Optimal probabilistic forecasts and their application in finance

A thesis submitted for the degree of
Bachelor of Commerce (Honours)
by

Yuru Sun

Student ID: 27514307

Supervisors:
Professor Gael Martin
Dr. Ruben Loaiza-Maya

Department of Econometrics and Business Statistics

Monash University
Australia

2020

Abstract

This paper conducts numerical and empirical analyses to explore the impact of the form and degree of model misspecification on optimal probabilistic forecasting method which is recently proposed to conduct a forecast using scoring rules instead of the conventional likelihood function in the frequentist econometrics. In particular, the role of optimal forecasts in predicting S&P 500 returns, the Value-at-Risk and the VIX Volatility Index. We find that the performance of optimal probabilistic forecasts is insensitive to the form of misspecification but improves as the degree of misspecification increases. After the Global financial crisis, predictions that are optimal according to the focused scoring rule that rewards predictive accuracy in a tail are substantially more accurate at predicting observations in said tail than the alternatives, including likelihood-based predictions.

Contents

1. Introduction

This paper explores the question: 'When do optimal probabilistic forecasts work', with particular attention given to the usefulness of optimal probabilistic forecasts in financial applications.

Probabilistic forecasts can provide complete information about future uncertainty, which can be more valuable than point and interval forecasts to forecasters. However, previous approaches to prediction - including probabilistic methods - typically assume that the predictive model correctly specifies the process that has generated the observations. In practice, the assumed model underpinning the likelihood function will almost certainly differ from the unknown true data generating process (DGP). This model misspecification problem has been an ongoing issue for conventional likelihood-based prediction.

As an alternative to likelihood-based prediction, scoring rules have been proposed as a means of producing probabilistic forecasts. A variety of alternative proper scoring rules allows users to produce probabilistic predictions that are designed to perform well according to the forecasting metric that is important to the problem at hand (Gneiting & Raftery, 2007). Some recent research shows that forecasts produced using a given scoring rule can yield better out-of-sample accuracy - measured by that scoring rule - than conventional likelihood-based forecasts, in particular in the presence of model misspecification; see Opschoor, van Dijk & van der Wel (2017), Loaiza-Maya, Martin & Frazier (2019) and Loaiza-Maya et al. (2020). Among all these investigations, it is the censored likelihood score (CLS) or the focused score (FSR) proposed originally by Diks, Panchenko & van Dijk (2011), that captures our attention, since its 'focusing' feature has great potential in financial risk management. In other words, CLS/FSR allows a forecaster to focus on predicting any particular region of interest more accurately and, therefore, is expected to make an important contribution to risk prediction in financial settings.

The 'optimal' probabilistic forecast has been discussed in Loaiza-Maya et al. (2020). It refers to a predictive probability distribution that is optimal according to a user-specified scoring rule. Whilst some empirical investigations were undertaken, these were primarily illustrative, with there being much scope left for exploring the performance of optimal methods in forecasting different types of financial measures, and under different assumed models.

This research paper extends the work of Loaiza-Maya et al. (2020) by investigating the effect of the form and degree of model misspecification on optimal forecast performance in particular financial settings; in particular by conducting an empirical analysis of how optimal forecasts perform in Value-at-Risk (VaR), Volatility Index (VIX) and portfolio optimization applications. Importantly, the empirical dataset used extends over a period of time that precedes the 2008 global financial crisis (GFC) and that includes the latest period in which COVID-19 has had an impact on financial markets.

The paper proceeds as follows. Section 2 explains the basic idea about how scoring rules are applied in producing and evaluating density forecasts, and provides the definitions of some commonly used scoring rules. Section 3 focuses on investigating the effects of different forms and degrees of model misspecification on the performance of optimal forecasts, using data simulated from the (generalized) autoregressive conditional heteroscedasticity ((G)ARCH)

and inversion copula models. The assumed predictive model is fixed as ARCH(1) for the purpose of manipulating the degree of model misspecification. Although the simulation design is introduced in this section, it forms the foundation of methodologies used in all experiments conducted in other sections of this paper. In Section 4, we emulate the simulation exercise on an empirical example for financial returns of the S&P 500 index, including the prediction of the Value-at-Risk (VaR). The results illustrate the practical contributions to financial risk management. The empirical analysis of the VIX is provided in Section 5, where the predictive model is the heterogeneous autoregressive-realized volatility (HAR-RV) model with different error term specifications. We conclude and discuss any possible caveats in Section 6.

2. Scoring rules in prediction

2.1 Overview and notation

Optimal probabilistic forecasts discussed in this paper refer to the forecasts produced from a model that is 'optimized' based on a user-specified proper scoring rule will perform the best out-of-sample – according to that same score (Loaiza-Maya et al.,2020).

Scoring rules are a type of criterion function that can be optimized to produce an optimal estimator and thus, a probabilistic prediction derived from this optimal estimator and an assumed model. They also play the role of encouraging forecasters to make careful and honest assessments in elicitation of subjectivist Bayesians, and evaluating and ranking competitive probabilistic forecasts, based on the predictive distribution and materialized events or values, by assigning a numerical score [Garthwaite et al., 2005, Gneiting and Raftery, 2007]. Therefore, the crucial importance of the propriety of scoring rules must be emphasized for their usage because improper scoring rules will assign a higher average score to an incorrect density forecast (Gneiting & Raftery, 2007). Consequently, forecasts based on improper scoring rules will not be 'optimal'.

Suppose P and Q are predictive distributions and Q is the best forecast given all the available information. S(P,Q) denotes the expected value of S(P,$\cdot$) under Q. A scoring rule is said to be proper if $S(Q,Q) \geq S(P,Q)$ for all P and Q, and is strictly proper if $S(Q,Q) = S(P,Q)$ only happens when P = Q (Gneiting & Raftery, 2007).

In terms of producing probabilistic forecasts, scoring rules can replace the conventional likelihood function in frequentist and Bayesian predictions; see Loaiza-Maya, Martin & Frazier (2019) and Loaiza-Maya et al. (2020). Supposing the scoring rule is positively oriented, an estimator $\hat{\theta}_n$ obtained by maximizing a scoring rule $S_n(\theta)$ is said to be 'optimal' based on this scoring rule.

$$\hat{\theta}_n = arg\max_{\theta} S_n(\theta) \tag{1}$$

Under certain conditions, including that the scoring rule is 'proper', $\hat{\theta}_n \to \theta_0 \; as \; T \to \infty$, where θ_0 is the true parameter and T is the total sample size of a time series variable y_t. The

predictions conditional on this optimal estimator, $p(y_{T+1}|M, \hat{\theta}_n, y_{1:T})$, is said to be produced from a model that is 'optimized' according to a proper scoring rule.

In addition, in terms of evaluating predictions, it is usually the average score ($\bar{S}_n$) that is used to directly compare predictive accuracy among predictions produced by optimizing different scoring rules (Gneiting & Reftery, 2007).

$$\bar{S}_n = \frac{1}{T}\sum_{t=1}^{T} S_n(p_{t+1}, y_t) \tag{2}$$

For a positively oriented score, for example, a higher value will be assigned to a better forecast between two competing candidates, on the condition the scoring rule being proper.

2.2 Some commonly used scoring rules

A variety of scoring rules have been developed to tackle different problems. Some commonly used proper scoring rules are adopted in the analyses in this paper, such as the logarithmic score (LS), the continuously ranked probability score (CRPS) and the censored likelihood score (CLS).

The logarithmic score is defined as (3) where p_t is the predictive density.

$$S_{LS}(p_{t-1}, y_t) = log[p_{t-1}(y_t)] \tag{3}$$

It is a local strictly proper scoring rule, which means it will assign a higher score to the correct probabilistic forecast, and it is superior to quadratic and spherical scoring rules when the rank ordering is important or the impact of the nonlinear utility function used is a concern for forecasters (Bickel, 2007). Maximum likelihood estimation (MLE), as a flexible and asymptotically optimal method in econometrics, is based on the logarithmic score so that it is included as a benchmark in the following numerical and empirical investigations of optimal forecasts' performance.

$$\hat{\theta}_{MLE} = arg\max_{\theta}\{log[p(\boldsymbol{y}|\theta)]\} \tag{4}$$

However, the logarithmic score is criticized for its unboundedness and its local property. Bernardo (1979) states "locality requires the utility of probabilistic influence to depend only upon the probability density of the true state". Gneiting & Raftery (2007) also argue that the logarithmic score is insensitive to distance and will not reward predictions that are close to but not identical to the materialized event. Therefore, Gneiting & Raftery (2007) propose the continuously ranked probability score, which is sensitive to distance and is defined as in (5)

$$CRPS(P_{t-1}, x_t) = -\int_{-\infty}^{\infty}[P(y) - I(y \geq x_t)]^2 dy \tag{5}$$

where P is the cumulative distribution function, I is the indicator function, and x is the materialized event.

The formula can be simplified to (6) if the prediction distribution is Gaussian with mean = μ and variance = σ^2.

$$CRPS(N_{t-1}, x_t) = \sigma[\frac{1}{\sqrt{\pi}} - 2\phi(\frac{x_t - \mu}{\sigma}) - \frac{x_t - \mu}{\sigma}(2\Phi(\frac{x_t - \mu}{\sigma}) - 1)] \tag{6}$$

where ϕ and Φ are the probability density function and cumulative distribution function of the Gaussian predictive distribution.

CRPS is defined as a negatively oriented scoring rule, but it can be easily transformed to a positively oriented form as shown in (7) for convenient comparison among scoring rules in the following sections.

$$CRPS^*(N_{t-1}, x_t) = \sigma[-\frac{1}{\sqrt{\pi}} + 2\phi(\frac{x_t - \mu}{\sigma}) + \frac{x_t - \mu}{\sigma}(2\Phi(\frac{x_t - \mu}{\sigma}) - 1)] \tag{7}$$

Both LS and CRPS are used for producing and evaluating the entire predictive densities. In terms of accurately predicting a certain region of a distribution, they are typically used together with weighted likelihood (Gneiting & Ranjan, 2011). Diks, Panchenko & van Dijk, (2011) propose the censored likelihood score, which allows users to assess forecasts only on a region (regions) of interest instead of using weights to emphasize a particular part of the entire density forecast. Moreover, it can be easily used to combine density forecasts and yield better predictive accuracy. It is defined as in (8)

$$S_{CLS}(p_{t-1}, y_t) = I(y_t \in A_t) log[p_{t-1}(y_t)] + I(y_t \in A_t^c) log[\int_{A_t^c} p_{t-1}(s)ds] \tag{8}$$

where p_t is the predictive probability density function, A_t is the region of interest and A_t^c is the complement of A_t. Opschoor, van Dijk & van der Wel (2017) prove that weighted density forecasts based on optimizing CLS outperform those on LS and CRPS.

3. Numerical investigation of optimal predictions

3.1 Simulation design

Optimal forecasts can outperform predictions produced by conventional methods out of sample; see Opschoor, van Dijk & van der Wel (2017), Loaiza-Maya, Martin & Frazier (2019) and Loaiza-Maya et al.(2020); but the underlying reasons driving this phenomenon are still under investigation. Loaiza-Maya, Martin & Frazier (2019) propose a new method, focused Bayesian prediction (FBP), which replaces the conventional likelihood function with the censored likelihood score. In their simulated and empirical analysis, Focused Bayes outperforms exact Bayes which uses the logarithmic score for updating the prior probability and they also point out that model misspecification plays a role in the performance of Focused Bayes since focusing incorrectly can harm.

Loaiza-Maya et al.(2020) extend the above discussion to frequentist probability forecasting. They address questions about when we can/cannot benefit from optimal probabilistic forecasts out of sample. Loaiza-Maya et al.(2020) introduce the concepts of 'coherence' and

'strict coherence' for conveniently documenting the optimal forecasts' performance. Coherence means that the optimal probabilistic forecast based on a given score is superior, or at least performs the same as, alternative forecasts according to the same score. Strict coherence happens when the optimal prediction is strictly preferable given that score. Let $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ be the optimizers based on scoring rules: S_1 and S_2. The predictive density $S_n(P_{\hat{\boldsymbol{\theta}}_n}^{t-1}, y_t)$ is said to be coherent if

$$\frac{1}{\tau}\sum_{t=T-\tau+1}^{T} S_1(P_{\hat{\boldsymbol{\theta}}_1}^{t-1}, y_t) \geq \frac{1}{\tau}\sum_{t=T-\tau+1}^{T} S_1(P_{\hat{\boldsymbol{\theta}}_2}^{t-1}, y_t) \tag{9}$$

$$\frac{1}{\tau}\sum_{t=T-\tau+1}^{T} S_2(P_{\hat{\boldsymbol{\theta}}_1}^{t-1}, y_t) \leq \frac{1}{\tau}\sum_{t=T-\tau+1}^{T} S_2(P_{\hat{\boldsymbol{\theta}}_1}^{t-1}, y_t) \tag{10}$$

It is said to be strictly coherent if (9) and (10) are strict inequalities.

Building on the previous research in this field, we conduct a numerical analysis in this section to further investigate the effects of the form and degree of model misspecification on the performance of optimal forecasts. In order to set the scene for the financial application analysis in following sections of this paper, we simulate a time series variable y_t that mimics the behavior of financial returns and volatility. Specifically, GARCH models are used to capture the volatility clustering and serial dependence usually observed in empirical stock returns, and the negative marginal skewness will be incorporated by using an inversion copula model, with the degree of marginal skewness controlled by the shape parameter. Copulas are functions used to describe the dependence among random variables, which "reweight" the marginal densities to produce a joint density that captures the dependence among random variables (Ref: Yanqin). It can be applied to conditional distributions for forecasting purposes as shown in (11) where X_t and Y_t are random variables with conditional marginal densities denoted by F_t and G_t, $\mathcal{F}_t$ is an information set available at time t, and H_t denotes the conditional joint distribution of X_t and Y_t (Ref: Yanqin).

$$H_t(x, y|\mathcal{F}_{t-1}) = C_t(F_t(x|\mathcal{F}_{t-1}), G_t(y|\mathcal{F}_{t-1})|\mathcal{F}_{t-1}) \tag{11}$$

The copula-based model used in this section is the stochastic volatility inversion copula. It allows us to keep the dependence structure of a state space model and provides the flexibility of using an arbitrary marginal distribution that is not allowed if we use state space models alone (Ref: M).

In order to manipulate the degree and form of model misspecification, we fix the underlying model to a simple and clearly misspecified ARCH(1) model and then adjust degrees of freedom and shape parameters of the true DGP in Scenario (i) and (ii).

Specific simulation scenarios are listed in the following table (t_ν indicates a Student-t distribution with ν degrees of freedom).

MLE is a frequently used estimation method in frequentist prediction. It often works with the log-likelihood, which corresponds to the logarithmic scoring rule. Therefore, we use the predictions produced by MLE (LS) as the benchmark of the following comparisons. Besdies, the tails of a density are important for risk management in finance so that we decide to

Simulation design		
	Scenario (i)	**Scenario (ii)**
True DGP	$y_t = \sigma_t \epsilon_t$ $\sigma_t^2 = 1 + 0.2y_{t-1}^2 + 0.7\sigma_{t-1}^2$ $\epsilon_t \sim (\frac{\nu-2}{\nu})^{0.5} * t_\nu$ $\nu \in (3, 12, 10000)$	Inversion copula Shape parameter = 0, -3, -5
Assumed model	$y_t = \mu + \sigma_t \epsilon_t$ $\sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2$ $\epsilon_t \sim N(0,1)$	$y_t = \mu + \sigma_t \epsilon_t$ $\sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2$ $\epsilon_t \sim N(0,1)$

apply the focused scoring rule to the 10%, 20%, 80% and 90% tails of the distribution, corresponding to the risks of long and short portfolios. Certain aspects of the methodology used in Loaiza-Maya et al., (2020) and Loaiza-Maya, Martin & Frazier, (2019) are adopted here. We conduct each simulation scenario as follows:

Let $P^{t-1}_{\hat{\boldsymbol{\theta}}^{[i]}_1}$ be the one-step-ahead prediction based on the assumed model, and $p(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta})$ ($P(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta})$) be the predictive probabilistic (cumulative probabilistic) distribution at time t.

1. Generate T = 6000 observations for y_t from the true DGP
2. Use $y_{1:1000}$ to estimate $\hat{\boldsymbol{\theta}}$ in the assumed predictive model based on the positively oriented score S_i where $\hat{\boldsymbol{\theta}} = \{\hat{\mu}, \hat{\alpha}_0, \hat{\alpha}_1\}$

$$\hat{\boldsymbol{\theta}}^{[i]} := arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \bar{S}_i(\boldsymbol{\theta}) \tag{12}$$

$$\bar{S}_i(\boldsymbol{\theta}) := \frac{1}{T - (\tau + 1)} \sum_{t=2}^{T-\tau} S(P^{t-1}_{\boldsymbol{\theta}}, y_t) \tag{13}$$

3. Produce the one-step-ahead predictive $P^{t-1}_{\hat{\boldsymbol{\theta}}^{[i]}_1}$, and compute the out-of-sample score using S_j, where S_i and S_j refer to (14) (15) (16):[1]

$$S_{LS}(P^{t-1}_{\boldsymbol{\theta}}, y_t) = lnp(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) \tag{14}$$

$$S^*_{CRPS}(P^{t-1}_{\boldsymbol{\theta}}, y_t) = \int_{-\infty}^{+\infty} [P(y|\mathcal{F}_{t-1}, \boldsymbol{\theta}) - I(y \geq y_t)]^2 dy \tag{15}$$

$$S_{FSR}(P^{t-1}_{\boldsymbol{\theta}}, y_t) = lnp(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) I(y_t \in A) + [ln \int_{A^c} p(y|\mathcal{F}_{t-1}, \boldsymbol{\theta}) dy] I(y_t \in A^c) \tag{16}$$

4. Expand estimation window by one observation and repeat step 2-3 with $\tau = T - 1000$ times and compute the average scores:

$$\bar{S}_j(\hat{\boldsymbol{\theta}}^{[i]}) = \frac{1}{\tau} \sum_{t=T-\tau+1}^{T} S_j(P^{t-1}_{\hat{\boldsymbol{\theta}}^{[i]}}, y_t) \tag{17}$$

[1]The CRPS in (15) is transformed to a positively-oriented score for the purpose of convenient comparison among all scores.

3.2 Simulation results

3.2.1 Average out-of-sample scores

The first column of each of the following tables presents the labels for the S_i that we used to produce predictions and the third-row shows S_j, which is used for the forecast's evaluation. The bolded numbers are the largest values of $\bar{S}_j(\hat{\boldsymbol{\theta}}^{[i]})$ in each column. We use all positively oriented scoring rules, and therefore, the column maximum(s) indicate(s) the optimal prediction(s) based on the evaluating scoring rule indicated by the column name.

Table 1: the true DGP is ARCH(1)

	Average out-of-sample scores					
In-sample optimizers	LS	CRPS	FSR10	FSR20	FSR80	FSR90
LS	**-1.510**	**-0.624**	**-0.376**	**-0.602**	**-0.593**	**-0.363**
CRPS	**-1.510**	**-0.624**	**-0.376**	**-0.602**	**-0.593**	**-0.363**
FSR10	-1.512	-0.625	-0.377	**-0.602**	-0.595	-0.365
FSR20	-1.514	-0.625	-0.377	**-0.602**	-0.597	-0.366
FSR80	-1.518	-0.626	-0.381	-0.609	-0.594	**-0.363**
FSR90	-1.516	-0.626	-0.380	-0.608	-0.594	**-0.363**

Table 2: the true DGP is GARCH(1,1) with degree of freedom = 12

	Average out-of-sample scores					
In-sample optimizers	LS	CRPS	FSR10	FSR20	FSR80	FSR90
LS	**-2.489**	**-1.641**	-0.508	-0.825	-0.813	-0.494
CRPS	-2.492	**-1.641**	-0.510	-0.827	-0.814	-0.496
FSR10	-2.538	-1.715	**-0.506**	-0.824	-0.860	-0.533
FSR20	-2.504	-1.662	**-0.506**	**-0.823**	-0.828	-0.505
FSR80	-2.496	-1.653	-0.512	-0.832	**-0.812**	**-0.493**
FSR90	-2.519	-1.684	-0.529	-0.852	-0.813	**-0.493**

Table 3: the true DGP is GARCH(1,1) with degree of freedom = 3

	Average out-of-sample scores					
In-sample optimizers	LS	CRPS	FSR10	FSR20	FSR80	FSR90
LS	**-2.335**	-1.248	-0.568	-0.873	-0.892	-0.574
CRPS	-2.452	**-1.233**	-0.625	-0.929	-0.967	-0.654
FSR10	-2.752	-2.120	**-0.520**	-0.843	-1.311	-0.960
FSR20	-2.472	-1.519	-0.528	**-0.834**	-1.045	-0.704
FSR80	-2.489	-1.532	-0.725	-1.049	**-0.841**	-0.526
FSR90	-2.736	-2.093	-0.957	-1.287	-0.842	**-0.513**

Table 4: the true DGP is copula with shape parameter = 0

	Average out-of-sample scores					
In-sample optimizers	LS	CRPS	FSR10	FSR20	FSR80	FSR90
LS	**-1.402**	-0.562	**-0.360**	**-0.585**	**-0.582**	**-0.346**
CRPS	**-1.402**	**-0.561**	-0.361	**-0.585**	**-0.582**	**-0.346**
FSR10	-1.417	-0.566	**-0.360**	-0.586	-0.594	-0.356
FSR20	-1.407	-0.563	**-0.360**	**-0.585**	-0.586	-0.349
FSR80	-1.411	-0.564	-0.367	-0.593	**-0.582**	**-0.346**
FSR90	-1.454	-0.575	-0.397	-0.630	-0.584	**-0.346**

Table 5: the true DGP is copula with shape parameter = -3

	Average out-of-sample scores					
In-sample optimizers	LS	CRPS	FSR10	FSR20	FSR80	FSR90
LS	**-1.405**	**-0.559**	-0.403	-0.653	-0.523	-0.316
CRPS	-1.408	**-0.559**	-0.411	-0.659	-0.523	-0.318
FSR10	-1.431	-0.565	**-0.394**	**-0.642**	-0.562	-0.357
FSR20	-1.454	-0.573	**-0.394**	**-0.642**	-0.586	-0.382
FSR80	-1.760	-0.593	-0.723	-1.016	**-0.507**	**-0.303**
FSR90	-2.140	-0.628	-0.986	-1.360	-0.512	**-0.303**

Table 6: the true DGP is copula with shape parameter = -5

	Average out-of-sample scores					
In-sample optimizers	LS	CRPS	FSR10	FSR20	FSR80	FSR90
LS	**-1.407**	-0.559	-0.414	-0.669	-0.501	-0.307
CRPS	-1.413	**-0.558**	-0.427	-0.680	-0.500	-0.310
FSR10	-1.454	-0.570	**-0.401**	-0.654	-0.567	-0.377
FSR20	-1.482	-0.581	**-0.401**	**-0.653**	-0.598	-0.410
FSR80	-2.412	-0.628	-1.230	-1.660	**-0.463**	**-0.277**
FSR90	-3.603	-0.686	-1.972	-2.673	-0.475	**-0.277**

The results from Table 1-3 show one particular type of model misspecification. That is, the assumed predictive model, ARCH(1), cannot capture the long term volatility clustering feature of the data generated from GARCH(1,1) models, i.e. the true DGP. Hence, there is misspecification. When the degrees of freedom parameter in the error term of GARCH(1,1) is large enough, the ARCH(1) model is correctly specified as shown in Table 1. As the degrees of freedom decreases, the degree of model misspecification increases.

From Table 1, the correct specification case, we could observe coherence and strict coherence in Table 3 when the degree of model misspecification is high. These results are consistent

with conclusions from Loiaza-Maya et al. (2020). It simply means that we can improve prediction accuracy by using FSR in the presence of model misspecification under certain conditions. That is, we expect to gain more from optimal forecasts while the degree of model misspecification is high.

Table 4-6 show another type of model misspecification where the ARCH(1) model cannot capture the asymmetric feature of the data generated from copula models (the true DGP). From Table 4-6, as the shape parameter decreases, the degree of negative marginal skewness increases, and thus the degrees of model misspecification increases. The results give similar conclusions to Table 1-3. It proves that optimal forecasts perform in the same way even when we have different types of model misspecification.

3.2.2 Trace plots

The predictive accuracy of tails is important for financial risk management. Therefore, we provide score trace plots to exhibit the comparison of MLE, CRPS and FSR's performance in predicting the 10 percentile and 90 percentile in Figure 1 and 2 so that we can evaluate their performance in predicting tails. The results in Section 3.2.1 shows that when the degree of model misspecification is high, FSR can perform the best compared with MLE and CRPS in predicting tails. However, only the average score values at the end of each iteration of the simulation are recorded in the tables, while trace plots provide a dynamic view of how optimal forecasts perform.

In the left panel of each trace plot, it shows the performance of each score evaluated by FSR 10%. That is, if the average score value of score A used for estimation is higher than score B, then score A predicts the 10 percentile of the distribution more accurately than score B. The right panel shows the same thing but is based on performance of predicting 90 percentile. From Figure 1 and 2, we could see that FSR almost makes no difference if the degree of model misspecification is very small compared with other scoring rules, but as it increases, FSR starts to outperform CRPS and MLE in predicting tails. The more misspecified the model is, the more benefits can we gain from using FSR. It is beneficial for econometricians to improve forecast accuracy without working so hard on finding a 'correct' model when in practice, a correct model does not exist.

Figure 1: the true DGP is (G)ARCH

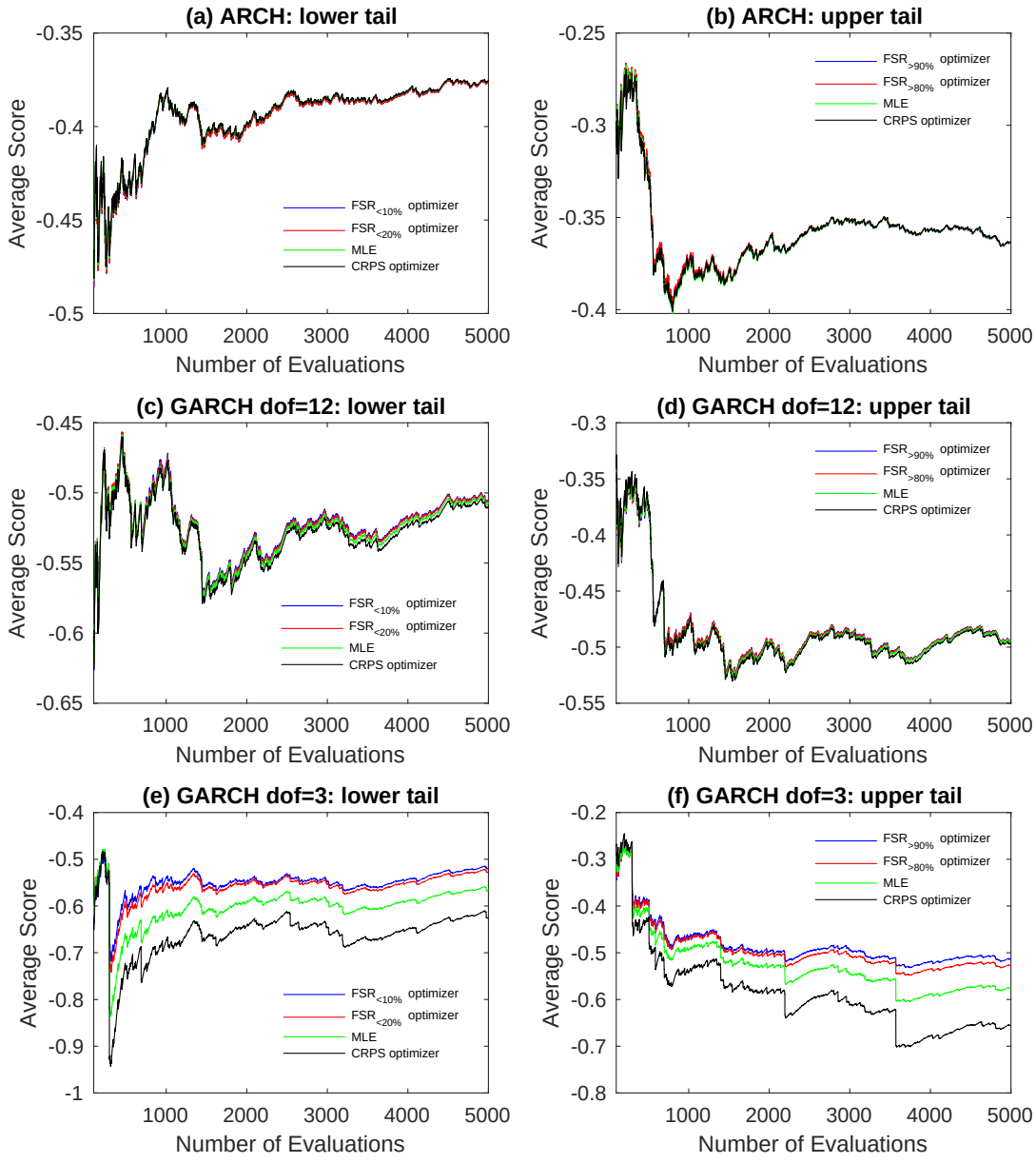


Figure 2: the true DGP is copula

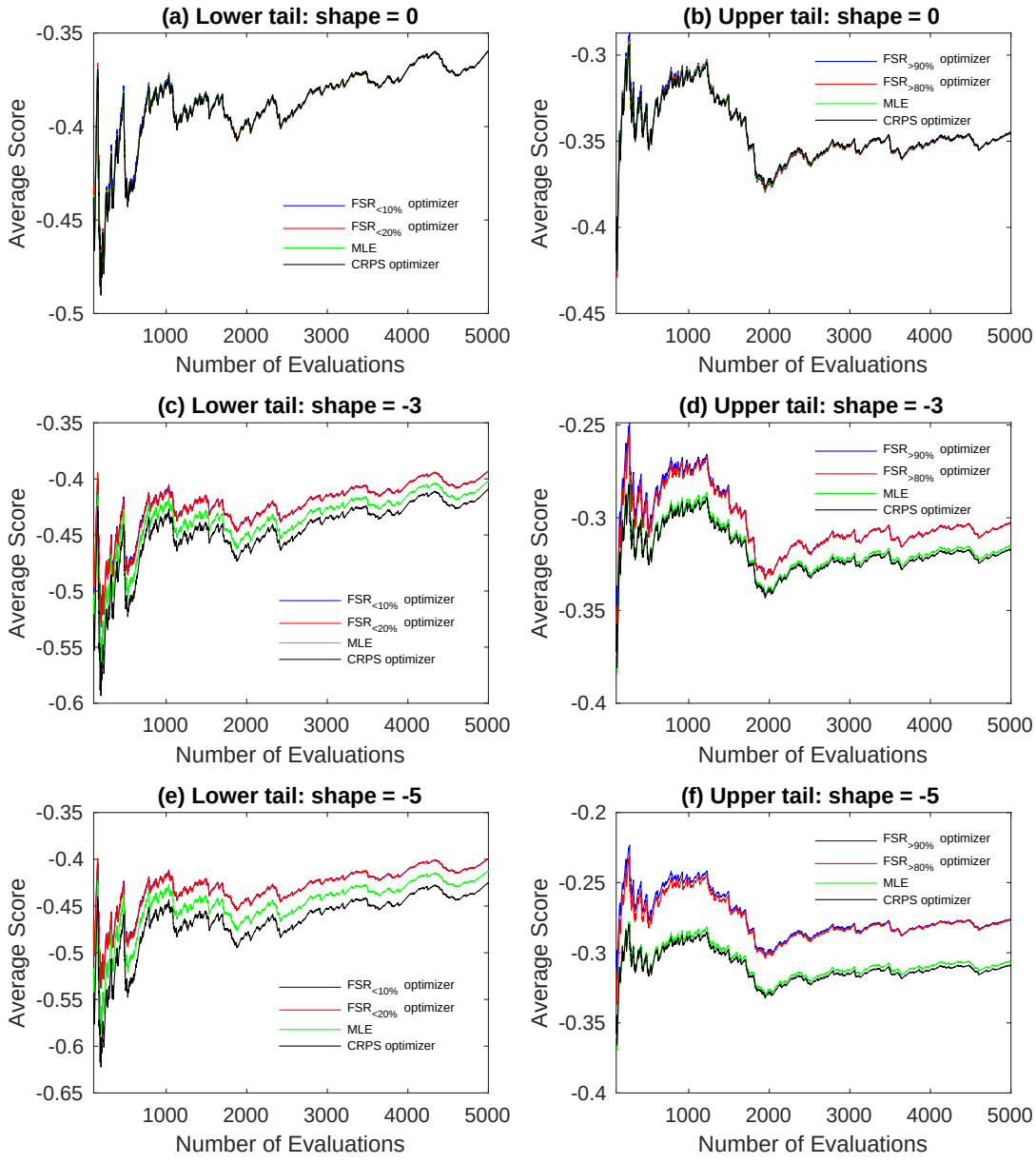
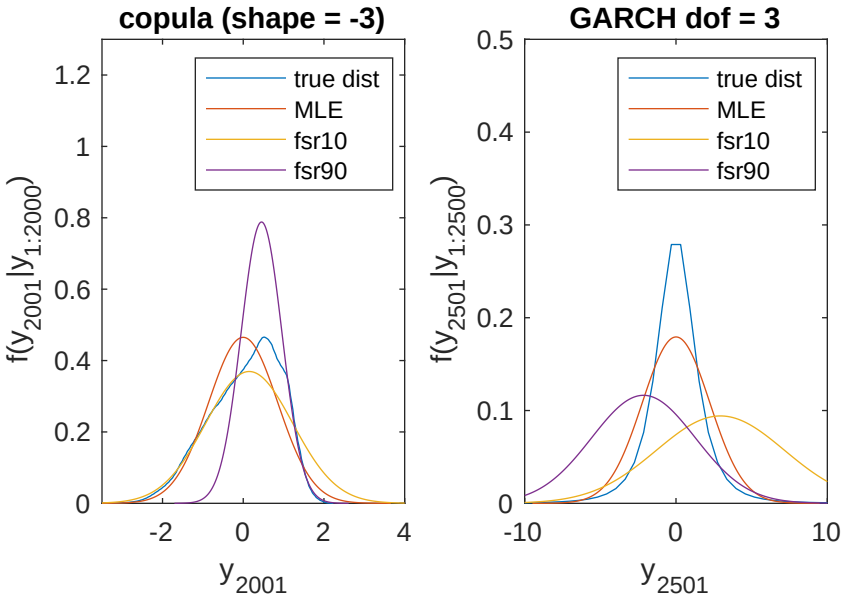


Figure 3: predictive density



Moreover, the predictive density plots show that predictions obtained from optimizing the focused score aim to match the shape of tails of the true conditional predictive distributions. It indicates that the shape of the true conditional predictive distributions plays a role in deciding how much we can gain by using optimal forecasts. One must note that in practice, the shape of the true DGP is unknown, but it still is a vital reason why the performance of optimal forecasts varies.[2]

4. Empirical analysis: financial returns

4.1 Overview and preliminary diagnostics

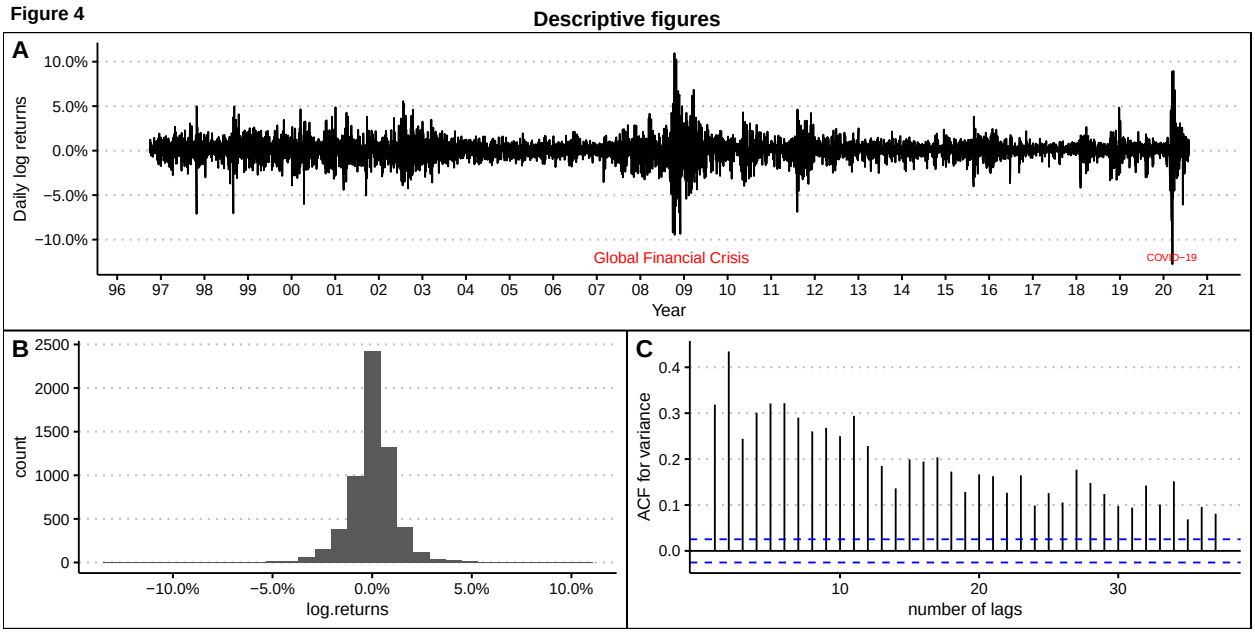
From the results in Section 3.2, we can conclude that optimal forecasts based on the focused score can improve the prediction accuracy in tails and its effect is clearest when the degree of model misspecification is high and does not vary much when we have different types of model misspecification. In Section 4, we apply the same method adopted in Section 3 to an empirical setting and investigate how it performs in practice.

The data used in the analysis is the continuously compounded daily log returns of S&P 500 listed in U.S.A financial market over a broad time period from 27 Sep,1996 to 30 July,2020. It aims to include two very volatile periods, the GFC and the recent COVID-19 pandemic. The initial training sample size and out-of-sample observations are kept consistent with what is in Section 3 with initial training sample size = 1000 and out-of-sample evaluations = 5000.

[2]The predictive density plots in Appendix 1 show the changes of optimal forecasts over time/iterations.

Table 7: Descriptive statistics

Stock	Min	Median	Mean	Max	Skewness	Kurtosis	JB.Test	LB.Test
S&P500	-0.1276522	0.0006406	0.0002591	0.0060242	-0.3954386	13.37497	28270	61.489



The descriptive statistics show that the distribution of S&P 500 log returns is negatively skewed and has fat tails and strong serial correlation in its volatility. Besides, during the GFC and COVID-19 pandemic periods, the volatility is very high due to high uncertainty.

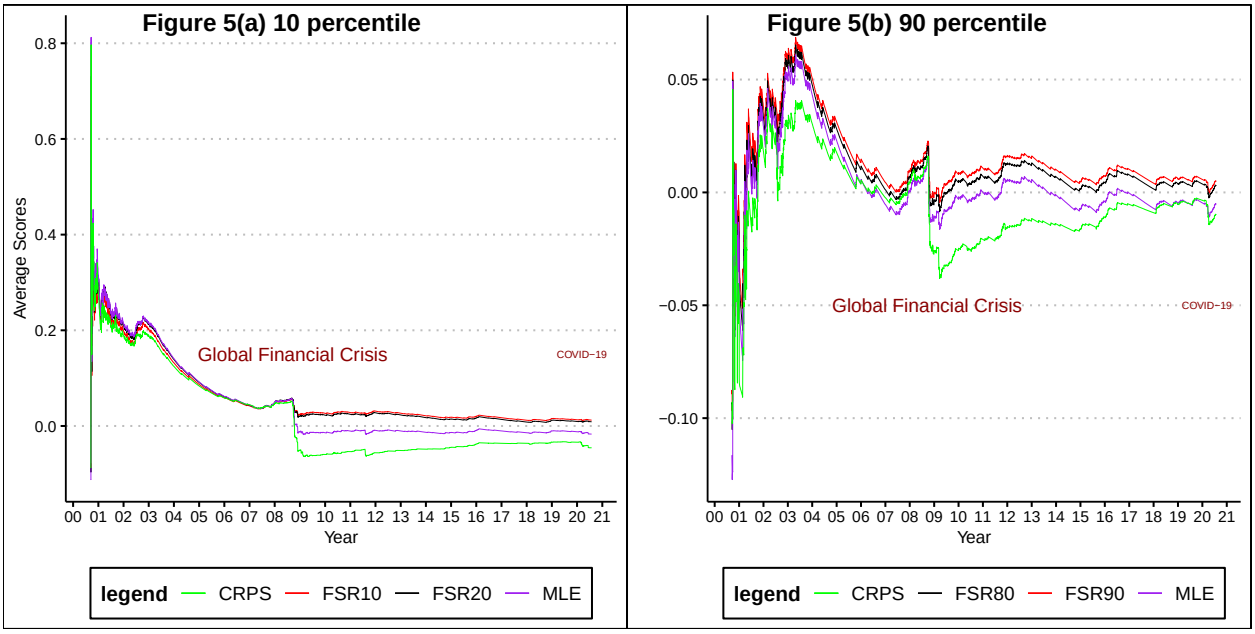
4.2 Empirical results: predictive distributions for returns

We still use ARCH(1) model in Section 3 as the assumed predictive model since it is a reasonable model to predict stock returns and volatility. More importantly, it is misspecified with a high degree of model misspecification. Therefore, based on the results shown in Section 3, we should expect that predictions produced by optimizing FSR do a better job in predicting two tails of the distribution. The results of Table 8 support our conjecture. However, the difference between using FSR10(80) and FSR20(90) is very small.

Table 8: Average scores for S&P 500

	Average out-of-sample scores					
In-sample optimizers	LS	CRPS	FSR10	FSR20	FSR80	FSR90
LS	**3.044**	**-0.006**	-0.017	0.123	0.169	-0.005
CRPS	3.030	**-0.006**	-0.046	0.098	0.166	-0.010
FSR10	2.701	-0.009	**0.012**	**0.163**	-0.168	-0.313
FSR20	2.852	-0.008	0.009	0.161	-0.025	-0.178
FSR80	2.964	-0.007	-0.062	0.052	**0.183**	0.003
FSR90	2.847	-0.008	-0.150	-0.051	0.182	**0.005**

Unlike in the simulation example, the score trace plots now show how different scoring rules perform at time points that correspond to some actual historical events. In Figure 5, one interesting phenomenon is that after the GFC, the focused score performs much better than CRPS and MLE compared with before, in terms of predicting both upper and lower tails. And this 'jump' effect continues until the very recent date. It indicates that optimal probabilistic forecasts might predict financial markets better in and after times of stock market turbulence, which could represent a great contribution to the existing financial analyses of risks and returns.



4.3 Empirical results: prediction of Value-at-Risk

4.3.1 Overview and preliminaries

Density forecasts have great practical value in finance. They form the foundation of risk management, such as Value-at-Risk and are useful for asset allocation and derivative pricing.

VaR is a commonly used risk measurement in finance. It can be obtained from a probability density function of returns over a chosen investment horizon. It is the p-quantile (q_p) of a predictive distribution of a portfolio's profit/loss over a holding period, according to a given confidence interval α (Dowd, 2005).

$$VaR = -q_p p = 1 - \alpha \tag{18}$$

It indicates that investors will not lose more than the amount as shown in VaR. It represents the largest possible investment loss in a given investment horizon with the given α confidence level. There are different methods to estimate VaR, but it is clear that predictive density plays an important role here. Therefore, a more accurate prediction can contribute to improving the accuracy of VaR prediction.

Opschoor, van Dijk & van der Wel (2017) combine density forecasts using focused scoring rules, and the results show weighted density forecasts based on optimizing the focused score outperform those based on optimizing CRPS or LS and improve the accuracy of 99% Value-at-Risk estimates.

We illustrate how optimal forecasts perform in estimating VaR by using predictions produced in simulation (Section 3) first, and expect that optimal forecasts based on FSR would give us a better prediction in tails. Specifically, VaR predictions at {10%, 20%, 80%, 90%} are conducted, since they correspond to the 10% and 20% expected loss of long and short portfolios. We assess VaR predictive accuracy by using the VaR backtesting method. The specific steps are as below:

1. Using the sequence of predictive densities (p) produced by S_i in Section 3:

- $S_i = \{LS, CRPS, FSR10, FSR20, FSR80, FSR90\}$;
- Conditional out-of-sample predictives: $p_{S_i}(y_{1001}|y_{1000})...p_{S_i}(y_{5000}|y_{4999})$;

2. Construct the VaR at {10%, 20%, 80%, 90%} using predictive densities from Step 1 for each scoring rule;

3. Compare the true values of $y_{1001}...y_{5000}$ with VaR_t and calculate the proportion of exceedances ($y_t < VaR_t$);

The closer the proportion of exceedances, calculated in Step 3, to nominal VaR levels, the more accurate are VaR predictions and thus, the more accurate are the optimal probabilistic forecasts. An accurate prediction of the p% VaR is observed if the proportion of exceedances equals p%.

Starting from the two different forms of model misspecification investigated in simulation analysis (Section 3), the values showed in the Table 9 and 10 refer to the out-of-sample proportion of exceedances over the nominal VaR level indicated by the column name. The focused score can provide the most accurate VaR prediction in the upper tail, while in the lower tail, it is not short of advantages by too much compared with CRPS. This conclusion

does not vary much when we have different forms of model misspecification as illustrated in the Table 9 and 10.

Table 9: the true DGP is copula (shape = -5)

	Out-of-sample exceedances			
Optimizers	VaR at 10%	VaR at 20%	VaR at 80%	VaR at 90%
LS	0.1176	**0.2018**	0.7784	0.9338
CRPS	0.1372	0.2210	0.8036	0.9450
FSR10	**0.0992**	0.1914	0.9178	0.9938
FSR20	0.1026	**0.2018**	0.9520	0.9984
FSR80	0.3502	0.4276	**0.7968**	0.8884
FSR90	0.4360	0.5130	0.8300	**0.9024**

Table 10: the true DGP is GARCH with degree of freedom = 3

	Out-of-sample exceedances			
Optimizers	VaR at 10%	VaR at 20%	VaR at 80%	VaR at 90%
LS	0.0728	0.1418	0.8710	0.9336
CRPS	**0.1062**	**0.1858**	0.8188	0.8960
FSR10	0.1070	0.3604	0.9858	0.9928
FSR20	0.0828	0.2162	0.9644	0.9802
FSR80	0.0182	0.0404	**0.7918**	0.9200
FSR90	0.0084	0.0158	0.6526	**0.8994**

4.3.2 Empirical results: S&P 500

The preliminary results obtained from the VaR predictions using simulated data in Section 4.3.1 help us understand what to expect in empirical VaR analysis. We now implement the VaR analysis to continuously compounded daily log returns of S&P 500 in U.S.A financial market. Learning from the trace plots in Section 4.2, the 2008 GFC is the turning point of the optimal forecasts' performance, especially for the focused scores. Therefore, we perform the VaR analysis over two different time periods, before the GFC and after the GFC. The results from Table 11 and 12 show that the focused score perform well enough in terms of estimating VaR for both long and short portfolios of S&P 500, but also CRPS is a very robust scoring rule which can provide relatively accurate predictions. Moreover, in Table 12, the focused score seems to perform better in 20% and 80% tails which have higher probability mass than 10% and 90% tails. If we think of the marginal distribution of S&P 500 estimated by histogram in Figure 4.B, the above results could be a consequence of the very sparse marginal distribution, which increases the difficulty of 'focusing' on the 10% and 90% tails.

Table 11: VaR for S&P 500 before the GFC

	Out-of-sample exceedances			
Optimizers	VaR at 10%	VaR at 20%	VaR at 80%	VaR at 90%
LS	0.0968	0.1752	0.8572	0.9260
CRPS	0.1176	**0.1944**	0.8340	**0.9108**
FSR10	0.1104	0.2524	0.9708	0.9868
FSR20	**0.1012**	0.2200	0.9500	0.9784
FSR80	0.0556	0.1112	0.8236	0.9196
FSR90	0.0348	0.0708	**0.7908**	0.9124

Table 12: VaR for S&P 500 after the GFC

	Out-of-sample exceedances			
Optimizers	VaR at 10%	VaR at 20%	VaR at 80%	VaR at 90%
LS	0.0540	0.1044	0.8916	0.9608
CRPS	**0.0760**	0.1364	0.8528	**0.9352**
FSR10	0.0692	**0.2192**	0.9936	0.9976
FSR20	0.0564	0.1544	0.9844	0.9948
FSR80	0.0252	0.0504	0.8356	0.9516
FSR90	0.0120	0.0280	**0.7908**	0.9424

5. Empirical analysis: VIX

5.1 Background and notation

Time-varying volatility is an important part in finance modelling and an accurate prediction will give investors a better understanding of the risks they take on. GARCH-type models used in Section 3 and 4 are conditionally deterministic. That is, σ_t^2 is a deterministic function of given past returns.

$$r_t = ln(\frac{P_t}{P_{t-1}}) \tag{19}$$

$$r_t = \mu + \sigma_t \epsilon_t \tag{20}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 (r_{t-1} - \mu)^2 + \beta_1 \sigma_{t-1}^2 \tag{21}$$

where P_t is the stock price and σ_t is the volatility.

GARCH models also neglect the fact that volatility in stock markets and derivative markets can have long memory, which means it slowly reverts to its long-run mean. They record changes usually in daily frequency and do not exploit the information in intraday data.

Therefore, it is better off if we model volatility in a continuous-time model where lnP_t and σ_t are allowed to vary continuously over time.

The realized volatility (RV) approach to modelling volatility allows us to exploit the information in intraday data, consider long memory feature, sudden jumps in the market and market microstructure. When the price follows the process as (22):

$$d\ln(P_t) = \mu_t dt + \sigma_t dw_t \tag{22}$$

RV is a direct estimate of integrated volatility (IV) for lnP_t based on continuously recorded observations of P_t over the day.

$$RV_t = \sum_{j=1}^{1/\Delta t} r^2_{t-1+j\Delta t} \xrightarrow{p} \int_{t-1}^{t} \sigma_s^2 ds = IV_t \ as \ \Delta t \to 0 \tag{23}$$

When considering the sudden jumps, that is, the price follows the process in (24). RV is also a consistent estimate of quadratic variation (QV) where κ_s represents the sudden jump on day t and q_t is the jump occasions.

$$d\ln(P_t) = \mu_t dt + \sigma_t dw_t + \kappa_t dq_t \tag{24}$$

$$RV_t = \sum_{j=1}^{1/\Delta t} r^2_{t-1+j\Delta t} \xrightarrow{p} \int_{t-1}^{t} \sigma_s^2 ds \ + \sum_{s=1}^{q_t} \kappa_s^2 = QV_t = IV_t + \sum_{s=1}^{q_t} \kappa_s^2 \ as \ \Delta t \to 0 \tag{25}$$

In addition to the fact that volatility is a direct measure of the risk of portfolios, it is also important for pricing derivatives, such as options, since it is the only unknown parameter to estimate. The famous Black-Scholes model assumes a constant volatility over time which does not hold in practice. Implied volatility allows us to incorporate the jumps and continuous variation of volatility over time. In other words, implied volatility is an estimate of QV extracted from option prices.

VIX index estimates QV that is implied by option prices under a risk-neutral (24) process, using a finite number of strike prices of S&P 500 index (Chicago Board Options Exchange, 2020). It acknowledges the occurrence of jumps and the fact that assumptions of the Black-Scholes model do not hold in practice. Since both the VIX and RV are estimations of QV, it is reasonable to model the VIX with a HAR-RV model which is usually used for modelling RV in the literature (Andersen, Bollerslev & Diebold, 2007; Corsi, 2009; Martin, Reidy & Wright, 2009; Maneesoonthorn, Martin, Forbes & Grose, 2012).

5.2 Preliminary diagnostics

We collect daily VIX index data starting from 27 Aug,1996 to 30 July,2020. And the descriptive statistics of $log(VIX_t)$ are shown in Table 13 and Figure 6:

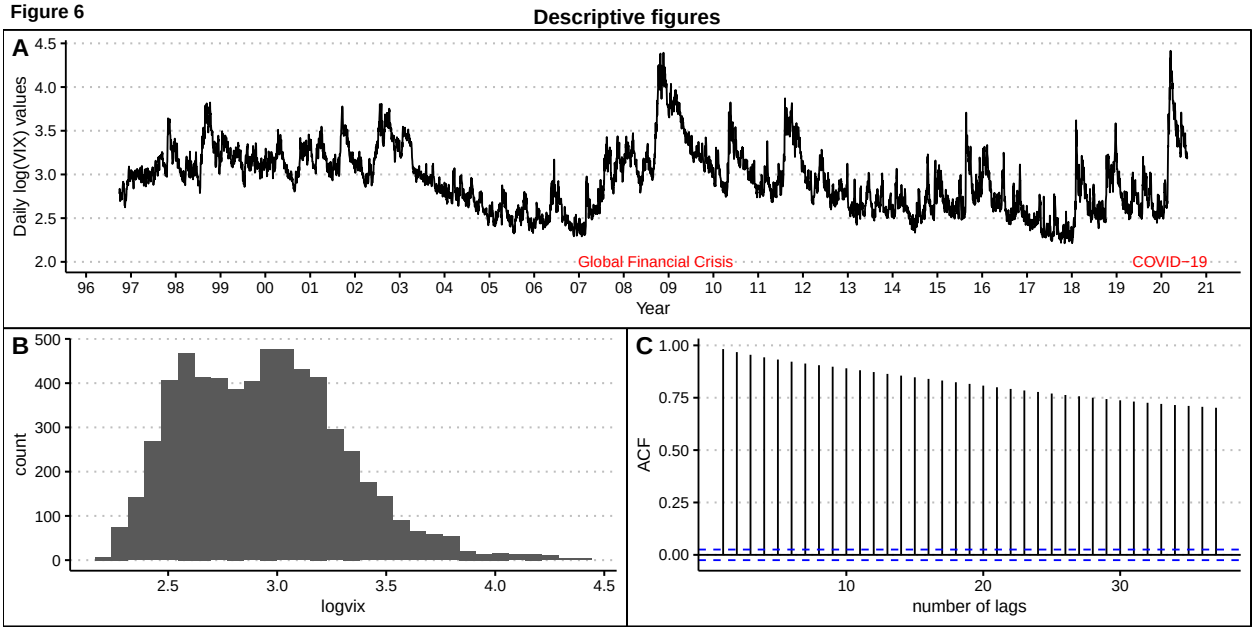
Table 13 and Figure 6 show that $log(VIX_t)$ is volatile and shows strong persistence over time. It is positively skewed and the autocorrelation of its volatility is slowly decaying. From the descriptive statistics, we can see that the features of the VIX is similar to what we have

observed from S&P 500 log returns. Similarly, there are unusual jumps happened during the GFC and the COVID-19. Thus, we expect that the focused score could outperform MLE and CRPS in predicting tails as in Section 4.2 and significantly improve accuracy after the GFC.

Table 13: Descriptive statistics

Series	Min	Median	Mean	Max	Skewness	Kurtosis	JB.Test	LB.Test
Log(VIX)	2.21266	2.916418	2.935766	4.415099	0.6044239	3.363097	410.77	11765

Figure 6



5.3 Model specification

Proposed by Corsi (2009), the HAR-RV model is a simple additive linear model which takes lagged squared returns as regressors. It does not belong to the class of long memory models, but it is able to produce the volatility persistence that is almost indistinguishable from what observed in financial markets through the simple autoregressive-type structure.

We design the VIX analysis with different error term specifications of the HAR-RV model:

1. $z_{t+1} \sim N(0,1)$ and constant σ
2. $z_{t+1} \sim Student - t(0,1,\nu)$ and constant σ [3]
3. $z_{t+1} \sim N(0,1)$ and time-varying σ_{t+1} following ARCH process

[3]We specify z_{t+1} as standardised Student-t distribution so that σ^2 can reflect the variance of the predicted VIX series. The CRPS is removed in Specification 2 and 4, since there is no closed form for this score.

4. $z_{t+1} \sim Student - t(0, 1, \nu)$ and time-varying σ_{t+1} following ARCH process [4]

$$log(VIX_{t+1}) = \beta_0 + \beta_1 log(VIX_t) + \beta_2 log(VIX_{t-5,t}) + \beta_3 log(VIX_{t-22,t}) + \sigma z_{t+1} \quad (26)$$

$$log(VIX_{t-5,t}) = \frac{1}{5}[log(VIX_t) + ... + log(VIX_{t-4})] \quad (27)$$

$$log(VIX_{t-22,t}) = \frac{1}{22}[log(VIX_t) + ... + log(VIX_{t-21})] \quad (28)$$

$$log(VIX_{t+1}) = \beta_0 + \beta_1 log(VIX_t) + \beta_2 log(VIX_{t-5,t}) + \beta_3 log(VIX_{t-22,t}) + \sigma_{t+1} z_{t+1} \quad (29)$$

$$\sigma^2_{t+1} = \alpha_0 + \alpha_1 e^2_t \quad (30)$$

By using the same methodology in Section 3 with initial sample size = 1000 and out-of-sample size $\tau = T - 1000 = 5000$, we produce and evaluate predictive densities with the assumed model as HAR-RV. The results are shown in the following sections.

5.4 Empirical results

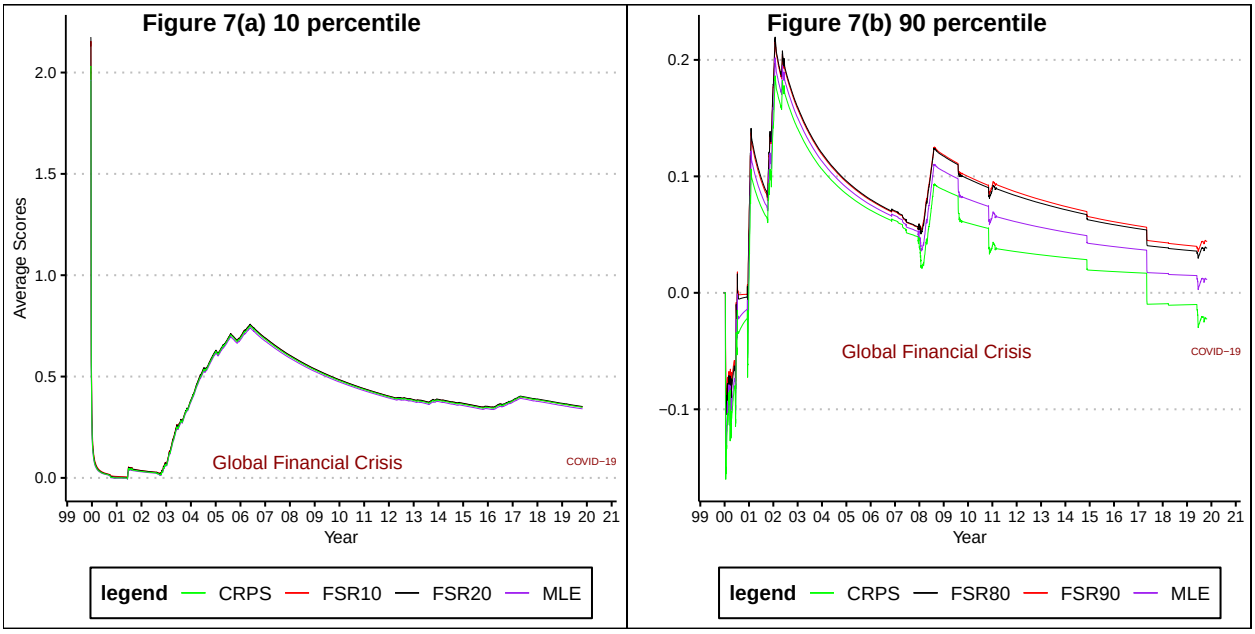
5.4.1 Results of specification 1

Section 5.4.1 provides the result of Specification 1 where the predictive density of the VIX is normal distribution with a constant variance. In Table 14, FSR performs better in predicting upper tail as expected and does a relatively good job in predicting lower tail as well, even though CRPS ranks the best in terms of forecasting 20 percentile of the distribution. The score trace plot, Figure 7, shows the predictions of optimizing four scoring rules evaluated by FSR10 (left panel) and FSR90 (right panel). It indicates how accurate these predictions are, according to their performance in 10 percentile and 90 percentile. In the lower tail, it seems that there is not much difference among four predictions, but zooming in, we can see that FSR is at the top of other scores, which means it can still provide the most accurate prediction. However, we only gain very little from using it. On the contrary, in the upper tail, FSR has a much bigger effect on improving accuracy after the GFC and this effect continues. One possible explanation could be that both LS and CRPS can predict the lower tail better if there is higher probability mass at that area, then the benefit of using FSR is not significant. Considering the marginal distribution shown in Figure 6.B of $log(VIX_t)$, we observe much more observations in the lower end than in the upper end so that it might be the reason why we observe the little improvement in predicting the lower tail in Figure 7(a).

[4] Specification 1 and 2 are based on equations (26)-(28); Specification 3 and 4 are based on equations (28)-(30)

Table 14: Average scores for the VIX

	Average out-of-sample scores					
In-sample optimizers	LS	CRPS	FSR10	FSR20	FSR80	FSR90
LS	1.228	**-0.037**	0.341	0.563	0.078	0.012
CRPS	1.151	**-0.037**	0.348	**0.568**	0.036	-0.022
FSR10	0.556	-0.041	0.350	0.549	-0.330	-0.316
FSR20	0.934	-0.038	**0.352**	0.565	-0.105	-0.135
FSR80	**1.233**	-0.038	0.297	0.517	0.108	0.038
FSR90	1.196	-0.039	0.269	0.479	**0.113**	**0.044**

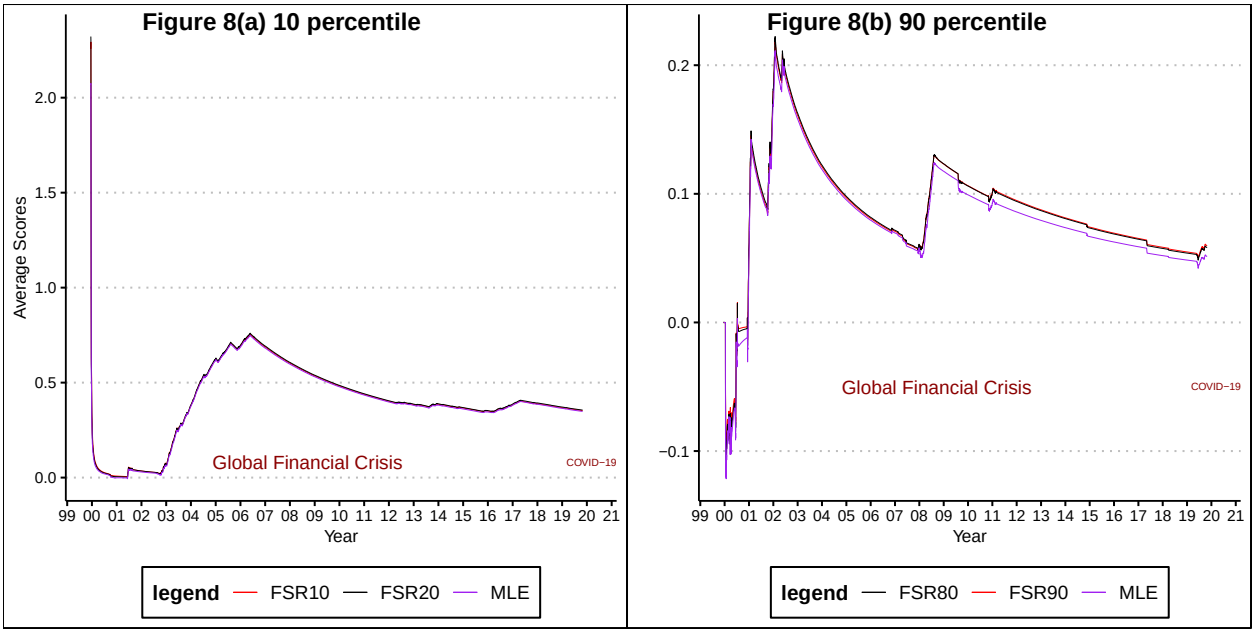


5.4.2 Results of specification 2

The numerical results in Section 3 show that the more misspecified the model is, the more benefits are gained in improving prediction accuracy. According to the marginal distribution estimated by histogram in Figure 6.B, the conditional prediction distribution of the VIX is likely to be positively skewed and has fatter tails than the normal distribution. Moving from Specification 1 to Specification 2, fatter tails are allowed in the predictive distribution and thus, the assumed model is less misspecified. Table 15 and Figure 8 show consistent performance of optimal forecasts compared with our investigations of simulated data and S&P 500 returns. That is, improvement of prediction accuracy by using FSR is still observed in the upper tail but the amount of improvement decreases as the degree of misspecification decreases.

Table 15: Average scores for the VIX Student-t error

	Average out-of-sample scores				
In-sample optimizers	LS	FSR10	FSR20	FSR80	FSR90
LS	**1.323**	0.348	0.579	0.125	0.051
FSR10	0.887	0.352	0.564	-0.133	-0.153
FSR20	1.233	**0.355**	**0.580**	0.072	0.013
FSR80	1.313	0.322	0.552	0.133	0.058
FSR90	1.281	0.299	0.523	**0.134**	**0.060**



5.4.3 Results of specification 3 & 4:

Table 16: Average scores for the VIX ARCH normal error

	Average out-of-sample scores					
In-sample optimizers	LS	CRPS	FSR10	FSR20	FSR80	FSR90
LS	**1.259**	**-0.036**	0.342	0.566	0.104	0.033
CRPS	1.204	**-0.036**	0.349	**0.571**	0.077	0.012
FSR10	0.660	-0.041	0.350	0.552	-0.251	-0.251
FSR20	1.055	-0.037	**0.352**	0.569	-0.015	-0.061
FSR80	1.249	-0.037	0.301	0.524	0.120	0.049
FSR90	1.201	-0.039	0.268	0.480	**0.121**	**0.050**

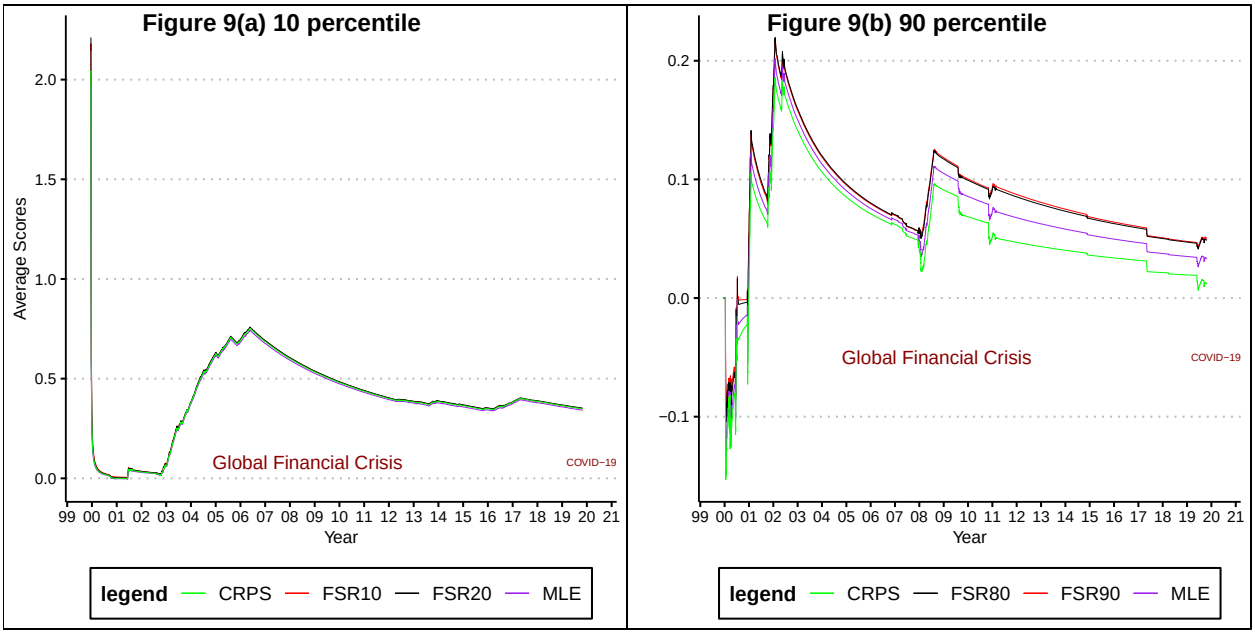
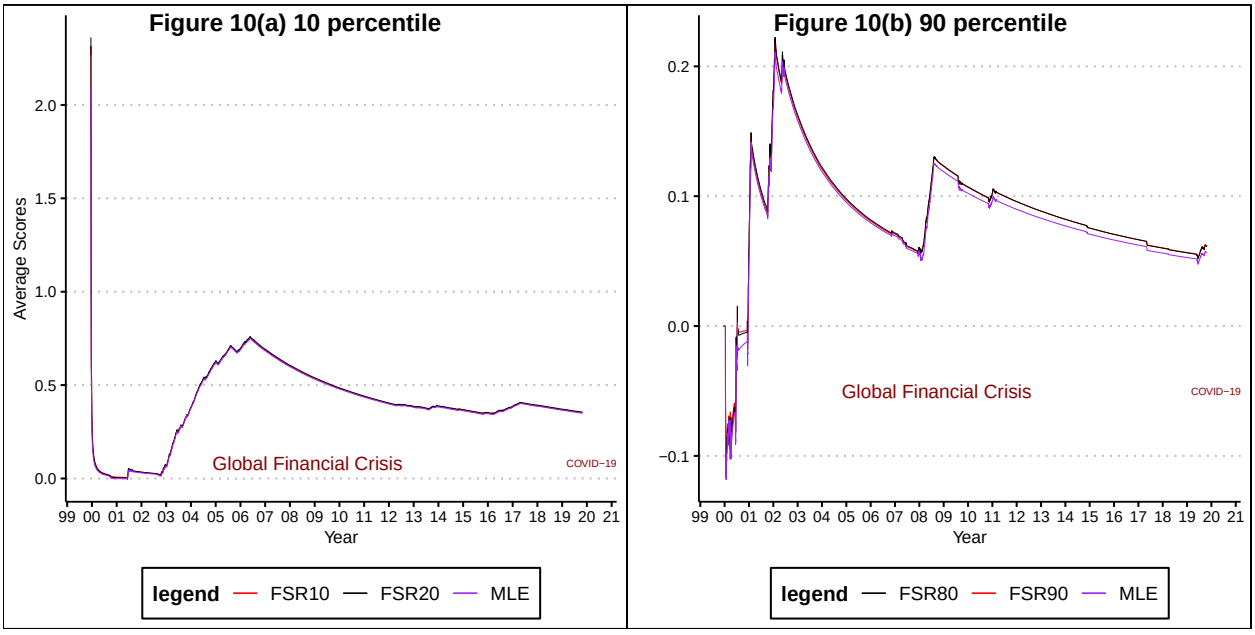


Table 17: Average scores for the VIX ARCH Student-t error

	Average out-of-sample scores				
In-sample optimizers	LS	FSR10	FSR20	FSR80	FSR90
LS	**1.334**	0.349	**0.582**	0.130	0.056
FSR10	0.923	0.352	0.565	-0.105	-0.130
FSR20	1.263	**0.355**	**0.582**	0.090	0.029
FSR80	1.320	0.325	0.556	**0.136**	0.061
FSR90	1.290	0.305	0.529	**0.136**	**0.062**



Specification 3 and 4 move from the constant variance of the predictive distribution to the time-varying variance, which is better specified since strong serial correlation of the volatility is observed in the VIX series as shown in Figure 6.C. Table 16 and Figure 9 illustrate the optimal forecast performance in Model specification 3. The amount of improvement decreased very slightly, compared with results of Specification 1 which also specifies a normal error term. Similar conclusions can be drawn from comparing Table 17 and Figure 10 with the results of Specification 2.

The above comparisons indicate that the amount of improvement of prediction accuracy might mainly depend on how the shape of predictive density matches the shape of true density. Moreover, the significant improvement of prediction accuracy in the upper tail brings great values in financial applications. Firstly, high values of the VIX indicates that the stock market is expected to be more volatile (high volatility), which is an opportunity for either long or short position investors to gain profits or hedge risks, since no one can make money if the stock price doesn't change at all. Besides, investors are typically risk averse, which makes the higher risk (upper tail) is more a concern or an interest to them, compared with the lower risk (lower tail). At last, CBOE provides futures contracts written on the VIX, and higher volatility indicates higher price of futures contracts.

6. Conclusions

According to the numerical and empirical results, the benefits of optimal forecasts can be observed when we have model misspecification and their performance does not vary between two different types of model misspecification investigated in this paper. By mainly focusing on investigating the tail performance of optimal forecasts, the results show that the more misspecified the model is, the more benefits can be gained from using the focused score in improving the tail prediction accuracy in comparison with the conventional likelihood-based counterparts. This result is particularly in evidence after the stock market turbulence in both financial returns and the VIX index. What is interesting is that significant improvement of prediction accuracy is observed in both lower and upper tails of S&P 500 log returns but in the VIX index, it only shows up in the upper tail.

One possible conjecture is that there are more observations in the lower tail of the VIX so that MLE or CRPS can still perform well and thus, less benefits are gained by using FSR. However, it is impossible to know what the shape of the true conditional predictive distribution looks like, but due to the connection between marginal density and conditional density, this might be a reasonable conjecture to make. Nevertheless, these results bring great values to financial applications. For the financial returns, optimal forecasts can give better VaR predictions and for the VIX index, the upper tail, which indicates higher volatility in the market, is more of an interest to investors in terms of making profits, hedging risks, and trading of derivatives.

...

References

Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100 (470):680–701, 2005.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.