Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 7854512, 11 pages
https://doi.org/10.1155/2021/7854512

Research Article

Improved ACD-Based Financial Trade Durations Prediction Leveraging LSTM Networks and Attention Mechanism

Yong Shi,[1,2,3,4] Wei Dai,[1,2,3] Wen Long,[1,2,3] and Bo Li[1,2,3]

[1]*School of Economics and Management, University of Chinese Academy of Sciences, No. 80 of Zhongguancun East Street Haidian District, Beijing 100190, China*
[2]*Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, No. 80 of Zhongguancun East Street, Haidian District, Beijing 100190, China*
[3]*Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, No. 80 of Zhongguancun, East Street, Haidian District, Beijing 100190, China*
[4]*College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA*

Correspondence should be addressed to Wen Long; longwen@ucas.ac.cn

Received 11 August 2020; Revised 11 December 2020; Accepted 16 January 2021; Published 30 January 2021

Academic Editor: Bekir Sahin

The liquidity risk factor of security market plays an important role in the formulation of trading strategies. A more liquid stock market means that the securities can be bought or sold more easily. As a sound indicator of market liquidity, the transaction duration is the focus of this study. We concentrate on estimating the probability density function $p(\Delta t_{i+1} \mid G_i)$, where Δt_{i+1} represents the duration of the $(i+1)$-th transaction and G_i represents the historical information at the time when the $(i+1)$-th transaction occurs. In this paper, we propose a new ultrahigh-frequency (UHF) duration modelling framework by utilizing long short-term memory (LSTM) networks to extend the conditional mean equation of classic autoregressive conditional duration (ACD) model while retaining the probabilistic inference ability. And then, the attention mechanism is leveraged to unveil the internal mechanism of the constructed model. In order to minimize the impact of manual parameter tuning, we adopt fixed hyperparameters during the training process. The experiments applied to a large-scale dataset prove the superiority of the proposed hybrid models. In the input sequence, the temporal positions which are more important for predicting the next duration can be efficiently highlighted via the added attention mechanism layer.

1. Introduction

Market liquidity refers to the degree to which an asset can be bought and sold easily for a fair price [1]. In other words, market liquidity can be regarded as the speed at which transactions can be concluded while maintaining a basically stable price [1]. Therefore, market liquidity risk is one of the most common factors considered by security investors especially by high-frequency traders in building a trading strategy.

With the rapid development of computer storage technology, transaction by transaction financial trading data is accessible to researchers. Let t_i stand for the time at which the $i-$ thtrade occurs so that the duration between the $(i+1)$-th and i-th trade is $\Delta t_{i+1} = t_{i+1} - t_i$, which can directly measure the transaction speed of financial trading. The autoregressive conditional duration (ACD) model proposed by Engle and Russell has been the primary framework used for analyzing trading durations of ultrahigh-frequency (UHF) data, which are irregularly time-spaced and convey meaningful information [2]. In ACD models, the transaction duration is decomposed into the multiplicative product of two components: the conditional (expected) duration and the unexpected duration. The expected component is the portion of transaction duration that is linearly conditional on past durations, whereas the unexpected duration is the fraction of duration beyond that which could be predicted from past durations and is usually characterized by an exponential distribution.

Based on the work of Engle and Russell [2], many works tried to improve the ability of capturing the relation between the conditional duration and the lagged durations. For example, the logarithmic version of ACD model was provided in [3], the threshold autoregressive conditional duration model was proposed in [4], the asymmetric autoregressive conditional duration model was put forward in [5], and the smooth transition ACD model and the time-varying ACD model were introduced in [6]. There are also many other works focusing on choosing a suitable distribution to characterize the unexpected duration. The distributions which have been applied to the ACD models include the generalized Gamma distribution in [7], generalized F distribution in [8], the mixture of two exponential distributions in [9], the regime-switching Pareto distribution in [10], and the mixture of an exponential and a generalized beta of type 2 (GB2) distribution in [11]. Like many other statistical models, the ACD family models require strong assumptions which are difficult to satisfy in realistic situations [12].

In recent years, machine learning methods have been widely applied to image identification and natural language processing problems. Compared with traditional statistical models, machine learning methods have looser model assumptions and better generalization ability. The artificial neural network (ANN), inspired by the biological neural network, is one of the most widely used machine learning methods. According to Universal Approximation theorem [13], feedforward neural networks can approximate a Borel measurable function to any desired degree of accuracy if sufficiently many hidden units with arbitrary squashing functions are provided. Recurrent neural networks (RNNs) are a family of specially designed artificial neural network networks capable of extracting temporal information via the cycle architecture [14]. For the development in optimization techniques and computation hardware, RNNs have been widely used in many different domains recently [15]. To solve the vanishing/exploding gradient problem of simple RNNs, Hochreiter S. proposed the long short-term memory (LSTM) neural networks which can help us to utilize a longer sequence of historical information [16]. Although having the merit of strong fitting ability, LSTMs cannot provide probabilistic output compared with ACD family models.

Inspired by the work from Kristjanpoller and Minutolo [17], we propose a new architecture called LSTM-ACD to predict the UHF transaction durations by combing the ANN networks and ACD framework. We take a fully data-driven approach to extend the mean equation of classic ACD models while retaining the probabilistic inference ability. In addition, attention layer is added into our model to make a visualization of the proposed network and to improve the interpretability. The proposed architecture is applied to real-world stock duration datasets. The result shows that the proposed model produces more accurate estimation and prediction, outperforming the traditional ACD family models.

The rest of this paper is organized as follows: Section 2 introduces the methodology in detail, while Section 3 contains the experiment design and the corresponding results in this study. Section 4 concludes this paper and points out the possible direction of future research.

2. Methodology

In Section 2, the ACD framework is integrated with LSTM networks to propose a new LSTM-ACD model for predicting the trading durations of UHF data. This section is organized as follows. Section 2.1 introduces the classic ACD model. Section 2.2 describes the proposed LSTM-ACD architecture in detail. In addition the attention mechanism layer is utilized to unveil the internal mechanism of the proposed model.

2.1. Traditional ACD Family Models. A classic ACD model assumes that the durations are conditionally exponentially distributed with a mean that follows an ARMA process [2]. As shown in (1), the duration Δt_i between the i-th and $(i-1)$-th trade is the multiplicative product of μ_i and ε_i, which represents expected and unexpected portion of the transaction duration, respectively. In the conditional mean equation, μ_i linearly depends on the lagged durations and the lagged terms of itself. p and q in formula (2) represent the lagged order:

$$\Delta t_i = \mu_i \varepsilon_i, \tag{1}$$

$$\mu_i = \omega + \sum_{j=1}^{p} \alpha_j \Delta t_{i-j} + \sum_{j=1}^{q} \beta_j \mu_{i-j}. \tag{2}$$

By adding the lagged terms of error term ε_i as the independent variables, Hautsch proposed the Additive and Multiplicative ACD (AMACD) [18] model, as follows:

$$\mu_i = \omega + \sum_{j=1}^{p} \alpha_j x_{i-j} + \sum_{j=1}^{r} \nu_j \varepsilon_{i-j} + \sum_{j=1}^{q} \beta_j \mu_{i-j}. \tag{3}$$

A major limitation of basic ACD model and AMACD model is the assumption that the variables in the conditional mean equation behave in strict stationarity and linearity, but the duration sequences are usually in a nonlinear or non-stationary state. Based on power transformation, a nonlinear Box–Cox ACD model was proposed in [19], as follows:

$$\mu_i^{\delta_1} = \omega + \sum_{j=1}^{p} \alpha_j \varepsilon_{i-j}^{\delta_2} + \sum_{j=1}^{r} \beta_j \mu_{i-j}^{\delta_1}. \tag{4}$$

This paper also extends the linear conditional mean equation by LSTM networks to propose a new LSTM-ACD framework due to the strong fitting ability of deep learning techniques.

2.2. The Proposed Attention-LSTM-ACD Model

2.2.1. LSTM-ACD Model. It has been generally known that the LSTM cell is able to store information over longer time range compared with simple RNNs. As depicted in Figure 1,

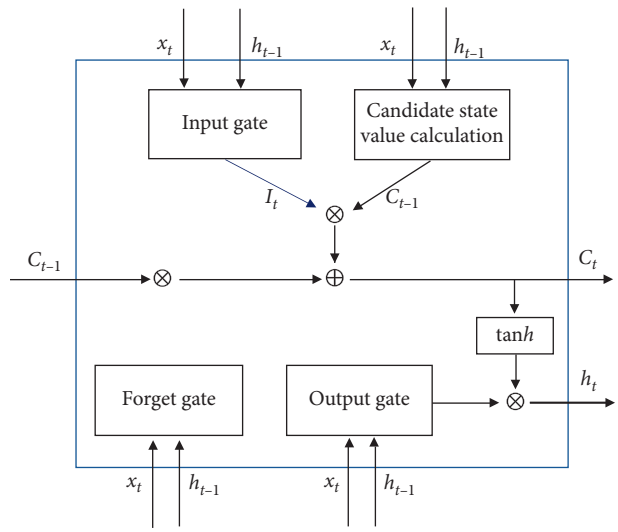

Figure 1: Structure of LSTM cell.

the information flow propagating across timesteps is controlled by three LSTM gates: the forget gate, the input gate, and the output gate.

Assuming that W_f, W_i, W_o, and W_c represent the LSTM weight matrices, and b_f, b_i, b_c, and b_o represent the bias vectors. The input vector, output vector, and cell state vector at time t are denoted as x_t, h_t, and C_t, respectively [20]. The operating process of a LSTM cell can be mathematically described as follows:

$$\begin{aligned}
\text{forget gate: } f_t &= \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right),\\
\text{input gate: } i_t &= \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right),\\
\text{output gate: } o_t &= \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right),\\
\text{candidate state values calculation: } \widetilde{C}_t &= \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right),\\
C_t &= f_t \times C_{t-1} + i_t \times \widetilde{C}_t,\\
h_t &= o_t \cdot \tanh\left(C_t\right).
\end{aligned} \tag{5}$$

As a type of RNNs specially designed to avoid the exponentially fast decaying factor, the LSTM networks can effectively prevent the gradient vanishing/explosion problem. Due to their ability to learn long-term dependencies, LSTMs are particularly suitable for financial prediction problems. Hence, we have the conjecture that extending the linear mean equation to LSTM network will improve the ability of extracting long-term dependencies for duration sequence. To verify this hypothesis, we take the Δt_{i-1} and $\ln \widehat{\mu}_{i-1}$ as the input for the LSTM cell at the time point of i-th transaction where Δt_{i-1} is the duration of last transaction and $\ln \widehat{\mu}_{i-1}$ is the logarithmic value of the output of the proposed LSTM-ACD model at time $i-1$. To retain the ability of probabilistic inference, the objective function is still the log likelihood function of $\Delta t_i = \mu_i \varepsilon_i$ which follows an exponential distribution. The log-likelihood function can be mathematically described as follows:

$$l = \sum \ln \frac{1}{\widehat{\mu}_i} \exp\left(-\frac{1}{\widehat{\mu}_i} \Delta t_i\right), \tag{6}$$

$$\ln \widehat{\mu}_i = \varphi\left(\Delta t_{i-1}, \ln \widehat{\mu}_{i-1}, h_{i-1}\right), \tag{7}$$

where φ represents a mapping from $\Delta t_{i-1}, \ln \widehat{\mu}_{i-1}, h_{i-1}$ to $\ln \widehat{\mu}_i$ by a LSTM cell.

2.2.2. Visualization and Promotion by Attention Mechanism. Attention mechanism was firstly proposed to improve the image processing accuracy by mimicking the perceptual system of human beings [21]. In the work of [22], attention mechanism was introduced to extend the basic encoder-decoder architecture and enhance the interpretability on the task of machine translation. Unlike the sequence-to-sequence modelling in sentence translation, the problem we focus on in this paper is to predict the financial duration one-step ahead. The attention weights which help automatically search for import hidden

states of the sequence-to-one LSTM architecture can be calculated by the following formulas:

$$e_{i-k} = v_\alpha^T \tanh\left(w_\alpha h_{i-k}\right),$$
$$\alpha_{i-k} = \frac{\exp\left(e_{i-k}\right)}{\sum_{k=1}^{T} \exp\left(e_{i-k}\right)}, \tag{8}$$

where h_{i-k} represents the hidden state lagged k timesteps and α_{i-k} represents the attention weight of h_{i-k}. The w_α and v_α are parameter matrices in the attention mechanism. By allocating different attention weights for different hidden states, a new vector c_i is produced as the input of a feedforward network f for predicting the target variable y_i:

$$c_i = \sum_{k=1}^{T} \alpha_k h_{i-k},$$
$$y_i = f\left(c_i\right). \tag{9}$$

In this study, the attention layer is integrated with LSTM to characterize the dynamics of $\ln \mu_i$ in the abovementioned mean equation of ACD model. The proposed Attention-LSTM-ACD model can be described by the following equations:

$$h'_{i-k} = \mathrm{LSTM}\left(h'_{i-k-1}, s'_{i-k-1}, \Delta t_{i-k-1}, \ln \widehat{\mu}_{i-k-1}\right),$$
$$c'_i = \sum_{k=1}^{T'} \alpha'_k h'_{i-k},$$
$$\ln \widehat{\mu}_i = f'\left(c'_i\right), \tag{10}$$

where $s_{i-k-1}{}'$ represents the cell state of LSTM lagged $k+1$ timesteps. Figure 2 shows the Attention-LSTM-ACD model in more detail.

3. Experiment

3.1. Data Description

3.1.1. Data Source. The Shenzhen Stock Exchange 100 Index (SZSE 100) is the first index designed for reflecting the multiple level market conditions of Chinese stock market. The constituent stocks of SZSE 100 represent the core high-quality assets in the Shenzhen A-share market, with strong growth, low valuation, and high investment value. In this paper, we collect duration data of the first 100,000 transactions which has excluded the transactions during premarket opening session, for each constituent stock from SZSE 100. The readers can acquire the data from the Transend DataBase System of Wind Information Co., Ltd (https://www.wind.com.cn/). Since the stock Tianjin Zhonghuan Semiconductor Co., Ltd., which is coding in 002129.SZ has no transactions during 2017, we totally have 99 stocks listed in SZSE COMP on December 31st, 2016, as our research dataset, which sums to 9900,000 transactions.

3.1.2. Data Characteristics. As the box plots in Figure 3 demonstrate, transaction durations of each constituent stock from SZSE 100 Index reveal a very long tail compared with the interquartile range. The large amount of data located in the tail means the existence of liquidity risk.

To further dig the dynamic characteristics of the duration sequence, the averaged coefficients of autocorrelation function (acf) and partial correlation function (pacf) coefficients are plotted. As shown in the following Figure 4, we can see that time series duration data show a longer memory in that both acfcoefficients and pacf coefficients decay very slowly as the lagged term increases. Hence, the higher complexity of the UHF duration data requires a forecasting algorithm with strong fitting ability.

3.2. Evaluation Criteria

*3.2.1. Mean Absolute Error (***MAE***) and Mean Squared Logarithmic Error (***MSLE***).* As two frequently used metrics, MAE and MSLE are both used to directly evaluate the performance of duration prediction and can be calculated by the following formulas:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left|\mathrm{duration}_i^{\mathrm{forecast}} - \mathrm{duration}_i^{\mathrm{real}}\right|,$$
$$\mathrm{MSLE} = \frac{1}{N} \sum_{i=1}^{N} \left(\log_e\left(1 + \mathrm{duration}_i^{\mathrm{forecast}}\right) - \log_e\left(1 + \mathrm{duration}_i^{\mathrm{real}}\right)\right)^2. \tag{11}$$

A smaller MAE or MSLE means that we have a more precise forecast of the transaction duration.

3.2.2. Performance Measure for Quantile Prediction. To evaluate the forecasting performance of quantile points, we utilize the loss function in quantile regression minimization problems [23]. Let $\{TaR_i: i = 1, \ldots, h\}$ be the prediction quantile points of the probability level α, x_i be the realistic duration of the $i-th$ transaction, and the I represent an indicative function, and the performance measure $QL_{\alpha,t}$ [11] can be calculated as follows:

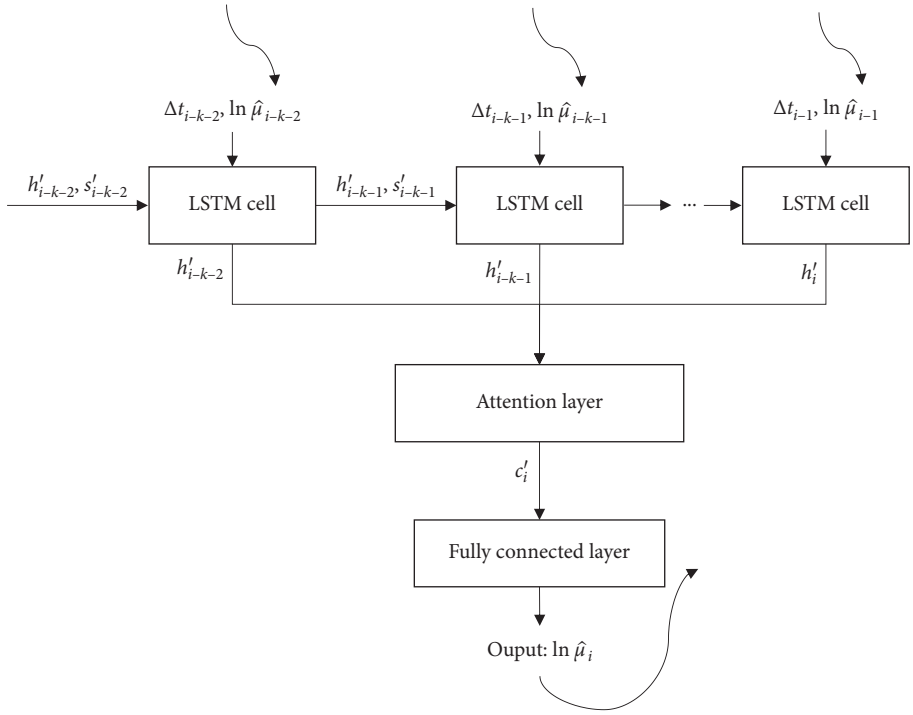

Figure 2: Architecture of Attention-LSTM-ACD model.

$$QL_{\alpha,t} = \sum_{i=T+1}^{T+h} \left(x_i - TaR_{i,\alpha}\right)\left[\alpha - I\left(x_i < TaR_{i,\alpha}\right)\right]. \quad (12)$$

3.2.3. Experiment Models. In Section 2, we have created a new framework for the one-step ahead prediction. The sequence of 50 lagged durations (1 feature, 50 timesteps) is firstly chosen as the input data, and we hence construct the LSTM-ACD model and Attention-LSTM-ACD model, which are presented in Sections 2.2.1 and 2.2.2, respectively. The only difference between the two models is the attention layer. To further utilize the information of transaction by transaction data, one-dimensional duration feature is extended to multidimensional feature vector by adding the transaction volume and transaction type information. And then, two other models are constructed, named as the Attention-LSTM-ACD (M) model and the LSTM-ACD (M) model. The experiments will be performed with the following seven models: the ACD model, the AMACD model, the BACD model, the LSTM-ACD model, the Attention-LSTM-ACD model, the Attention-LSTM-ACD (M) model, and the LSTM-ACD (M) model.

3.3. Training. During the training process, configurations are determined with as few exogenous inputs as possible because of the various drawbacks of manual tuning. We adopt fixed hyperparameters including learning rate, number of neurons of each layer, batch size, and timesteps for each constituent stock of SZSE 100.

3.3.1. Generation of Training Sets, Validation Sets, and Test Sets. As mentioned above, the sample used in this study is the 100,000 durations in 2017 for each stock collected from SZSE 100. We select the last 30% of data as the test set, while the remaining data are divided into training set and validation set according to the ratio of 8 : 2.

3.3.2. Training Process. During the experiment, fixed hyperparameter combination is selected for each model based on LSTM-ACD framework. Table 1 lists the hyperparameters used in our experiment. The attention size represents the height of the tensor w_α in formula (6). The initial learning rate is 0.5, and it is reduced by 50% after 1000 training steps. Besides the selection of hyperparameter combination, the remaining parameters of the proposed hybrid models are learned by taking advantage the early-stopping technique to avoid the overfitting problem. We evaluate model performance on the validation set every 100 training steps, and the early-stopping patience represents the number of times that there is continuously no improvement in the log likelihood function calculated on the validation set.

In this paper, the four models based on LSTM-ACD framework are coded in Tensorflow1.0, and the traditional ACD family methods are modelling by the ACDm package based on *R* language.

3.4. Experiment Results

3.4.1. Comparison of Different Models in **MAE** *and* **MSLE.** The out-of-sample forecasting errors of the seven types of experiment models are calculated. Table 2 lists the average

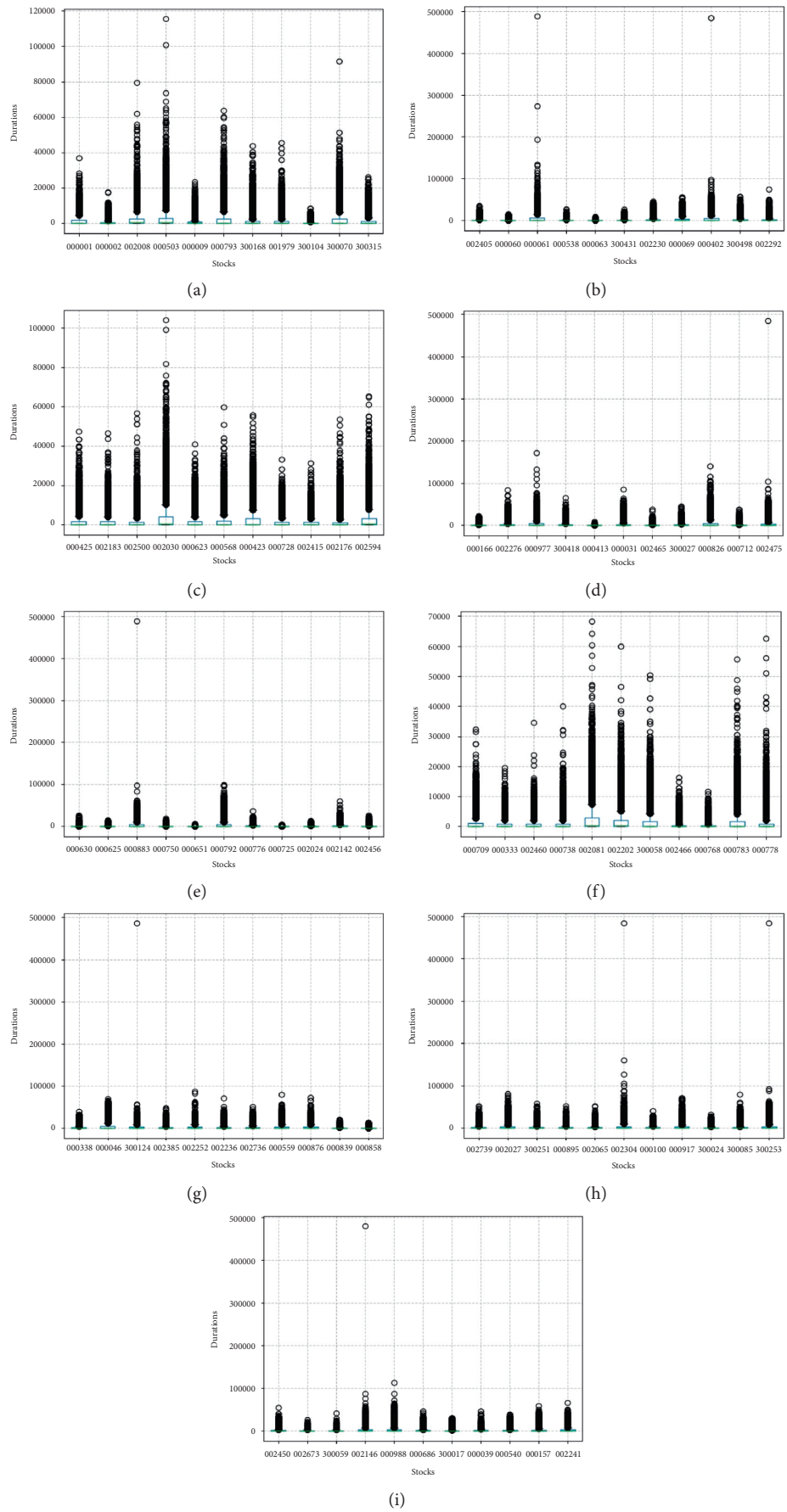

Figure 3: Box plots of durations of 99 stocks from SZSE 100 (minimum time unit: millisecond). The 99 stocks are listed in the x-axis, and the y-axis represents the duration dimension.

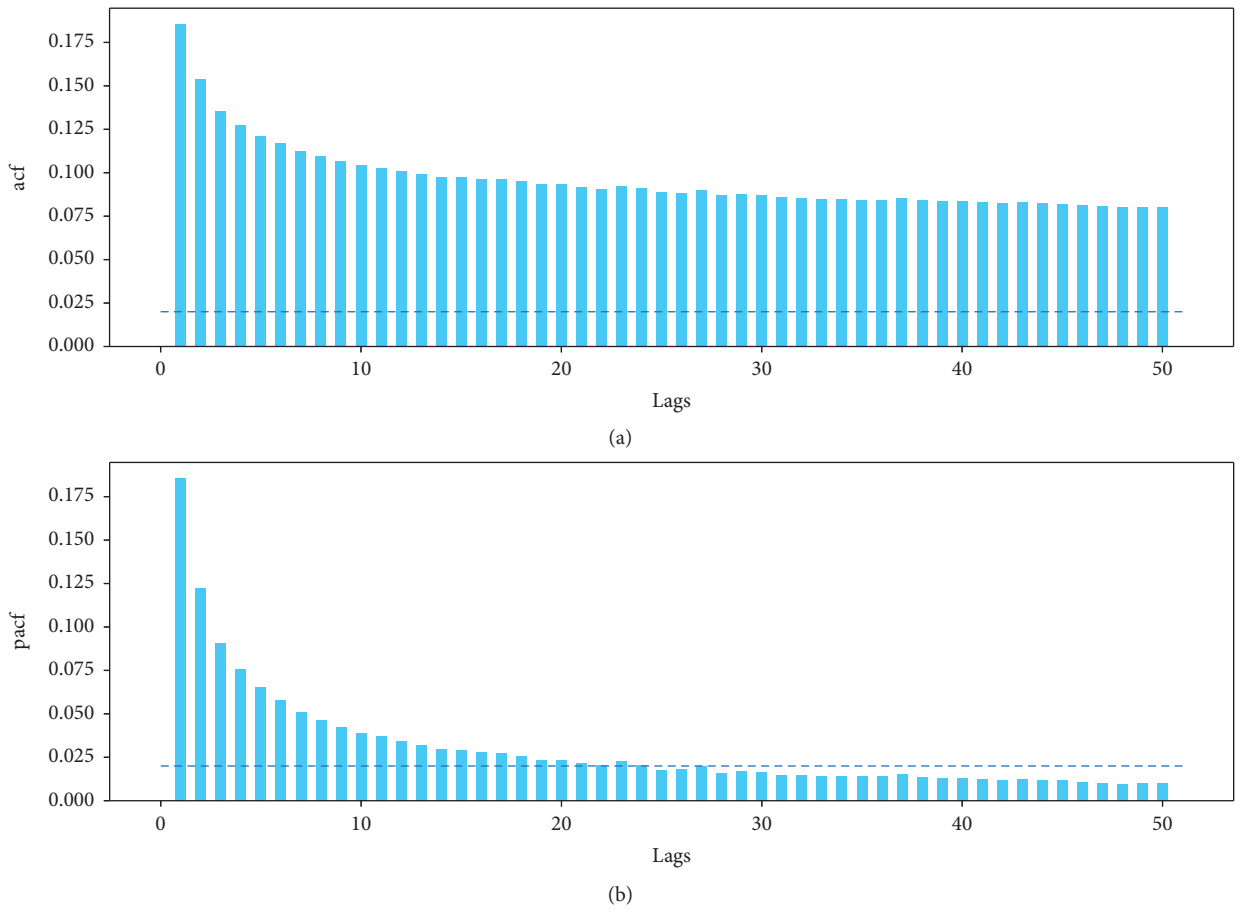

Figure 4: Averaged (a) acf and (b) pacf.

Table 1: The hyperparameters of each model.

	Attention-LSTM-ACD (M) and LSTM-ACD (M)	Attention-LSTM-ACD and LSTM-ACD
Input layer	3 features, 50 timesteps	1 feature, 50 timesteps
LSTM layer	5 hidden neurons	5 hidden neurons
Attention size	2 (for model Attention-LSTM-ACD (M))	2 (for model Attention-LSTM-ACD)
Fully connected layer	2 hidden neurons	2 hidden neurons
Batch size	300	300
Start learning rate	0.5	0.5
Decay steps	1000	1000
Decay rate	50%	50%
Early stopping patience	10	10

Table 2: The average MAE and MSLE on SZSE 100 Index constituent stocks of each model.

	Average MAE	Average MAE_{lagged}	Difference	Average MSLE	Average $MSLE_{lagged}$	Difference
Attention-LSTM-ACD(M)	2.0264	1.9762	0.0502	0.7088	0.6892	0.0195
LSTM-ACD(M)	1.7990	1.7602	0.0388	0.5947	0.5677	0.0270
Attention-LSTM-ACD	1.9758	1.9367	0.0390	0.6935	0.6629	0.0306
LSTM-ACD	1.8964	1.8368	0.0596	0.6520	0.6043	0.0477
ACD	1.8641	1.6612	0.2030	0.6285	0.5184	0.1100
BACD	—	—	—	0.6980	0.5928	0.1052
AMACD	1.8149	1.6446	0.1702	0.6007	0.4988	0.1019

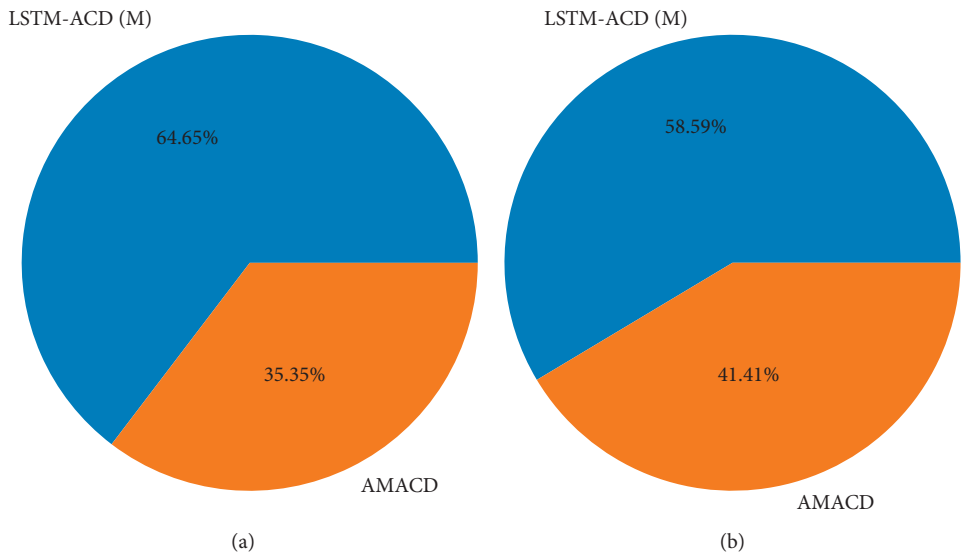

Figure 5: The contrasts between LSTM-ACD (M) and AMACD in (a) MAE and (b) MSLE (the proportion of each slice in a pie chart represents the quantity of stocks on which the corresponding model performs better than the other).

MAE and average MSLE in the test sets when the seven models are applied to SZSE 100 Index constituent stocks, respectively. The LSTM-ACD (M) model has the minimum average MAE and average MSLE, while the remaining three models based on LSTM-ACD framework all perform a bit worse than the traditional ACD family models in the two metrics. As mentioned above, uniform hyperparameter combination is chosen when applying the hybrid models. If we select different hyperparameters when focusing on different stocks, the performance of these hybrid models will be much better. In BACD model (see equation (4)), $\varepsilon_i = (x_i/\mu_i)$ and we calculate $\widehat{\varepsilon}_i$ by $(\text{Duration}_i^{\text{real}}/\text{Duration}_i^{\text{forecast}})$ during the process of iterated prediction. The estimation value of δ_1 could be much smaller than δ_2, so we will get an extremely large forecasted value if $\text{Duration}_i^{\text{real}}$ is much larger than $\text{Duration}_i^{\text{forecast}}$. This situation appears when we are applying BACD model to some stocks and results in extremely large average MAE and average quantile loss. Hence, the average MAE and average quantile loss of BACD model are not listed in this paper.

As the AMACD model places second in both average MAE and average MSLE, we compare LSTM-ACD (M) model with AMACD model in detail. We can see from Figure 5 that the LSTM-ACD (M) is also superior to the AMCD model on more stocks in both metrics of MAE and MSLE.

In addition, we calculate the MAE and MSLE one-step lagged of each model for the durations by the following formulas:

$$\begin{aligned}\text{MAE}_{\text{lagged}} &= \frac{1}{N}\sum_{i=1}^{N}\left|\text{Duration}_i^{\text{forecast}} - \text{Duration}_{i-1}^{\text{real}}\right|,\\ \text{MSLE}_{\text{lagged}} &= \frac{1}{N}\sum_{i=1}^{N}\left(\log_e\left(1+\text{Duration}_i^{\text{forecast}}\right) - \log_e\left(1+\text{Duration}_{i-1}^{\text{real}}\right)\right)^2.\end{aligned} \tag{13}$$

Suppose there are two models A and B for the same duration prediction task. We get $\text{MAE}^A_{\text{lagged}}, \text{MAE}^A$ for model A and get $\text{MAE}^B_{lagged}, \text{MAE}^B$ for model B. If $\text{MAE}^A = \text{MAE}^B$ and $\text{MAE}^A_{\text{lagged}} < \text{MAE}^B_{\text{lagged}}$, we can deduce that model B is more likely to extract the long-term dependency in the time series data because model A utilizes a higher proportion of $\text{Duration}_{i-1}^{\text{real}}$ value to forecast $\text{Duration}_i^{\text{real}}$. In an extreme situation that $\text{MAE}^A = \text{MAE}^B$ and $\text{MAE}^A_{\text{lagged}} = 0 < \text{MAE}^B_{\text{lagged}}$, obviously, model A just uses $\text{Duration}_{i-1}^{\text{real}}$ as the value of $\text{Duration}_i^{\text{forecast}}$. By an extension of this logic, if $\text{MAE}^A - \text{MAE}^A_{\text{lagged}} > \text{MAE}^B - \text{MAE}^B_{\text{lagged}}$, we can also deduce that model B has a stronger ability in capturing the long-term dependency relationship. Similarly, a smaller $\text{MSLE} - \text{MSLE}_{\text{lagged}}$ also indicates a higher ability in modelling long-term dependency. The results in columns 4 and 7 of Table 2 show that the average $\text{MAE}_{\text{lagged}}$ and average $\text{MSLE}_{\text{lagged}}$ of each traditional ACD model are significantly smaller than the average MAE and average MSLE, respectively. It means the four models based

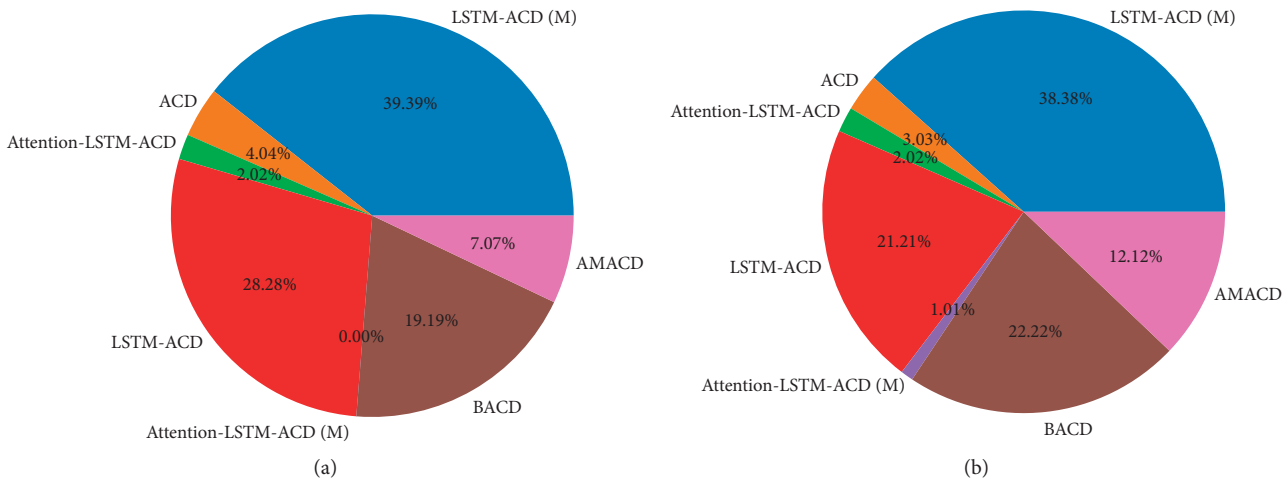

Figure 6: Detailed comparison of the 7 models in MAE and MSLE (the size of each pie slice represents the quantity of stocks on which the corresponding model achieves the best performance). (a) Minimum MAE. (b) Minimum MSLE.

on LSTM-ACD framework are superior to the traditional ACD family models in reflecting the long-term relationship of the sequential data.

Further detail for *MAE* and *MSLE* is provided in Figure 6. It can be seen that the LSTM-ACD (M) and LSTM-ACD are both superior to the previous ACD family models on more stocks in the metric of MAE. Moreover, LSTM-ACD (M) also has the minimum MSLE on more stocks than all other models. BACD model and LSTM-ACD model performs best on 21.21% and 22.22% of the assets, respectively, in terms of MSLE. Take the unstable performance of BACD model into consideration, the LSTM-ACD model still places second overall.

3.4.2. Comparison of Different Models in Quantile Forecasts. Table 3 lists the quantile forecast measure QL of different upper quantile levels for the seven models. It can be found that the Attention-LSTM-ACD (M) model is the supreme one at all three quantile levels among the models based on LSTM-ACD framework. In terms of the Attention-LSTM-ACD model, it also provides a better quantile forecasting than LSTM-ACD model at all quantile levels. Although our proposed models are not superior to the traditional ACD family models in extreme quantile forecasting, these results still indicate that the attention layer can improve the accuracy in conditional distribution forecasting. The QL of BACD model is not presented in Table 3 due to the extreme large values of QL. Besides, the performance of our proposed models can be further improved by selecting different hyperparameters for different stocks.

3.4.3. Attention Weights of Different Lag Orders. This section makes visualization for the Attention-LSTM-ACD model and Attention-LSTM-ACD (M) model. As can be seen in Table 4 and Figure 7, the weights learned by the attention layer in both the two models decrease exponentially with the increase in lag order. This means that the closer transaction

Table 3: Quantile loss for the models at different upper quantile levels.

Model	Upper quantile level		
	0.1	0.05	0.01
Attention-LSTM-ACD (M)	23323.42	17772.65	8964.29
LSTM-ACD (M)	24252.67	18974.48	10296.03
Attention-LSTM-ACD	24423.39	19732.88	11523.59
LSTM-ACD	24691.75	19866.13	11676.77
ACD	20040.66	14566.03	6484.51
BACD	—	—	—
AMACD	20084.59	14713.16	6663.78

Table 4: Average weights of the Attention-LSTM-ACD (M) model and the Attention-LSTM-ACD model on SZSE 100 Index constituent stocks.

Attention-LSTM-ACD (M)		Attention-LSTM-ACD	
Lag order	Weight	Lag order	Weight
Lag 1	**0.034797791**	Lag 1	**0.078697926**
Lag 2	0.028296111	Lag 2	0.034143126
Lag 3	0.024825037	Lag 3	0.027711418
Lag 4	0.023690568	Lag 4	0.025372255
Lag 5	0.022575405	Lag 5	0.023460835
Lag 6	0.022012244	Lag 6	0.022699354
Lag 7	0.02148028	Lag 7	0.021109585
Lag 8	0.020997523	Lag 8	0.020365109
Lag 9	0.020678233	Lag 9	0.02006378
Lag 10	0.020491562	Lag 10	0.01944402
.....	.....	.....	.....
Lag 41	0.018665664	Lag 41	0.017536173
Lag 42	0.018792729	Lag 42	0.017302261
Lag 43	0.018596823	Lag 43	0.017404814
Lag 44	0.018732648	Lag 44	0.017530387
Lag 45	0.018695344	Lag 45	0.017503974
Lag 46	0.018753901	Lag 46	0.017434779
Lag 47	0.018770904	Lag 47	0.017377114
Lag 48	0.01898069	Lag 48	0.017801802
Lag 49	0.018924108	Lag 49	0.017549126
Lag 50	0.018977175	Lag 50	0.01779628

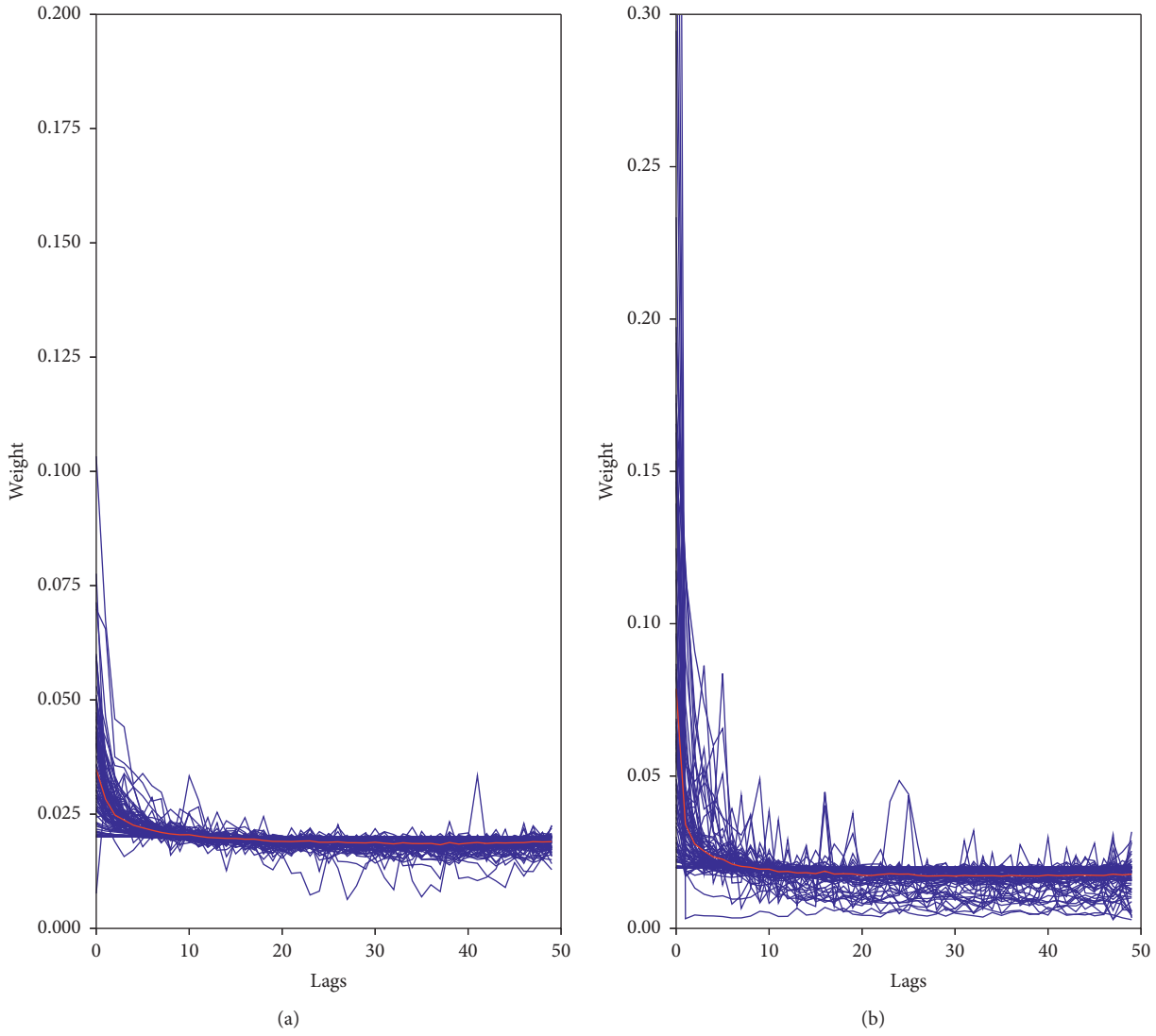

Figure 7: Attention weights of the (a) Attention-LSTM-ACD (M) model and (b) Attention-LSTM-ACD model of different lags on SZSE 100 Index constituent stocks (each blue line represents the attention weight sequence of a stock for the corresponding model, and each red line represents the average attention weights on SZSE 100 Index constituent stocks for the corresponding model.

has a more important effect on the current duration, which is consistent to our intuition.

4. Conclusion and Future Research

In this paper, we review the studies of transaction duration modelling based on ACD framework and find that these studies can be classified into two categories: (a) propose a new nonlinear equation form to describe the dynamics of conditional (expected) duration; (b) choose a more flexible distribution for the unexpected portion of the duration.

This study constructs a new framework for transaction duration modelling from the perspective of extending the mean equation of ACD model by machine learning methods. Firstly, we build a LSTM-ACD model by combining the LSTM networks with classic ACD model to characterize the complexity of the conditional mean process while retaining the advantage of providing probabilistic output. And then, attention layer is added to construct the Attention-LSTM-ACD model with the ability of unveiling importance of each hidden state in the LSTM networks.

Our proposed new framework is applied to a large-scale dataset. The fixed hyperparameters are chosen for all constituent stocks of SZSE 100 Index to reduce the impact of manual tuning, and the parameters (and consequently the underlying distributions) are learned via maximize the log-likelihood function. The results show that LSTM-ACD (M) model can present highest accuracy on the task of forecasting on real-world financial datasets among all the presented models. Although Attention-LSTM-ACD model and Attention-LSTM-ACD (M) model could not provide a more accurate performance inMAEmetric, the attention layer vividly depicts the importance of different temporal points of the input sequence and outperforms the corresponding LSTM-ACD model and LSTM-ACD (M) model in QL loss metric, respectively. In addition, the average MAE_{lagged} of

ACD model is significantly smaller than the average MAE, which means the predictions of the LSTM-ACD framework models to some extent convey more meaningful information. As a suitable chosen residual distribution does matters, the exponential distribution used in our framework can be extended to more flexible distributions in future research.

Data Availability

The data used to support the findings of this study can be purchased from Shanghai Wind Information Co., Ltd. (https://www.wind.com.cn/).

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

Yong Shi managed resources, supervised the study, and obtained funding acquisition. Wei Dai did formal analysis, analyzed using software, visualized the study, and wrote the original draft. Wen Long conceptualized the study, prepared methodology, reviewed and edited the study, and validated the study. Bo Li was responsible for data curation and investigated the study. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 71932008 and 71771204) and the Fundamental Research Funds for the Central Universities.

References

[1] J. Mueller, "Learn about financial liquidity," 2020.
[2] R. F. Engle and J. R. Russell, "Autoregressive conditional duration: a new model for irregularly spaced transaction data," *Econometrica*, vol. 66, no. 5, p. 1127, 1998.
[3] H. Giot, "The logarithmic ACD model: an application to the bid-ask quote process of three NYSE stocks," *Annales d'Économie et de Statistique*, vol. 60, no. 60, p. 117, 2000.
[4] M. Y. Zhang, J. R. Russell, and R. S. Tsay, "A nonlinear autoregressive conditional duration model with applications to financial transaction data," *Journal of Econometrics*, vol. 104, no. 1, pp. 179–207, 2001.
[5] L. Bauwens and P. Giot, "Asymmetric ACD models: introducing price information in ACD models," *Empirical Economics*, vol. 28, no. 4, pp. 709–731, 2003.
[6] M. Meitz and T. Teräsvirta, "Evaluating models of autoregressive conditional duration," *Journal of Business & Economic Statistics*, vol. 24, no. 1, pp. 104–124, 2006.
[7] L. Asger, "A generalized gamma autoregressive conditional duration model," 1999.
[8] N. Hautsch, *The Generalized F ACD Model*, Mimeo, University of Konstanz, London, UK, 2001.
[9] G. De Luca and G. M. Gallo, "Mixture processes for financial intradaily durations," *Computational Statistics & Data Analysis*, vol. 8, no. 2, 2004.
[10] G. De Luca and P. Zuccolotto, "Regime-switching Pareto distributions for ACD models," *Computational Statistics & Data Analysis*, vol. 51, no. 4, pp. 2179–2191, 2006.
[11] R. P. Yatigammana, J. S. K. Chan, and R. H. Gerlach, "Forecasting trade durations via ACD models with mixture distributions," *Quantitative Finance*, vol. 19, no. 12, pp. 2051–2067, 2019.
[12] R. Luo, W. Zhang, X. Xu, and J. Wang, "A neural stochastic volatility model," 2018.
[13] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
[14] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015.
[15] D. T. Tran, A. Iosifidis, J. Kanniainen, and M. Gabbouj, "Temporal attention-augmented bilinear network for financial time-series data analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1407–1418, 2019.
[16] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 2, pp. 107–116, 1998.
[17] W. Kristjanpoller and M. C. Minutolo, "Forecasting volatility of oil price using an artificial neural network-GARCH model," *Expert Systems with Applications*, vol. 65, pp. 233–241, 2016.
[18] N. Hautsch, *Econometrics of Financial High-Frequency Data*, Springer, Berlin, Germany, 2012.
[19] N. Hautsch, "Assessing the risk of liquidity suppliers on the basis of excess demand intensities," *Applied Sciences*, vol. 1, no. 2, 2003.
[20] F. Rundo, "Deep LSTM with reinforcement learning layer for financial trend prediction in FX high frequency trading systems," *Applied Sciences*, vol. 9, no. 20, p. 4460, 2019.
[21] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," 2014.
[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016.
[23] R. Koenker and G. Bassett, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.
[24] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Breakthroughs in Statistics*, S. Kotz and N. L. Johnson, Eds., pp. 610–624, Springer, New York, NY, USA, 1992.
[25] R. Y.-T. Chou, "Forecasting financial volatilities with extreme values: the conditional autoregressive range (CARR) model," *Journal of Money, Credit, and Banking*, vol. 37, no. 3, pp. 561–582, 2005.
[26] R. G. Donaldson and M. Kamstra, "An artificial neural network-GARCH model for international stock return volatility," *Journal of Empirical Finance*, vol. 4, no. 1, pp. 17–46, 1997.
[27] R. F. Engle, "The econometrics of ultra-high-frequency data," *Econometrica*, vol. 68, no. 1, pp. 1–22, 2000.