Journal of Economic Dynamics & Control 109 (2019) 103771

Contents lists available at ScienceDirect

Journal of Economic Dynamics & Control

journal homepage: www.elsevier.com/locate/jedc

Capturing deep tail risk via sequential learning of quantile dynamics☆

Qi Wu*, Xing Yan

School of Data Science & CityU-JD Digits Joint Laboratory in Financial Technology and Engineering, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 2 July 2019
Revised 14 September 2019
Accepted 25 September 2019
Available online 27 September 2019

Keywords:
Dynamic quantile modeling
Parametric quantile functions
Time-varying higher-order conditional moments
Asymmetric heavy-tail distribution
Long short-term memory
Machine learning
Neural network
VaR Forecasts
Financial risk management

ABSTRACT

This paper develops a conditional quantile model that can learn long term and short term memories of sequential data. It builds on sequential neural networks and yet outputs interpretable dynamics. We apply the model to asset return time series across eleven asset classes using historical data from the 1960s to 2018. Our results reveal that it extracts not only the serial dependence structure in conditional volatility but also the memories buried deep in the tails of historical prices. We further evaluate its Value-at-Risk forecasts against a wide range of prevailing models. Our model outperforms the GARCH family as well as models using filtered historical simulation, conditional extreme value theory, and dynamic quantile regression. These studies indicate that conditional quantiles of asset return have persistent sources of risk that are not coming from those responsible for volatility clustering. These findings could have important implications for risk management in general and tail risk forecasts in particular.

1. Introduction

Many institutional decisions in financial risk management rely on good forecasts of the conditional distributions of asset returns, especially their left tails. These decisions include the setting of bank capital requirement (Committee, 2016) and the posting of collateral in lending and clearing (BIS, 2013). What keeps risk managers awake at night, however, are not daily price fluctuations but (unexpected) downfalls of unusual magnitudes. The concern is that they may trigger systemic spirals that bring down the system. Researchers have conducted in-depth research on what constitutes a good measurement of tail risk (Kou and Peng, 2016; Kou et al., 2013). Meanwhile, it is of broad interest to be able to forecast into the far-left tail in any risk class and for any asset type. However, unusual movements of asset prices do not happen often. By definition, a 1%-quantile-breaching event occurs about twice a year. It is thus useful to first look at what could cause them before discussing how to forecast them using historical data.

The most recognized causes for price drawdowns of unusual magnitudes are extreme events with a broad market impact. Examples include the 1987 stock market crash, the burst of the dot com bubble in 2000, the Lehman default, and the ensuing 2008 financial crisis. However, idiosyncratic shocks such as short selling (Geraci et al., 2018), fire sales

☆ This research is supported in part by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) Projects 14211316 and 14206117.
* Corresponding author.
E-mail addresses: qiwu55@cityu.edu.hk, qw2107@columbia.edu (Q. Wu), xingyan4@cityu.edu.hk (X. Yan).

https://doi.org/10.1016/j.jedc.2019.103771
0165-1889/

(Cont and Wagalath, 2016), and flash crashes (Kirilenko et al., 2017) can cause similar damages. When these events occur, markets destabilize rapidly, either through positive feedback between drying up of market liquidity and funding liquidity (Brunnermeier and Pedersen, 2008) or arising from complex interactions among participants (Easley et al., 2011). Even if there are no severe external shocks, studies show that accumulations of ordinary-sized shocks could still give rise to large price swings. A case in point is the widely-observed empirical regularity called volatility clustering (Cont, 2007). Glasserman and Wu (2018) shows that the self-exciting nature of volatility can turn conditionally light-tailed innovations into unconditionally heavy-tail ones.

If we allow data to speak, there is no reason to believe that asset returns have memories only in the lower moments such as volatility. On the contrary, it is likely that markets remember those more extreme episodes as well, though occurred less commonly, and leave footprints in the motions of price skewness, price kurtosis, and so on. This thinking motivates us to find out whether real data exhibits different degrees of serial dependence at different probability levels in the conditional distribution of asset returns. If the conditional distribution of asset return indeed exhibits distinct lengths of memory at different probability levels, risk managers would be better informed when they forecast. In other words, the predictions of near-extreme price movements should improve if one captures the driving forces of tail events beyond those responsible for volatility clustering.

To do so, one needs a dynamic model that exhibits different lengths of memory at different levels of probability in the conditional distribution. Allowing data to speak demands a non-parametric component in the model, which differs from the parametric ideas used in constructing models such as the GARCH family, as well as those models based on the Extreme Value Theory. This is the reason why we take a semi-parametric machine learning approach in this paper. The model we construct inherits the flexibility of a recurrent neural network (RNN) in capturing nonlinear dynamics and memory effects of different lengths. We also overcome the black-box shortcoming of typical end-to-end neural networks by adding to the output layer a parametrization of quantile function so that the "learning" is statistically interpretable. By formulating the model training as a quantile regression problem, our earlier studies show that serial dependencies of conditional quantiles at low probability levels indeed contain risk factors that are independent of those driving the second moments (Yan et al., 2018).

1.1. Methodology review

In discrete time, forecasting conditional quantiles of asset returns is to find the time-t conditional α-quantile q_t of a scalar Y_t where $P(Y_t \leq q_t|\mathcal{I}_{t-1}) = \alpha$, given the probability α: $0 < \alpha < 1$ and the information set $\mathcal{I}_{t-1}$ up to time $t-1$. There are three approaches to this problem: fully parametric, non-parametric, and semi-parametric, depending on how one approaches the conditional distribution of Y_t: $F(y, \theta_t|\mathcal{I}_{t-1})$ where θ_t is the unknown parameter of F, which controls the dynamics of the conditional distribution.

Fully parametric models are interpretable ones. They assume functional forms for the conditional distribution F, and specify parametric dynamics of how F evolves through time. The leading parametric models used to model discrete-time asset return dynamics are the GARCH family. However, a GARCH model assumes no extra risk factors for the dynamics of tail behavior other than those driving the clustered conditional volatility. The conditional kurtosis as well as conditional skewness are constant over time. Within the parametric family, an important attempt to improve GARCH is to allow time-varying higher conditional moments. It starts with Hansen (1994) and later followed by Rockinger and Jondeau (2002), León et al. (2005) and Bali et al. (2008). These improvements allow parameters controlling the tail heaviness in $F(y, \theta_t|\mathcal{I}_{t-1})$ to be time-varying (autoregressive structures), but not giving them extra risk drivers. In other words, the autoregressive structures for different moments share one stochastic term, i.e., the innovation term.

By contrast, non-parametric models do not assume particular functional forms. Instead, kernel methods are used for estimating the conditional distribution. The limitation is the lack of structural interpretation. Historical Simulation (HS) is one of the leading non-parametric methods for Value-at-Risk (VaR) forecasts and is based on the rectangular kernel. In practice, the filtered version, called the filtered Historical Simulation (FHS), is used more often to accommodate data that are not independent and identically distributed (i.i.d.) (Barone-Adesi et al., 1999). In a FHS model, a location-scale model such as GARCH is used to pre-filter the data. One shortcoming of the FHS approach is that its forecasts are very sensitive to the length of historical data, e.g., whether to include the event of the 1987 stock market crash.

Semi-parametric models strike a balance between flexibility and structure. They typically focus solely on particular quantile levels of interest, leaving the rest of the conditional distribution unspecified. Two semi-parametric approaches are considered the best: the conditional Extreme Value Theory (CEVT), and the dynamic quantile regression (DQR). The CEVT approach (Mc Neil and Frey, 2000) specifies the tail distribution as one of three possible forms for the limiting distribution of extreme order statistic. The question is whether distributions of the extreme are good proxies of near-extreme quantiles at probability levels that are small, e.g., 1% or 0.1%, but not 10^{-10}. For this reason, Mc Neil and Frey (2000) proposed to fit a GARCH model first to the data and then apply extreme distributions to the standardized residuals, which are assumed to be i.i.d. This is in the same spirit of the FHS approach. On the other hand, the DQR approach cares less about parametrizing the tail distribution. Instead, it focuses on the dynamics of specific quantiles. A notable example is the conditionally autoregressive VaR (CAViaR) model in Engle and Manganelli (2004). The common issue with DQR models is that they are prone to the problem of quantile crossing.

1.2. Our contribution

We take a semi-parametric approach in constructing a dynamic return model, recognizing that parametrization allows interpretability whereas data-driven demands flexibility. In particular, our contributions are three folds.

- **i)** *A novel parametric construction of conditional quantile function.* In our setup, we construct a novel quantile function of conditional asset return directly, making sure it not only covers a wide range of tail heaviness but also treats the left and right tails separately. It is fully parametric and it allows one to avoid the difficult choice of which conditional density is better. This parametric quantile function is parsimonious and yet flexible enough to model asymmetric heavy tails.
- **ii)** *A machine learning approach to estimate the parameter dynamics.* The non-parametric component of our approach lies in the way we specify the dynamics of the quantile function's parameters. We use a sequential neural network, called Long Short Term Memory (LSTM), to learn them from historical data. The advantage of this approach, comparing to the GARCH family, is that it can capture a) unspecified dynamics that is potentially nonlinear in a given data set as well as b) memories that could exhibit different lengths at different probability levels. This approach is also flexible to expand information sets. The training of the neural network can also be easily cast into a quantile regression formulation.
- **iii)** *Captures high-order memories beyond volatility persistence.* Through a combination of simulation experiment and comprehensive testing on real data, we show that our model captures well long memory in the dynamics of the conditional distribution of asset returns. We focus on the higher moment dynamics and the left-tail quantile dynamics. For conditional skewness and kurtosis, our model discovers the existence of long memory that is independent of that responsible for volatility persistence. Similarly, the left-tail quantiles, e.g., the 0.01-quantile, also exhibit very long memories that differ from the dynamics of volatility clustering, and an ARMA model can numerically explain these extra serial dependencies. All of these imply that there exist extra risk factors that drive the dynamics of high moments or tail-side quantiles.
- **iv)** *Improves VaR forecast for a wide range of asset classes.* At last, across a wide range of asset classes including international equity indexes, exchange rates, and commodities, our method performs quite well in out-of-sample VaR forecasts comparing to the GARCH family, and models using FHS, CEVT, and CAViaR. All these findings have important implications for asset pricing and risk management.

We organize the rest of the paper as follows. We introduce some related works in Section 2. In Sections 3 and 4, we introduce the LSTM neural network and describe our proposed model respectively. A simulation experiment is conducted in Section 5, followed by the information of some real-world data we collect in Section 6. In Section 7, we present the empirical findings by our model about the tail dynamics or quantile dynamics learned from data. In Section 8, we perform the comparisons of several models on VaR forecasts through statistical backtesting and loss function. At last, we conclude our paper and give some discussions about future works.

2. Related works

Let P_t be the asset price at time t and $r_t = \log(P_t) - \log(P_{t-1})$ be the asset return over the period $t-1$ to t. The models we are going to compare assume r_t follows the following process (Engle and Patton, 2001):

$$r_t = \mu_t + \sigma_t \varepsilon_t, \quad \varepsilon_t \sim \text{ i.i.d. } F(\cdot), \tag{1}$$

where μ_t and σ_t are conditional mean and conditional volatility respectively, and F is the cumulative distribution function of ε. A complete model should specify the distribution F and how μ_t and σ_t depend on past information set. For example, GARCH(1,1) model Bollerslev (1986); Engle (1982) chooses a standard normal distribution for ε and evolves σ_t according to

$$\sigma_t^2 = \beta_0 + \beta_1(\sigma_{t-1}\varepsilon_{t-1})^2 + \beta_2\sigma_{t-1}^2. \tag{2}$$

Extensions like EGARCH (Nelson, 1991) and GJR-GARCH (Glosten et al., 1993) make reasonable changes to the above equation. Alternatives can also be made in the choice of the distribution F, e.g., the heavy-tailed t-distribution or the (generalized) skewed-t distribution proposed by Hansen (1994). However, the degrees of the freedom which represent the tail heaviness and the degree of asymmetry are constant over time. The motivation for using asymmetric heavy-tailed distribution is that numerous studies have highlighted the empirical facts of heavy tails and asymmetry of financial returns, see Cont (2001) for a summary. While the heavy tail phenomenon is a consensus, the evidence of asymmetry is not statistically as strong as the heavy tailness. Although the asymmetry is moderate, it cannot be ignored. For example, Chen et al. (2001) and Albuquerque (2012) documented some facts about the skewness of single stock and stock index, pictured that stock index returns are negatively skewed while single stock returns are positively skewed, and proposed theories for explaining this.

The conditional mean μ_t is often modeled in a linear auto-regressive way: $\mu_t = \gamma_0 + \gamma_1 r_{t-1}$. Throughout the paper, we will denote GARCH-type models with t-distribution innovation and linear auto-regressive conditional mean as AR-GARCH-t, AR-EGARCH-t, *etc*. Unless specifically stated, in this paper we set the orders of the GARCH terms and ARCH terms both to be 1 for all GARCH-type models, like GARCH(1,1) described above.

Besides the GARCH family, FHS and CEVT methods that filter the original return series by a GARCH first can also be written in Eq. (1). The differences are in the descriptions of ε. An FHS model uses the samples of $\{\varepsilon_t\}$ (the residuals after filtering) to estimate the empirical quantiles of ε non-parametrically. The CEVT approach would model ε with a non-parametric kernel distribution for the interior and an extreme value distribution for the tail sides, such as the generalized Pareto distribution.

Eq. (1) gives us an important indication that for different probability level α, the α-quantile of r_t, denoted as $q_t(\alpha)$, is linearly related to the α-quantile of ε:

$$q_t(\alpha) = \mu_t + \sigma_t F^{-1}(\alpha). \tag{3}$$

If we fix α and look at the temporal behavior of $q_t(\alpha)$, we can find that $\{q_t(\alpha)\}$ behaves proportionally to the volatility series $\{\sigma_t\}$ (plus $\{\mu_t\}$). Thus it generates very similar quantile dynamics for different α. This limitation holds for the GARCH family as well as models based on FHS and CEVT and may restrict the model to express the real data. There has been literature dealing with this issue. In Hansen (1994), Rockinger and Jondeau (2002), León et al. (2005), Bali et al. (2008), authors use skewed heavy-tailed innovation distribution and let the skewness and kurtosis be time-varying. The dynamics of the skewness and kurtosis are like linear autoregression too, just as the volatility. The difficulties are in the choice of an appropriate probability density and the complexity of its mathematical form.

In CAViaR (Engle and Manganelli, 2004), Engle and Manganelli also criticize that in GARCH, the negative extremes follow the same process as the volatility. Their point is that different quantiles may have very different dynamics. They thus propose to model the quantile dynamics separately for different α, instead of specifying the full conditional distribution. The fitting of CAViaR is done through quantile regression with the loss function: $L_\alpha(r, q) = (\alpha - I(r < q))(r - q)$. However, since different quantiles are estimated separately, CAViaR may suffer from the common issue of quantile regression, i.e., the quantile crossing. That is the possible occurrence of $q_t(\alpha_1) > q_t(\alpha_2)$ when $\alpha_1 < \alpha_2$.

Another big family of models that are related to conditional distribution forecasts is stochastic volatility (SV) models. Some comparisons between GARCH-type and SV models were made in Taylor (1994), Fleming and Kirby (2003), Carnero et al. (2004), Franses et al. (2007). SV models are applied in situations when volatility contains extra risk driver. In continuous time, if driven by Brownian Motion, they are Markovian, which is essentially different from GARCH family and our proposed model and may not be suitable for modeling serial dependence of volatility. What are comparable with our model and are consistent with the focus of this paper, are long-memory volatility models driven by, e.g., fractional Brownian Motion or Hawkes process, and preferably in discrete time.

2.1. Quantile regression

For two variables x and y, quantile regression aims to estimate the α-quantile q of the conditional distribution $p(y|x)$, see Koenker and Bassett (1978), Koenker and Hallock (2001) for a general background. To do this, without making any assumption on $p(y|x)$, a parametric function $q = f_\theta(x)$ is chosen, for example, a linear one $q = w^\top x + b$. Note that q is an unobservable quantity, a specially designed loss function between y and q makes the estimation feasible in quantile regression:

$$L_\alpha(y, q) = \begin{cases} \alpha|y - q| & y > q \\ (1-\alpha)|y - q| & y \le q \end{cases}. \tag{4}$$

Then we minimize the expected loss in a traditional regression way to get the estimated parameter $\hat{\theta}$:

$$\min_\theta \mathbf{E}[L_\alpha(y, f_\theta(x))]. \tag{5}$$

Given a dataset $\{x_i, y_i\}_{i=1}^N$, the empirical average loss $\frac{1}{N}\sum_{i=1}^N L_\alpha(y_i, f_\theta(x_i))$ is minimized instead. When we want to estimate multiple conditional quantiles $q_1, q_2, \ldots, q_K$ for different probability levels $\alpha_1 < \alpha_2 < \ldots < \alpha_K$, K different parametric functions $q_k = f_{\theta_k}(x)$ are chosen and the losses are summed up to be minimized simultaneously:

$$\min_{\theta_1,\ldots,\theta_K} \frac{1}{K}\frac{1}{N}\sum_{k=1}^K\sum_{i=1}^N L_{\alpha_k}(y_i, f_{\theta_k}(x_i)). \tag{6}$$

Because θ_j and θ_k are estimated separately in the optimization, this combination is prone to the issue of quantile crossing, i.e., for some x and $\alpha_j < \alpha_k$, it is possible that $f_{\theta_j}(x) > f_{\theta_k}(x)$ which contradicts the probability theory. To overcome this issue, additional constraints on the monotonicity of the quantiles can be added to the optimization to ensure non-crossing (Takeuchi et al., 2006). Another simpler solution is post-processing, i.e., sorting or rearranging the original estimated quantiles to be monotone (Chernozhukov et al., 2010). Another shortcoming of the conventional quantile regression is that the number of parameters grows with the size of the set of α, i.e., K. For a more elaborate description of a distribution, large K is necessary in some cases.

3. Long short-term memory

Since the AI player AlphaGo beat the top human player in Go matchup, deep learning (Le Cun et al., 2015) has become more and more popular in many other fields rather than just computer science and engineering. In the financial industry,

leading investment banks and hedge funds start to invite AI experts from the computer science field to their newly formed AI research groups, to help apply machine learning in all business units like trading execution, risk management, and portfolio management. To economists and management scientists, machine learning, including deep learning, can provide a new class of applied econometric or statistical approaches (Mullainathan and Spiess, 2017) that may have significant advantages over traditional models because of their data-driven property and their fewer restrictions on model assumptions. There are possibilities that machine learning can generate new insights into traditional research problems in finance.

With many parameters, deep learning models can extract complex nonlinear relationships contained in rich data, despite the lack of interpretability to some extent. The linearity assumption in many econometric models may not reflect the real word and is not sufficient when there is a large amount of data. So, sometimes we would rather sacrifice some interpretability in exchange for the model capability and performance improvement. Recent advances in optimization algorithms and computing hardware make the efficient fitting of deep models possible. Moreover, counterintuitively, many parameters do not necessarily lead to overfitting because of some useful machine learning techniques one can adopt. Deep learning has achieved breakthroughs in many application areas like computer vision, machine translation, and bioinformatics. However, successful applications in financial context are less reported.

The most common deep learning architecture is the deep neural network, in which nonlinearity is represented by compositions of the chosen nonlinear activation functions. For modeling time series data, in deep learning, Long Short-term Memory (LSTM) is a popular sequential neural network model designed for capturing both long-term and short-term dependencies or complex dynamics in sequences. Recently remarkable successes have been made in its applications like speech recognition, machine translation, protein structure prediction, *etc*. So it is a natural choice for us to model the dynamics of the conditional distribution of financial asset return series. Mathematically, LSTM is a highly composite nonlinear function that maps a sequence of vectors $x_1, \ldots, x_n$ to another sequence of vectors $y_1, \ldots, y_n$ (or to just one vector y), through hidden state vectors $h_1, \ldots, h_n$. Examples include machine translation from a Chinese sentence to an English sentence and the classification of a music clip to its genre.

Before describing the full mathematics of it, we first introduce the simple recurrent neural network (RNN), which is the understructure of LSTM and has the form:

$$h_j = \sigma_h(W_h x_j + U_h h_{j-1} + b_h), \tag{7}$$

$$y_j = \sigma_y(W_y h_j + b_y), \tag{8}$$

for $j = 1, \ldots, n$. W_h, U_h, b_h, W_y, b_y are the parameters need to be learned from the data and σ_h, σ_y are nonlinear activation functions. It models a nonlinear functional relationship between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. One can stack this structure multiple times to get a multi-layered or deep RNN, i.e., obtain layer k's hidden state vectors $h_1^k, \ldots, h_n^k$ through $h_1^{k-1}, \ldots, h_n^{k-1}$: $h_j^k = \sigma_h(W_h^k h_j^{k-1} + U_h^k h_{j-1}^k + b_h^k)$.

LSTM extends this understructure and has the equations:

$$f_j = \sigma_g(W_f x_j + U_f h_{j-1} + b_f), \tag{9}$$

$$i_j = \sigma_g(W_i x_j + U_i h_{j-1} + b_i), \tag{10}$$

$$o_j = \sigma_g(W_o x_j + U_o h_{j-1} + b_o), \tag{11}$$

$$g_j = \sigma_h(W_g x_j + U_g h_{j-1} + b_g), \tag{12}$$

$$c_j = f_j * c_{j-1} + i_j * g_j, \tag{13}$$

$$h_j = o_j * \sigma_h(c_j), \tag{14}$$

where $*$ represents element-wise multiplication of two vectors. At last, the output y_j is any chosen nonlinear function of h_j, like:

$$y_j = \sigma_h(W_y h_j + b_y). \tag{15}$$

All W, U, b are parameters that need to be learned and σ_g, σ_h are nonlinear activation functions, which are chosen as the S-shaped logistic function and tanh function respectively in this paper. In the case of outputting only one vector y, one can use the average of all hidden state vectors $\frac{1}{n}\sum h_j$ or just the last one h_n, e.g., $y = \sigma_h(W_y h_n + b_y)$, like in our Eq. (22). The logistic function and tanh function are the two most popular activation functions in all deep neural network models, playing the role of nonlinear transformation which maps the support of arguments to a bounded domain. Multiple compositions of these activation functions can approximate complex nonlinear relationships between input vectors $x_1, \ldots, x_n$ and output vectors $y_1, \ldots, y_n$ or y.

3.1. The long memory

To illustrate the flexible serial dependence structure that LSTM can represent, especially the long memory, and the comparison to simple RNN, we examine the derivative of hidden state h_j with respect to the former D input x_{j-D}, and see how it varies as D becomes large. To make analysis easier, we suppose all vectors and matrix degenerate to scalars in the network. For the simple RNN:

$$\frac{\partial h_j}{\partial x_{j-D}} = \frac{\partial h_j}{\partial h_{j-1}}\frac{\partial h_{j-1}}{\partial x_{j-D}} = \cdots = \frac{\partial h_{j-D}}{\partial x_{j-D}}\prod_{k=1}^{D}\frac{\partial h_{j-k+1}}{\partial h_{j-k}} \tag{16}$$

$$= (1-h_{j-D}^2)W_h U_h^D \prod_{k=1}^{D}(1-h_{j-k+1}^2). \tag{17}$$

The usual situation is $|U_h|<1$. Now $|\partial h_j/\partial x_{j-D}| \le |W_h||U_h|^D$ is exponentially decaying with respect to D, suggesting a short memory of the network only. If $|U_h|>1$, the derivative may explode, which is another known issue of simple RNN.

For LSTM, to make things simple, we replace $h_j = o_j * \sigma_h(c_j)$ by $h_j = c_j$, and set f_j and i_j to be constant (do not change according to j, this can be done by setting W_f, U_f, W_i, and U_i to be 0). Now $h_j = f_j h_{j-1} + i_j g_j$, and

$$\frac{\partial h_j}{\partial x_{j-D}} = \frac{\partial h_j}{\partial h_{j-1}}\frac{\partial h_{j-1}}{\partial x_{j-D}} = \cdots = \frac{\partial h_{j-D}}{\partial x_{j-D}}\prod_{k=1}^{D}\frac{\partial h_{j-k+1}}{\partial h_{j-k}} \tag{18}$$

$$= \frac{\partial h_{j-D}}{\partial x_{j-D}}\prod_{k=1}^{D}\left(f_{j-k+1} + i_{j-k+1}(1-g_{j-k+1}^2)U_g\right). \tag{19}$$

One can treat $f_{j-k+1}, i_{j-k+1} \in (0,1)$ be the weights of 1 and $(1-g_{j-k+1}^2)U_g$ respectively. The term in the multiple product is a re-balance between 1 and $(1-g_{j-k+1}^2)U_g$ no matter $|U_g|<1$ or $|U_g|>1$. If f_{j-k+1} is close to 1 and i_{j-k+1} is close to 0, the vanishing or exploding derivative will not appear when D grows. This suggests a long memory of the LSTM network. Moreover, when f_{j-k+1}, i_{j-k+1} are set to be varying, it allows more flexible memory structure of LSTM.

4. The LSTM-HTQF model

We first describe the proposed parametric quantile function, then use it to model the conditional distribution $p(r_t|\mathcal{I}_{t-1})$ of financial return series and show how to model its dependence on past information set. We then complete the proposed method by fitting the model with quantile regression formulation.

4.1. A Novel heavy-tailed quantile function

There are three common ways to fully express a continuous distribution, through probability density function (PDF), cumulative distribution function (CDF), or quantile function. To model financial data, people pay much attention to how to choose an appropriate parametric PDF that is consistent with the empirical facts of financial returns, like heavy tails. As far as we know, no literature does it with an appropriate CDF or quantile function. In this paper, we design a parsimonious parametric quantile function that allows varying heavy tails with intuitive parameters.

Our idea starts with the Q-Q plot, which is a popular method to determine whether a set of observations follows a normal distribution. The theory behind this is quite simple: the α-quantile of a normal distribution $\mathcal{N}(\mu,\sigma^2)$ is $\mu+\sigma Z_\alpha$, where Z_α is the α-quantile of the standard normal one. When α takes different values in (0,1), their Q-Q plot forms a straight line. If the Q-Q plot yields an inverted S shape, it indicates that the corresponding distribution is heavy-tailed (see Fig. 1 (a) for an example of the Q-Q plot of t-distribution with 2 degrees of freedom against $\mathcal{N}(0,1)$).

We construct a parsimonious parametric quantile function, as a function of Z_α, to let it have a controllable-shape Q-Q plot against the standard normal distribution. Specifically, the up tail and down tail of the inverted S-shaped Q-Q plot are controlled by two parameters respectively. Our proposed heavy-tailed quantile function (abbreviated as HTQF) has the following form:

$$Q(\alpha|\mu,\sigma,u,v) = \mu + \sigma Z_\alpha\left(\frac{e^{uZ_\alpha}}{A} + \frac{e^{-vZ_\alpha}}{A} + 1\right), \tag{20}$$

where μ, σ are location and scale parameters respectively, A is a relatively large positive constant; $u \ge 0$ controls the up tail of the inverted S shape, i.e., the right tail of the corresponding distribution; $v \ge 0$ controls the down tail, i.e., the left tail of the corresponding distribution. The larger u or v, the heavier the tail. When $u = v = 0$, the HTQF degenerates to the quantile function of a normal distribution.

To understand these, suppose that in Eq. (20), Z_α is first multiplied by a simpler factor $f_u(Z_\alpha) = e^{uZ_\alpha}/A + 1$, then multiplied by σ and added by μ (for simplicity one can set $\mu = 0$ and $\sigma = 1$). The factor f_u is a monotonically increasing and

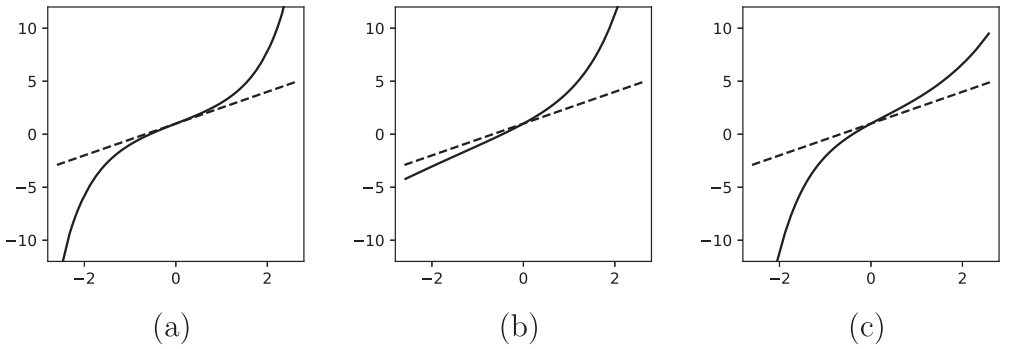

Fig. 1. Q-Q plots against $\mathcal{N}(0,1)$: (a) $t(2)$; (b) HTQF with $u = 1.0$ and $v = 0.1$; (c) HTQF with $u = 0.6$ and $v = 1.2$. For all three distributions, $\mu = 1$ and $\sigma = 1.5$. For HTQF, $A = 4$.

convex function of Z_α, and satisfies $f_u \to 1$ as $Z_\alpha \to -\infty$. So $Z_\alpha f_u(Z_\alpha)$ will exhibit the up tail of the inverted S only. The same analysis applies to $Z_\alpha f_v(Z_\alpha) = Z_\alpha(e^{-vZ_\alpha}/A + 1)$ too. Thus, $Z_\alpha(f_u(Z_\alpha) + f_v(Z_\alpha))/2$ exhibits the whole inverted S-shaped Q-Q plot (we replace $2A$ by A in Eq. (20)). The roles of A are to let $f_u(0)$ and $f_v(0)$ be close to 1, and to ensure the HTQF is monotonically increasing with Z_α. Fig. 1 (b) and (c) show the Q-Q plots of HTQF with different values of u and v against $\mathcal{N}(0,1)$. They exhibit different degrees of tailedness and the tails can flexibly change according to u and v. In addition, an HTQF with fixed parameters is the quantile function of a unique probability distribution because its inverse function exists and is a CDF. Please refer to the proof in the Appendix A.

4.2. The LSTM-HTQF model

For the distribution $p(r_t|\mathcal{I}_{t-1})$, different from GARCH-type models, we do not make assumptions on the PDF of it. Instead, we assume its quantile function being an HTQF, denoted by $Q(\alpha|\mu_t, \sigma_t, u_t, v_t)$, where μ_t, σ_t are time-varying parameters representing the location and scale; and u_t, v_t control the shapes of left tail and right tail of the corresponding distribution.

We assume the time-t parameters μ_t, σ_t, u_t, v_t are functions of past return history. To model that, we select a subsequence of fixed length from $r_{t-1}, r_{t-2}, \ldots$ to construct a feature vector sequence, and apply an LSTM to learn the mapping between the feature vectors and the HTQF parameters. LSTM (Hochreiter and Schmidhuber, 1997) is a popular and powerful sequential neural network model in machine learning, and is a natural choice in our method (see Lipton et al., 2015 for a comprehensive review of LSTM). In detail, a fixed length L is chosen, and then a feature vector sequence of length L is constructed from $r_{t-1}, \ldots, r_{t-L}$:

$$x_1^t, \ldots, x_L^t = \begin{bmatrix} r_{t-L} \\ (r_{t-L} - \bar{r}_t)^2 \\ (r_{t-L} - \bar{r}_t)^3 \\ (r_{t-L} - \bar{r}_t)^4 \end{bmatrix}, \ldots, \begin{bmatrix} r_{t-1} \\ (r_{t-1} - \bar{r}_t)^2 \\ (r_{t-1} - \bar{r}_t)^3 \\ (r_{t-1} - \bar{r}_t)^4 \end{bmatrix}, \tag{21}$$

where $\bar{r}_t = \frac{1}{L}\sum_{i=1}^{L} r_{t-i}$. The intuition behind this construction is straightforward, which is to extract information contained in raw quantities associated with the first, second, third, and fourth central moments of past L samples. We believe the high-order moments in the past contain direct information for future conditional distribution, especially for future left/right tail heaviness. This construction will make the neural network extract useful information more easily comparing to only including the first two moments. After this construction, we model the four HTQF parameters μ_t, σ_t, u_t, v_t as the output of an LSTM when feeding input $x_1^t, \ldots, x_L^t$:

$$[\mu_t, \sigma_t, u_t, v_t]^\top = \tanh(W^o h_t + b^o), \qquad h_t = \mathrm{LSTM}_\Theta(x_1^t, \ldots, x_L^t), \tag{22}$$

where Θ is the LSTM parameters, h_t is the last hidden state vector of LSTM. W^o, b^o are the output layer parameters.

At last, for fitting our model, we select K fixed probability levels $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_K < 1$ and minimize the average quantile regression loss between r_t and its conditional quantiles $Q(\alpha_k|\mu_t, \sigma_t, u_t, v_t)$ over all k and t, like in traditional quantile regression:

$$\min_{\Theta, W^o, b^o} \frac{1}{K}\frac{1}{T-L}\sum_{k=1}^{K}\sum_{t=L+1}^{T} L_{\alpha_k}(r_t, Q(\alpha_k|\mu_t, \sigma_t, u_t, v_t)). \tag{23}$$

Combine Eqs. (20)–(23) to complete our proposed LSTM-HTQF model and its fitting. After fitting, for subsequent out-of-sample series $\{r_{t'}\}_{t'>T}$, the time-varying HTQF parameters $\mu_{t'}, \sigma_{t'}, u_{t'}, v_{t'}$ can be calculated directly from historical returns

with the learned model parameters $\hat{\Theta}, \hat{W}^o, \hat{b}^o$. So the full conditional distribution at time t' can be estimated, as well as the conditional quantiles or any moments of interest. Then we can analyze the dynamics of the conditional distribution learned. Besides, our paper also focuses on the quantile or VaR forecasting. To evaluate the performance on the out-of-sample set, one can obtain the quantile sequence forecasted and apply some common VaR backtesting procedures as suggested by the regulatory authorities, or evaluate through a loss function.

4.3. Remarks

One of the advantages of the LSTM-HTQF model is that the proposed HTQF is more intuitive to understand and flexible enough to model asymmetric dynamics of heavy tails. It has a simpler mathematical form comparing to the distributions in EVT and the PDFs in GARCH-type models, such as the generalized skewed t-distribution in Hansen (1994) and Bali et al. (2008). Besides, the LSTM in our model is data-driven and can learn nonlinear dependence and long memory on past information set from the data while the linear auto-regressive FHS, CEVT, CAViaR, and GARCH family may not. Furthermore, from another perspective, comparing to traditional quantile regression, our model overcomes the issue of quantile crossing since the HTQF is monotonically increasing with α.

Generally, the feature vector sequence $x_1^t, x_2^t, \ldots, x_L^t$ can be designed to contain any information that is related to the conditional distribution of r_t or is helpful to the prediction, like trading volume, related assets, or fundamentals. To keep consistency with GARCH family and other related models, and to ensure the fairness of the comparisons, we construct $x_1^t, x_2^t, \ldots, x_L^t$ only from past returns $r_{t-1}, r_{t-2}, \ldots$. In real applications of our method, more information can be included in the feature vector sequence.

Our method is widely applicable in quantile prediction or time series modeling in many other non-financial fields. Time series data exhibiting time-varying asymmetrical tail behavior and nonlinear serial dependence of conditional distribution, e.g., hydrologic data, internet traffic data, or electricity price and demand, is most suited. One can also change the standard normal distribution in the Q-Q plot to other baseline distribution, i.e., replace Z_α in HTQF in Eq. (20) by other quantile function, to let the HTQF have a controllable-shaped Q-Q plot against the specified distribution, like exponential one or lognormal one, the choice of which relies on specific domain knowledge.

5. Simulation studies

The purpose of the simulation experiment is to verify whether our method can learn the serial dependence of conditional distribution whose higher moments exhibit strong time-varying effects. Similar to a GARCH specification, we generate a simulated time series in discrete time according to

$$r_t = \mu_t + \sigma_t z_t, \tag{24}$$

$$\mu_t = 0.052 + 0.172 r_{t-1}, \tag{25}$$

$$\sigma_t^2 = 0.293 + 0.161 (\sigma_{t-1} z_{t-1})^2 + 0.575 \sigma_{t-1}^2, \tag{26}$$

but with a skew-t distributed innovation where λ, η are parameters controlling the skewness and kurtosis (degrees of freedom):

$$z \sim \text{ skew-}t(\lambda, \eta).$$

Its density is given by:

$$f(z|\lambda,\eta) = c\left(1 + \frac{2(g-\rho^2)\left(z+\rho/\sqrt{g-\rho^2}\right)^2}{(\eta+1)\left(1+\lambda\ \text{sign}\left(z+\rho/\sqrt{g-\rho^2}\right)\right)^2}\right)^{-\frac{\eta+1}{2}}, \text{ where} \tag{27}$$

$$c = \left(\frac{2(g-\rho^2)}{\eta+1}\right)^{1/2} B\left(\frac{\eta}{2}, \frac{1}{2}\right)^{-1} \tag{28}$$

$$\rho = 2\lambda \left(\frac{\eta+1}{2}\right)^{1/2} B(\frac{\eta-1}{2}, 1) B\left(\frac{\eta}{2}, \frac{1}{2}\right)^{-1} \tag{29}$$

$$g = (1+3\lambda^2)\frac{\eta+1}{2} B\left(\frac{\eta}{2}, \frac{1}{2}\right)^{-1} B\left(\frac{\eta-2}{2}, \frac{3}{2}\right) \tag{30}$$

and $B(x,y) := \int_0^1 t^{x-1}(1-t)^{y-1} dt$ is the Beta function.

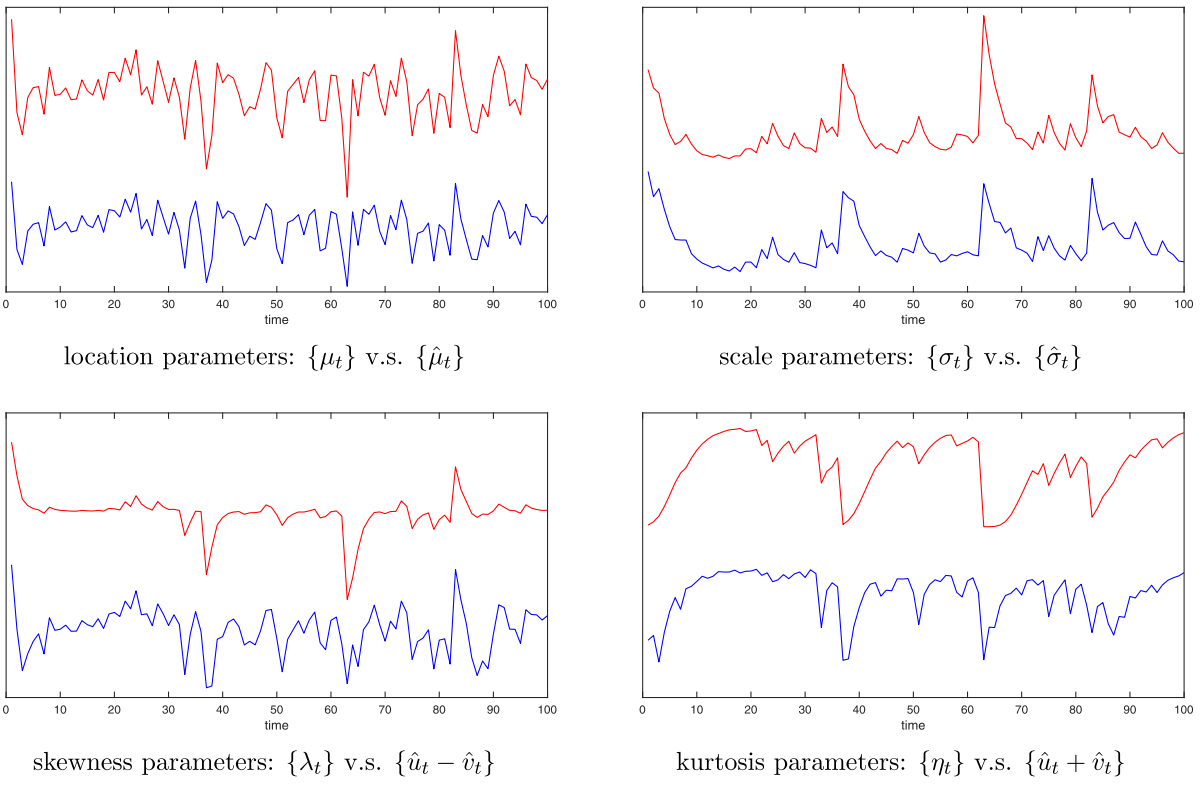

Fig. 2. Comparisons between the true parameters (red lines) of the simulated time series and the forecasted HTQF parameters (blue lines) by our model on the out-of-sample set. The linear correlation coefficients between the four pairs of parameters are 0.9780, 0.9104, 0.8014, and −0.7867 respectively. Linear transformations are made before plotting. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To impose parametric dynamics to the skewness parameter λ_t and the kurtosis parameter η_t so that their evolutions are known as a prior, we modify the setup in Bali et al. (2008) as such:

$$\lambda_t = -1 + 2/(1 + \exp(-\tilde{\lambda}_t)), \text{where} \tag{31}$$

$$\tilde{\lambda}_t = -0.038 + 0.076 z_{t-1}^3 + 0.463 \tilde{\lambda}_{t-1}; \tag{32}$$

$$\eta_t = 2 + 2\exp(3 - \tilde{\eta}_t), \text{where} \tag{33}$$

$$\tilde{\eta}_t = 0.136 + 0.057 z_{t-1}^4 + 0.717 \tilde{\eta}_{t-1}\,. \tag{34}$$

We generate 30,000 data points in total and treat the last one-tenth as the out-of-sample data set. Another one-tenth are extracted to form the validation set, which is used to stop the fitting process when the loss on this set begins to increase, to prevent overfitting. In LSTM we set $L = 25$, $H = 8$ and in HTQF we set $A = 4$ without change. $K = 21$ probability levels are chosen into the α set for model fitting: $[\alpha_1, \ldots, \alpha_{21}] = [0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99]$.

After fitting, we forecast the conditional distribution for every day on the out-of-sample set through forecasting the four HTQF parameters and compare them to the true parameters that generate the time series: $\{\mu_t\}$, $\{\sigma_t\}$, $\{\lambda_t\}$, $\{\eta_t\}$. Denoting the forecasted HTQF parameters as $\{\hat{\mu}_t\}$, $\{\hat{\sigma}_t\}$, $\{\hat{u}_t\}$, $\{\hat{v}_t\}$, we use $\{\hat{u}_t - \hat{v}_t\}$ as the proxy of skewness and $\{\hat{u}_t + \hat{v}_t\}$ as the proxy of kurtosis. We use these proxies because they are intuitive and, the true skewness or kurtosis may not exist for some λ_t and η_t (λ_t and η_t are proxies too). We plot in Fig. 2 the pairs of the true/forecasted location, scale, skewness, and kurtosis parameters respectively, where the red lines are the true parameters and the blue ones are the forecasted. Linear transformations are made before plotting, to let them be in similar ranges. One can see that the forecasted parameters are highly linearly correlated to the true ones, which means that our model has successfully learned the temporal behavior of the conditional distribution of r_t. The linear correlation coefficients between the four pairs of parameters are 0.9780, 0.9104, 0.8014, and −0.7867, respectively. The negative sign is because the heavier the tail, the bigger $\hat{u}_t$ or $\hat{v}_t$, but the smaller η_t.

Table 1
The whole period and the out-of-sample period of every asset return time series. We roll forward 250 days every time and re-fit the model.

Asset	Start date	End date	Out-of-sample start date	Out-of-sample set length	Total length
NASDAQ 100	1985-10-01	2018-07-02	1996-08-26	5500	8257
HSI	1986-12-31	2018-06-29	1998-03-26	5000	8031
Nikkei 225	1965-01-05	2018-07-02	1975-11-17	10,500	13,785
FTSE 100	1983-12-30	2018-08-16	1994-11-17	6000	8767
DAX	1987-12-30	2018-07-02	1998-10-20	5000	7873
USDEUR	1975-01-02	2018-07-09	1985-07-11	8500	11,162
USDGBP	1971-01-04	2018-07-09	1981-02-20	9750	12,394
USDJPY	1971-01-04	2018-07-09	1981-02-24	9750	12,395
USDAUD	1971-01-05	2018-07-09	1981-12-09	9500	12,111
Crude Oil	1983-03-31	2018-07-13	1993-10-25	6250	8906
Gold	1979-12-27	2018-07-13	1990-09-18	7250	9995

6. Empirical studies

6.1. Daily return data

We select 11 representative assets coming from 3 different asset classes including equity indexes, foreign exchange rates, and spot commodities. They are equity index NASDAQ 100, HSI, Nikkei 225, FTSE 100, DAX, exchange rates of USD to EUR/GBP/JPY/AUD, and spot crude oil/gold. For every asset, we collect the time series of its daily returns that has the maximum possible length. The start date, the end date, and the length of each time series are shown in Table 1. We adopt a rolling-window forecast setting that every time 250 days are used as the out-of-sample set for forecasting and then these 250 days are included into the in-sample set and the model is re-fitted. The in-sample set always starts from the beginning of the time series. In Table 1, we also list the start date and the length of the whole out-of-sample set for every time series. All returns are calculated by $r_t = \ln(P_t/P_{t-1})$, where P_t is the price or rate at time t. Each time series is normalized to have zero sample mean and unit sample variance.

Whenever the model is fitted, a quarter subset is extracted from the in-sample set to form the validation set, which is a machine learning technique used for selecting hyper-parameters, and for stopping the optimization iterations when the loss on the validation set begins to increase, to prevent overfitting. Our model has two hyper-parameters: the length L of past series $r_{t-1}, \ldots, r_{t-L}$ on which time-t HTQF parameters μ_t, σ_t, u_t, v_t depend, and the hidden state dimension H of the LSTM. In our experiments, we find that values greater than $L = 100$ and $H = 16$ do not increase the performance evidently for most of the assets while the performance of smaller L or H such as $L = 25$ or $H = 4$ is noticeably inferior. Hence, we set $L = 100$ and $H = 16$ uniformly for every asset in the following empirical studies to avoid the problem of fine-tuning for each individual asset. The A in the HTQF is set to be 4. We choose $K = 21$ probability levels into the α set for model fitting: $[\alpha_1, \ldots, \alpha_{21}] = [0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99]$, in the optimization in Eq. (23).

To accelerate the training of neural networks, we use the commonly used mini-batch technique, which avoids performing the gradient descent over the full data set. We first divide the in-sample set randomly into five equal parts called five batches. Then, for every five iterations, we apply the gradient descent over each one of the five batches in turn to update the neural network parameters. This is feasible because the loss function that needs to be minimized in Eq. (23) is a sum of the loss for every data point, or for every time t, with shared parameters. In every iteration, only a small part of all t is used to calculate the gradient and update the model parameters. For more about mini-batch, please see Keskar et al. (2016). We adopt Adam optimization algorithm in Kingma and Ba (2014) with suggested default setting implemented by TensorFlow (Abadi et al., 2016). The training for one asset takes a few hours using a desktop PC with two 3.30 GHz quad-core CPUs.

6.2. High-frequency data

In addition to the daily return data collected above, we also collect some high-frequency data from the Hong Kong stock market. Specifically, 5-min returns of 8 component stocks contained in the HSI index are obtained by us. They come from 4 different sectors and are the most liquid blue-chip stocks with the largest market capitalization, e.g., HSBC, Tencent, AIA, *etc*. We use the exchange codes like 0005.HK to denote them. Their return time series all start from May 27, 2015 and end on April 17, 2019 with over 60,000 observations. The first 5-min return of each day is calculated using the close price of the first 5 min and the close price of the previous day. All prices we obtain are trade prices.

We set the first two-thirds of each return time series as the in-sample set and the remaining one third as the out-of-sample set without rolling. However, like the daily data, a quarter subset is extracted uniformly from the in-sample set to form the validation set for our model. All other experiment settings are the same as the daily data case. However, one issue of using high-frequency data is the intra-day seasonal effect. For example, the intra-day volatility is highest near the market open and close, and forms a U shape in the whole trading day. To neutralize the seasonal effect, every return in a 5-min bin is divided by the standard deviation of all returns in that bin across all trading days, assuming their mean is 0. We only use the data after this pre-processing.

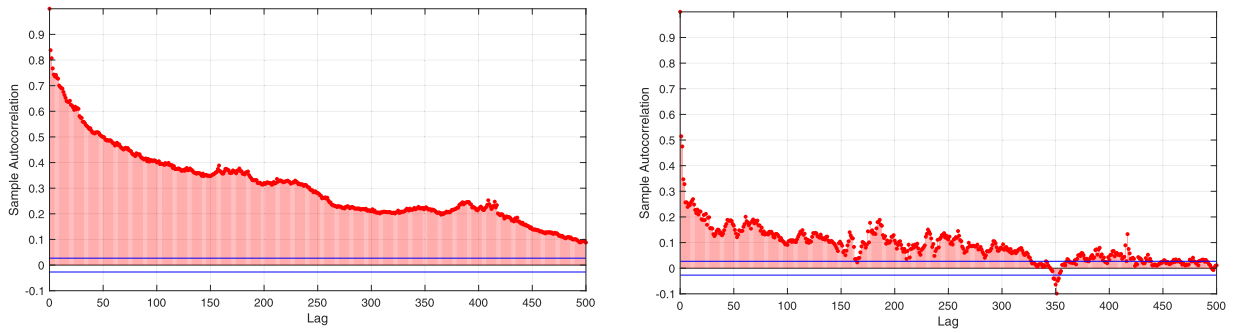

Fig. 3. Autocorrelation functions of the volatility and skewness sequences forecasted by our model on NASDAQ 100 out-of-sample set. The skewness is regressed to the mean and volatility first and the residual sequence is used instead.

7. Tail dynamics learned

Because our adoption of LSTM does not assume any particular parametric form for the conditional distribution dynamics, in this section, we examine what dynamics our model has learned from the daily return data. These dynamics include the dynamics of moments and quantiles of the conditional distribution.

7.1. Moment dynamics

On the out-of-sample set $\{r_{t'}\}$, our model predicts the HTQF parameters $\hat{\mu}_{t'}, \hat{\sigma}_{t'}, \hat{u}_{t'}, \hat{v}_{t'}$ for every t', which completely determine the conditional distribution of $r_{t'}$ through its quantile function $Q(\alpha|\hat{\mu}_{t'}, \hat{\sigma}_{t'}, \hat{u}_{t'}, \hat{v}_{t'})$. Hence, we can randomly sample from this distribution and calculate the sample moments for estimating the true moments of the distribution. Then we examine how these moments behave temporally to see what dynamics our model has learned from the data. The sampling is quite simple as we only need to sample a number z from the standard normal and then put it into the HTQF:

$$\hat{\mu}_{t'} + \hat{\sigma}_{t'} z \left(\frac{e^{\hat{u}_{t'} z}}{A} + \frac{e^{-\hat{v}_{t'} z}}{A} + 1 \right). \tag{35}$$

This is a sample of a random variable that follows the distribution whose quantile function is $Q(\alpha|\hat{\mu}_{t'}, \hat{\sigma}_{t'}, \hat{u}_{t'}, \hat{v}_{t'})$. We take the equity index NASDAQ 100 as an example. For every t', after sampling 100,000 times, we calculate the sample mean, volatility, skewness, and kurtosis respectively. They are the estimations of the conditional mean, volatility, skewness, and kurtosis of $r_{t'}$. Thus we obtain four sequences of moments for the out-of-sample set.

Next, we examine the length of dependence or memory of these moment sequences through the autocorrelation function. To examine the volatility dynamics learned by our model, we plot in Fig. 3 the autocorrelation function of the volatility sequence forecasted by our model on NASDAQ 100 out-of-sample set. One can see that our model generates very persistent volatility whose autocorrelation function is slowly decaying. In the second subplot of Fig. 3, we show the memory length of the skewness after linearly regressing the skewness to the mean and volatility to remove correlations, and obtain the residuals. This subplot shows the autocorrelation function of the residual sequence, which also exhibits a long memory that is significant for even 300 days' lag, despite the relatively small magnitude comparing to volatility memory.

From these observations, we can conclude that the skewness itself has dynamics that differs from that of the volatility, which is implicitly proved by the skewness memory length after we remove the mean and volatility. It suggests that existing models may miss extra risk drivers and new stochastic terms would be needed to drive the skewness process. Because the kurtosis may not exist at some time points, we cannot show its memory length through the autocorrelation function. However, in the next section, we will examine the memories of the left-tail quantiles instead.

7.2. Quantile dynamics

Our LSTM-HTQF model predicts the conditional quantiles for all probability levels too. We analyze the length of dependence or memory of the forecasted quantile sequence again through the autocorrelation function, also taking NASDAQ 100 index as an example. Note that the quantile must be greatly correlated to the mean and volatility. So, on the out-of-sample set of NASDAQ 100, we linearly regress the forecasted quantile to the forecasted mean and volatility (calculated through sampling) and analyze the autocorrelation function of the residual sequence. This can reveal the independent dynamics of the quantile that separates from the mean and volatility.

We choose four left-tail quantiles of probability levels 0.01, 0.05, 0.1, 0.15, and plot the autocorrelation functions of their residuals obtained by regression in Fig. 4. One can see very long dependences in all four quantile sequences, proving that the quantile itself has independent dynamics. Besides, as the probability level increases from 0.01 to 0.15, the magnitude

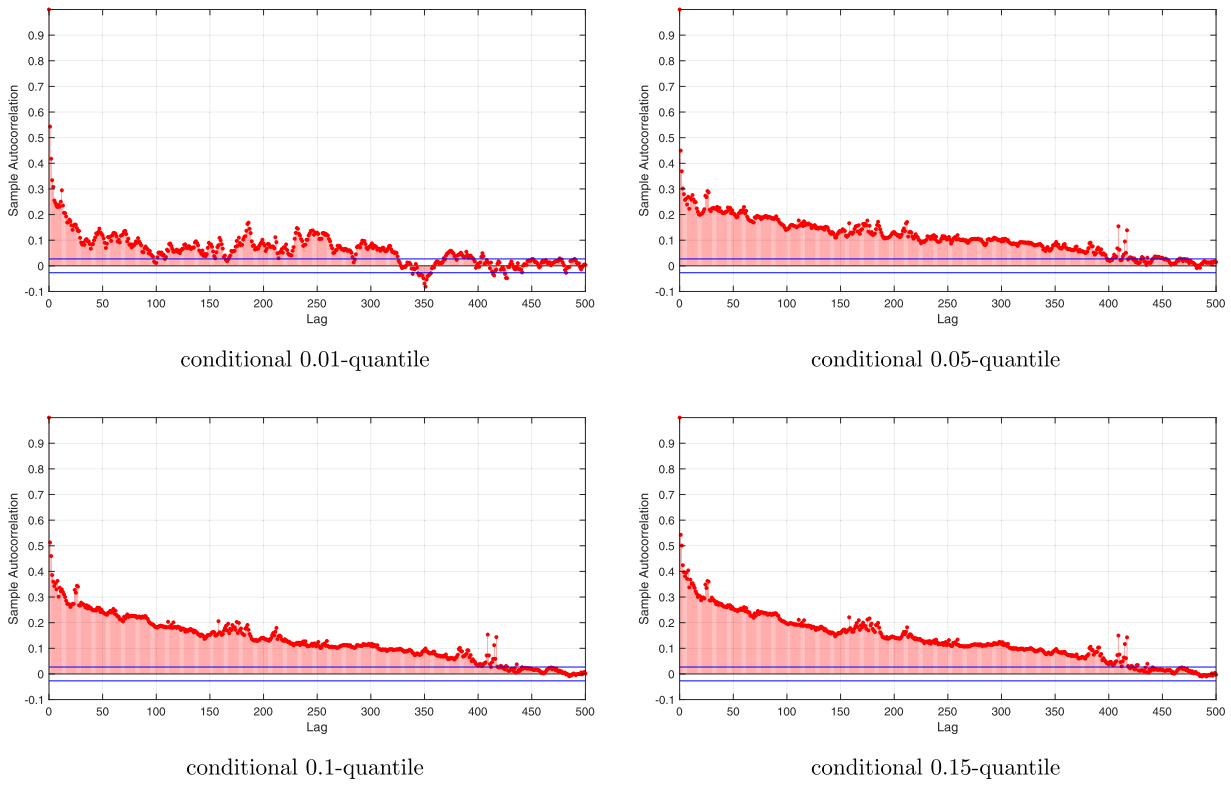

Fig. 4. Autocorrelation functions of the forecasted quantile sequences (regressed to mean and volatility) on the out-of-sample set of NASDAQ 100.

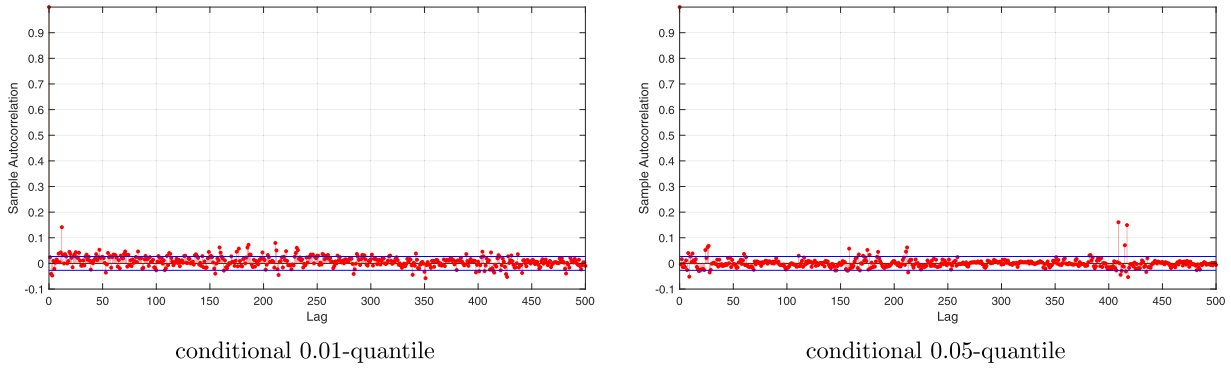

Fig. 5. Fit an ARMA(2,2) to the regressed 0.01 and 0.05-quantile sequences of NASDAQ 100 out-of-sample set and plot the autocorrelation functions of the ARMA residuals. The fitted ARMA equation for 0.01-quantile is $y_t = -0.0000 - 0.1502y_{t-1} + 0.7814y_{t-2} + \varepsilon_t + 0.5512\varepsilon_{t-1} - 0.3888\varepsilon_{t-2}$, and for 0.05-quantile is $y_t = -0.0000 + 1.5439y_{t-1} - 0.5457y_{t-2} + \varepsilon_t - 1.2713\varepsilon_{t-1} + 0.2958\varepsilon_{t-2}$.

of the dependence becomes larger and larger. This provides evidence that different quantiles may have different dynamics according to probability levels.

To figure out what dynamics the regressed quantile sequences most likely follow, we fit an ARMA model to them respectively. Then we check the autocorrelation function of the ARMA residual sequence to see if any serial dependence exists. After fitting an ARMA(2,2), the ARMA residuals of the quantile sequences (regressed to mean and volatility) for all probability levels have no serial dependence at all, as indicated by their autocorrelation functions in Fig. 5 (taking 0.01 and 0.05-quantiles as examples). The autocorrelations almost always lie between the significance bounds. Fitting an ARMA(1,1) cannot generate ARMA residuals that have autocorrelation functions like this. It suggests that the regressed quantile sequences have dynamics that are very close to ARMA(2,2). The fitted ARMA equations are also provided in the caption of Fig. 5. At last, the conclusions made in this section hold not only for NASDAQ 100 but also for all other assets in Table 1.

Table 2
Unconditional coverage test for $\alpha = 0.01$ VaR forecasts. The test statistic shown below has an asymptotically chi-square distribution with one degree of freedom. The threshold for rejecting the null hypothesis with 95% confidence level is 3.8415. The superscript * represents the threshold is exceeded (the number is also displayed in gray). In the parentheses, we report the number of quantile violations given by each model against the ideal number of violations.

(a)

Method\Asset	NASDAQ 100	HSI	Nikkei 225	FTSE 100	DAX
AR-GARCH-t	1.4133 (64/55)	**0.0201** (51/50)	2.3475 (121/105)	19.4220* (97/60)	1.2297 (58/50)
AR-EGARCH-t	2.4726 (67/55)	0.1855 (47/50)	1.8077 (119/105)	22.4351* (100/60)	5.2762* (67/50)
AR-GJR-t	1.4133 (64/55)	1.3672 (42/50)	1.3358 (117/105)	10.0344* (86/60)	0.3150 (54/50)
FHS	**0.0183** (56/55)	0.3321 (46/50)	2.9420 (88/105)	2.6612 (73/60)	0.9471 (57/50)
CEVT	3.3793 (42/55)	6.6343* (33/50)	6.5507* (80/105)	**0.0167** (61/60)	3.1719 (38/50)
CAViaR-s	0.3012 (51/55)	2.6444 (39/50)	0.1520 (109/105)	3.0717 (74/60)	3.1542 (63/50)
CAViaR-a	4.3063* (71/55)	0.7579 (44/50)	6.4844* (132/105)	1.3008 (69/60)	4.6992* (66/50)
LSTM-t	1.7354 (65/55)	0.0203 (49/50)	1.3358 (117/105)	15.6991* (93/60)	0.9471 (57/50)
LSTM-HTQF	1.2363 (47/55)	**0.0201** (51/50)	**0.0874** (102/105)	**0.0169** (59/60)	**0.1855** (47/50)

(b)

Method\Asset	USDEUR	USDGBP	USDJPY	USDAUD	Oil	Gold
AR-GARCH-t	1.3802 (96/85)	1.1037 (108/98)	65.4023* (187/98)	93.5326* (203/95)	2.3735 (75/62)	0.1735 (69/72)
AR-EGARCH-t	0.9981 (76/85)	57.9293* (33/98)	24.5914* (53/98)	88.2707* (20/95)	**0.0041** (62/62)	5.8437* (53/72)
AR-GJR-t	1.3802 (96/85)	1.5544 (110/98)	64.0866* (186/98)	95.0794* (204/95)	3.1644 (77/62)	**0.1680** (76/72)
FHS	**0.0479** (83/85)	6.2406* (74/98)	8.6883* (70/98)	7.9435* (69/95)	0.5037 (57/62)	1.0480 (64/72)
CEVT	4.6473* (66/85)	19.9744* (57/98)	156.1795* (5/98)	171.4737* (2/95)	1.8898 (52/62)	4.6548* (55/72)
CAViaR-s	4.0048* (104/85)	**0.0026** (98/98)	**0.0026** (97/98)	7.1128* (122/95)	0.0367 (61/62)	6.4948* (52/72)
CAViaR-a	2.8692 (101/85)	**0.0026** (98/98)	0.0234 (96/98)	9.8478* (127/95)	0.5037 (57/62)	5.2306* (54/72)
LSTM-t	1.1448 (95/85)	6.8108* (73/98)	4.0800* (118/98)	**3.2725** (78/95)	0.8751 (70/62)	1.6153 (62/72)
LSTM-HTQF	0.1873 (89/85)	2.2938 (83/98)	0.1254 (101/98)	**3.2725** (78/95)	3.6933 (48/62)	0.6069 (66/72)

8. VaR forecasts

In this section, we quantitatively compare the VaR forecasts using our model and using some well-known competing models on the out-of-sample sets of the assets mentioned in Section 6, through some statistical backtesting procedures of VaR forecasts, as well as the loss function criteria. Our competing models are mainly GARCH family, from which we select some popular ones for comparisons: AR-GARCH-t, AR-EGARCH-t, and AR-GJR-GARCH-t. The orders of the GARCH terms and ARCH terms are all 1. We also compare with some other popular models for forecasting VaR, including FHS, CEVT, symmetric CAViaR, and asymmetric CAViaR. We adopt the rolling-window setting for all models.

Table 3
Independence test for $\alpha = 0.01$ VaR forecasts. The test statistic shown below has an asymptotically chi-square distribution with one degree of freedom. The threshold for rejecting the null hypothesis with 95% confidence level is 3.8415. The superscript * represents the threshold is exceeded (the number is also displayed in gray).

(a)

Method\Asset	NASDAQ 100	HSI	Nikkei 225	FTSE 100	DAX
AR-GARCH-t	1.4910	1.0514	1.4565	0.1131	0.1418
AR-EGARCH-t	3.5885	0.8922	1.5696	0.8962	**0.0115**
AR-GJR-t	1.4910	0.7117	1.6878	0.4150	0.2512
FHS	0.2705	0.8544	4.0652*	1.0576	0.1660
CEVT	0.9379	**0.4386**	4.9888*	0.2006	1.0885
CAViaR-s	0.4544	0.6133	2.2140	**0.0083**	0.0507
CAViaR-a	**0.0075**	0.7814	**0.0727**	0.0508	0.0186
LSTM-t	1.4112	0.9701	6.3550*	0.1561	4.6768*
LSTM-HTQF	3.3355	1.0514	0.8317	0.2551	0.5307

(b)

Method\Asset	USDEUR	USDGBP	USDJPY	USDAUD	Oil	Gold
AR-GARCH-t	2.1936	0.4603	0.8649	3.8543*	1.0573	**0.1578**
AR-EGARCH-t	1.3715	0.2242	1.0843	4.5127*	0.2070	0.6854
AR-GJR-t	2.1936	0.4014	0.8301	3.9227*	0.9098	1.3137
FHS	1.6371	1.1320	1.0125	1.0098	1.0494	0.2778
CEVT	1.0331	0.6705	0.0051	**0.0008**	0.8727	0.5950
CAViaR-s	0.0645	0.8242	0.8673	0.1131	1.2027	0.7336
CAViaR-a	0.4559	0.8242	**0.0032**	0.0523	1.0494	0.6393
LSTM-t	**0.0038**	0.3069	1.3523	4.6918*	**0.0560**	5.6604*
LSTM-HTQF	2.9766	**0.1099**	2.4897	0.1754	0.7498	5.0284*

In order to improve the flexibility of our HTQF for better performance, we can replace Z_α in the HTQF by other baseline quantile function, e.g., the one of t-distribution, and re-define the HTQF as:

$$Q(\alpha|\mu,\sigma,u,v) = \mu + \sigma Z_\alpha^v \left(\frac{e^{uZ_\alpha^v}}{A} + \frac{e^{-vZ_\alpha^v}}{A} + 1 \right), \tag{36}$$

where Z_α^v is the quantile function of t-distribution with v degrees of freedom. When $v = +\infty$, Z_α^v becomes the quantile function of a standard normal and the above definition is equivalent to the original one. In the forecasting, in every rolling, we select the best v from the set $\{4, 6, 8, 10, +\infty\}$ for our model using the validation set. We treat it as a hyper-parameter and choose the one that yields the minimum loss on the validation set. It is not easy to train the model if we set v to be time-varying and to be the neural network output, because generally neural network training needs symbolic derivatives of the loss function with respect to model parameters.

To figure out whether LSTM and HTQF both help in the modeling, we implement a degenerated model of our LSTM-HTQF, in which we set $u = v = 0$ and also use Z_α^v as the baseline quantile function. Now HTQF becomes the quantile function of a t-distribution, and the degrees of freedom v is obtained by applying an AR-GARCH-t model first. This new model denoted as LSTM-t is very similar to AR-GARCH-t. They use the same innovation distribution, with the only difference that the dependence on past information set is modeled by an LSTM now, instead of the linear autoregressive way. It is also different from LSTM-HTQF that the tail heaviness does not vary with time. We compare it with AR-GARCH-t to see whether LSTM helps to model the dependence, and compare it with LSTM-HTQF to see whether HTQF helps to model the conditional distribution well.

8.1. Backtesting

On daily return data, the evaluation of VaR forecasts is done through backtesting. Consider a sequence of realized returns or observations $\{r_{t'}\}$ on out-of-sample set and a sequence of VaR forecasts $\{q_{t'}\}$ for a fixed probability level α by any model. In order to implement the testing procedure, we need the definition of hitting sequence of quantile violations:

Table 4
Conditional coverage test for $\alpha = 0.01$ VaR forecasts. The test statistic shown below has an asymptotically chi-square distribution with two degree of freedom. The threshold for rejecting the null hypothesis with 95% confidence level is 5.9915. The superscript * represents the threshold is exceeded (the number is also displayed in gray).

(a)

Method\Asset	NASDAQ 100	HSI	Nikkei 225	FTSE 100	DAX
AR-GARCH-t	2.9042	1.0714	3.8041	19.5351*	1.3715
AR-EGARCH-t	6.0612*	1.0777	3.3773	23.3313*	5.2877
AR-GJR-t	2.9042	2.0790	3.0237	10.4494*	**0.5663**
FHS	**0.2888**	1.1866	7.0073*	3.7188	1.1132
CEVT	4.3172	7.0729*	11.5395*	**0.2173**	4.2603
CAViaR-s	0.7556	3.2578	2.3660	3.0799	3.2050
CAViaR-a	4.3138	1.5394	6.5571*	1.3516	4.7178
LSTM-t	3.1467	**0.9905**	7.6909*	15.8552*	5.6239
LSTM-HTQF	4.5718	1.0714	**0.9191**	0.2721	0.7162

(b)

Method\Asset	USDEUR	USDGBP	USDJPY	USDAUD	Oil	Gold
AR-GARCH-t	3.5737	1.5639	66.2672*	97.3869*	3.4308	**0.3313**
AR-EGARCH-t	2.3696	58.1535*	25.6757*	92.7834*	**0.2110**	6.5291*
AR-GJR-t	3.5737	1.9558	64.9167*	99.0021*	4.0742	1.4817
FHS	1.6851	7.3726*	9.7008*	8.9533*	1.5532	1.3258
CEVT	5.6804	20.6448*	156.1846*	171.4745*	2.7625	5.2498
CAViaR-s	4.0693	**0.8268**	0.8699	7.2259*	1.2393	7.2285*
CAViaR-a	3.3251	**0.8268**	**0.0266**	9.9001*	1.5532	5.8698
LSTM-t	**1.1485**	7.1177*	5.4323	7.9643*	0.9311	7.2756*
LSTM-HTQF	3.1639	2.4037	2.6152	**3.4479**	4.4431	5.6353

$$I_{t'} = \begin{cases} 1 & \text{if } r_{t'} < q_{t'} \\ 0 & \text{if } r_{t'} \geq q_{t'} \end{cases}. \tag{37}$$

Ideally, $\{I_{t'}\}$ should be an i.i.d. Bernoulli distribution sequence with parameter α. To test that, we use three likelihood ratio tests.

The Kupiec's unconditional coverage test (Kupiec, 1995) checks if the unconditional distribution of $\{I_{t'}\}$ is the Bernoulli distribution, i.e., if the proportion of quantile violations is equal to α. The test statistic is given by:

$$\mathrm{LR}_{\mathrm{uc}} := -2\ln\left[\frac{(1-\alpha)^{n_0}\alpha^{n_1}}{(1-n_1/(n_0+n_1))^{n_0}(n_1/(n_0+n_1))^{n_1}}\right], \tag{38}$$

where n_0, n_1 are the number of zeros and ones respectively in the hit sequence $\{I_{t'}\}$.

Christoffersen's independence test (Christoffersen, 1998) checks if $I_{t'}$ is independent of $I_{t'-1}$, i.e., current violation (or not) is independent of previous violation (or not). Its test statistic reads:

$$\mathrm{LR}_{\mathrm{ind}} := -2\ln\left[\frac{(1-p)^{n_{00}+n_{10}}p^{n_{01}+n_{11}}}{(1-p_0)^{n_{00}}p_0^{n_{01}}(1-p_1)^{n_{10}}p_1^{n_{11}}}\right], \tag{39}$$

where n_{ij} is the number of observations of i followed by j in the hit sequence, and $p_0 = n_{01}/(n_{00}+n_{01})$, $p_1 = n_{11}/(n_{10}+n_{11})$, $p = (n_{01}+n_{11})/(n_{00}+n_{01}+n_{10}+n_{11})$.

A mixed conditional coverage test jointly checks these two null hypotheses: $\mathrm{LR}_{\mathrm{cc}} := \mathrm{LR}_{\mathrm{uc}} + \mathrm{LR}_{\mathrm{ind}}$. $\mathrm{LR}_{\mathrm{uc}}$ and $\mathrm{LR}_{\mathrm{ind}}$ are asymptotically chi-square distributed with 1 degree of freedom and $\mathrm{LR}_{\mathrm{cc}}$ are asymptotically chi-square distributed with 2 degrees of freedom. The VaR forecast model fails (hypothesis is rejected) if they exceed critical values under a certain confidence level, say 95% chosen by us in this paper. One can refer to Dias (2013) for details of the three tests. We report the statistics of these tests for $\alpha = 0.01$ VaR forecasts given by 9 models on the 11 representative assets in following Tables 2–4.

In the unconditional coverage test in Table 2, our LSTM-HTQF model performs quite well on all 11 assets without a single rejection, while other models all receive at least 3 rejections. Especially, on equity indexes, the test statistic given by our model is very close to 0, which means that the proportion of quantile violations is close to α and the tail risk is accurately forecasted. FHS method also performs well on equity indexes, but not on exchange rates. The performances of LSTM-t are

Table 5
The loss function between return realization and quantile forecasted reported below is defined in quantile regression. The loss is the sum for three different probability levels $\alpha = 0.01, 0.05, 0.1$ and is the average over the entire out-of-sample set. So this is a comparison of the three corresponding VaR forecasted. We highlight the model of the best performance with bold number.

Method Asset	0001.HK	0002.HK	0005.HK	0016.HK	0700.HK	0939.HK	0941.HK	1299.HK
AR-GARCH-*t*	0.101767	0.105293	0.090912	0.099954	0.109387	0.097239	0.10071	0.100917
AR-EGARCH-*t*	0.101757	0.105238	0.091302	0.100054	0.109297	0.097777	0.100662	0.100935
AR-GJR-*t*	0.101745	0.105404	0.090915	0.09995	0.109119	**0.097195**	0.100692	0.100816
FHS	0.101767	0.105274	0.090819	0.099961	0.109295	0.097283	0.100717	0.100912
CEVT	0.101753	0.105309	0.091063	0.100033	0.109257	0.097284	0.100718	0.100949
CAViaR-s	0.102273	0.106078	0.091906	0.100618	0.108616	0.099017	0.101241	0.101016
CAViaR-a	0.102101	0.105523	0.091809	0.100326	0.10847	0.099088	0.101139	0.100921
LSTM-*t*	0.101692	**0.104768**	0.09072	0.099233	0.108184	0.097621	0.100102	0.100742
LSTM-HTQF	**0.101470**	0.104943	**0.090436**	**0.098924**	**0.107948**	0.097660	**0.099885**	**0.100728**

mixed too. In the independence test in Table 3, there are few rejections overall and it seems the majority of models have acceptable performances. In the conditional coverage test in Table 4, we also receive no rejection while others receive at least two. In summary, our model beats all competing models in some cases and generates comparable forecast results in other cases. Besides, the performances of LSTM-*t* are good in some cases, but not as good as LSTM-HTQF, which indicates that both LSTM and HTQF contribute to the performance improvement of forecasts.

8.2. Loss function

On high-frequency data described in Section 6.2, the loss function in quantile regression is used for evaluating VaR forecasts instead of backtesting since it is more comprehensive to compare by different criteria. The loss function is given by $L_\alpha(r, q) = (\alpha - I(r < q))(r - q)$ where r is the realization of return in out-of-sample set and q is the quantile/VaR forecasted. We choose three different probability levels $\alpha = 0.01, 0.05,$ and 0.1 for VaR, sum their loss functions together, and take the average over the entire out-of-sample set. The final losses of all methods on the 5-minute return data of the eight stocks are reported in Table 5.

It is shown that on 6 out of 8 assets, our LSTM-HTQF model outperforms all the remaining models on the loss, verifying its significant power in high-frequency situation. Besides, LSTM-*t* model also performs better than GARCH family and other competitors, indicating that high-frequency data may have very long memory that is only captured by LSTM. Moreover, the improvements from LSTM-*t* to LSTM-HTQF verify our proposed HTQF is flexible and capable enough. All of these are consistent with the intuition and empirical evidence that high-frequency data has heavier tails and longer memory than low-frequency data. And the rich of the high-frequency data makes our model's ability fully released.

9. Conclusions

To summarize, we proposed a novel parametric HTQF to represent the asymmetric heavy-tailed conditional distribution of financial return series. The dependence of HTQF's four parameters on past information set was modeled by a non-parametric data-driven machine learning approach, the sequential neural network LSTM. The training of our LSTM-HTQF model was casted into a quantile regression formulation. After learning from data, our model captured the dynamics of the conditional distribution. We examined the dynamics of the higher moments and tail-side quantiles, which all show quite long memories that are independent of that responsible for volatility clustering. This imply that there exists extra risk factors that drive the dynamics of higher moments or tail-side quantiles. Besides, our method can forecast conditional VaR with better accuracy on a wide range of assets, comparing to some popular existing models.

In the future, more advanced models that can learn more elaborately the dynamics of the conditional distribution of financial time series are necessary, e.g., improving the flexibility of the HTQF or modifying the way how LSTM is used may be needed. Moreover, the feature vector sequence that is fed into the LSTM can contain more information than just the past return history. All of these aim at learning more details of the conditional distribution dynamics and producing more accurate VaR forecasts.

Appendix A. The proof of the existence of HTQF's unique probability distribution

The proof idea is to show that HTQF is continuously differentiable, is strictly monotonically increasing over (0,1), and approaches $-\infty/+\infty$ as α tends to 0/1. So the inverse function of HTQF exists and is a cumulative distribution function.

The HTQF has the specification:

$$Q(\alpha|\mu, \sigma, u, v) = \mu + \sigma Z_\alpha \left(\frac{e^{uZ_\alpha}}{A} + \frac{e^{-vZ_\alpha}}{A} + 1 \right) = \mu + \sigma g(Z_\alpha), \tag{40}$$

where $g(x) = x(\frac{e^{ux}}{A} + \frac{e^{-vx}}{A} + 1)$. Z_α is the quantile function of the standard normal distribution, so we only need to prove that $g(x)$ is continuously differentiable, is strictly monotonically increasing over $(-\infty, +\infty)$, and approaches $-\infty/+\infty$ as x tends

to $-\infty/+\infty$. Obviously $g(x)$ is continuously differentiable and $\lim_{x\to-\infty/+\infty} g(x) = -\infty/+\infty$. To prove the monotonicity, we calculate the derivative of $g(x)$:

$$g'(x) = \left(\frac{e^{ux}}{A} + \frac{e^{-vx}}{A} + 1\right) + x\left(u\frac{e^{ux}}{A} - v\frac{e^{-vx}}{A}\right) \tag{41}$$

$$= \frac{e^{ux}}{A}(1+ux) + \frac{e^{-vx}}{A}(1-vx) + 1 \tag{42}$$

$$= \frac{1}{A}h(ux) + \frac{1}{A}h(-vx) + 1. \qquad (h(x) = e^x(1+x)) \tag{43}$$

Next we prove $h(x) \geq -1$, $\forall x$. This is equivalent to $1 + x \geq -e^{-x}$, or $1 + x + e^{-x} \geq 0$, $\forall x$. A simple monotonic analysis on the function $1 + x + e^{-x}$ can reveal that its global minimum is reached at $x = 0$, so $1 + x + e^{-x} \geq 2 \geq 0$. So, $h(x) \geq -1$ and

$$g'(x) \geq -\frac{1}{A} - \frac{1}{A} + 1. \tag{44}$$

If we choose $A \geq 3$, then $g'(x) \geq -\frac{1}{3} - \frac{1}{3} + 1 = \frac{1}{3} > 0$ holds for all x. So $g(x)$ is strictly monotonically increasing and our proof is completed.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: a system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–283.

Albuquerque, R., 2012. Skewness in stock returns: reconciling the evidence on firm versus aggregate returns. Rev Financ. Stud. 25 (5), 1630–1673.

Bali, T.G., Mo, H., Tang, Y., 2008. The role of autoregressive conditional skewness and kurtosis in the estimation of conditional var. J. Bank. Financ. 32 (2), 269–282.

Barone-Adesi, G., Giannopoulos, K., Vosper, L., 1999. Var without correlations for portfolios of derivative securities. J. Futures Mark. 19 (5), 583–602.

BIS I., Margin requirements for non-centrally cleared derivatives, https://www.bis.org/bcbs/publ/d317.htm, 2013.

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

Brunnermeier, M.K., Pedersen, L.H., 2008. Market liquidity and funding liquidity. Rev. Financ. Stud. 22 (6), 2201–2238.

Carnero, M.A., Peña, D., Ruiz, E., 2004. Persistence and kurtosis in garch and stochastic volatility models. J. Financ. Econ. 2 (2), 319–342.

Chen, J., Hong, H., Stein, J.C., 2001. Forecasting crashes: trading volume, past returns, and conditional skewness in stock prices. J. Financ. Econ. 61 (3), 345–381.

Chernozhukov, V., Fernández-Val, I., Galichon, A., 2010. Quantile and probability curves without crossing. Econometrica 78 (3), 1093–1125.

Christoffersen, P.F., 1998. Evaluating interval forecasts. Int. Econ. Rev. (Philadelphia) 841–862.

Committee, B., et al., 2016. Minimum capital requirements for market risk. Consultative Document.

Cont, R., 2001. Empirical properties of asset returns: stylized facts and statistical issues. Quant. Financ. 1 (2), 223–236.

Cont, R., 2007. Volatility clustering in financial markets: empirical facts and agent-based models. In: Long Memory in Economics. Springer, pp. 289–309.

Cont, R., Wagalath, L., 2016. Fire sales forensics: measuring endogenous risk. Math. Financ. 26 (4), 835–866.

Dias, A., 2013. Market capitalization and value-at-risk. J. Bank. Financ. 37 (12), 5248–5260.

Easley, D., De Prado, M.L., O'Hara, M., 2011. The microstructure of the flash crash: flow toxicity, liquidity crashes and the probability of informed trading. J. Portf. Manag. 37 (2), 118–128.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econ.: J. Econ. Soc. 987–1007.

Engle, R.F., Manganelli, S., 2004. Caviar: conditional autoregressive value at risk by regression quantiles. J. Bus. Econ. Stat. 22 (4), 367–381.

Engle, R.F., Patton, A.J., 2001. What good is a volatility model? Quant. Financ. 1, 237–245.

Fleming, J., Kirby, C., 2003. A closer look at the relation between garch and stochastic autoregressive volatility. J. Financ. Econ. 1 (3), 365–419.

Franses, P.H., Van Der Leij, M., Paap, R., 2007. A simple test for garch against a stochastic volatility model. J. Financ. Econ. 6 (3), 291–306.

Geraci, M., Garbaravicius, T., Veredas, D., 2018. Short selling in extreme events. J. Financ. Stab 39, 90–103.

Glasserman, P., Wu, Q., 2018. Persistence and procyclicality in margin requirements. Manag. Sci 64 (12), 5705–5724.

Glosten, L.R., Jagannathan, R., Runkle, D.E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. J. Financ. 48 (5), 1779–1801.

Hansen, B.E., 1994. Autoregressive conditional density estimation. Int. Econ. Rev. (Philadelphia) 705–730.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Keskar N.S., Mudigere D., Nocedal J., Smelyanskiy M. and Tang P.T., 2016. On large-batch training for deep learning: generalization gap and sharp minima. arXiv:1609.04836.

Kingma D.P. and Ba J., 2014 Adam: a method for stochastic optimization. arXiv:1412.6980.

Kirilenko, A., Kyle, A.S., Samadi, M., Tuzun, T., 2017. The flash crash: high-frequency trading in an electronic market. J. Financ. 72 (3), 967–998.

Koenker, R., Bassett Jr., G., 1978. Regression quantiles. Econ.: J. Econ. Soc. 33–50.

Koenker, R., Hallock, K.F., 2001. Quantile regression. J. Econ. Perspect. 15 (4), 143–156.

Kou, S., Peng, X., 2016. On the measurement of economic tail risk. Oper. Res. 64 (5), 1056–1072.

Kou, S., Peng, X., Heyde, C.C., 2013. External risk measures and basel accords. Math. Oper. Res. 38 (3), 393–417.

Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. J. Deriv. 3 (2), 73–84.

Le Cun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436.

León, A., Rubio, G., Serna, G., 2005. Autoregresive conditional volatility, skewness and kurtosis. Q. Rev. Econ. Financ. 45 (4–5), 599–618.

Lipton Z.C., Berkowitz J. and Elkan C., 2015. A critical review of recurrent neural networks for sequence learning. arXiv:1506.00019.

Mc Neil, A.J., Frey, R., 2000. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. J. Empir. Financ. 7 (3–4), 271–300.

Mullainathan, S., Spiess, J., 2017. Machine learning: an applied econometric approach. J. Econ. Perspect. 31 (2), 87–106.

Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: a new approach. Econ.: J. Econom. Soc. 347–370.

Rockinger, M., Jondeau, E., 2002. Entropy densities with an application to autoregressive conditional skewness and kurtosis. J. Econ. 106 (1), 119–142.

Takeuchi, I., Le, Q.V., Sears, T.D., Smola, A.J., 2006. Nonparametric quantile estimation. J. Mach. Learn. Res. 7, 1231–1264.

Taylor, S.J., 1994. Modeling stochastic volatility: a review and comparative study. Math. Financ. 4 (2), 183–204.

Yan, X., Zhang, W., Ma, L., Liu, W., Wu, Q., 2018. Parsimonious quantile regression of financial asset tail dynamics via sequential learning. Neural Information Processing Systems (NIPS).