第 50 卷第 12 期 中国科学技术大学学报 Vol. 50, No. 12
2020 年 12 月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Dec. 2020

Received: 2020-02-03; Revised: 2020-06-11 doi:10.3969/j.issn.0253-2778.2020.12.005

L_p quantile regression with realized measure

TANG Li, CHEN Yu*

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

* Corresponding author. E-mail: cyu@ustc.edu.cn

Abstract: A new financial risk model named L_p quantile regression with a realized measure (realized L_p quantile) was proposed. The realized measure and L_p quantiles were combined and L_p quantile were added to the measurement equation. The realized L_p quantile model is a generic model that includes realized quantile model and expectile model. An asymmetric exponential power distribution (AExpPow) was proposed to derive the formula of log-likelihood. And a simulation was conducted to justify the validity of the log-likelihood. Finally an empirical study was conducted to justify the validity of the realized L_p quantile. And some conclusions were drawn as follows: differfent power indices suit different data and different time-frequencies suit different realized measures, and higher frequency is not always better.

Keywords: realized measure; realized L_p quantile regression; asymmetric exponential power distribution

CLC number: F830.9;F064.1 **Document code**: A

2010 Mathematics Subject Classification: 62J02

1 Introduction

In recent decades, quantitative financial risk measurements have become more and more fundamental in investment decisions, capital allocation, and regulation. Among them, quantiles are one of the most popular measurements. It is the minimizer of an asymmetric linear loss function. Koenker and Bassett[1] exploited this property to propose quantile regression. Quantiles satisfy the property of robustness but lack sensitiveness. If we modify the loss function to be minimized, we will get different statistical functions. Therefore, different generalized quantiles are created. Newey, Powell[2] and Efron[3] proposed expectile regression, one of the generalized quantiles, which switches asymmetric linear loss function to asymmetric least square. Expectile is sensitive but lacks the property of robustness. Chen[4] proposed the L_p quantile that minimizes asymmetric power function, combining the property of quantile and expectile. L_p quantile is both sensitive and robust. Breckling and Chamber[5] considered a generic asymmetric loss function including the class of M-quantile.

Volatility estimation plays a significant role in almost all quantitative financial risk measurements, including the estimation of VaR, expectile, and L_p quantile. Parkinson[6], Garman and Klass[7] considered the daily high-low range as an improved volatility estimator compared to the daily return. Since Engle[8] introduced auto-regressive conditionally heteroskedasticity (ARCH) model, different volatility estimators have been proposed in the past decades. Bollerslev[9] proposed generalized ARCH (GARCH) in 1986, which is a big step for volatility estimation. Since high-frequency intra-day data is available to us now, we can calculate realized estimators more precisely. Realized estimators include realized variance (RV) [10] and realized range (RR) [11], etc. Regarding the volatility modeling, Hansen et al.[12] proposed a volatility framework named realized-GARCH, which incorporates a measurement equation that connects the realized estimators to return equation. Bee et al.[13] extended realized GARCH to realized quantile. Gerlach et al.[14] introduced CARE model with realized estimators, which was an extension of expectile estimation.

In this paper, we combine realized measures with L_p quantile regression to propose a generic framework named realized L_p quantile regression, which is analogous to realized quantile. realized L_p quantile regression adds a measurement equation that links the latent conditional L_p quantile with realized measures into the conventional L_p quantile model. Additionally we find that minimizing the L_p loss function is equal to maximizing the likelihood when adopting asymmetric exponential power distribution. To evaluate the forecast performance, we adopt the L_p loss function as the

Citation: TANG Li, CHEN Yu. L_p quantile regression with realized measure. J. Univ. Sci. Tech. China, 2020, 50(12): 1478-1487.

penalty function. We find $p=1.2$ and $p=1.5$ are the best indices, neither $p=1$ (quantile) nor $p=2$ (expectile). It means different indices are applicable to different p, but $p=1.2$ and $p=1.5$ are more likely to be accepted. And the frequency of the realized measures around 2 to 5 min is more acceptable.

The paper is organized as follows: Section 2 presents the model. The realized measures will be introduced in Section 3. In Section 4, we will introduce asymmetric exponential power distribution and the likelihood adopting asymmetric exponential power distribution. Simulation and empirical study are discussed in Section 5 and Section 6, respectively. Finally, Section 7 concludes the paper.

2 Quantile, expectile and L_p quantile with realized measures

Bee et al.[13] proposed a realized quantile model. Let r_t be the portfolio return at time t, x_t be a realized measure observable at time t and θ be the probability associated with the quantile regression model. And let $(\beta(\theta), \gamma(\theta))$ be a vector of parameters associated respectively with past conditional quantiles and the realized measures. The general structure can be written as the following system of equations:

$$r_t = q_t^\theta + \epsilon_t^\theta \tag{1}$$

$$q_t^\theta = f(q_{t-1}^\theta, \cdots, q_{t-p}^\theta, x_{t-1}, \cdots, x_{t-q}; \beta(\theta), \gamma(\theta)) \tag{2}$$

$$\begin{aligned} x_t = {} & \omega(\theta) + \phi(\theta) q_t^\theta + \tau_1(\theta) z_t^\theta + \\ & \tau_2(\theta)[(z_t^\theta)^2 - 1] + u_t \end{aligned} \tag{3}$$

where ϵ_t^θ is such that given the information to time $t-1$, the θ quantile of ϵ_t^θ is equal to 0; $z_t^\theta = r_t / q_t^\theta$, $u_t \sim N(0, \sigma_u^2)$. The function $\tau_1(\theta) z_t^\theta + \tau_2(\theta)[(z_t^\theta)^2 - 1]$ is called the leverage function because it captures the dependence between return and future volatility, according to Ref. [12]. The equations (1)-(3) are called the return equation, the quantile equation and the measurement equation, respectively.

The model with a linear specification is defined by the following quantile and measurement equation:

$$q_t^\theta = \beta_1 + \beta_2 q_{t-1}^\theta + \beta_3 x_{t-1} \tag{4}$$

$$x_t = \xi + \phi q_t^\theta + \tau_1 z_t^\theta + \tau_2[(z_t^\theta)^2 - 1] + u_t \tag{5}$$

where z_t and u_t share the same meaning mentioned above. $\beta_1, \beta_2, \beta_3, \xi, \phi, \tau_1, \tau_2$, and σ_u are the parameters to be estimated.

The model can adopt the quantile regression model as loss function to estimate parameters. Consider the return equation only, the loss function can be written as follows:

$$\rho_t^\theta(r_t, q_t) = |\theta - I\{r_t < q_t\}| \, |r_t - q_t| \tag{6}$$

Based on this model and loss function, we can estimate quantiles more precisely, and this is a good way to incorporate high-frequency data into models. What's more, we can forecast returns.

Bee et al.[13] proposed quasi-maximum likelihood to estimate the parameters. The logarithm of the tick-exponential density is proportional to the function ρ_t^θ, which means minimizing ρ_t^θ is equal to maximizing the likelihood. According to Ref. [15], the quantile regression minimization of expression (6) is equivalent to maximizing likelihood based on the asymmetric Laplace density. Gerlach et al.[14] proposed the model that switches the power index of the loss function (6) from 1 to 2. The new loss function is

$$\rho_t^\theta(r_t, q_t) = |\theta - I\{r_t < q_t\}| \, |r_t - q_t|^2 \tag{7}$$

This is exactly the expectile regression. And they found that by adopting asymmetric Gaussian density, the maximization of the likelihood function is equivalent to minimizing the loss function (7).

In this paper, we let p represent the power index of the loss function, where $1 \leqslant p \leqslant 2$. see Eq. (8). In addition, we combine Eq. (8) with return equation (1), quantile equation (4), and measurement equation (5). The total structure is named realized L_p quantile regression.

$$\left.\begin{array}{l} \rho_t^\theta(r_t, q_t) = |\theta - I\{r_t < q_t\}| \, |r_t - q_t|^p, \\ 1 \leqslant p \leqslant 2 \end{array}\right\} \tag{8}$$

Realized L_p quantile regression is a generic model including quantile regression and expectile regression. When $p=1$, the model is quantile regression. When $p=2$, the model is expectile regression.

We find minimizing Eq. (8) is equivalent to maximizing the likelihood function when adopting asymmetric exponential power distribution. More details will be discussed in Section 3.

3 Realized measures

This section introduces different volatility estimators, especially the realized variance (RV) and realized range (RR). Since we concentrate on the comparison of realized measures with different time-frequencies, we adopt RV and RR to be our realized measures.

Let H_t, L_t, and C_t be the daily high, daily low, and closing prices in day t respectively. The daily return is the difference between the consecutive log daily closing prices, which is

$$r_t = \ln C_t - \ln C_{t-1} \tag{9}$$

Then the daily range (DR) proposed by Ref. [8] is calculated as follows:

$$\mathrm{DR}_t^2 = \frac{(\ln H_t - \ln L_t)^2}{4 \ln 2} \tag{10}$$

where 4ln2 scales DR_t^2 to be an unbiased return variance estimator. Supposing that day t is divided into N equally sized intervals of length Δ, we have the subscription of each intra-day set $\Theta = 0, 1, 2, \cdots, N$ and can calculate the high-frequency volatility measures. For day t, denote the i^{th} interval closing price as $P_{t-1+i\Delta}$. Then $H_{t,i}=\sup_{(i-1)\Delta<j<i\Delta}P_{t-1+j}$ and $L_{t,i}=\inf_{(i-1)\Delta<j<i\Delta}P_{t-1+j}$ represent the high and low prices during this interval. The RV proposed by Ref. [10] is calculated as follows:

$$RV_t^{\Delta}=\sum_{i=1}^{N}[\ln(P_{t-1+i\Delta})-\ln(P_{t-1+(i-1)\Delta})]^2 \quad (11)$$

Further, Ref. [11] developed the realized range (RR), which sums the intra-day range.

$$RR_t^{\Delta}=\frac{\sum_{i=1}^{N}(\ln H_{t,i}-\ln L_{t,i})^2}{4\ln 2} \quad (12)$$

In this paper, we use RV and RR with different frequencies such as 1, 2, 3, 4, 5, 10, and 20 min to be the realized measures. And we will select which measure performs best.

4 Asymmetric exponential power distribution and likelihood

4.1 Asymmetric exponential power distribution

Ref. [16] proposed an exponential power distribution. We will modify the distribution to an asymmetric distribution so the kernel of a probability density function (PDF) for the asymmetric exponential power distribution random variable is exactly the loss function of L_p quantile regression. Minimizing the loss function of L_p quantile regression is equivalent to maximizing the likelihood function when adopting our asymmetric exponential power distribution. Now we introduce an exponential power distribution and our asymmetric exponential power distribution. The PDF of exponential power distribution is

$$f(x\mid\sigma,p)=\frac{1}{2\sigma^{1/p}\Gamma(1+1/p)}\exp\left(-\frac{|x|^p}{\sigma}\right) \quad (13)$$

where σ is the scale factor, and p is the power index. Then, we will modify the distribution. For simplicity we let the scale parameter σ be 1.

The modified distribution is called asymmetric exponential power distribution, denoted by AExpPow (α,q,p), and the PDF is as follows:

$$f(x\mid\alpha,q,p)=2\left(\frac{\Gamma(1+1/p)}{|\alpha-1|^{\frac{1}{p}}}+\frac{\Gamma(1+1/p)}{\alpha^{\frac{1}{p}}}\right)^{-1}\cdot \exp(-|x-q|^p|\alpha-I(x<q)|) \quad (14)$$

where q is the mode, α is the shape parameter and p is the power index. Let $p=1$, the PDF be an asymmetric Laplace distribution, which can be used for quantile regression. Let $p=2$, the PDF is an asymmetric Gaussian distribution, which can be used for expectile regression.

Fig. 1 shows AExpPow (α,q,p) with different settings. We fix $\alpha=0.1$ and $q=0$ to observe the impact of the changes of p. We can see that the graph is asymmetric whatever p is. The left tails are all thinner

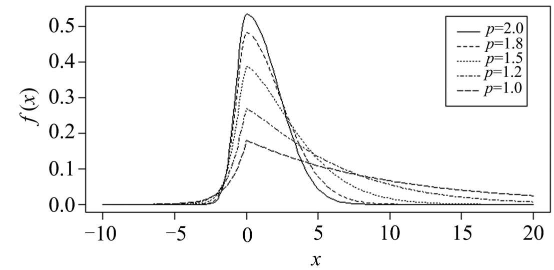

Fig. 1 Asymmetric exponential power distribution with different p from 1 to 2, where we set $q=0$ and $\alpha=0.1$.

than the right tails. The higher the p, the higher the peak and the thinner the right tail. When $p=2$, $q=0$ and $\alpha=0.5$, we get symmetric standard Gaussian distribution mentioned in Ref. [14].

4.2 Realized log-likelihood

After introducing asymmetric exponential power distribution, we can rewrite our return equation (1) for knowing the distribution of ϵ_t^{θ} as follows:

$$r_t=q_t^{\theta}+\epsilon_t^{\theta},\ \epsilon_t^{\theta}\sim \mathrm{AExpPow}(\alpha,0,p) \quad (15)$$

where ϵ_t^{θ} is an independent and identically distributed process with mode 0, shape parameter α, and power index p. The corresponding pseudo-log-likelihood based on a sample $r_1,r_2\cdots r_n$ from Eq. (15) is equivalent to

$$L(r;\delta)=-\sum_{t=1}^{n}|r_t-q_t^{\theta}|^p|\alpha-I(r_t<q_t^{\theta})| \quad (16)$$

where δ represents all the parameters needed to be estimated.

Adopting the asymmetric exponential power distribution, we see that the quasi-log-likelihood of Eq. (16) has a similar form to Eq. (8). Minimizing Eq. (8) is equivalent to maximizing Eq. (16).

Eq. (16) is a part of the full log-likelihood function. It can be used as a score function to compare forecast results with different power index p since it can exclude the influence of measurement equation. While the realized quantile framework has a quantile equation (4) and a measurement equation (5) with $u_t\sim N(0,\sigma_u^2)$, the full log-likelihood function equals to the sum of log-likelihood of return equation, $L(r;\delta)$, and log-likelihood of the measurement equation $L(x|r;\delta)$, where

$$u_t=x_t-\xi+\phi q_t^{\theta}+\tau_1 z_t^{\theta}+\tau_2[(z_t^{\theta})^2-1].$$

Therefore, the full log-likelihood is as follows:

$$L(r,x;\delta)=L(r;\delta)+L(x\mid r;\delta)=\underbrace{-\sum_{t=1}^{n}|r_t-q_t^{\theta}|^p|\alpha-I(r_t<q_t^{\theta})|}_{L(r;\delta)}+\underbrace{\left(-\frac{1}{2}\sum_{t=1}^{n}(\ln(2\pi)+\ln(\sigma_u^2)+u_t^2/\sigma_u^2)\right)}_{L(x|r;\delta)} \quad (17)$$

Given q_1, as x_t can be observed, q_t can be written as a formulation of x_t and δ by the iteration of Eq. (4). Then $L(r;\delta)$ is an expression only consisting of δ. So is $L(x|r;\delta)$. Then we use the optim function of R to solve the log-likelihood. And the method of the optim function we used is "L-BFGS-B".

In Section 6, we will use Eq. (17) to estimate the parameters of empirical data and forecast returns by the estimated parameters. Then we will use Eq. (16) as the score function to find for the best results.

5 Simulation study

A simulation study is conducted to illustrate whether the maximum likelihood approach can estimate the parameters well. Then we compare the simulation performance of different power index p.

We simulate 500 datasets from return equation (18), quantile equation (19) and measurement equation (20) for each ϵ_t^θ following different distributions including normal distribution, student t and different AExpPow distributions with different power index p. Each dataset consist of 4000 data of r and x.

$$r_t = q_t^\theta + \epsilon_t^\theta \tag{18}$$

$$q_t^\theta = -0.023 + 0.6 q_{t-1}^\theta - 0.17 x_{t-1} \tag{19}$$

$$x_t = 0.1 - 0.76 q_t^\theta + 0.02 z_t^\theta + 0.02[(z_t^\theta)^2 - 1] + u_t,\ u_t \sim N(0, 0.03^2) \tag{20}$$

Then we use different L_p quantile regression models to estimate parameters with these datasets respectively. We consider the mean and the standard error to compare the bias and the precision respectively. Some estimation results are summarised in Tabs. 1−5, where bold one represents the best result in each row.

Tab. 1 Simulation results with ϵ_t^θ following student t(8) distribution, n=500.

parameters	true	p=1		p=1.2		p=1.5		p=1.8		p=2	
		mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
β_1	−0.023	−0.0292	0.0753	**−0.0240**	**0.0040**	−0.0325	0.1702	−0.0257	0.0305	−0.0243	0.0111
β_2	0.6	0.5705	0.0568	**0.5746**	**0.0180**	0.5710	0.0626	0.5716	0.0558	0.5740	0.0184
β_3	−0.17	−0.1597	0.0247	**−0.1601**	**0.0183**	−0.1578	0.0330	−0.1579	0.0214	−0.1577	0.0202
ξ	0.1	0.0503	0.9926	**0.0987**	**0.0070**	0.0974	0.0747	0.0921	0.1071	0.0958	0.0629
ϕ	−0.76	−0.8877	0.9099	**−0.8443**	**0.0752**	−0.8529	0.1972	−0.8657	0.2491	−0.8605	0.1550
τ_1	0.02	**0.0172**	0.0038	0.0171	**0.0010**	0.0154	0.0369	0.0170	0.0010	0.0168	0.0024
τ_2	0.02	0.0167	0.0105	0.0159	0.0024	**0.0189**	0.0693	0.0157	0.0025	0.0155	**0.0023**
σ_u	0.03	0.0397	0.0289	**0.0384**	**0.0095**	0.0570	0.4140	0.0395	0.0239	0.0395	0.0188

Tab. 2 Simulation results with ϵ_t^θ following student t(5) distribution, n=500.

parameters	true	p=1		p=1.2		p=1.5		p=1.8		p=2	
		mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
β_1	−0.023	−0.0377	0.2301	−0.0286	0.0532	**−0.0251**	**0.0106**	−0.0280	0.0460	−0.0345	0.0594
β_2	0.6	0.5754	0.0759	0.5742	0.0790	**0.5780**	**0.0354**	0.5716	0.1046	0.5724	0.0932
β_3	−0.17	−0.1648	0.0293	−0.1655	0.0246	**−0.1665**	**0.0214**	−0.1642	0.0232	−0.1634	0.0271
ξ	0.1	0.0948	0.0819	0.0911	0.1379	**0.1007**	**0.0119**	0.0897	0.1473	0.0721	0.5899
ϕ	−0.76	−0.8099	0.1301	−0.8262	0.3365	**−0.8077**	**0.0822**	−0.8429	0.4321	−0.8281	0.2486
τ_1	0.02	0.0162	0.0387	0.0183	0.0078	**0.0184**	0.0032	0.0181	**0.0019**	0.0135	0.0975
τ_2	0.02	**0.0182**	0.0095	0.0178	0.0039	0.0176	0.0032	0.0174	**0.0031**	**0.0218**	0.0981
σ_u	0.03	0.0472	0.0664	0.0451	0.0476	0.0427	0.0365	**0.0425**	**0.0339**	0.0594	0.3606

Tab. 3 Simulation results with ϵ_t^θ following normal distribution, $n=500$.

parameters	true	$p=1$		$p=1.2$		$p=1.5$		$p=1.8$		$p=2$	
		mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
β_1	−0.023	−0.0232	0.0035	−0.0228	0.0061	**−0.0230**	0.0035	−0.0229	**0.0035**	−0.0229	0.0035
β_2	0.6	0.5699	**0.0194**	0.5700	0.0214	0.5703	0.0197	0.5700	0.0198	**0.5698**	0.0199
β_3	−0.17	**−0.1536**	0.0162	−0.1535	0.0195	−0.1525	0.0164	−0.1518	0.0162	−0.1513	**0.0161**
ξ	0.1	**0.0964**	**0.0058**	0.0962	0.0060	0.0962	0.0059	0.0962	0.0059	0.0961	0.0059
ϕ	−0.76	**−0.8869**	**0.0674**	−0.8896	0.0689	−0.8927	0.0694	−0.8966	0.0687	−0.8999	0.0691
τ_1	0.02	**0.0159**	0.0005	0.0156	0.0052	0.0158	0.0005	0.0158	**0.0005**	0.0157	0.0005
τ_2	0.02	**0.0143**	0.0017	0.0142	0.0018	0.0142	0.0017	0.0140	0.0017	0.0139	**0.0017**
σ_u	0.03	0.0360	0.0057	0.0373	0.0283	**0.0359**	0.0057	0.0360	**0.0045**	0.0360	0.0045

Tab. 4 Simulation results with ϵ_t^θ following AExpPow distribution with $p=1$, $n=500$.

parameters	true	$p=1$		$p=1.2$		$p=1.5$		$p=1.8$		$p=2$	
		mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
β_1	−0.023	−0.0214	0.0032	**−0.0218**	0.0109	−0.0212	**0.0031**	−0.0211	0.0031	−0.0216	0.0169
β_2	0.6	**0.5870**	**0.0143**	0.5848	0.0517	0.5868	0.0146	0.5865	0.0149	0.5843	0.0486
β_3	−0.17	**−0.1539**	0.0205	−0.1530	0.0211	−0.1525	0.0202	−0.1511	**0.0197**	−0.1491	0.0205
ξ	0.1	0.0944	0.0034	**0.0947**	0.0077	0.0942	0.0033	0.0940	**0.0033**	0.0926	0.0263
ϕ	−0.76	**−0.8746**	0.0762	−0.8774	0.0766	−0.8827	0.0763	−0.8905	**0.0756**	−0.9045	0.1293
τ_1	0.02	**0.0164**	**0.0007**	0.0163	0.0023	0.0163	0.0007	0.0162	0.0007	0.0159	0.0038
τ_2	0.02	**0.0153**	0.0029	0.0152	0.0029	0.0151	0.0028	0.0148	0.0027	0.0144	**0.0026**
σ_u	0.03	0.0367	**0.0074**	0.0376	0.0230	0.0366	0.0074	**0.0366**	0.0075	0.0377	0.0278

Tab. 5 Simulation results with ϵ_t^θ following AExpPow distribution with $p=2$, $n=500$.

parameters	true	$p=1$		$p=1.2$		$p=1.5$		$p=1.8$		$p=2$	
		mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
β_1	−0.023	−0.0264	0.0768	**−0.0229**	**0.0053**	−0.0261	0.0935	−0.0255	0.03667	−0.0364	0.0781
β_2	0.6	0.5885	0.0377	0.5873	0.0564	0.5865	0.0619	**0.5914**	**0.0373**	0.5846	0.0662
β_3	−0.17	−0.1687	0.1162	**−0.1610**	**0.0189**	−0.1505	0.0541	−0.1004	0.1771	−0.0324	0.2898
ξ	0.1	0.0947	0.0491	**0.0965**	0.0070	0.0960	**0.0043**	0.0934	0.0876	0.1137	0.3772
ϕ	−0.76	**−0.8131**	**0.0712**	−0.8434	0.4006	−0.8578	0.0975	−0.9520	0.2562	−1.1885	0.8190
τ_1	0.02	0.0131	0.1100	**0.0177**	**0.0018**	0.0174	0.0210	0.0252	0.0864	0.0302	0.3277
τ_2	0.02	**0.0174**	0.0068	0.0171	**0.0029**	0.0157	0.0044	0.0130	0.0130	0.0076	0.0182
σ_u	0.03	0.0419	0.0625	**0.0401**	**0.0265**	0.0404	0.0377	0.0623	0.1692	0.1345	0.9715

In these tables, the bold parameters are preferred to others for both bias (mean) and precision (standard deviation, SD). We can see that all models generate close to unbiased and quite reasonably precise parameter estimation. It provides an evidence that maximum likelihood approach is a good approach to estimate the parameters in this model when adopting the asymmetric exponential power distribution.

However, there are still some small differences between models with different power index. When considering bias, we count the estimated parameters closest to the true value, which are bold in the mean column. When ϵ_t^θ follows the normal distribution, power index $p=1$ is preferred than other power indices, with 5 of 8 parameters outperforming others. When ϵ_t^θ follows student t(8) distribution, power index $p=1.2$ is preferred, with 6 parameters estimated better than others. When ϵ_t^θ follows student t(5) distribution, power index $p=1.5$ is preferred, 6 of 8 parameters are estimated better. When ϵ_t^θ follows AExpPow distribution with $p=1$, power index $p=1.2$ is preferred. When ϵ_t^θ follows AExpPow distribution with $p=2$, power index $p=1$ is preferred. All the standard deviations are at an acceptable low level.

We can conclude that in simulation, datasets with different shapes need different power indices to model. We need to select the best one by estimation results before forecast. However, the simulation also indicates that lower power indices such as $p=1$ and $p=1.2$ are more preferable than higher power indices.

6 Empirical study

6.1 Data description

All market data including daily open, daily close, daily high, daily low prices as well as 1-minute open, 1-minute closing 1-minute high and 1-minute low price data are downloaded from Bloomberg. We collect the S&P 500 Index to represent the market index. The time range is from January 2008 to May 2019, with a total of 2853 trading days. The RV5min data for AEX, FTSE, GDAXI, and N225 come from the Oxford-Man Institute " Realized Library", ranging from January 2001 to October 2019.

The daily return is calculated using daily price data by Eq. (9), which includes overnight jumps. We plot the daily return of S&P 500 Index in Fig. 2. Fig. 2 exhibits the biggest fluctuant of returns that occurred at

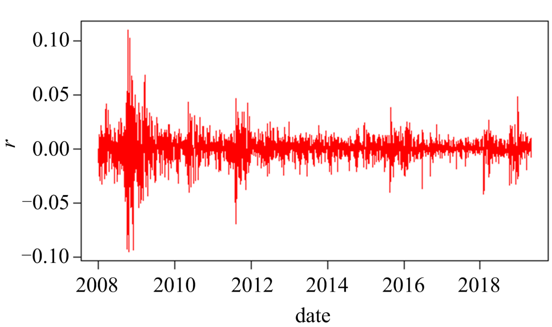

Fig. 2 Daily return of S&P 500 Index from January 2008 to May 2019.

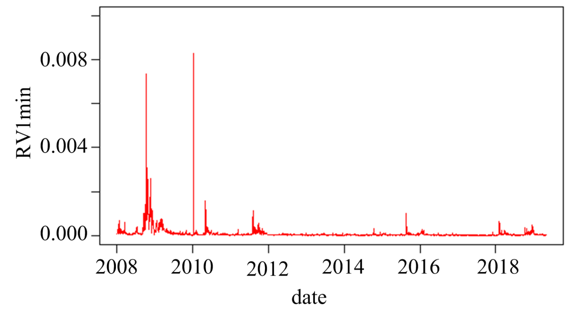

Fig. 3 Value of RV1min at different times.

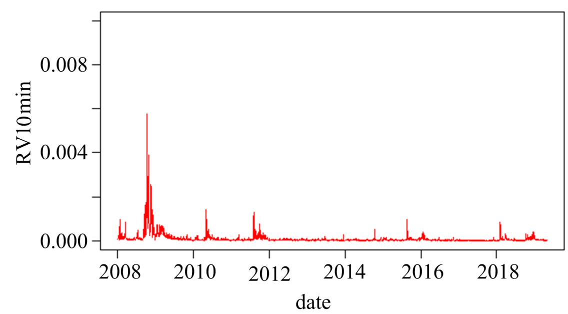

Fig. 4 Value of RV10min at different times.

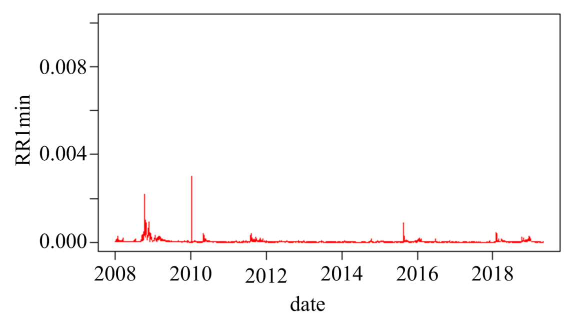

Fig. 5 Value of RR1min at different times.

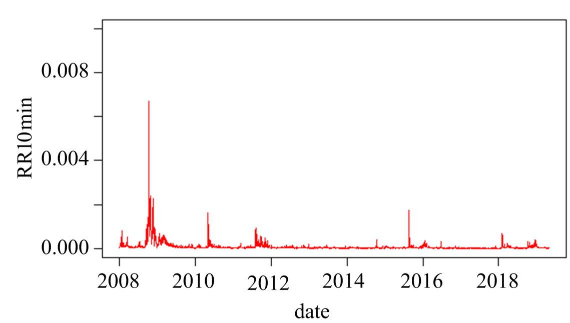

Fig. 6 Value of RR10min at different times.

the end of the year 2008, which is exactly the global financial crisis period.

We adopt RV and RR as our realized measures, which are calculated by Eqs. (11) and (12) with different frequencies. We use time-frequencies of 1, 2, 3, 4, 5, 10, and 20 min for comparison to select the most proper time-frequency. Fig. 3–Fig. 6 show the

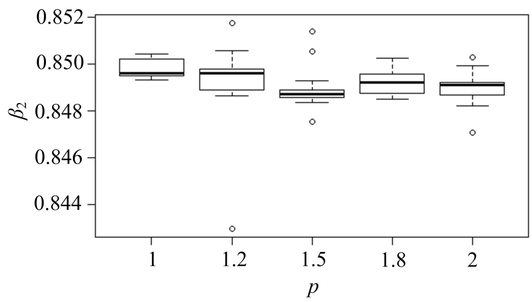

Fig. 7 Boxplot of β_2 for different realized measures with different p.

RV with 1 and 10 min and RR with 1 min and 10 min.

Different realized measures provide different patterns. We can see that RV1min has two much higher peaks in 2008 and 2010 than the other measures. In addition to different realized measures, they have different peaks at different times. That means they can capture different information.

In the next subsection, we will use 2200 days' daily data for estimation, approximately 9 years. The remaining 653 observations are reseverd for the out-of-sample evaluation. And we compare them with the out-of-sample data by using Eq. (16) as our score function. The lowest score provides the best prediction.

6.2 In-sample parameters estimation

In this subsection, we will use in-sample data to fit Eqs. (1), (4) and (5) to estimate β_1, β_2, β_3, ξ, ϕ, τ_1, τ_2, and σ_u. We use MLE to estimate our parameters. The likelihood function is Eq. (17). We choose 5 different power indices to estimate the parameters, which are 1, 1.2, 1.5, 1.8, and 2. All the computations are done with optim function of R.

We choose $\alpha = 0.1$, which produces 10%-L_p quantile. When $p=1$, the result is 10%-quantile, and when $p=2$, the result is 10%-expectile.

The value of β_2, see Fig. 7, is around 0.85 in all realized measures with different p. It is very close to 1, which means the L_p quantile is mostly determined by its previous value, that is highly persistent. The parameter ϕ, see Fig. 8, is around -0.92 regardless of power index p and realized measures, which is nearly -1 and is negative. That means that realized measures are influenced by the quantile of the same period to large extent and their correlation is negative.

The parameter β_3, see Fig. 9, is significant regardless of the realized measure used and the power index. They almost have the same value of -0.1, which means the previous realized measure also contributes to L_p quantile to some extent but negatively.

Another interesting finding is that the leverage parameters τ_1 and τ_2, see Fig. 10, are always significant. The $\tau_1 z_t + \tau_2 (z_t^2 - 1)$ represents a leverage function as we mentioned in Section 2. The significant parameters value mean they are indispensable. Adding them to the measurement equation can improve performance. However, while τ_1 and τ_2 are very small in most circumstances, their contribution to realized measure x_t and L_p quantile is very small. Nevertheless, they are not ignorable.

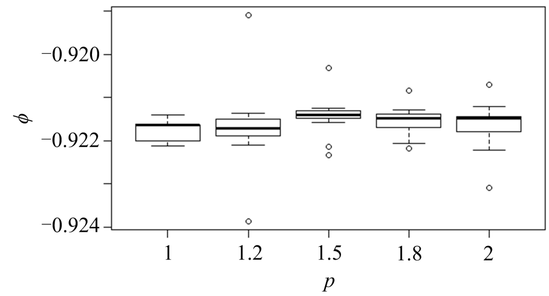

Fig. 8 Boxplot of ϕ for different realized measures with different p.

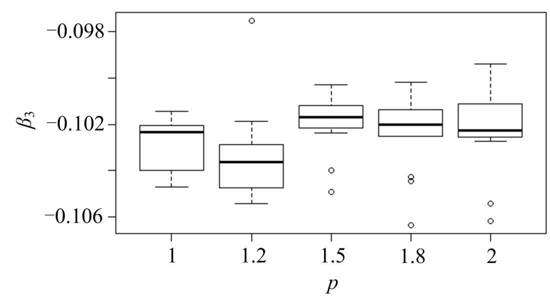

Fig. 9 Boxplot of β_3 for different realized measures with different p.

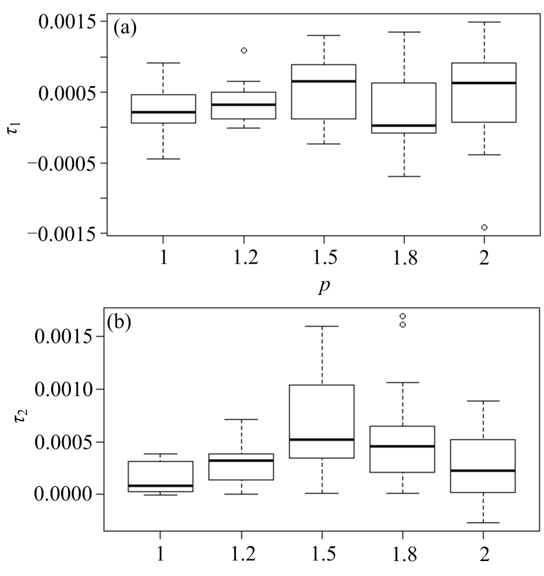

Fig. 10 Boxplot of τ_1 and τ_2 for different realized measures with different p.

The parameters estimated are very close regardless of power index p and realized measures. In the next subsection, we will compare the prediction results with different power index p and realized measures.

6.3 Out-of-sample forecast

In this section, we compare the out-of-sample forecasting performance of L_p regression with different power indices and different realized measures. We consider the loss function of Eq. (16) introduced in Section 4, but with an adaption. The score function is

$$L(r;\theta) = \sum_{t=1}^{n} | r_t - q_t^\theta || \alpha - I(r_t < q_t^\theta) | \quad (21)$$

The parameters are estimated in the previous subsection by 2200 in-sample-data. We use the estimated parameters to forecast 653 returns. Then we compare the forecast return to the real data to get the absolute error. By using Eq. (21) as score function, we get different absolute errors in the different models, which we call them scores. The scores with different power index and different realized measures are shown in Tab. 6. The lower the scores, the better the model.

The boxplot of scores by power index p is shown in Fig. 11. Though through Tab. 6 we see that the best three are $p=1.5$ with RV4min, $p=1.8$ with RV3min and $p=2$ with RV3min, Fig. 11 shows that these three results are all outliers. Fig. 11 shows that $p=1$ and $p=1.2$ outperform others on average. The scores in $p=1$ and $p=1.2$ are very close, but when we consider the extreme value, $p=1$ is better than $p=1.2$ here. However it seems the results are influenced by the data we use. Other datasets may support others. So, we can only say that $p=1$ is preferred in our results by S&P 500 Index.

Tab. 6 Scores for different power indices and different realized measures.

	$p=1$	$p=1.2$	$p=1.5$	$p=1.8$	$p=2$
RV1min	1.751 0	2.039 0	2.507 1	2.367 1	2.155 1
RR1min	2.064 9	2.840 9	3.115 4	2.112 1	2.805 5
RV2min	1.981 1	1.584 1	2.620 6	2.513 8	2.377 3
RR2min	2.640 1	2.870 7	3.033 5	3.189 3	2.776 1
RV3min	2.028 7	3.532 9	2.246 2	1.051 2	1.060 1
RR3min	2.615 1	2.084 9	3.076 8	2.855 9	2.827 8
RV4min	2.011 2	1.927 4	1.047 0	2.360 4	2.289 5
RR4min	2.719 7	2.645 6	3.177 7	2.971 4	1.997 4
RV5min	1.140 0	1.881 1	2.478 4	2.008 2	1.191 8
RR5min	2.625 5	2.127 7	2.083 7	2.318 9	2.796 2
RV10min	2.014 6	1.770 3	2.462 1	2.202 4	2.143 4
RR10min	2.050 1	2.658 8	2.762 2	2.896 9	3.537 5
RV20min	1.063 3	2.048 7	2.418 2	2.378 4	2.379 9
RR20min	2.072 9	2.081 6	2.703 9	2.399 0	2.384 6

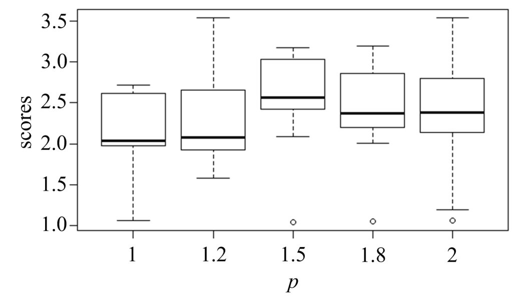

Fig. 11 Scores classified by power index p.

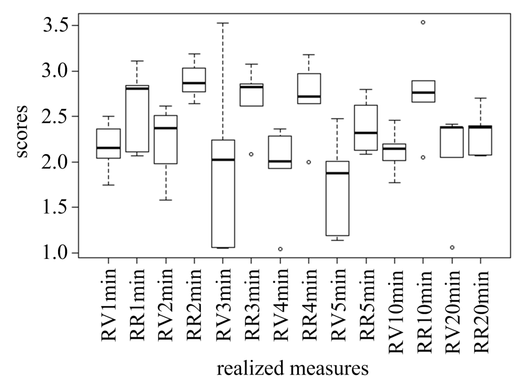

Fig. 12 Scores classified by realized measures.

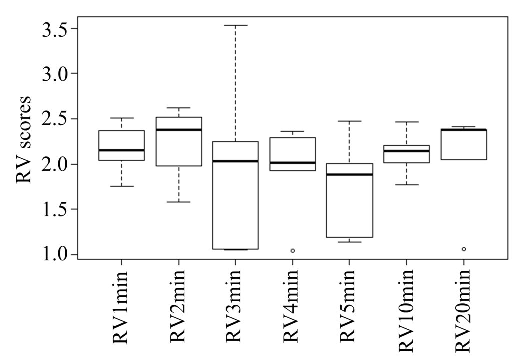

Fig. 13 Scores of RV with different time-frequency.

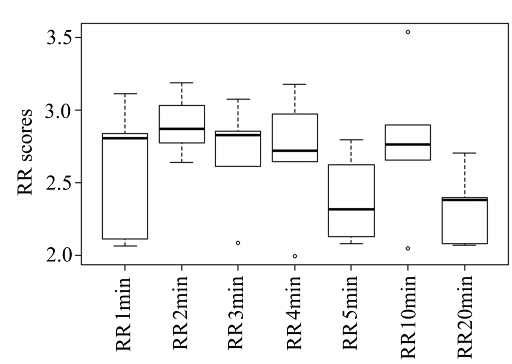

Fig. 14 Scores of RR with different time-frequency.

Then we compare different realized measures. Fig. 12 shows that RV measures are better than RR measures on average. Though the best result of RV20min is a very good result, the boxplots shows it is an outlier. By the boxplot, RV3min and RV5min are preferred. It also indicates that high time-frequency is not recommended when calculating RV and RR. Because too frequent data includes too big micro noise. In a word, from Fig. 12 we can conclude that RV measures are preferred to RR measures overall.

To acquire more detailed information, we plot boxplot for RV measures and RR measures separately in Fig. 13 and Fig. 14 respectively.

In Fig. 13, we find that RV3min and RV5min perform best. The performance of RV with time frequency 1 min is not as good as 3 min and 5 min. This means that not the most frequent realized measures produce the best results. The high frequency realized measures may contain much more micro noise that needs to be ignored in modeling.

The boxplot of RR measures is shown in Fig. 14. The results seem to be different from those of RV measures. The best 3 results are RR with time frequency 1, 5 and 20 min, where RR20min is the best one. The best time-frequencies of RV measures are bigger than those of RR measures, which means RV measures can accept more micro noise data. The most suitable time-frequencies are not the same as those of RV measures, which means different realized measures have different time-frequencies to suit for the best results.

In conclusion, this is not a simple case where the higher or the lower the power index p, the better the measure. Different datasets need different power indices. RV measures are better than RR measures in our empirical study. And we find that different time-frequency realized measures are suitable for different data. Moderate frequency is better.

6.4 More indices

In subsection 6.3, we draw a conclusion that the power index should be moderate. In this subsection, we exam that different power indices are suitable for different market indices. We use AEX, FTSE, GDAXI, and N225 as our new datasets ranging from 2000-01-01 to 2019-10-08. The first 3000 observartions (in sample) are used to estimate parameters, and the rest 2014 data are out-of-sample for examination. We adopt RV5min as the realized measure. The scores are also calculated by Eq. (21).

Tab. 7 shows that $p=1.2$ is the best power index for AEX, GDAXI, N225, and SPX, while $p=1.5$ is best for FTSE, with $p=1.2$ the second best. We can conclude that different indices need different power indices. We should try different power indices for better estimation. And the preferred power index may be located at around 1.2 to 1.5.

Tab. 7 Scores for different power indices for different indices.

	$p=1$	$p=1.2$	$p=1.5$	$p=1.8$	$p=2$
AEX	202.50	**24.06**	24.53	24.91	24.29
FTSE	15.10	14.02	**4.78**	14.10	13.33
GDAXI	13.10	**12.98**	13.13	13.30	13.30
N225	21.41	**20.44**	21.43	22.15	21.73
SPX	13.41	**12.71**	13.15	13.70	114.81

7 Conclusions

In this paper, we propose an L_p quantile regression model with realized measures, in which a measurement equation incorporates intra-day and high-frequency volatility. This is a generic model including realized quantile and expectile models. We can use it to model different data with different power indices.

We also develop an asymmetric exponential power distribution. We find that when adopting our asymmetric exponential power distribution, maximizing likelihood function is equivalent to minimizing the loss function of L_p quantile regression. This is a generic model including realized quantile model and realized expectile model. We can useit to model different data with different power index p.

The simulation results show that all the power indices p in 1 to 2 perform well. The empirical results show that both q_t and x_t are self-correlated and the leverage function is of significance in measurement equation. In addition, in our empirical study, RV is preferred to RR overall. We also find that different frequency data suits different realized measures, and that higher frequency is not always better.

However in this paper, power indices are some fixed parameters. In the next stage, power indices should be variables to be estimated for different data.

Acknowledgements

The work is supported by the National Key Research and Development Plan (2016YFC0800100), the NNSF of China (71771203).

Conflict of interest

The authors declare no conflict of interest.

Author information

TANG Li received his master degree from University of Science and Technology of China in 2020. His research interests focus on risk management.

CHEN Yu (corresponding author) received her Ph. D. degree in probability and statistics from University of Science and Technology of China in 2006. She is an associate professor of Department of Statistics and Finance, School of Management, University of Science and Technology of China. Her research

interests include network risk analysis, extreme value theory, and high-frequency data analysis.

References

[1] Koenker R, Bassett G. Regression quantiles. Econometrica, 1978, 46(1): 33-50.
[2] Newey W K, Powell J L. Asymmetric least squares estimation and testing. Econometrica, 1987, 55 (4): 819-847.
[3] Efron B. Regression percentiles using asymmetric squared error loss. Statist. Sinica, 1991, 1(1): 93-125.
[4] Chen Z. Conditional L_p-quantiles and their application to the testing of symmetry in non-parametric regression. Statist. Probab. Lett., 1996, 29(2): 107-115.
[5] Breckling J, Chambers R. M-quantiles. Biometrika, 1988, 75(4): 761-771.
[6] Parkinson M. The extreme value method for estimating the variance of the rate of return. Journal of Business, 1980, 53 (1): 61-65.
[7] Garman M B, Klass M J. On the estimation of security price volatilities from historical data. The Journal of Business, 1980, 53: 67-78.
[8] Engle R F. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflations. Econometrica, 1982, 50: 987-1007.
[9] Bollerslev T. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 1986, 31: 307-327.
[10] Andersen T G, Bollerslev T. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International Economic Review, 1998, 39: 885-905.
[11] Martens M, Van Dijk D. Measuring volatility with the realized range. Journal of Econometrics, 2007, 138 (1): 181-207.
[12] Hansen P R, Huang Z, Shek H H. Realized GARCH: A joint model for returns and realized measures of volatility. Journal of Applied Econometrics, 2011, 27 (6): 877-906.
[13] Bee M, Dupuis D J, Trapin L. Realized extreme quantile: A joint model for conditional quantiles and measures of volatility with EVT refinements. Journal of Applied Econometrics, 2018, 33(3): 398-415.
[14] Gerlach R, Walpole D, Wang C. Semi-parametric Bayesian tail risk forecasting incorporating realized measures of volatility. Quantitative Finance, 2017, 17(2): 199-215.
[15] Koenker R, Machado J A. Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association, 1999, 94 (448): 1296-1310.
[16] Lee A, Caron F, Doucet A, et al. A hierarchical Bayesian framework for constructing sparsity-inducing priors. https://arxiv. org / abs/ 1009. 1914.

基于已实现波动率的 L_p 分位数回归

汤李,陈昱*

中国科学技术大学管理学院统计与金融系,安徽合肥 230026

摘要: 提出了一种基于已实现波动率的 L_p 分位数回归模型,这是一种新的金融风险模型. 基于已实现波动率的 L_p 分位数回归模型将已实现波动率与 L_p 分位数回归结合起来,并且将 L_p 分位数加入模型的度量等式中. 该模型是囊括基于已实现波动率的分位数回归模型和基于已实现波动率的 Expectile 回归模型的更为一般的模型. 通过非对称幂指数分布(AExpPow)导出模型的对数似然函数,并且通过模拟证实了所提出的对数似然函数的正确性. 最后通过实证研究证实基于已实现波动率的 L_p 分位数回归模型的有效性,得出如下结论:不同的幂指数 p 适用于不同的数据,不同的时间频率适用于不同的已实现波动率,而不是时间频率越高越好.

关键词: 已实现波动率;基于已实现波动率的 L_p 分位数回归;非对称幂指数分布