The Annals of Applied Statistics
2017, Vol. 11, No. 4, 1901–1911
https://doi.org/10.1214/17-AOAS1041F
Main article: https://doi.org/10.1214/17-AOAS1041

REJOINDER: "ELICITABILITY AND BACKTESTING: PERSPECTIVES FOR BANKING REGULATION"

BY NATALIA NOLDE AND JOHANNA F. ZIEGEL

University of British Columbia and University of Bern

We would like to thank the discussants for interesting and insightful contributions. The discussants raised a number of diverse points, related to both theory underlying backtesting methodologies as well as to practical implications for banking regulation.

Robust traditional and comparative backtests. Chen Zhou clarifies the relation between the notion of identifiability of a risk measure and the ability to perform traditional backtests in the form of conditional calibration tests. We fully agree with him that in the absence of an identification function it is still possible to perform traditional backtests by assuming common properties of the conditional distributions across time. In our work, we have entirely focused on *robust* backtests as Zhou has phrased it, where robustness refers to robustness with respect to model uncertainty.

We would like to add that the same clarifications are in order for comparative backtests. Both elicitability and identifiability are only meaningful concepts when stated with respect to which class of distributions $\mathcal{P}$ they hold; cf. Definitions 1 and 2. Broadly speaking, the smaller the class $\mathcal{P}$, the weaker the condition for existence of an identification function or a strictly consistent scoring function for a given functional T. Let us give the following simple example: Suppose that $\mathcal{P}_s$ is a class of symmetric distributions. Then, for each $P \in \mathcal{P}_s$, the mean and the median coincide. Therefore, all consistent scoring functions for the median are also consistent scoring functions for the mean *relative to* $\mathcal{P}_s$, and the same holds for the respective identification functions. Relative to a class $\mathcal{P}_c$ of distribution functions such that all distributions have the same α-quantile, say $\mathrm{VaR}_\alpha(P) = c$ for all $P \in \mathcal{P}_c$, ES is identifiable and elicitable. Strictly consistent scoring functions can be obtained by setting $r_1 = c$ in equation (2.4). Similarly, the second component of the identification function at (2.7) with $r_1 = c$ identifies ES_α relative to $\mathcal{P}_c$. This is reflected in the ES backtest given by Zhou: The assumptions on the data-generating process allow to estimate c well enough that asymptotically we can work as if c was known.

Hajo Holzmann and Berhard Klar suggest comparative backtests for the entire tail of the P&L distribution instead of a specific risk measure; let us term them

Received August 2017.

distributional comparative backtests. We appreciate their suggestion and we agree that such an approach is natural given that forecasts of risk measures are often preceded by an estimate of the whole P&L distribution. [Examples of exceptions for VaR forecasts are the CAViaR approach by Engle and Manganelli (2004) and the feedback algorithm of Davis (2016); for ES forecasts, see Patton, Ziegel and Chen (2017).] One may argue in line with Gneiting and Katzfuss (2014) that forecasting (and evaluating) the entire P&L distribution (or its tail) may have the merit of providing a more complete assessment of the risk. The approach takes into account that predictive accuracy of internal models should be the main concern to regulators and banks. However, it has to be acknowledged that some tasks such as regulatory capital calculation still require a single number rather than a distribution. Performing backtests at the level of the P&L distribution has the advantage that, subsequently, conditionally on passing the backtest, the predictive distribution may be used to calculate various measures of risk or other functionals of interest that may be distinct, for example, for internal decision making versus regulatory requirements.

The distributional comparative backtests are based on proper scoring rules, in particular, weighted versions of the continuous ranked probability score (CRPS) [Gneiting and Ranjan (2011), Holzmann and Klar (2016). Once a specific proper scoring rule is chosen, different P&L distribution forecasts are compared using Diebold–Mariano tests [Diebold and Mariano (1995)], just as we have suggested for the comparative backtests for specific elicitable risk measures. Traffic light matrices are therefore also easily constructed for distributional comparative backtests. The proposed distributional comparative backtests are robust with respect to model uncertainty as discussed by Zhou.

We would like to add that (robust) distributional traditional backtests, that is, traditional backtests for (the tail of) the P&L distribution, are known in the forecasting literature as tests for calibration of probabilistic forecasts. One of the most prominent examples are tests for uniformity and independence of probability integral transform (PIT) values going back to Diebold, Gunther and Tay (1998); see also Gneiting, Balabdaoui and Raftery (2007), Gneiting and Ranjan (2013), Gordy, Lok and McNeil (2017), Strähl and Ziegel (2017).

Marie Kratz notes that a simple implicit (traditional) backtest for ES is the approach to test VaR at several levels simultaneously [Kratz, Lok and McNeil (2016)]. While we feel that it is debatable whether this backtest should be termed a backtest *for ES*, we greatly appreciate her pointing out this test. In our terminology, this is a simple conditional calibration test for the vector of risk measures

$$\Theta = (\mathrm{VaR}_{\alpha_1}, \ldots, \mathrm{VaR}_{\alpha_J}), \tag{1}$$

where $\alpha_j = \alpha + ((j-1)/J)(1-\alpha)$, $j = 1, \ldots, J$. As pointed out by Kratz and Davis, VaR plays a special role for one-period ahead forecasts in that it does not require an asymptotic test as in (2.11) due to the uniformity and independence of

PIT values also quoted above. Backtesting VaR at J different levels is a middle ground between the proposal of Holzmann and Klar to assess the whole (tail of the) P&L distribution versus backtesting only one specific risk measure.

A simulation study. It is straightforward to perform also comparative backtests for Θ at (1). Complementing their results, we report results of a small simulation study with the same setting as in Kratz, Lok and McNeil (2016) (Section 3.3), but with a focus on comparative backtesting. We consider Θ as in (1) with $\alpha = 0.975$ and $J \in \{4, 8, \infty\}$. The case $J = \infty$ corresponds to assessing the whole tail of the P&L quantile function beyond the level α as suggested by Holzmann and Klar. For $J \in \{4, 8\}$, we use the following scoring function to assess performance of forecasting procedures

$$S(r_1, \ldots, r_J, x) = \frac{1}{J} \sum_{j=1}^{J} S_j^{(h)}(r_j, x), \tag{2}$$

where r_j denotes a forecast for VaR_{α_j} and $S_j^{(h)}$ is a consistent scoring function for VaR_{α_j} with $h = 1$ corresponding to the standard 1-homogeneous case in equation (2.19) and $h = 0$ to the 0-homogeneous case in equation (2.20). The scoring function in (2) relates to the quantile-weighted CRPS from Gneiting and Ranjan (2011) as considered by Holzmann and Klar (in the case $J = \infty$) with the left Riemann sum approximation to the integral

$$\begin{aligned} \mathrm{QCRPS}(F, x; \alpha) &= \frac{1}{1-\alpha} \int_{\alpha}^{1} S_{\beta}^{(h)}\big(F^{-1}(\beta), x\big)\, d\beta \\ &\approx \frac{1}{1-\alpha} \sum_{j=1}^{J} S_j^{(h)}\big(F^{-1}(\alpha_j), x\big)(\alpha_{j+1} - \alpha_j) \\ &= \frac{1}{J} \sum_{j=1}^{J} S_j^{(h)}\big(F^{-1}(\alpha_j), x\big). \end{aligned} \tag{3}$$

The out-of-sample size $n = 2000$ is used to evaluate average scores. The data are generated from a GARCH(1, 1) process:

$$X_t = \sigma_t Z_t, \qquad \sigma_t^2 = 2.18 \times 10^{-6} + 0.109 X_{t-1}^2 + 0.890 \sigma_{t-1}^2, \tag{4}$$

where innovations $\{Z_t\}_{t \in \mathbb{Z}}$ form an i.i.d. sequence of standardized Student t-distributed random variables with 5.06 degrees of freedom. A moving estimation window of size 500 is used to produce one-step ahead forecasts of Θ using the same methods as in Kratz, Lok and McNeil (2016):

- "hs": historical simulation with VaR_{α_j} values given by empirical quantiles;
- "arch.t": an ARCH(1) filter fitted assuming Student t innovations;

- "garch.n": a GARCH(1, 1) filter fitted assuming normally distributed innovations;
- "garch.hs": a GARCH(1, 1) filter fitted assuming Student t-distributed innovations and using empirical estimates to estimate quantiles based on realized innovations;
- "garch.evt": a GARCH(1, 1) filter fitted assuming Student t-distributed innovations and using EVT methodology to estimate quantiles based on realized innovations;
- "garch.t": a GARCH(1, 1) filter fitted assuming Student t-distributed innovations and using model-based quantile estimates;
- "oracle": based on the knowledge of the data-generating process (the true model).

Table 1 summarizes the average scores using the combined score in (2) for different values of J, the number of quantiles being evaluated, and the 1- and 0-homogeneous individual scoring functions. The corresponding traffic light matrices at the test level $\eta = .10$ are displayed in Figure 1 for $\mathrm{VaR}_{0.99}$ corresponding to $J = 1$, and in Figure 2 for $J > 1$ with the starting level $\alpha = 0.975$. Note that the case $J = \infty$ corresponds to the quantile-weighted CRPS in (3).

Comparative backtests clearly distinguish between the true model (oracle), good models (garch.t, garch.hs, garch.evt) and poor models (garch.n, arch.t and hs) as grouped by Kratz, Lok and McNeil (2016). For a single VaR ($J = 1$ case) and the entire distributional tail past $\alpha = 0.975$ level ($J = \infty$ case), rankings within each group (see Table 1) are consistent with our intuition, although, in the case of the true model and good models, results are generally not significant. The historical simulation approach is consistently ranked the lowest among the other methods considered, followed by the arch.t method; both methods are in the red region against the other methods for $J > 1$. When considering multiple VaR levels, the results suggest a change in ranking for some of the methods. In particular, the "garch.hs" method, which is based on a correctly specified GARCH filter and takes an empirical estimate for the quantile of innovations, outperforms its direct EVT-based and fully parametric counterparts, "garch.evt" and "garch.t" ($J = 4$ case). As we consider the entire tail, there is a good discrimination between the good and poor models, and, as Holzmann and Klar point out, forecasting procedures based on a flexible parametric method (in our setting, the correctly specified "garch.t" method) outperform the EVT-based approach ("garch.evt").

Different methods may perform better or worse at different risk measure levels, and averaging over an arbitrarily chosen set of these levels may lead to either inconclusive assessments or give preference to a method that performs well for that particular selection of risk measure levels. From this perspective, it may be better to either focus on forecasting and backtesting for a specific targeted risk measure level or follow the approach advocated by Holzmann and Klar to consider the entire distributional tail giving equal importance to all quantile levels.

TABLE 1

Average scores and corresponding ranks (in brackets) based on the scoring function in (2) for a simulated series coming from a GARCH(1, 1) *process with Student's t innovations in (4). The case* $J = 1$ *is for* $\mathrm{VaR}_{0.99}$*, with the other cases as defined in (1) for the starting level* $\alpha = 0.975$*. Average scores are evaluated based on* 2000 *verifying observations*

	$J = 1$		$J = 4$		$J = 8$		$J = \infty$	
Method	$10^3 \times \bar{S}^{(1)}_{0.99}$	$10 \times \bar{S}^{(0)}_{0.99}$	$10^3 \times \bar{S}^{(1)}$	$10 \times \bar{S}^{(0)}$	$10^3 \times \bar{S}^{(1)}$	$10 \times \bar{S}^{(0)}$	$10^3 \times \bar{S}^{(1)}$	$10 \times \bar{S}^{(0)}$
hs	0.763 (7)	−0.284 (7)	1.332 (7)	−0.747 (7)	1.280 (7)	−0.697 (7)	−36.064 (7)	−0.361 (7)
arch.t	0.657 (6)	−0.299 (6)	1.195 (6)	−0.775 (6)	1.145 (6)	−0.723 (6)	−37.746 (6)	−0.377 (6)
garch.n	0.594 (4)	−0.312 (5)	1.042 (4)	−0.818 (5)	0.999 (4)	−0.763 (5)	−39.487 (5)	−0.395 (5)
garch.hs	0.597 (5)	−0.312 (4)	1.040 (3)	−0.822 (1)	0.997 (2)	−0.767 (1)	−39.895 (4)	−0.399 (4)
garch.evt	0.588 (3)	−0.316 (3)	1.046 (5)	−0.819 (4)	1.005 (5)	−0.764 (4)	−39.900 (3)	−0.399 (3)
garch.t	0.579 (2)	−0.317 (2)	1.038 (2)	−0.820 (3)	0.997 (3)	−0.765 (3)	−40.038 (2)	−0.400 (2)
oracle	0.578 (1)	−0.318 (1)	1.037 (1)	−0.821 (2)	0.995 (1)	−0.766 (2)	−40.134 (1)	−0.401 (1)

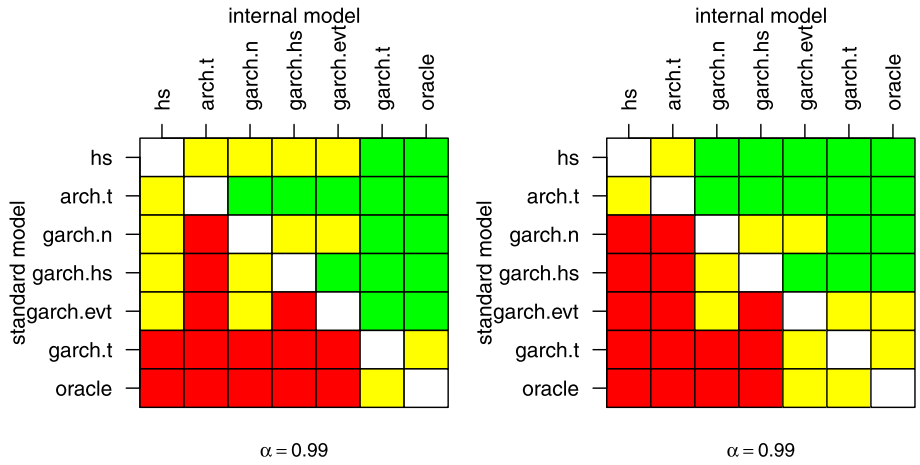

FIG. 1. *Traffic light matrices for forecasts of* $\mathrm{VaR}_{0.99}$ *at the test confidence level* $\eta = 0.10$. *The left panel is based on the standard* 1*-homogeneous scoring function* [*equation* (2.19)] *and the right panel uses the* 0*-homogeneous scoring function* [*equation* (2.20)]. *The data-generating process is given at* (4) *with a moving estimation window of* 500 *and the out-of-sample size to evaluate average scores of* 2000.

Backtests and incentives. Patrick Schmidt makes a strong point for backtests that are sensitive with respect to increasing information sets; that is, comparative backtests rather than calibration tests as backtesting should incentivize the development of accurate and informative risk models.[1] We fully agree with his argument and find that the idea of introducing a cost function for acquiring information (and incorporating it optimally into the risk model) is illustrative.

Formally, one could state his idea as follows. Suppose we are at timepoint $t-1$ and there exists an increasing sequence of σ-algebras of information $(\mathcal{A}_{t-1,k})_{k\in\mathbb{N}}$, where $\mathcal{A}_{t-1,1} = \sigma(X_1, \ldots, X_{t-1})$ and $\mathcal{A}_{t-1,\infty} := \bigcup_{k\in\mathbb{N}} \mathcal{A}_{t-1,k}$ is such that X_t is $\mathcal{A}_{t-1,\infty}$-measurable, or, in other words, $\sigma(X_1, \ldots, X_t) \subseteq \mathcal{A}_{t-1,\infty}$. Having access to the information in $\mathcal{A}_{t-1,k}$ comes at the cost $c_{t-1,k}$, where $\lim_{k\to\infty} c_{t-1,k} = \infty$. If we were able to pay infinitely much, then we could have access to $\mathcal{A}_{t-1,\infty}$ and know the value of X_t already at time point $t-1$ having removed all randomness in our prediction problem. Clearly, this is never possible. Therefore, we need to find a compromise between the cost of using and acquiring information and the resulting forecast accuracy. However, it is important to note that independently of how much we are willing to invest in information, that is, which $\mathcal{A}_{t-1,k}$ we base our predictions on, we can issue a calibrated forecast that passes all conditional calibration

[1] We believe that when he refers to "unconditional calibration test" he is actually considering simple conditional calibration tests such as the current Basel VaR backtest. These tests are not conditional on extra information, but they are conditional on past realizations of the asset or portfolio like the Basel test for VaR exceedances.

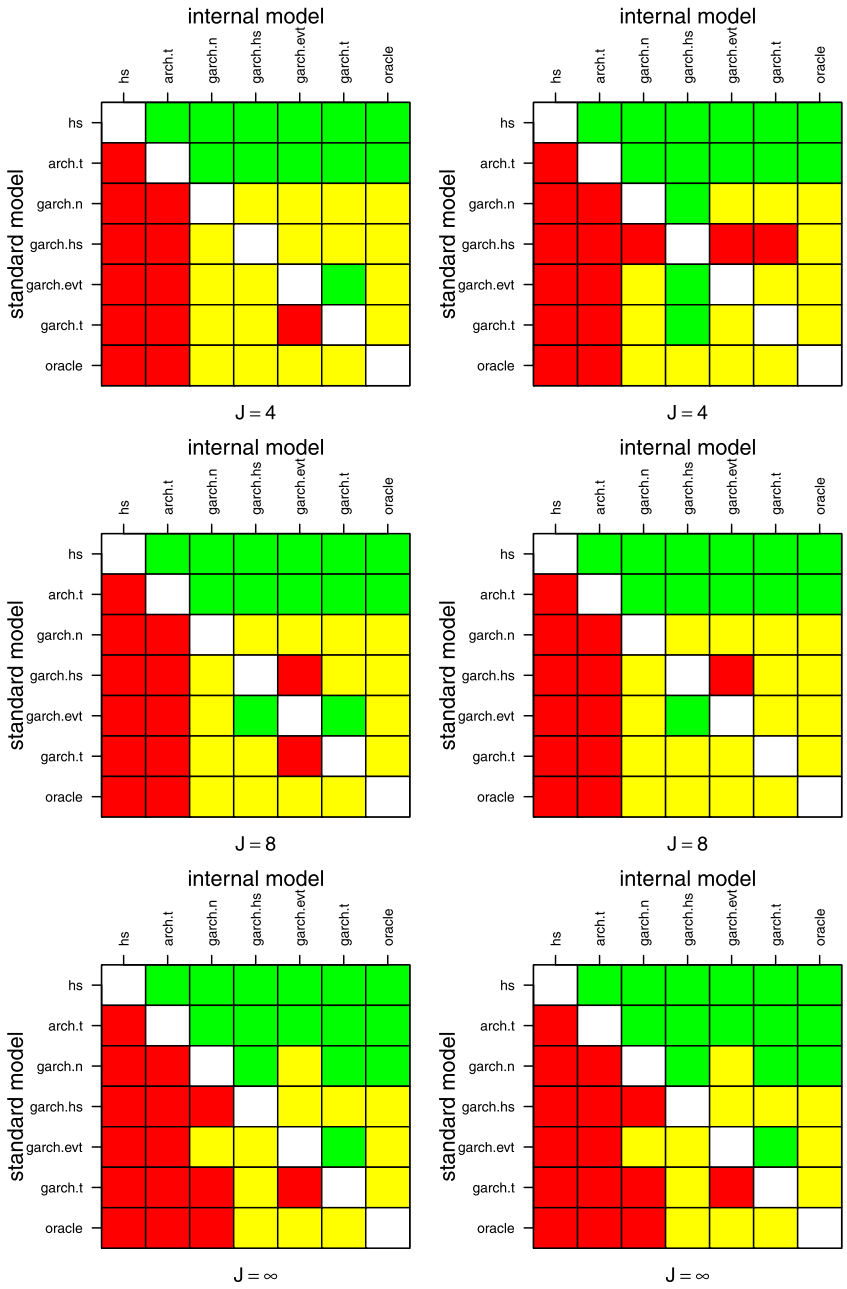

FIG. 2. *Traffic light matrices for forecasts of* Θ *in* (1) *at the test confidence level* $\eta = 0.10$. *The left panel is based on the standard* 1*-homogeneous scoring function* [*equation* (2.19)] *and the right panel uses the* 0*-homogeneous scoring function* [*equation* (2.20)]. *The data-generating process is given at* (4) *with a moving estimation window of* 500 *and the out-of-sample size to evaluate average scores of* 2000.

tests (simply because there is even a calibrated forecast based on $\mathcal{A}_{t-1,1}$). These tests are completely agnostic about the amount of information we use (as long as we use this information correctly). This is in stark contrast to comparative backtests where we get a lower score the more information we incorporate (correctly) [Holzmann and Eulert (2014)].

Examples of additional information are intra-daily data or high-frequency information. Models using such information are given in Shephard and Sheppard (2010). Bee, Dupuis and Trapin (2016) combine such models with EVT for estimation of tail-risk measures. They assess the quality of the models using traditional backtests, while Ziegel et al. (2017) provide a comparative analysis of a model for (VaR, ES) using intra-daily data versus a GARCH model using daily data suggesting superiority of the more informative model.

Backtesting as indication of future performance and choice of a scoring function. Kratz raises the issue of using results of "scoring," which is one ingredient of comparative backtesting, as an indicator of future forecasting performance. We note that scoring and backtesting are done on a test set or out of sample.[2] The choice of a test set is indeed important and in general should be sufficiently large so as to cover different "regimes" of a financial time series. It is true that scoring and corresponding rankings of forecasting methods may be unstable and change over time. This is the case when the test set size is small and the test set is not representative of a "typical" behavior of a given time series. Both traditional and comparative backtests will be misleading in such situations. Some illustrations of this phenomenon are provided in Section D of the Online Supplement on backtesting with a small out-of-sample size.

Concerning the choice of a "right" scoring function, as long as the scoring function is consistent for a given functional, the resulting comparative backtests will favor forecast accuracy; however, the choice of a scoring function does affect finite sample size properties of the underlying Diebold–Mariano tests, and so taking a scoring function with higher power for the Diebold–Mariano test will lead to conclusive assessments more often. Kratz points out cases where there are differences in ranking based on two different scoring functions. One example is the middle panel for expectiles in Table 3. The corresponding traffic light matrices in Figure 4 confirm that the pairwise differences between different methods are statistically insignificant, with the exception of the "n-FP" method which is significantly inferior in performance to all the other methods, and both scoring functions are in agreement. Overall, there is no inconsistency in results of comparative backtests

[2]From this perspective, there is an important difference in ranking performance of investment funds and scoring forecasting procedures. The former is usually done in-sample in that investment strategies are optimized on the same set on which they are evaluated. As scoring is done out of sample, it is more akin to the idea of cross-validation and, in fact, could be repeated on multiple test sets to check stability of results.

as well as rankings when statistical significance of score differences is taken into account.

For the specific choice of the scoring function, there is some evidence that more powerful tests are achieved when the homogeneity index is lower within the family of homogeneous generalized piecewise linear scoring functions for a single VaR [Agyeman (2017), Chapter 5.3]. This suggests the 0-homogeneous scoring function at (2.20) as a better alternative from the power perspective, although the classical choice at (2.19) is widely accepted. For the ES, similar to the VaR, consensus seems to emerge on the 0-homogeneous choice (2.24) [Dimitriadis and Bayer (2017), Patton, Ziegel and Chen (2017), Taylor (2017)].

Kratz also notes that "the scoring functions seem to be more sensitive to the estimation method than to the model." To elaborate on this point, recall that scoring and comparative backtesting assess forecasting procedures as a whole without a distinction between model specification and estimation. In fact, backtesting principles apply even if predictions come from an "expert" without reference to a model or past data, and so indeed numerical illustrations in the paper and in the rejoinder allow to judge the interplay between model specification (such as model dynamics and assumptions on the innovation distribution) and the method for estimating the tail [EVT versus (filtered) historical simulation versus a fully parametric treatment]. This, for example, reveals how semiparametric and nonparametric methods have a better ability to cope with partial model misspecification than their fully parametric counterparts.

Finally, we agree with Kratz that the choice of test functions in conditional calibration tests is an open problem that requires further investigation.

Some more technical aspects. We are grateful to Mark Davis for elaborating on the relation of our work to his 2016 manuscript and for bringing in some of the technical points related to the conditions required to obtain asymptotics for test statistic T_1. (He points out that the conditions could be debated in the context of financial time series with regard to the degree of stationarity they require.)

Zhou mentions in his Introduction that it is somewhat contradictory that, on the one hand, "For $k = 1$, identifiability implies elicitability under some additional assumptions," whereas, on the other hand, Acerbi and Szekely (2014) argue that only identifiability is of concern for traditional backtests, and therefore ES can be backtested despite not being elicitable. We are happy to be given a chance to clarify that both statements are correct without causing any contradiction. ES is 2-identifiable jointly with VaR, and 2-identification functions for (VaR,ES) are used in many common backtests [Acerbi and Szekely (2014), McNeil and Frey (2000)]. However, ES is neither 1-identifiable nor 1-elicitable (with respect to reasonably large classes of distributions). It is currently unclear whether higher order identifiability implies elicitability (under some additional assumptions).

In his Section 3, Zhou hints at the possibility that the identification function can be chosen time-varying (as long as it is predictable). This is certainly the case, and is equally true for the consistent scoring function for comparative backtests. We did not follow this route in our paper as we felt that it introduces additional degrees of freedom without a clear benefit.

REFERENCES

ACERBI, C. and SZEKELY, B. (2014). Backtesting expected shortfall. *Risk Mag*. **December** 76–81.

AGYEMAN, J. (2017). On the choice of scoring functions for forecast comparisons. Master's thesis, Univ. Bristish Columbia. Available at https://open.library.ubc.ca/cIRcle/collections/ubctheses/24/items/1.0344014.

BEE, M., DUPUIS, D. J. and TRAPIN, L. (2016). Realizing the extremes: Estimation of tail-risk measures from a high-frequency perspective. *J. Empir. Finance* **36** 86–99.

DAVIS, M. H. A. (2016). Verification of internal risk measure estimates. *Stat. Risk Model.* **33** 67–93. MR3574946

DIEBOLD, F. X., GUNTHER, T. A. and TAY, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *Internat. Econom. Rev.* **39** 863–883.

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 253–263.

DIMITRIADIS, T. and BAYER, S. (2017). A joint quantile and expected shortfall regression framework. Preprint, arXiv:1704.02213.

ENGLE, R. F. and MANGANELLI, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *J. Bus. Econom. Statist.* **22** 367–381. MR2091566

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 243–268. MR2325275

GNEITING, T. and KATZFUSS, M. (2014). Probabilistic forecasting. *Ann. Rev. Stat. Appl.* **1** 125–151.

GNEITING, T. and RANJAN, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econom. Statist.* **29** 411–422. MR2848512

GNEITING, T. and RANJAN, R. (2013). Combining predictive distributions. *Electron. J. Stat.* **7** 1747–1782. MR3080409

GORDY, M. B., LOK, H. Y. and MCNEIL, A. J. (2017). Spectral backtests of forecast distributions with applications to risk management. Preprint, arXiv:1708.01489.

HOLZMANN, H. and EULERT, M. (2014). The role of the information set for forecasting—with applications to risk management. *Ann. Appl. Stat.* **8** 595–621. MR3192004

HOLZMANN, H. and KLAR, B. (2016). Weighted scoring rules and hypothesis testing. Preprint, arXiv:1611.07345.

KRATZ, M., LOK, Y. and MCNEIL, A. (2016). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. Preprint, arXiv:1611.04851.

MCNEIL, A. J. and FREY, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *J. Empir. Finance* **7** 271–300.

PATTON, A., ZIEGEL, J. F. and CHEN, R. (2017). Dynamic semiparametric models for expected shortfall (and value-at-risk). Preprint, arXiv:1707.05108.

SHEPHARD, N. and SHEPPARD, K. (2010). Realising the future: Forecasting with high-frequency-based volatility (heavy) models. *J. Appl. Econometrics* **25** 197–231. MR2758633

STRÄHL, C. and ZIEGEL, J. (2017). Cross-calibration of probabilistic forecasts. *Electron. J. Stat.* **11** 608–639. MR3619318

TAYLOR, J. W. (2017). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *J. Bus. Econom. Statist*. To appear.

ZIEGEL, J. F., KRÜGER, F., JORDAN, A. and FASCIATI, F. (2017). Murphy diagrams for evaluating forecasts of expected shortfall. Preprint, arXiv:1705.04537.

DEPARTMENT OF STATISTICS
UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, BRITISH COLUMBIA V6T 1Z4
CANADA
E-MAIL: natalia@stat.ubc.ca

INSTITUTE OF MATHEMATICAL STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF BERN
CH-3012 BERN
SWITZERLAND
E-MAIL: johanna.ziegel@stat.unibe.ch