Discussion of "Elicitability and backtesting: Perspectives for banking regulation"

Hajo Holzmann[1] **and Bernhard Klar**[2]

[1]: *Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, Germany.*
[2]: *Institut für Stochastik, Karlsruher Institut für Technologie (KIT), Germany.*

Abstract

In our discussion of the insightful paper by Nolde and Ziegel, we further investigate comparative backtests based on consistent scoring functions. We use Diebold-Mariano tests in pairwise comparisons instead of mere rankings in terms of average scores, and illustrate the use of weighted proper scoring rules, which address the quality of forecasts of the full loss distribution in its upper tail rather than some specific risk measure such as the Value at Risk. Overall, at lower levels up to 95%, these allow for better discrimination between competing forecasting methods.

Keywords. backtesting, forecasting, risk management, scoring rule

Nolde and Ziegel have provided us with an insightful paper on backtesting of risk measure forecasts. They shed light on the role of identification functions for traditional, calibration-oriented backtests (see also Davis, 2016), discuss the role of one-sided tests in this framework and propose conditional tests for calibration which are expected to be more powerful than their standard unconditional competitors.

Further, together with Fissler et al. (2016) they make a strong case for the potential usefulness of comparative backtests for regulatory purposes, with the recommendation for regulators to set-up one flexible yet reasonably simple, benchmark forecasting model against which any internal model needs to be tested.

Comparative backtests are based on consistent scoring functions that are tailored to the risk measure at hand. Recently the existence of consistent and strictly consistent scoring functions has been intensively investigated, notably by Gneiting (2011), who characterizes the scoring functions for Value at Risk (VaR) and expectiles, and shows non-existence for the expected shortfall, and by Fissler and Ziegel (2016), who consider VaR and expected shortfall jointly and show existence of and characterize scoring functions for this bivariate functional. In the present paper, Nolde and Ziegel contribute by classifying homogeneous scoring functions, that is scoring functions with certain scaling properties, which is practically relevant due to potentially distinct scales of risks measure forecasts.

In terms of methodology for forecast generation, the authors consider forecasting daily log return series with AR-GARCH models, and focus on the modelling of the distribution of the residuals, where they advocate either flexible parametric models such as the skew-t distribution, or semiparametric models that built on the generalized Pareto distribution from extreme-value theory for the tails of the distribution of the residuals.

[1]Address for correspondence: Prof. Dr. Hajo Holzmann, Philipps-Universität Marburg, Fachbereich Mathematik und Informatik, Hans-Meerwein-Straße. D-35032 Marburg, Germany, Email: holzmann@mathematik.uni-marburg.de, Fon: + 49 6421 2825454

Our discussion will be concerned with the two latter aspects. Forecasts of a risk measure are typically preceded, as in the forecasting framework by Nolde and Ziegel, by an estimate of the whole loss distribution. For regulatory purposes and in particular the evaluation of capital requirements, it is finally the value of the risk measure itself that matters. However, the overall quality of the forecast of the loss distribution, in particular in its upper tail, is also of interest, and rankings of distinct forecasting schemes that rely directly on the loss distribution thus do not depend on the choice of the risk measure, be it VaR, expected shortfall or expectile.
The appropriate tools for comparing distribution forecasts are proper scoring rules (Gneiting and Raftery 2007), and we shall restrict ourselves to weighted versions of the continuous-ranked probability score (CRPS) from Gneiting and Ranjan (2011) and Holzmann and Klar (2017), which we recall below.
The quantile-weighted version of the CRPS (QCRPS) from Gneiting and Ranjan (2011) takes into account the full forecast distribution above a specified α-quantile, and we shall compare test results using score differences for distinct forecasts based on the QCRPS with those based on scores for the VaR or the pair VaR/expected shortfall for the same level α. Probability-weighted versions of the CRPS from Gneiting and Ranjan (2011) (termed twCRPS) and Holzmann and Klar (2017) (wsCRPS) address the quality of the forecast of the (conditional) loss distribution above a certain fixed threshold r, say a $r = 1\%$ or a $r = 2\%$ loss, rather than above the conditional α-quantile as for the QCRPS. Here, we investigate test results for score differences from twCRPS and wsCRPS on the one side, and the QRCPS on the other side, where the level α in the QCRPS is chosen approximately equal to the level of the unconditional distribution function of the observations at the threshold r.
The forecasting methods that we consider are basically those from Nolde and Ziegel, which are applied to time series of daily log returns $y_t = \ln(P_t/P_{t-1})$, where P_t is the closing price on day t, adjusted for dividends and splits for the four stock-indices S&P500, DAX, Nikkei 225 and NASDAQ, running from January 1, 2009 until December 31, 2016, giving a total of about 2000 observations for each series. The data is publicly available and has been downloaded from `http://finance.yahoo.com`. In our analysis, we also restrict ourselves to one-step ahead forecasts, but do not merely list the rankings in terms of average scores, but rather report values of the corresponding t-statistics from Diebold-Mariano (1995) tests for pairwise comparisons to assess whether observed score differences are statistically significant.

Methodology

For a distribution function F (the forecast), a real number x (the following observation) and a threshold quantile value $\alpha \in (0, 1)$, the quantile-weighted CRPS from Gneiting and Ranjan (2011) is defined by

$$\mathrm{QCRPS}(F, x; \alpha) = \frac{1}{1-\alpha} \int_\alpha^1 QS_\beta\big(F^{-1}(\beta), x\big)\, \mathrm{d}\beta,$$

where F^{-1} is the quantile function of F and $\mathrm{QS}_\beta(q, x) = 2\,\big(1_{x<q} - \beta\big)(q - x)$ is the 1-homogeneous version of the scoring function for the β-quantile. Note that other versions of the quantile score, e.g. the 0-homogenous score from (2.20) in Nolde and Ziegel would be possible as well in the definition of the QCRPS. The threshold-weighted CRPS from Gneiting and Ranjan (2011) with threshold parameter r is given by

$$\mathrm{twCRPS}(F, x; r) = \int_r^\infty \big(F(z) - 1\{x \le z\}\big)^2\, \mathrm{d}z,$$

and the variant introduced in Holzmann and Klar (2017) is defined by

$$\begin{aligned}&\mathrm{wsCRPS}(F, x; r)\\ &= 1\{x > r\}\left[F(r)^2 + \int_r^\infty \Big(\frac{F(z) - F(r)}{1 - F(r)} - 1\{x \le z\}\Big)^2 \,\mathrm{d}z\right] + 1\{x \le r\}\big(1 - F(r)\big)^2.\end{aligned}$$

Properties of these scoring rules are detailed in Holzmann and Klar (2017). For the Value at Risk we also use the 1-homogeneous scoring function for the sake of comparison, and for the pair VaR/expected shortfall we use the version in (2.23) in Nolde and Ziegel.
Concerning the time series modelling of the log-return series, we restrict ourselves to GARCH(1,1)-models without autoregressive component but including a constant intercept. We consider the fully parametric approaches with normal, t- and skew-t-distributed innovations, fitted by maximum likelihood, as well as a semiparametric, extreme-value based method. This is, however, only used in connection with residuals obtained from normal-GARCH pseudo maximum likelihood estimates for the GARCH coefficients, which are known to work quite generally and are, in contrast to other parametric specifications such as t-residuals, robust to misspecification (Straumann, 2005). Similar to McNeil and Frey (2000) and Nolde and Ziegel, we set the cut-off for the peak-over-threshold method to include the largest 10% of the observations. Above the 0.9 quantile, the parametric generalized Pareto fit for the distribution function is used, see p. 282 in McNeil and Frey (2000), below, the empirical distribution is employed. We use one-step-ahead density forecasts with a rolling window scheme for parameter estimation using R and the R package rugarch (R Core Team 2016, Ghalanos, 2014). The length of the estimation window is set to 500 observations, so that the number of out-of-sample observations is about 1500 for each of the four series.
Maximum-likelihood fitting for peaks over threshold modelling using the generalized Pareto distribution is done using the R library evd (Stephenson, 2002).

Case study

Let us first consider the test results for score differences based on the QCRPS, as compared to those for the score for VaR as well as for the pair VaR/expected shortfall. We use the three levels $\alpha = 0.95$, $\alpha = 0.975$ and $\alpha = 0.99$. The results are contained in Table 1, where the entries n-t, n-st and t-st stand for normal vs. t, normal vs. skew-t, and t vs. skew-t distributions for the innovations. Analogous notation is used for the comparison of parametric models against the extreme value theory (evt) based semiparametric models.
Unsurprisingly, for testing the normal against any of the other models, the values of the t-statistics based on any score are highest overall (hence show evidence against the normal model). Among these, the t-statistics based on the VaR-score are the smallest for all three levels, while at level 0.95, the largest values arise from using the QCRPS, in particular for the NASDAQ series but also for the other series. In contrast, for the high 0.99 level, largest values of the t-statistics are obtained based on the score for the pair VaR/expected shortfall, while for the 0.975 level, QCRPS and VaR/expected-shortfall-based values are comparable. Overall, various values of the t-test statistics are at around 2, showing statistical significance at a 5% level (without multiple testing correction).
Values of the Diebold-Mariano statistic for testing t against skew-t models are always positive (in favour of skew-t), with the highest values based on the score for VaR/expected shortfall, though even then they are only rarely significant (e.g. S&P 500, levels 0.95 and 0.975, NASDAQ for level 0.975). When testing the t-model against the extreme value approach, the

			n-t	n-st	t-st	n-evt	t-evt	st-evt
		QCRPS	1.09	1.67	1.69	1.13	1.02	-0.12
	$\alpha = 0.95$	VaR	-2.08	0.90	1.68	0.25	0.74	-0.29
		VaR/ES	-1.37	1.44	1.94	1.11	1.42	0.27
		QCRPS	1.59	1.70	1.53	1.55	1.25	0.17
S&P 500	$\alpha = 0.975$	VaR	0.67	1.66	1.80	1.09	1.12	-0.05
		VaR/ES	1.46	2.05	1.99	1.68	1.61	0.33
		QCRPS	1.59	1.52	0.97	1.86	0.60	-0.03
	$\alpha = 0.99$	VaR	1.41	1.46	1.33	1.49	1.36	0.72
		VaR/ES	1.89	1.91	1.59	2.02	1.68	0.83
		QCRPS	1.92	1.88	1.33	1.59	1.08	0.88
	$\alpha = 0.95$	VaR	-0.42	0.78	1.29	0.97	0.95	0.77
		VaR/ES	0.66	1.68	1.50	1.44	1.12	0.89
		QCRPS	2.26	1.97	0.98	1.84	0.98	0.81
DAX	$\alpha = 0.975$	VaR	1.75	1.90	1.58	1.67	1.25	0.98
		VaR/ES	2.11	2.15	1.66	1.94	1.41	1.15
		QCRPS	1.65	1.38	0.21	1.55	0.04	-0.11
	$\alpha = 0.99$	VaR	2.00	1.73	0.40	1.81	0.78	0.87
		VaR/ES	2.18	1.92	0.62	2.00	0.96	0.96
		QCRPS	1.46	1.62	1.15	1.23	0.84	0.50
	$\alpha = 0.95$	VaR	-0.67	-0.23	0.56	0.69	0.80	0.84
		VaR/ES	0.13	1.03	0.94	1.33	1.18	1.20
		QCRPS	1.98	1.85	1.20	1.53	0.65	-0.06
Nikkei 225	$\alpha = 0.975$	VaR	0.50	0.98	0.88	0.72	0.62	0.29
		VaR/ES	1.71	1.68	1.16	1.41	0.98	0.69
		QCRPS	2.12	2.06	1.28	1.90	-0.14	-1.06
	$\alpha = 0.99$	VaR	1.85	1.85	1.47	1.82	1.33	0.90
		VaR/ES	2.24	2.21	1.63	2.16	1.19	0.27
		QCRPS	2.01	2.03	1.73	1.52	1.10	0.49
	$\alpha = 0.95$	VaR	-0.51	1.25	1.53	0.65	0.71	0.07
		VaR/ES	0.60	1.88	1.86	1.36	1.17	0.54
		QCRPS	2.17	2.05	1.60	1.85	1.28	0.74
NASDAQ	$\alpha = 0.975$	VaR	1.52	1.90	1.79	1.46	1.21	0.67
		VaR/ES	2.15	2.36	2.08	1.99	1.60	1.08
		QCRPS	1.99	1.81	0.77	1.89	-0.23	-0.86
	$\alpha = 0.99$	VaR	2.28	2.01	1.27	1.96	1.31	1.06
		VaR/ES	2.56	2.30	1.37	2.29	1.44	1.10

Table 1: t-statistics for Diebold-Mariano test based on QCRPS, VaR and the pair VaR/expected shortfall for equal predictive accuracy. Positive values indicate superiority of forecasts from the second method, while negative values indicate superiority of forecasts from the first method.

		n-t	n-st	t-st	n-evt	t-evt	st-evt
S&P 500	twCRPS	-0.08	1.38	1.73	0.73	0.69	-0.60
	wsCRPS	2.12	2.30	1.54	1.06	0.20	-0.74
	QCRPS	-1.27	1.05	1.75	0.74	1.17	-0.09
DAX	twCRPS	0.74	1.53	1.04	1.36	0.50	0.29
	wsCRPS	2.27	2.71	1.83	1.86	1.00	0.70
	QCRPS	-0.11	0.79	1.25	0.80	0.64	0.41
Nikkei 225	twCRPS	2.23	2.64	0.83	2.70	-0.37	-0.90
	wsCRPS	2.76	3.06	0.99	1.48	-0.19	-0.82
	QCRPS	0.57	1.48	0.89	0.81	0.10	-0.38
NASDAQ	twCRPS	0.33	1.75	1.52	1.32	0.83	0.03
	wsCRPS	2.43	2.34	1.31	0.92	-0.01	-0.09
	QCRPS	-0.63	1.32	1.83	0.99	1.03	0.18

Table 2: t-statistics for Diebold-Mariano test based on twCRPS (with $r = 1$), wsCRPS (with $r = 1$) and QCRPS (with $\alpha = 0.85$) for equal predictive accuracy. Positive values indicate superiority of forecasts from the second method, while negative values indicate superiority of forecasts from the first method.

results are similar, but even without multiple testing correction there are no longer significant results in this case. Finally, when testing the skew-t model against the extreme value based approach, there are mainly positive values (in favour of extreme value), which are, however, even smaller than when testing against the t distribution. In summary, the light-tailed normal model for the innovations can often be rejected, for level 0.95 particularly based on the QCRPS, and for the high level 0.99 on the basis of expected shortfall. Since nearly all entries in the corresponding columns are positive, the t model seems inferior overall to the skew-t and extreme-value based method. The extreme-value based method and the skew-t method perform similarly, with slight but non-significant evidence that the extreme-value method is best if the target is the expected shortfall at high levels.
Finally, let us extract from Table 1 the results for comparing normal against t distributed innovations for the S&P 500 at various levels, which are notable since Diebold-Mariano statistics are positive for QCRPS but negative for VaR and expected shortfall at level 0.95. The following tables shows the results, also for the additional levels 0.9 and 0.85.

α	0.85	0.90	0.95	0.975	0.99
QCRPS	-1.27	-1.05	1.09	1.59	1.59
VaR	-0.78	-1.52	-2.08	0.67	1.41
VaR/ES	-1.07	-1.54	-1.37	1.46	1.89

For all scores, the values are negative for moderate levels, but positive for high levels, with the transition for the QCRPS at 0.9-0.95, and for the VaR and the VaR/ES at 0.95-0.975. Thus, in the far tail, the t distribution is always superior, while the normal distribution is preferable for moderate quantiles, but the emphasis given by the distinct scores to moderate and high quantiles varies.

Second, let us compare the test results for score differences based on the QCRPS at the more moderate level of $\alpha = 0.85$ with those for the twCRPS and the wsCRPS for threshold $r = 1$, which are displayed in Table 2. The following table shows the percentages of observations for

the four series of negated log returns which fall above 1, which correspond reasonably well to the choice of $1 - \alpha = 0.15$ for the QCRPS.

	S&P 500	DAX	Nikkei 225	NASDAQ
perc. above 1	12%	18%	19%	14%

As a first observation none of the values of the t-statistics based on the QCRPS (and neither for scores based on VaR and expected shortfall, results not displayed) at level 0.85 are significant at a 5% test level - even without correction for multiple testing.
In contrast, for the probability-weighted versions of the CRPS, the wsCRPS gives significant evidence for the superiority of the parametric t- and skew-t models as compared to the normal model for each of the four indices. The twCRPS appears to be somewhat less conclusive than the wsCRPS for these comparisons, and gives significant results only in case of the Nikkei index.
There is no significant evidence in favour of the extreme-value model in all but one of the pairwise comparisons. Further, whereas all entries of the comparisons between skew-t and the other parametric models are positive, this is not the case for the extreme-value based method. Hence, the latter seems to be inferior to the skew-t model, even if the direct competition between the two methods remains inconclusive.

Concluding remarks

Weighted scoring rules allow for comparative backtests of loss-distribution forecasts that do not rely on a particular choice of a risk measure, but rather depend on the overall quality of the loss-distribution forecast in its upper tail. For a univariate time series of portfolio log returns, our recommendation would be to test a standardized benchmark procedure such as the GARCH(1,1)-model with skew-t-distributed innovations against an internal model based on the quantile-weighted version of the continuous-ranked probability score from Gneiting and Ranjan (2011), which takes into account the full loss distribution above a specified quantile of level α, with the choice of $\alpha = 0.95$.
Often, the loss-distribution forecasts that precede the risk measure forecasts are themselves preceded by forecasts of various risk factors which enter into the portfolio, see McNeil et al. (2005). Thus, backtesting could and should address forecasts of these multidimensional quantities as well. Comparative backtests based on risk factor distribution forecasts do neither depend on a risk measure nor on the composition of the portfolio. Both generating and backtesting such multivariate distribution forecasts provide additional challenges. For the latter, multivariate versions of the wsCRPS from Holzmann and Klar (2017) could be used to emphasize regions of interest of risk factors that generate large losses. Setting up some benchmark, flexible and yet sufficiently simple multidimensional forecasting mechanism for risk factors appears to be a formidable challenge.

Additional references

Ghalanos, A. (2014). rugarch: Univariate GARCH models. *R package version 1.3-5.*

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359-378.

Holzmann, H. and Klar, B. (2017). Focusing on regions of interest in forecast evaluation. Submitted.

R Core Team (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*

Stephenson, A. G. (2002). evd: Extreme Value Distributions. *R News* **2**, 31-32.

Straumann, D. (2005). *Estimation in Conditionally Heteroscedastic Time Series Models.* Lecture Notes in Statistics, Springer.