The Annals of Applied Statistics
2017, Vol. 11, No. 4, 1894–1900
https://doi.org/10.1214/17-AOAS1041E
Main article: https://doi.org/10.1214/17-AOAS1041

DISCUSSION OF "ELICITABILITY AND BACKTESTING: PERSPECTIVES FOR BANKING REGULATION"

BY MARIE KRATZ

ESSEC Business School, CREAR Risk Research Center

The discussion focuses on four points in the context of Basel 3. The first concerns the choice of test functions in the calibration tests. Then we discuss the interpretation of the "internal model," as well as the choice of risk measures. Last, we consider the score difference stationarity, an important issue in practice.

1. Context. Since the seminal paper of Artzner et al. (1999), there has been a large literature on risk measures, their properties, and their impact on practice and regulation. In the paper by Nolde and Ziegel (2017), the authors review the specific properties of elicitability and backtestability of risk measures and their impacts in view of model validation and banking regulation.

The two most popular risk measures used in financial institutions and regulation are Value at Risk (VaR) and Expected Shortfall (ES). VaR has been the dominant risk measure in banking regulation, although it is not a coherent risk measure for all distributions, unlike ES [see Acerbi and Tasche (2002), Tasche (2002)]. For the VaR, a direct backtest is possible and the most popular procedure is based on the so-called violation process. In practice, it consists of replacing VaR at the $\alpha\%$ level by its estimates and checking that this process behaves like independent and identically distributed Bernoulli random variables with violation (success) probability close to $1 - \alpha$. If the proportion of VaR violations is not significantly different from $1 - \alpha$, the estimation/prediction method is concluded to be reasonable [see, e.g., Christoffersen (2003), Christoffersen and Pelletier (2004) to test the independence assumption, Campbell (2006) for a review on VaR backtesting procedures, and Emmer, Kratz and Tasche (2015) for further references on the topic].

As a lesson of the last financial crisis, and notwithstanding the absence of an obvious direct backtest method for ES, it has been discussed and then decided that a 10-day ES at the 97.5% level would replace the daily 99% VaR under Basel 3 [see Basel Committee on Banking Supervision (2016)], ES being a more tail-sensitive measure of risk, and hence being more adequate for assessing extreme risks.

This decision caused some debate among all actors: academics, professionals at banks, and regulators, especially in terms of backtesting. First Gneiting (2011) raised an issue with direct backtesting of ES estimates because ES is not elicitable,

Received May 2017; revised June 2017.
Key words and phrases. Regulation, risk measure, scoring function.

unlike VaR, an optimal point forecast elicited by the weighted absolute error scoring function. Nevertheless, if not elicitable, then ES has been proved to be conditionally elicitable [Emmer, Kratz and Tasche (2015)] (which corresponds to the step procedure often done in practice) as well as jointly elicitable with VaR [Fissler and Ziegel (2016)]. Then Acerbi and Székely (2014) pointed out that elicitability (or lack of elicitability) is rather relevant for comparing the forecast performance of different estimation methods. Currently, a consensus has emerged that the problem of comparing the forecasting performance of different estimation methods (which requires the property of "elicitability") is distinct from the problem of checking whether a single series of *ex ante* estimates of a risk measure is consistent with a series of *ex post* realizations of profit and loss (*i.e.*, "backtestability"). Moreover, there are reasonable approaches to backtest ES, some of them being developed recently to answer the concerns of banks; the simplest one, based on the violation process, suggests an implicit backtest for ES via simultaneous backtests of VaR at four levels [see Kratz, Lok and McNeil (2016a, 2016b)].

In Nolde and Ziegel (2017), the authors suggest to use coherent and elicitable risk measures in order to adopt a two-stage framework. The first one would correspond to the current practice. If passing the first stage, then the second one would allow for a comparative backtest between standard and internal models.

2. The study. The paper is developed in two parts. The first one suggests a theoretical framework to revisit the state of the art in terms of risk measures and the concepts of elicitability, identifiability and dominance (conditional and on average), before proposing comparative backtests that would complement (or supplement?) traditional backtests when using elicitable risk measures. The second part illustrates partly those tests on simulated and real data. Three elicitable risk measures are considered, VaR, the pair (VaR, ES) and the expectile.

I will come back to the main issues, discussing their advantages and disadvantages, in view of both the theory and the respective applied results.

Calibration tests. In view of regulation, a one-sided conditional super-calibration test is proposed component-wise (when considering a multivariate risk measure) to simplify the problem. For the joint risk measure (VaR, ES), choosing a component-wise test implies the equivalence between a conditional sub-calibration test for (VaR, ES) and testing that the conditional VaR, ES, respectively, is at least as large as its optimal conditional prediction, which might appear a bit counter-intuitive. It illustrates how relations between super (or sub) calibration and over (or under) estimation depend on the identification functions (and the component-wise condition). It would be interesting to see if using the concept of conditionally elicitable risk measures [Emmer, Kratz and Tasche (2015)] instead might lead to a more direct relation.

For the VaR risk measure, the conditional calibration test for $\mathbf{h}_t = 1$ is close to the standard backtest for VaR based on the violation process, and does not require for one-step ahead forecasts an asymptotic test (under adequate conditions) as provided; the exact Rosenblatt–Diebold–Davis test gives indeed the result, moreover, with no condition. It is for multi-steps ahead forecasts that the proposed asymptotic test might be useful. Note that regulators do not ask for such forecasts, except for a qualitative evaluation of the risk evolution over the next three years.

In Table 2, we observe that H_0 is rejected more often with the two-sided test than with the one-sided [except for the (VaR, ES) at low threshold], as expected. General Conditional Calibration Tests (CCTs) are more conservative than simple ones. For VaR and Expectile, general tests are able to better discriminate between models than simple ones when considering the first two thresholds; for the Basel reference threshold, such an observation is more contestable as the general test might reject H_0 when it should not, especially when using the expectile. It might be the problem of using this specific risk measure at such a high threshold. For the bidimensional risk measure (VaR, ES), the decision of rejecting H_0 does not really depend on whether the test is simple or general. In Table 3, when using real data (Nasdaq), we observe similar results as those obtained with simulation data: the choice of the distribution (likelihood function) seems to have a lesser impact than the choice of the method at the second stage, which is not what we would like; again, the expectile does not give reliable results with the general CCT; for (VaR, ES), the general CCT rejects all methods using a normal likelihood, even with the EVT method.

In view of the obtained results, the main question remains *how to choose the predictable test functions* (using the available information at $t-1$) for the conditionally calibrated test to be effective on finite samples. Right now, the test functions have been deduced from a parallel made with existing tests on VaR (from Diebold–Davis) or on ES (from McNeil–Frey). It is always interesting to have a larger theoretical frame to cover specific cases, as suggested in this paper, but it still deserves further study to see if one can go beyond the existing, and especially improve current practice, in particular for regulation.

An alternative way for model validation. If looking at the size of a traditional test is necessary, then looking at its power is also crucial, especially for regulators, as emphasized in Kratz, Lok and McNeil (2016b). But it is true that the alternative may be quite broad. Hence it is a very good idea to look for an alternative way to compute the power, introducing as here a comparative backtest on the prediction between two models with two possible null hypotheses, permuting the role of each model. The authors develop theoretically this backtest having in view a comparison between the standard and internal models, but, when applying it numerically, come back to a comparison between classical trading book models calibrated with standard estimation methods, as done when computing the power for traditional tests considering various possible models for the alternatives [see, e.g., Costanzino and Curran (2015 or 2016) when looking at probabilistic forecasts for the tail, and, in the case of point forecasts in the tail, Kratz, Lok and McNeil (2016a or 2016b) where the same types of experiments as in this paper have been performed], and so the applications might have been chosen to better highlight the supplement of traditional backtests (defined with specific alternatives) with a comparative backtest.

Note also that considering the average $\Delta_n \bar{S}$ over time ($t = 1, \dots, n$) for comparative backtests reformulated according to Diebold and Mariano (1995), might raise an issue in practice, if the score differences are not first order stationary! We will come back later to this issue.

Standard and internal models. The standard model in a bank considers three types of risks, independently of each other: the credit risk, the market risk and the operational risk. When dealing with market risk, standard and internal models are based on the same principle, evaluating the VaR at 99% in Basel 2 and the ES 97.5% in Basel 3. It is essentially for the two other risks, credit and operational, that standard and internal models might differ, as the techniques used to measure them are usually more sophisticated with internal models than with standard ones.

Here the study is made on market risk (taking into account the daily assets), for which no real distinction is made between internal and standard models, and so it appears a bit artificial, if not misleading, to use this terminology in the various figures where only estimation methods of the innovation distribution are compared. Other types of tests could answer the concern when computing the power, with no need for the risk measure to be elicitable or for the data to satisfy some conditions.

Nevertheless, I find it very interesting to propose a comparative backtest between standard and internal models, as addressed theoretically in this paper when using elicitable risk measures. It may be seen as a first step of an alternative way for regulators and for the banks to check the relevance of their internal model; it certainly deserves much more investigation to make it a useful tool.

Scoring. In the numerical study, ranking between the methods via the average scores is given in Tables 1 (simulated data) and 3 (Nasdaq example). We observe that the ranking depends on the choice of the scoring function, with generally more stability for the VaR than for the two other risk measures. In Table 3, judging on the VaR and (VaR, ES), the scoring functions seem to be more sensitive to the estimation method than to the model.

Besides the problem of finding the "right" scoring function, the main general question concerns their stability with time (stationarity). Considering the mean over t instead of comparing $E[S(R_t) - S(R_t^*)]$ for each t might raise some issues in practice and weaken a decision tool. The question has been passed over when speaking about S-dominance on average and the hypothesis of the score differences being first order stationary.

For the scoring to be useful, it needs to be a good predictor of the future ranking/scoring. Otherwise it could be really misleading and would even induce large errors relying on an unstable decisional tool. To illustrate this question, we could draw the parallel, for instance, with investment funds. Ranking has been made in terms of the funds performance (not in terms of risk measures), but it has been observed that the ranking is not at all a good predictor for the future ranking. The fund exhibiting the best performance one year will not be ranked number one the following year.

Choice of risk measure. Besides the two "regulatory" risk measures, VaR and ES, the authors consider another risk measure, the expectile. It is a coherent and elicitable risk measure [see Bellini et al. (2014), Ziegel (2016), Chen (2013)] but not comonotonically additive [see Emmer, Kratz and Tasche (2015)]. Expectiles, besides appearing less attractive for practitioners because of a less intuitive underlying concept than the concepts for VaR or ES, clearly show limitations in this study for their use in practice, specifically from a regulatory perspective. There is no explicit estimator of the expectile, whatever is the estimation method (Full Parametric, Filtered Historical or EVT). Moreover, to have a comparable magnitude of risk as VaR at 99% or ES at 97.5%, expectiles require looking at the 99.855% level, which might make statistical studies quite nonrobust, as observed here. In Table 2, if the expectile presents the best results at the lowest threshold (96.561%) when using a two-sided general test, the Student distribution is more often rejected than the normal one at level 98.761% with this test, and, when looking at the "Basel reference threshold," it is worse. The true model is rejected when using a two-sided general test for the FP and FHS methods but not rejected with the EVT one. We could argue it makes sense at such a high level, but we observe the reverse phenomenon when using the misspecified likelihood, and so we cannot conclude to an over-sensitivity toward estimation methods rather than toward likelihoods at this high level. For the comparative backtest given in Table 1, the expectile behaves more or less as the other risk measures, exhibiting some instability depending on the choice of the scoring function. When considering the highest level, we observe issues if one uses the FHS method for the expectile with one given scoring function and for the VaR with both scoring functions. In the example on real data, the expectile gives poor results, with no help to take any decision most of the time. Figures 1 to 4 provide a visible way to judge risk measures in terms of predictability. All figures confirm that expectile is not of great help for decision making. Various reasons might explain this, as, for instance, the choice of the test function, but this does not favor the use of this risk measure in practice.

3. Conclusion. The authors have put forward the right theoretical questions and have come up with very interesting alternative backtests to compare models and methods. It is quite promising, even if the numerical study performed on simulated data and a real data example shows that there is still a long way to propose a practical tool. The authors themselves point that out.

For conditional calibration tests, the choice of the test functions is a bit tricky. Simulations and applications show that it still needs to be investigated to be effective in practice. I am in favor of rigorous but simple methods for practitioners, with not too many tuning parameters, in order to avoid most of the errors coming from using results relying on conditions not always easy to check out.

For the scoring functions as well, we observe some variability; further investigation is needed. But for the scoring to be useful, it has to be a good predictor of the future ranking/scoring. This is definitively the first problem to be looked at.

I would like to end this discussion with a general comment on how to push for the right incentive providing the right tools. The main issue is to optimize the capital, finding the right and fair amount for all actors: companies, shareholders, regulators, society at large. A company underestimating or overestimating her risks means that she does not manage them well, not balancing the opposite goals of the various actors. It is true that the role of regulators is to check that a company did not underestimate her risks, not to put in danger her clients and also, as a consequence, society (that has been often asked to pay for the errors made by banks). I believe their incentive is as well to push companies to evaluate their risks accurately to foster good risk management. The new regulatory rules are going in this direction, encouraging companies to develop their internal models for better understanding their risks. The primary goal of regulation is clearly to protect consumers, but it is also important for regulation to favor best practices and encourage the development of a healthy industry. Hence I am not convinced that checking conditional super-calibration (or sub-calibration) is key to risk management. If a one-sided test might look at first glance reasonable from a regulatory point of view, then a two-sided test is a useful test for companies, even if it has not been requested by regulators yet.

Acknowledgments. The author thanks an anonymous referee and the Editor for their careful review and insights.

REFERENCES

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

ACERBI, C. and TASCHE, D. (2002). On the coherence of expected shortfall. *J. Bank. Financ*. **26** 1487–1503.

ARTZNER, P., DELBAEN, F., EBER, J.-M. and HEATH, D. (1999). Coherent measures of risk. *Math. Finance* **9** 203–228. MR1850791

BASEL COMMITTEE ON BANKING SUPERVISION (2016). Minimum capital requirements for market risk. Bank of International Settlements.

BELLINI, F., KLAR, B., MÜLLER, A. and GIANIN, E. R. (2014). Generalized quantiles as risk measures. *Insurance Math. Econom*. **54** 41–48. MR3145849

CAMPBELL, S. D. (2006). A review of backtesting and backtesting procedures. *J. Risk* **9** 1–17.

CHEN, J. M. (2013). Measuring market risk under Basel II, 2.5, and III: VAR, stressed VAR, and expected shortfall. Available at SSRN: http://ssrn.com/abstract=2252463.

CHRISTOFFERSEN, P. (2003). *Elements of Financial Risk Management*. Academic Press, New York.

CHRISTOFFERSEN, P. and PELLETIER, D. (2004). Backtesting Value-at-Risk: A duration-based approach. *J. Financ. Econom*. **2** 84–108.

COSTANZINO, N. and CURRAN, M. (2015). Backtesting general spectral risk measures with application to expected shortfall. *J. Risk Model Validation* **9** 21–33.

COSTANZINO, N. and CURRAN, M. (2016). A simple traffic light approach to backtesting expected shortfall. Working paper.
DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 253–263.
EMMER, S., KRATZ, M. and TASCHE, D. (2015). What is the best risk measure in practice? A comparison of standard risk measures. *J. Risk* **18** 31–60.
FISSLER, T. and ZIEGEL, J. F. (2016). Higher order elicitability and Osband's principle. *Ann. Statist.* **44** 1680–1707. MR3519937
GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988
KRATZ, M., LOK, Y. and MCNEIL, A. (2016a). A multinomial test to discriminate between models. In *ASTIN 2016 Proceedings*. Available online.
KRATZ, M., LOK, Y. and MCNEIL, A. (2016b). Multinomial VaR Backtests: A simple implicit approach to backtesting expected shortfall. ESSEC Working Paper 1617, arXiv1611.04851v1.
NOLDE, N. and ZIEGEL, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Ann. Appl. Stat.* **11** 1833–1874.
TASCHE, D. (2002). Expected shortfall and beyond. In *Statistical Data Analysis Based on the L_1-Norm and Related Methods* (*Neuchâtel*, 2002) 109–123. Birkhäuser, Basel. MR2001308
ZIEGEL, J. F. (2016). Coherence and elicitability. *Math. Finance* **26** 901–918. MR3551510

ESSEC BUSINESS SCHOOL
CREAR RISK RESEARCH CENTER
AVENUE BERNARD HIRSCH BP 50105
CERGY-PONTOISE, 95021 CEDEX
FRANCE
E-MAIL: kratz@essec.edu