The Annals of Applied Statistics
2017, Vol. 11, No. 4, 1888–1893
https://doi.org/10.1214/17-AOAS1041D
Main article: https://doi.org/10.1214/17-AOAS1041

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

BY CHEN ZHOU

De Nederlandsche Bank and Erasmus University Rotterdam

1. Introduction. Nolde and Ziegel (2017) (NZ throughout) aim at evaluating the performance of risk forecasts. First, NZ focused on the traditional backtest, that is, backtesting whether a series of reported risk forecasts, usually obtained from one risk model, are valid. Second, NZ proposed the comparative test, that is, to compare the performance of two series of risk forecasts obtained from two different models. The main ideas behind constructing the traditional backtest and the comparative test are the concepts *identifiability* and *elicitability*, respectively.

The general perception that elicitability is equivalent to backtestability generated a serious concern for the regulators in practice. I have personally been consulted by regulatory policymakers about whether the nonelicitable expected shortfall (ES) would cause a problem for backtesting. Fortunately, Acerbi and Szekely (2014) calmed down such a concern by demonstrating that ES can be backtested. They claimed that elicitability is almost irrelevant for backtesting or, more precisely, model validation. Instead, elicitability is only relevant for model selection. NZ followed exactly this line of argument to construct the comparative test based on elicitability.

This discussion will, however, focus on the first issue: traditional backtest and identifiability. As stated in Section 2 in NZ, *"In fact, for* $k = 1$*, identifiability implies elicitability under some additional assumptions."* This means, for a single risk measure, if one intends to establish a traditional backtest as in NZ, the risk measure must be identifiable and consequently elicitable. Ironically, this brings back the elicitability concern, which somehow contradicts the statement in Acerbi and Szekely (2014).

This discussion aims to reconcile such a debate and fairly evaluate the role of identifiability in the traditional backtest. In Section 2, I will start from a regulator's perspective and discuss how to define "traditional backtestability." Then I will argue that identifiability is, to a certain extent, necessary for backtestability if no common property across the conditional distributions of future losses is assumed. However, with assuming some common properties across the conditional distributions of future losses, identifiability is not a necessary condition for a risk measure to be backtested. This will be discussed in Section 3. Section 4 concludes this discussion.

Received May 2017; revised May 2017.

2. Identifiability is necessary for backtestability. Let us start with considering only one risk measure $\Theta = \Theta(F)$ for distribution functions F in a proper class. At each time $t-1$, the bank must report a risk forecast for the next period, denoted as R_t. At time t, the loss is realized as X_t. With collecting the observations (R_t, X_t) at $t = 1, 2, \ldots, n$, the regulator intends to *backtest* whether the banks' reported risk forecasts are valid. The null hypothesis that the risk forecasts are conditionally valid can be written as

$$H_0 : R_t = \Theta(F_t) \qquad \text{for all } t = 1, 2, \ldots, n,$$

where F_t is the "conditional" distribution of X_t, conditioning on the information set $\mathcal{F}_{t-1}$.

A traditional backtest is to establish a test statistic $T(R_1, \ldots, R_n; X_1, \ldots, X_n)$ such that the (asymptotic) distribution of T can be derived under the null hypothesis. Hence "backtestability" refers to the possibility of finding such a test statistic and the corresponding asymptotic result.

Next, we recall the definition of identifiability. Under the null hypothesis, an identification function satisfies $\mathbb{E}(V(R_t, X_t)|\mathcal{F}_{t-1}) = 0$ for all t. Consequently, a traditional backtest based on identifiability is defined via the test statistic T_1 in (2.11) in NZ. In essence, $T_1 = T'(V_1, V_2, \ldots, V_n)$, where $V_t = V(R_t, X_t)$.

At this point, we already observe the difference between identifiability and backtestability. The definition of backtestability is broader: the regulator can make use of the information set $\{(R_t, X_t)\}_{t=1}^n$ in any form possible. By contrast, the backtest based on identifiability requires using the information set $\{(R_t, X_t)\}_{t=1}^n$ in a specific way; that is, R_t is only "compared" to X_t via the identification function V, not to any other observations at other time periods.

I would further argue that identifiability is, to a certain extent, necessary for backtestability if we assume no common knowledge across the conditional distributions at a different time point t. If all conditional distributions $\{F_t\}_{t=1}^n$ are different and there is no common property across them, it is not possible to use all observed losses $\{X_t\}_{t=1}^n$ to infer any information regarding a specific F_t. The only possible way to check the risk forecast R_t, is to compare it with X_t. Such a comparison is conducted by using an error (or utility) function V. In addition, it is not possible to adjust the error function V in each time period t to accommodate each specific F_t because no information regarding F_t is known. Consequently, the only sensible way to backtest all risk forecasts $\{R_t\}_{t=1}^n$ is to use a unified error function V and the observations $\{V(R_t, X_t)\}_{t=1}^n$. Testing the mean of this series is the most straightforward choice. Therefore, assuming identifiability is necessary for performing *any* traditional backtest in this case.

Here are a few practical examples that justify this argument. Example 1 in NZ shows the traditional backtest for Value at Risk (VaR). The test assumes no additional knowledge regarding the conditional distributions across time periods. Similarly, the backtesting procedures proposed in Acerbi and Szekely (2014) for jointly

backtesting the VaR and ES do not make such an assumption either. In all these examples, the backtested risk measure(s) are in fact identifiable. Even for the Acerbi and Szekely (2014) result, VaR and ES are jointly backtested and they are also jointly identifiable.

In the next section, I will discuss that identifiability is not necessary if some common properties across the conditional distributions in different time periods are assumed.

3. Identifiability is not necessary for backtestability. There are backtesting frameworks that assume common properties across the conditional distributions in different time periods, for example, example 2 in NZ [similar to the ES backtest in McNeil and Frey (2000)]. To jointly test the reported VaR and ES at the probability level v, the test statistic in (2.14) is approximately given as

$$T_4 = \frac{1}{n}\sum_{t=1}^{n} V_t(R_t, X_t),$$

where $R_t = (r_{1,t}, r_{2,t})$, $V_t(r_{1,t}, r_{2,t}, X_t) = \frac{1}{\sigma_t}\frac{1}{1-v}(X_t - r_{2,t})1_{\{X_t > r_{1,t}\}}$. Notice that V_t is varying over time. Therefore, this is not a test motivated by an identification function. By rewriting T_4 as

$$T_4 = \frac{1}{n}\sum_{t=1}^{n} V\left(\frac{r_{1,t}-\mu_t}{\sigma_t}, \frac{r_{2,t}-\mu_t}{\sigma_t}, \frac{X_t-\mu_t}{\sigma_t}\right),$$

where $V(r_1, r_2, z) = \frac{1}{1-v}(z - r_2)1_{\{z > r_1\}}$, the test can be regarded as jointly testing the VaR and ES of $\{Z_t := (X_t - \mu_t)/\sigma_t\}_{t=1}^n$, where the VaR and ES are correspondingly standardized using the predicted shift and scale μ_t and σ_t at each time point t. Here, the identification function V is the same across all time periods. The test is therefore based on the joint identification property of VaR and ES. Example 3 in NZ, though designed for expectiles, can be interpreted in a similar way.

In these examples, common properties regarding the standardized observations $\{Z_t\}_{t=1}^n$ are assumed. In the most convenient form, they are assumed to be i.i.d., though NZ commented that some weaker models might be considered. Such assumptions are essentially assuming common properties across the conditional distributions $\{F_t\}_{t=1}^n$. Since μ_t and σ_t are regarded as $\mathcal{F}_{t-1}$-measurable functions, with assuming stationarity in $\{Z_t\}_{t=1}^n$, all conditional distributions $\{F_t\}_{t=1}^n$ are only subject to a scale and shift difference.

Assuming that $\{Z_t\}_{t=1}^n$ are i.i.d., there is no need to require identifiability of a risk measure for the traditional backtest. The general intuition is that one can use all observations to infer the common distribution, and consequently test any risk measure defined as a functional of the distribution.

As an example, consider the nonidentifiable risk measure, ES. The ES at probability level v is backtestable *solely* when assuming that the underlying data-generating process follows a time series model such as the AR(1)–GARCH(1, 1) model.

Suppose the true data-generating process is given as $X_t = \mu_t + \sigma_t Z_t$, where $\{Z_t\}_{t=1}^n$ are i.i.d. random variables with a common distribution function G. Assume that the conditional shift and scale μ_t and σ_t can be predicted by $\hat{\mu}_t$ and $\hat{\sigma}_t$, respectively, where the two predictors are $\mathcal{F}_{t-1}$-measurable. In addition, assume that, as $n \to \infty$,

$$\sqrt{n} \sup_{1\le t\le n} (\hat{\mu}_t - \mu_t) = o_P(1) \quad \text{and} \quad \sqrt{n} \sup_{1\le t\le n} \left(\frac{\hat{\sigma}_t}{\sigma_t} - 1\right) = o_P(1). \tag{3.1}$$

Define $\hat{Z}_t = \frac{X_t - \hat{\mu}_t}{\hat{\sigma}_t}$ for $1 \le t \le n$. By ranking all $\hat{Z}_t$, we obtain $\hat{Z}_{1,n} \le \cdots \le \hat{Z}_{n,n}$. Denote $\hat{q} = \hat{Z}_{\lceil \nu n \rceil, n}$ as an estimator of the $\nu-$quantile of $\hat{Z}_t$. Finally, define

$$T = \frac{1}{\hat{s}} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(\frac{R_t - \hat{\mu}_t}{\hat{\sigma}_t} - \frac{1}{1-\nu} \hat{Z}_t 1_{\hat{Z}_t > \hat{q}} \right),$$

where

$$\hat{s} = \frac{1}{1-\nu} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(\hat{Z}_t - \frac{1}{n(1-\nu)} \sum_{s=1}^{n} \hat{Z}_s 1_{\hat{Z}_s > \hat{q}} \right)^2 1_{Z_t > \hat{q}}}.$$

Then under standard regularity conditions regarding the distribution function G and the null hypothesis H_0, the statistic T is asymptotically standard normal. The regularity conditions and the proof for this result are similar to those for the asymptotic normality of the nonparametric estimator of the ES; see, for example, Zwingmann and Holzmann (2016).[1]

I have two remarks on this test. First, notice that T is a function of $(R_1, \ldots, R_n)$ and $(X_1, \ldots, X_n)$, therefore, it is a traditional backtest. Although a nonparametric estimator of the VaR, $\hat{q}$, was used in the construction of the statistic T, it should be regarded as a VaR estimate produced by the regulator after obtaining the realizations of $X_1, \ldots, X_n$. It is not a VaR forecast produced by the bank at the time of reporting. In other words, the regulator did not backtest the VaR forecasts of the bank. Hence the test is a backtest for the ES *solely*, not a joint test for the VaR and ES.

Second, the assumptions regarding the model and the estimators are not very restrictive. Regarding the model assumption, it essentially assumes that all conditional distributions are only subject to a scale and shift difference. Most time series models satisfy this assumption, such as the AR(1)–GARCH(1, 1) model used in NZ. In addition, the essence of this assumption is that there are some common

[1]The essential construction of the test statistic is based on the nonparametric estimator of the ES using the observations $\{\hat{Z}_t\}_{t=1}^n$. Nevertheless, $\{\hat{Z}_t\}_{t=1}^n$ might not be i.i.d. observations drawn from the distribution G. Consequently, the proof relies on establishing the asymptotic property of the empirical process based on $\{\hat{Z}_t\}_{t=1}^n$. Details are available upon request.

properties across the conditional distributions of the future losses. Practically, this assumes that the regulator knows the essential commonality in the data-generating process. This assumption is subject to critiques regarding model uncertainty.

Regarding the condition (3.1), it assumes that the speed of convergence for the predictors are faster than $1/\sqrt{n}$ uniformly. This is not unrealistic in practice. Notice that here n is the time horizon used for the backtest, while the predictors are usually estimated using historical data with a longer horizon, say m observations. Typical estimators such as the maximum likelihood approach will guarantee that the speed of convergence for the predictors is $1/\sqrt{m}$. By assuming that $n/m \to 0$ as $n \to \infty$, the condition (3.1) is valid. In practice, for example, in the Basel traffic light system, n is set to the number of daily observations in one year, 250. In contrast, the estimation of time series models typically uses daily observations from at least a five-year horizon, that is, n/m is less than $1/5$. For example, the estimation horizon in Engle (2001) is set to ten years, while the estimation horizon in McNeil and Frey (2000) goes even beyond 12 years. In addition, Angelidis, Benos and Degiannakis (2004) show that the performance of the GARCH model in risk analysis is unsatisfactory with less than 1000 observations, that is, four years of daily observations.

To summarize, this example shows that, when assuming some common properties across the conditional distributions $\{F_t\}_{t=1}^n$, it is possible to backtest some risk measures that are not identifiable. In other words, identifiability is not necessary for backtestability.

4. Conclusion. The work by NZ associates identifiability and elicitability to the traditional backtest and the comparative test, respectively. While the latter association reflects the discussion in Acerbi and Szekely (2014), the former association may be subject to different model assumptions. Without making any further model assumption, the association is valid in the sense that identifiability is to a certain extent necessary for backtestability. If the regulator is willing to make model assumptions for the underlying data-generating process, the necessity breaks down.

Alternatively, this discussion can be viewed as a discussion on the robustness of backtesting methods. Recall that we assumed all conditional distributions $\{F_t\}_{t=1}^n$ are from a proper class. If the class is broad, that is, assuming no common knowledge, to establish a robust backtest, it is necessary that the risk measure being tested is identifiable. However, if the class is restrictive, that is, assuming some common knowledge regarding $\{F_t\}_{t=1}^n$, identifiability can be waived. In all, this is a trade-off between making assumptions on the data-generating process versus assumptions on the risk measure.

In practice, making assumptions on the data-generating process might be difficult due to model uncertainty. In contrast, assumptions on the risk measure can be studied ex ante in a theoretical way. This is exactly where risk theorists such as Natalia Nolde and Johanna Ziegel can help. It is important to get practitioners

such as regulators to understand the conclusions drawn by risk theorists. For example, the work by NZ should not be read as "nonidentifiable risk measures are not backtestable." Instead, it should be read as "identifiable risk measures can be backtested robustly."

Acknowledgments. The author is grateful to an anonymous referee and the editor Tilmann Gneiting for their constructive suggestions.

Views expressed are my own and do not reflect the official positions of De Nederlandsche Bank.

REFERENCES

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

ANGELIDIS, T., BENOS, A. and DEGIANNAKIS, S. (2004). The use of GARCH models in VaR estimation. *Stat. Methodol*. **1** 105–128.

ENGLE, R. (2001). GARCH 101: The use of ARCH/GARCH models in applied econometrics. *J. Econ. Perspect*. **15** 157–168.

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.

NOLDE, N. and ZIEGEL, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Ann. Appl. Stat*. **11** 1833–1874.

ZWINGMANN, T. and HOLZMANN, H. (2016). Asymptotics for the expected shortfall. arXiv preprint, arXiv:1611.07222.

ECONOMICS AND RESEARCH DIVISION
DE NEDERLANDSCHE BANK
1000 AB AMSTERDAM
THE NETHERLANDS
E-MAIL: c.zhou@dnb.nl
zhou@ese.eur.nl