18. Rosenow B (2002) Fluctuations and market friction in financial trading. Internat J Modern Phys C 13(03):419–425
19. Schlag C, Stoll H (2005) Price impacts of options volume. J Financ Mark 8(1):69–87
20. Silantyev E (2019) Order flow analysis of cryptocurrency markets. Digital Finance 1(1):191–218
21. Sirignano J, Cont R (2019) Universal features of price formation in financial markets: perspectives from deep learning. Quant Finance 19(9):1449–1459
22. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
23. Tashiro D, Matsushima H, Izumi K, Sakaji H (2019) Encoding of high-frequency order information and prediction of short-term stock price by deep learning. Quant Finance 19(9):1499–1506
24. Wang Q, Teng B, Hao Q, Shi Y (2021) High-frequency statistical arbitrage strategy based on stationarized order flow imbalance. Procedia Computer Science 187:518–523
25. Weber P, Rosenow B (2005) Order book approach to price impact. Quant Finance 5(4):357–364
26. Zhang Z, Zohren S, Roberts S (2019) Deeplob: Deep convolutional neural networks for limit order books. IEEE Trans Signal Process 67(11):3001–3012