Backtesting for Risk-Based Regulatory Capital *

Jeroen Kerkhof† and Bertrand Melenberg‡

May 2003

ABSTRACT

In this paper we present a framework for backtesting all currently popular risk measurement methods for quantifying market risk (including value-at-risk and expected shortfall) using the functional delta method. Estimation risk can be taken explicitly into account. Based on a simulation study we provide evidence that tests for expected shortfall with acceptable low levels have a better performance than tests for value-at-risk in realistic financial sample sizes. We propose a way to determine multiplication factors, and find that the resulting regulatory capital scheme using expected shortfall compares favorably to the current Basel Accord backtesting scheme.

Keywords: Risk management, capital requirements, Basel II, multiplication factors, and model selection.

JEL codes: C12, G18

*We thank John Einmahl, Hans Schumacher, Bas Werker, and an anonymous referee for constructive and helpful comments. Any remaining errors are ours.

†Department of Econometrics and Operations Research, and CentER, Tilburg University, and Product Development Group, ABN◇AMRO Bank, Amsterdam. E-mail: F.L.J.Kerkhof@CentER.nl, Phone: +31-13-4662134, Fax: +31-13-4663280.

‡Corresponding author; Department of Econometrics and Operations Research, Department of Finance, and CentER, Tilburg University. E-mail: B.Melenberg@CentER.nl.

Backtesting for Risk-Based Regulatory Capital

ABSTRACT

In this paper we present a framework for backtesting all currently popular risk measurement methods for quantifying market risk (including value-at-risk and expected shortfall) using the functional delta method. Estimation risk can be taken explicitly into account. Based on a simulation study we provide evidence that tests for expected shortfall with acceptable low levels have a better performance than tests for value-at-risk in realistic financial sample sizes. We propose a way to determine multiplication factors, and find that the resulting regulatory capital scheme using expected shortfall compares favorably to the current Basel Accord backtesting scheme.

Keywords: Risk management, capital requirements, Basel II, multiplication factors, and model selection.

JEL codes: C12, G18

I. Introduction

Regulators face the important but difficult task of determining appropriate capital requirements for regulated banks. Such capital requirements should protect the banks against adverse market conditions and prevent them from taking extraordinary risks (where, in this paper, we focus on market risk). At the same time, regulators should not prevent banks from practicing one of their core businesses, namely trading risk. The crucial ingredients in the process of risk based capital requirement determination are the use of a risk measurement method (to quantify market risk), a backtesting procedure, and multiplication factors, based on the outcomes of the backtesting procedure. Regulators apply multiplication factors to the risk measurement method they use in order to determine the capital requirements. The multiplication factors depend on the backtesting results, where a bad performance of the risk measurement method results in a higher multiplication factor. Consequently, to guarantee an appropriate process of capital requirement determination, regulators need an accurate backtesting procedure, combined with a suitable way of determining multiplication factors. Based on these requirements the regulators will assign the risk measurement method.

Since its introduction in the 1996 amendment to the Basel Accord (see Basel Committee on Banking Supervision (1996a) and Basel Committee on Banking Supervision (1996b)) the value-at-risk has become the standard risk measurement method. However, although the value-at-risk may be interesting from a practical point of view, it has a serious drawback: it does not necessarily satisfy the property of subadditivity, which means that one can find examples where the value-at-risk of a portfolio as a whole is higher than that of the sum of the value-at-risks of its mutually exclusive sub-portfolios. An alternative, practically viable risk measurement method that satisfies the subadditivity property (and other desirable properties [1]) is the expected shortfall. Currently, a debate is going on whether the use of expected shortfall should be recommended in Basel II. So far, it is not in Basel II due to the expected difficulties concerning backtesting (see

[1] Namely, translation invariance, monotonicity, and positive homogeneity. These three properties are also satisfied by value-at-risk.

Yamai and Yoshiba (2002)). Thus, although the value-at-risk does not necessarily satisfy the subadditivity property, it is still assigned by regulators, because of its perceived superior performance in case of backtesting.

Both the value-at-risk and the expected shortfall (as well as many other risk measurement methods) are level-based methods, meaning that one first has to choose a level; given this level, the risk depends on the corresponding left-hand tail of the profit and loss distribution. For the value-at-risk the Basel Committee chooses a level of 0.01, meaning that the value-at-risk is based on the 1% quantile of the profit and loss distribution. For the sake of comparison, one might be tempted to choose the same level for alternative risk measurement methods, like the expected shortfall, so that they are calculated based on the same left-hand tail of the profit and loss distribution. When the level in both cases equals 0.01 it seems obvious to expect that backtesting expected shortfall will be much harder than backtesting the value-at-risk, even without trying it out. However, comparing alternative risk measurement methods by equating their levels does not seem to be appropriate from the viewpoint of capital reserve determination. From that perspective it seems much better to choose the levels such that the risk measurement methods result in (more or less) the same quantiles of the profit and loss distribution. The 0.01-level of value-at-risk will then correspond to a higher level in case of the expected shortfall. But then it is no longer clear which method will perform better in backtesting. It is the aim of this paper to make this comparison.

The contribution of the paper is threefold. First, we provide a general backtesting procedure for a large class of risk measurement methods, which contains all major risk measurement methods used nowadays. In particular, as a result a test for expected shortfall is derived which appears to be new in the literature. Using the functional delta method we provide a framework that requires the regulator only to determine the influence function of the risk measurement method in order to determine the critical levels of the capital requirements table. We show that the present backtesting methodology in the Basel Accord is a special case. Furthermore, a simple method to incorporate estimation risk is presented. The fact that banks have time-varying portfolio sizes and

risk exposures complicates the use of standard statistical techniques. We deal with this issue using a standardization procedure based on the probability integral transform also used by Diebold et al. (1998) and Berkowitz (2001). The key idea of the standardization procedure is that banks should not only report whether or not the realized profit/loss is beyond the value-at-risk, but also which quantile of the predicted profit and loss distribution is realized. Second, we establish, via simulation experiments, that backtests for expected shortfall have a more promising performance than for value-at-risk, when the comparison is based on (more or less) equal quantiles instead of equal levels. In this way we provide evidence for a viable risk based regulatory capital scheme using expected shortfall with good backtesting properties. Finally, we suggest a general method to determine multiplication factors for the risk measurement methods using the backtest procedure developed.

The setup of the paper is as follows. In Section II we review the most popular risk measurement methods in current quantitative risk management. In Section III we present the standardization procedure in order to take account of the time-varying portfolio sizes and risk exposures. Section IV treats the backtesting of the Basel Accord, its generalization using the functional delta method, and the incorporation of estimation risk. Simulation experiments are presented in Section V. In Section VI a suggestion for determination of multiplication factors is given. Finally, Section VII concludes.

II. Risk measurement methods

A. Definitions and notation

Though risk profiles contain much relevant information for risk managers, they become unmanageable for large firms with many divisions and portfolios. Therefore, for risk management purposes, risk managers prefer low dimensional characteristics of the risk profiles. In order to compute these low dimensional characteristics they use a financial model $m = (\Omega, \mathbb{P})$, where Ω denotes the states of the world, and $\mathbb{P}$ the postulated

probability distribution.[2] A risk is defined as follows.[3]

Definition 1 Let a financial model m be given. A *risk* defined on m belongs to $\mathcal{R}(m)$, the set of random variables defined on Ω.

This definition, in which a "risk" is a random variable, follows the terminology of Artzner et al. (1999) and Delbaen (2000). Artzner et al. (1999) defined a risk measure for a particular financial model.

Definition 2 Let a financial model m be given. A *risk measure*, ρ, defined on m is a map from $\mathcal{R}(m)$ to $I\!R \cup \{\infty\}$.[4]

In order to allow for several financial models, we use a class of financial models denoted by $\mathcal{M}$. Each of these models defines a set of risks $\mathcal{R}(m)$. Following Kerkhof et al. (2002) we denote a mapping defined on $\mathcal{M}$ that assigns a risk measure defined on m for each $m \in \mathcal{M}$ by a *risk measurement method defined on* $\mathcal{M}$, RMM. The most well-known risk measurement method nowadays is the value-at-risk method which was supported by the Basel Committee in the 1996 amendment to the Basel Accord (see Basel Committee on Banking Supervision (1996a)).

Before coming to the formal definitions of the popular risk measurement methods we present the quantile definitions.

Definition 3 (Quantiles) Let $X \in \mathcal{R}(m)$ be a risk for model $m = (\Omega, \mathbb{P})$.

1. $Q_p(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \leq x) \geq p\}$ is the lower p-quantile of X.
2. $Q^p(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \leq x) > p\}$ is the upper p-quantile of X.

The definition of the value-at-risk method can then be given by

[2] Formally, a model is defined by $m = (\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is the information available.

[3] Formally, $\mathcal{R}(m)$ is defined as the space of all equivalence classes of real-valued measurable functions on $(\Omega, \mathcal{F})$.

[4] Including ∞ allows risks to be defined on more general probability spaces, see Delbaen (2000).

Definition 4 The *value-at-risk* method with *reference asset* N and *level* $p \in (0,1)$ assigns to a model $m = (\Omega, \mathbb{P})$ the risk measure VaR_m^p given by

$$\mathrm{VaR}_m^p : \mathcal{R}(m) \ni X \mapsto -Q^p\left(X/N_m\right) = Q_{1-p}\left(-X/N_m\right) \in I\!R \cup \{\infty\}, \tag{1}$$

where N_m denotes the reference asset in model m.

We use a reference asset N (for example, the money market account) to measure the losses in terms of money lost relative to the reference asset. This allows comparison of risk measures for different time horizons.

Since the introduction of value-at-risk by RiskMetrics (1996), the literature on value-at-risk has surged (see, for example, Risk Magazine (1996), Duffie and Pan (1997), and Jorion (2000) for overviews). Though value-at-risk is an intuitive risk measure, the reasoning behind it was more practical than theoretically grounded. Recently, Artzner et al. (1997) introduced the notion of coherent risk measures having the properties of translation invariance, monotonicity, positive homogeneity, and subadditivity. Their ideas were formalized in Artzner et al. (1999) and Delbaen (2000), amongst others. The value-at-risk method does not necessarily satisfy the relevant subadditivity property. This means that we can find examples where the value-at-risk of a portfolio is higher than that of the sum of the value-at-risks of a set of mutually exclusive sub-portfolios (see, for example, Artzner et al. (1999), Acerbi and Tasche (2002), and Tasche (2002)). A practically usable coherent risk measure is the expected shortfall as given in Acerbi and Tasche (2002).

Definition 5 The *expected shortfall method* with *reference asset* N and *level* $p \in (0,1)$ assigns to a model $m = (\Omega, \mathbb{P})$ the risk measure ES_m given by

$$\begin{aligned}\mathrm{ES}_m : \mathcal{R}(m) \ni X \mapsto -\frac{1}{p}\big(&\mathbb{E}X\mathbf{I}_{(-\infty, Q_p(X/N_m)]} \\ &+ Q_p\left(X/N_m\right)\left(p - \mathbb{P}\left(X/N_m \le Q_p\left(X/N_m\right)\right)\right)\big) \in I\!R \cup \{\infty\}.\end{aligned} \tag{2}$$

In case that $p = \mathbb{P}\left(X/N_m \leq Q_p\left(X/N_m\right)\right)$, the expected shortfall equals[5]

$$\mathrm{ES}_m(X) = -\frac{1}{p}\mathbb{E}\left[X\mathbf{I}_{(-\infty,Q_p(X/N_m)]}\right] = \mathbb{E}\left[X \mid X \leq Q_p\left(X/N_m\right)\right]. \tag{3}$$

Thus, informally, value-at-risk gives "the *minimum potential loss* for the worst $100p$ % cases"[6] while expected shortfall gives the "*expected potential loss* for the worst $100p$ % cases". Therefore, the expected shortfall takes the magnitude of the exceeding of the value-at-risk into account, while for value-at-risk the magnitude of exceeding is irrelevant.

B. Which levels?

Both the value-at-risk and expected shortfall risk measurement method are defined for arbitrary levels $p \in (0,1)$. This leaves the issue of the choice of p open. Since we are interested in protecting against adverse market conditions it is clear that p should be chosen small. But how small? For value-at-risk the most common choices are $p = 0.05$ or $p = 0.01$ (the level chosen by the Basel Committee). In combination with the current multiplication factors used by the Basel Committee, the 1% value-at-risk results in more or less satisfactory capital reserves. In order to get a risk based capital reserve scheme based on expected shortfall, we need to determine a level p for the expected shortfall. In most comparisons between value-at-risk and expected shortfall their levels are taken to be equal. This seems to lead to the general opinion that, although expected shortfall has nice theoretical properties, it is much harder to backtest than value-at-risk (see Yamai and Yoshiba (2002)), the main reason why expected shortfall is still absent in Basel II.[7] However, for capital reserve determination it seems to make sense to look at comparable quantiles instead of levels. For example, take the median shortfall, that is, take the median in the tail instead of the expectation. The median shortfall with level $2p$ corresponds to value-at-risk with level p. If we would compare the backtest results of the

[5]The additional term $Q_p\left(X/N_m\right)\left(p - \mathbb{P}\left(X/N_m \leq Q_p\left(X/N_m\right)\right)\right)$ is needed in order to make the expected shortfall coherent, see Acerbi and Tasche (2002).

[6]Most value-at-risk devotees prefer the alternative formulation of "the maximum loss in the 100(1-p)% best cases."

[7]We thank Jon Danielsson for pointing this out to us.

median shortfall and the value-at-risk with the same level, we probably find that value-at-risk has a better performance than median shortfall. But for a valid comparison, we should use the median shortfall with twice the level of value-at-risk, in which case we find equal performance. A similar reasoning applies to expected shortfall. In order to have a valid comparison of the backtest results we should look at the quantiles and not the levels. Doing this for the Gaussian distribution (as a reference distribution), we find $p = 0.025$ for the expected shortfall when $p = 0.01$ for value-at-risk.[8] In case of excess kurtosis we need to take a higher level for the expected shortfall for it to equal the 1% value-at-risk. Since, in practice, we usually encounter distributions with heavier tails than the Gaussian distribution, the level of 2.5% can be seen as a lower bound on the level for equal capital requirement.

III. Standardization procedure

Let $(h_t)_{t\in\mathcal{T}_T}$ with $\mathcal{T}_T = \{1, ..., T\}$ (the test period) be a time-series of (in our case daily) returns on a profit and loss account (P&L) of a bank. Usually, the sequence $(h_t)_{t\in\mathcal{T}_T}$ cannot be modelled appropriately as a sample from one single distribution, say F, due to the fact that banks change the composition of their portfolio frequently. In general, the risk profile (the distribution of the P&L) of the bank changes over time. Therefore, we allow $(h_t)_{t\in\mathcal{T}_T}$ to be drawn from a different (marginal) distribution each period, that is,

$$h_t \sim F_t \quad t \in \mathcal{T}_T. \tag{4}$$

A bank is required to report the riskiness of its portfolio every day by means of a risk measure $\rho(h_t)$, where $\rho(h_t)$ denotes the risk measure for period t using the information up to time $t-1$.[9] In order to compute these risk measures the bank uses a sequence of forecast distributions $(P_t)_{t\in\mathcal{T}_T}$, with corresponding densities $(p_t)_{t\in\mathcal{T}_T}$.

[8] Notice that for the value-at-risk at level $p = 0.01$ we have $-\Phi^{-1}(0.01) = 2.33$, while for the expected shortfall at level $p = 0.025$ we have $\Phi^{-1}(0.025) = -1.96$ and $-\mathbb{E}[X|X < -1.96] = \phi(-1.96)/\Phi(-1.96) = 2.34$ (see (3)), when X follows a standard normal distribution (where ϕ and Φ denote the density and distribution function of the standard normal distribution, respectively).

[9] It would be more appropriate to write $\rho_{t-1}(h_t)$, but we suppress the subscripts for notational convenience.

Often F_t is assumed to belong to a location-scale family; that is, it is assumed that the sequence $\{(h_t - \mu_t)/\sigma_t\}_{t\in\mathcal{T}_T}$ is identically distributed (see, for example, McNeil and Frey (2000) and Christoffersen et al. (2001)). However, this restricts the way in which the procedure takes portfolio changes of banks into account. In this set-up moments higher than two are only allowed to vary over time through the first two moments. More generally, we can use the probability integral transform (see, for example, Van der Vaart (1998)) to go from a non-identically distributed sequence $(h_t)_{t\in\mathcal{T}_T}$ to an identically distributed sequence $(y_t)_{t\in\mathcal{T}_T}$. This transform is defined as

$$y_t = G^{-1}\left(\int_{-\infty}^{h_t} p_t(u)\,du\right) = G^{-1}\left(P_t(h_t)\right), \qquad t \in \mathcal{T}_T, \tag{5}$$

In case $P_t = F_t$ for each $t \in \mathcal{T}_T$, the distribution of y_t equals G, otherwise, the distribution of y_t is equal to, say, Q_t, unequal to G (for at least one time period t). The following lemma (see special cases in Diebold et al. (1998) and Berkowitz (2001)) gives the density q_t of y_t.

Lemma 1 Let $f_t(\cdot)$ denote the density of h_t, $p_t(\cdot)$ the density corresponding to $P_t(\cdot)$, g the density associated with G, and $y_t = G^{-1}(P_t(h_t))$. If $\frac{dP_t^{-1}(G(y_t))}{dy_t}$ is continuous and nonzero over the support of h_t, y_t has the following density:

$$\begin{aligned} q_t(y_t) &= \left|\frac{dG^{-1}(P_t(h_t))}{dh_t}\right|^{-1} f_t(h_t) \\ &= g(y_t)\frac{p_t(h_t)}{f_t(h_t)}. \end{aligned} \tag{6}$$

Proof. Just apply the change of variables transformation to $y_t = G^{-1}(P_t(h_t))$ and the result follows. ■

In case the forecast distributions of the bank are correct, i.e., $P_t = F_t$, $t \in \mathcal{T}_T$, we have that $q_t(y_t) = g(y_t)$. Thus, under the hypothesis that $P_t = F_t$, $t \in \mathcal{T}_T$ we can go from a non-identically distributed sequence $(h_t)_{t\in\mathcal{T}_T}$ to an identically distributed sequence $(y_t)_{t\in\mathcal{T}_T}$ with distribution G. We denote this procedure as *standardization to*

G. For example, Berkowitz (2001), uses $G = \Phi$, the standard normal distribution, in order to use the Gaussian likelihood for his Likelihood Ratio tests.[10]

IV. Backtest procedure

After assigning a risk measurement method the regulator faces the important task of determining the quality of the models that the regulated banks use in order to compute the risk measure. One of the reasons that the value-at-risk approach is often preferred to the coherent risk measures is the fact that the quality of value-at-risk models seems more easily verifiable. Therefore, the choice of risk measurement method by the regulator is based on the tools available to the regulator to verify model quality. In order to motivate the regulated to improve their models, regulators often impose model reserves or multiplication factors (see, for example, the multiplication factors by the Basel Committee). In Section IV.A we review the backtest procedure of the Basel Committee. Then we provide an alternative and more general procedure, in Section IV.B ignoring estimation risk, and in Section IV.C taking estimation risk into account.

A. Backtest procedure of Basel Committee

In this section we briefly describe the backtest procedure used by the BIS for determining the multiplication factors for capital requirements. A full exposition can be found in the Basel Committee on Banking Supervision (1996b).

Banks need to produce T ($T = 250$ in the current BIS implementation) value-at-risk forecasts (1% value-at-risk in the current BIS implementation) $(\mathrm{VaR}_t)_{t\in\mathcal{T}_T}$, where VaR_t denotes the value-at-risk forecast for day t using the information up to time $t-1$. It is assumed that these value-at-risk forecasts $(\mathrm{VaR}_t)_{t\in\mathcal{T}_T}$ are such that the exceedances sequence $(e_t)_{t\in\mathcal{T}_T}$ consists of independent elements with a Bernoulli distribution with probability p, that is, Bern(p), where p denotes the quantile relevant to the value-at-risk method employed. The exceedances $(e_t)_{t\in\mathcal{T}_T}$ are defined by

[10]Notice, however, that when $P_t \neq F_t$, for at least one $t \in \mathcal{T}_T$, the standardization procedure will result in distributions Q_t, not necessarily equal for different $t \in \mathcal{T}_T$.

Table I
BIS multiplication factors

The table shows the plus factors (multiplication factor = 3 + plus factor) used by the BIS for capital requirements based on a sample of 250. Tables for other sample sizes can be constructed by letting the yellow zone start when the cumulative probability exceeds 95% and the red zone when it exceeds 99.99%.

zone	Number of exceedances	Plus factor	Cumulative probability
green zone	0	0,00	8,11
	1	0,00	28,58
	2	0,00	54,32
	3	0,00	75,81
	4	0,00	89,22
yellow zone	5	0,40	95,88
	6	0,50	98,63
	7	0,65	99,60
	8	0,75	99,89
	9	0,85	99,97
red zone	≥ 10	1,00	99,99

$$e_t = \mathbf{I}_{(-\infty, -\mathrm{VaR}_t)}(h_t), \quad t \in \mathcal{T}_T. \tag{7}$$

By definition we have that

$$\mathbb{P}(e_t = 1) = \mathbb{P}(h_t < -\mathrm{VaR}_t), \quad t \in \mathcal{T}_T. \tag{8}$$

If $-\mathrm{VaR}_t = F_t^{-1}(p)$, with F the cumulative distribution function of h_t, we have that $\mathbb{P}(e_t = 1) = p$ and, consequently, the distribution of e_t indeed follows a Bernouilli-distribution. Using the cumulative distribution of the binomial distribution one may then compute multiplication factors based on the number of exceedances. For completeness, we present Table 2 from Basel Committee on Banking Supervision (1996b) in Table I.

The capital requirement can then be computed as the product of the value-at-risk at time t, $\mathrm{VaR}_t^{0.01}$, multiplied by a multiplication factor, mf_t, that is determined by the

results of a backtest of model m on the previous T ($T = 250$ in Basel Accord) days,[11]

$$\mathrm{CR}_t = \mathrm{mf}_t \cdot \mathrm{VaR}_t^{0.01}. \tag{9}$$

The backtest procedure given by the Basel Committee described above has some serious shortcomings. It assumes that under the null hypothesis the exceedances $(e_t)_{t=1}^{T}$ are *i.i.d.* while empirical evidence shows a clustering phenomenon in the exceedances (see, for example, Berkowitz and O'Brien (2002)). However, in case of dependence, one could adapt the test procedure by applying, for instance, the Newey-West (1987) approach which allows for quite general forms of dependence over time. Another drawback is that the above procedure does not take estimation risk into account which manifests itself in the fact that $\mathrm{VaR}_t = \widehat{F}_t^{-1}(p)$ which is not necessarily equal to $F_t^{-1}(p)$. Due to the limited amount of data there is likely some inaccuracy in the estimate for the value-at-risk which in effect causes an estimation error in the exceedances (compare West (1996)). This issue is treated in Section IV.C. A final drawback is that by transforming the information of the distribution into one characteristic (exceeding of value-at-risk or not) we lose relevant information of the return distribution (see also Berkowitz (2001)). In Section V we see that the power of the test is affected by removing this information.

B. General backtest procedure

We assume given a sample of transformed data $(y_t)_{t\in\mathcal{T}_T}$ to which the standardization procedure, described in Section V has been applied; this yields observations drawn from actual distributions Q_t, some or all possibly unequal to the postulated standardized distribution G. In this subsection we refrain from possible estimation risk in estimating the distribution function. This will be discussed in the next subsection.

The null hypothesis $H_0 : Q_t = G$ can be tested against numerous alternatives. We shall formulate these alternatives under the additional assumption of stationarity, i.e.,

[11] Actually, the used value-at-risk is $\max\{\mathrm{VaR}_t^{0.01}, \frac{1}{60}\sum_{i=1}^{60}\mathrm{VaR}_{t-i}^{0.01}\}$ instead of $\mathrm{VaR}_t^{0.01}$ (see Basel Committee on Banking Supervision (1996b)). Furthermore, the multiplication factors are set every 3 months.

$Q_t = Q$.[12] For example, Berkowitz (2001) tests this hypothesis using a likelihood ratio (LR) test using the Gaussian likelihood ($H_1 : Q \neq G = \Phi$) and a censored Gaussian likelihood ($H_1 : Q_{(-\infty, Q^{-1}(p)]} \neq G_{(-\infty, G^{-1}(p)]}$).[13] Using the censored Gaussian likelihood has the advantage that it ignores model failures in the interior of the distribution: only the tail behavior matters.

Following this line of reasoning, we use risk measurement methods which focus by construction on the tail behavior to evaluate the null hypothesis. We do not directly care about conservative models, that is, the true risk $\varrho(Q)$ is smaller than or equal to $\varrho(G)$, the risk expected by our model. Since we do not want that the model underestimates the risk, the alternative is taken to be $H_1 : \varrho(Q) > \varrho(G)$.

In Section II, we defined risk measurement methods as functions of random variables (defined on a financial model $m = (\Omega, \mathbb{P})$) following the quantitative risk measurement literature. For the purpose of testing it is more convenient to define the risk measurement method as a functional, $\varrho : D_F \to \mathbb{R}$, of a distribution function to $\mathbb{R} \cup \infty$.[14] Thus, $\mathrm{RMM}_m(X) = \varrho(F)$ for risk X if F is the distribution function of X associated with model m.

If $\varrho : D_F \to \mathbb{R}$ is Hadamard differentiable on D_F, we can apply the functional delta method (see, for example, Van der Vaart (1998) Thm. 20.8)

$$\sqrt{T}\left(\varrho\left(Q_T\right) - \varrho\left(Q\right)\right) = \sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\psi_t\left(Q\right) + o_p\left(1\right), \qquad \mathbb{E}\psi_t\left(Q\right) = 0,\ \mathbb{E}\psi_t^2\left(Q\right) < \infty, \tag{10}$$

where Q_T denotes the empirical distribution of the random sample $(y_t)_{t \in \mathcal{T}_T}$ and $\psi_t(Q)$ denotes the influence function of the risk measurement method ϱ at observation t. As can easily be shown, the common risk measures such as value-at-risk and expected shortfall

[12] When presenting the test statistics, we maintain this assumption and implicitly assume that this stationarity is transferred in the risk measures $\varrho(Q_t)$. Notice, however, the testing procedure is more generally applicable than just for the case of stationarity.

[13] For distribution function F, $F_{(-\infty, F^{-1}(p)]}$ denotes the left tail of the distribution up to the p^{th} quantile.

[14] D_F denotes the space of all distribution functions, that is, all non-decreasing cadlag functions F on $[-\infty, \infty]$ with $F(-\infty) \equiv \lim_{x \to -\infty} F(x) = 0$ and $F(\infty) \equiv \lim_{x \to \infty} F(x) = 1$. D_F is equipped with the metric induced by the supremum norm.

are Hadamard differentiable.[15] We can then use the following test statistic:

$$S_T = \sqrt{T}\frac{\left(\varrho\left(Q_T\right) - \varrho\left(Q\right)\right)}{\sqrt{V}} \xrightarrow[H_0]{d} \mathcal{N}(0,1), \tag{11}$$

with $V = \mathbb{E}\psi_t^2\left(Q\right)$ and $\varrho(Q)$ evaluated under the null hypothesis, $Q = G$.[16] Some important examples are:

Example 1 (Value-at-risk) In the case of value-at-risk written as a function of the distribution function

$$\varrho(Q) = -Q^{-1}(p), \tag{12}$$

the influence function $\psi\left(Q\right)$ is given by

$$\psi_{\text{VaR}}\left(Q\right) = -\frac{p - \mathbf{I}_{(-\infty, Q^{-1}(p)]}\left(x\right)}{q\left(Q^{-1}\left(p\right)\right)}, \tag{13}$$

and

$$\mathbb{E}\psi_{\text{VaR}}^2\left(Q\right) = \frac{p\left(1-p\right)}{q^2\left(Q^{-1}\left(p\right)\right)}. \tag{14}$$

This leads to the following test statistic

$$S_{\text{VaR}} = \sqrt{T}q\left(Q^{-1}\left(p\right)\right)\frac{\left(\varrho\left(Q_T\right) - \varrho\left(Q\right)\right)}{\sqrt{p\left(1-p\right)}} \tag{15}$$

The critical value-at-risk levels for the yellow and red zones are given by

$$\begin{aligned} \text{VaR}_{yellow} &= \sqrt{\frac{z_{0.95}}{T}\frac{p\left(1-p\right)}{q^2\left(Q^{-1}\left(p\right)\right)}} + \text{VaR}\left(Q\right) \\ \text{VaR}_{red} &= \sqrt{\frac{z_{0.9999}}{T}\frac{p\left(1-p\right)}{q^2\left(Q^{-1}\left(p\right)\right)}} + \text{VaR}\left(Q\right), \end{aligned} \tag{16}$$

[15] For the value-at-risk, see, for example, Van der Vaart and Wellner (1996) Lemma 3.9.20. In case of the expected shortfall, the influence function is easily obtained by applying the chain rule for Hadamard differentiable functions to the quantile function and the mean, see, for example, Van der Vaart and Wellner (1996) Lemma 3.9.3.

[16] Under the assumption of stationarity, i.e., $Q_t = Q$, we could also evaluate V under the alternative as $V = \frac{1}{T}\sum_{t=1}^{T}\left(\psi_t\left(Q_T\right) - \frac{1}{T}\sum_{t=1}^{T}\psi_t\left(Q_T\right)\right)^2$. However, our simulation study indicates a much worse performance of the test statistics using this estimate than when evaluating V under the null.

where z_p denotes the p^{th} quantile of the standard Gaussian distribution.

Example 2 (Exceedances) In the case of the number of exceedances written as a function of the distribution function

$$\varrho(Q) = \mathbf{I}_{(-\infty, Q^{-1}(p)]}, \tag{17}$$

the influence function $\psi(Q)$ is given by

$$\psi_{\text{exc}}(Q) = p - \mathbf{I}_{(-\infty, Q^{-1}(p)]}(x), \tag{18}$$

and

$$\mathbb{E}\psi^2_{\text{exc}}(Q) = p(1-p). \tag{19}$$

This gives the following test

$$S_{\text{exc}} = \sqrt{T}\frac{(\varrho(Q_T) - \varrho(Q))}{\sqrt{p(1-p)}} \tag{20}$$

The critical numbers of exceedances for the yellow and red zones are given by

$$\begin{aligned} \text{Exc}_{yellow} &= \sqrt{z_{0.95} T p (1-p)} + pT \\ \text{Exc}_{red} &= \sqrt{z_{0.9999} T p (1-p)} + pT \end{aligned} \tag{21}$$

For the regular backtest size of 250, these critical values are equal to the exact setting of the binomial distribution used by the BIS.

Example 3 (Expected shortfall) In the case of ES written as a function of the distribution function

$$\varrho(Q) = -\int_{-\infty}^{Q^{-1}(p)} x dQ(x) + Q^{-1}(p)\left(p - \int_{-\infty}^{Q^{-1}(p)} dQ(x)\right), \tag{22}$$

the influence function $\psi(Q)$ is given by

$$\begin{aligned}\psi_{\mathrm{ES}}(Q) &= -\frac{1}{p}\left[\left(x-Q^{-1}(p)\right)\mathbf{I}_{(-\infty,Q^{-1}(p)]}(x)\right. \\ &\quad \left.+\psi_{\mathrm{VaR}}(Q)\left(p-\int_{-\infty}^{Q^{-1}(p)}dQ(x)\right)\right]-\mathrm{ES}(Q)+\mathrm{VaR}(Q) \qquad (23)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\psi_{\mathrm{ES}}^2(Q) &= \frac{1}{p}\mathbb{E}\left[X^2|X\leq Q^{-1}(p)\right]-\mathrm{ES}(Q)^2 \\ &\quad +2\left(1-\frac{1}{p}\right)\mathrm{ES}(Q)\,\mathrm{VaR}(Q)-\left(1-\frac{1}{p}\right)\mathrm{VaR}(Q)^2. \qquad (24)\end{aligned}$$

This leads to the following test statistic

$$S_{\mathrm{ES}}=\sqrt{T}\frac{(\varrho(Q_T)-\varrho(Q))}{\sqrt{\mathbb{E}\psi_{\mathrm{ES}}^2(Q)}} \qquad (25)$$

The critical ES levels for the yellow and red zones are given by

$$\begin{aligned}\mathrm{ES}_{yellow} &= \sqrt{\frac{z_{0.95}}{T}\mathbb{E}\psi_{\mathrm{ES}}^2(Q)^2}+\mathrm{ES}(Q) \\ \mathrm{ES}_{red} &= \sqrt{\frac{z_{0.9999}}{T}\mathbb{E}\psi_{\mathrm{ES}}^2(Q)^2}+\mathrm{ES}(Q) \qquad (26)\end{aligned}$$

We conclude this subsection by illustrating that the test statistics can easily be implemented for the Gaussian case $G=\Phi$, by presenting the outcomes of $\mathbb{E}\psi_t^2(G)$ in case of value-at-risk and expected shortfall. For this, let $\phi(x)$ denote the density function of the standard Gaussian $\mathcal{N}(0,1)$ distribution and z_p the p^{th} quantile of the standard normal distribution. The value-at-risk in case of a normal distribution $\mathcal{N}(0,1)$ is given by

$$\mathrm{VaR}_p(X)=z_p, \qquad (27)$$

and the expected shortfall is given by

$$\mathrm{ES}_p(X)=-\phi(z_p)/p. \qquad (28)$$

$\mathbb{E}\psi_t^2(\Phi)$ for value-at-risk and expected shortfall are then given by,

$$\mathbb{E}\psi_t^2(\Phi) = \frac{p(1-p)}{\phi(z_p)},$$

for value-at-risk and

$$\mathbb{E}\psi_t^2(\Phi) = 1 - z_p\frac{\phi(z_p)}{p} - \left(\frac{\phi(z_p)}{p}\right)^2 - 2\left(1-\frac{1}{p}\right)\frac{\phi(z_p)}{p}z_p - \left(1-\frac{1}{p}\right)z_p^2,$$

for expected shortfall.

C. Estimation risk

The backtesting procedures described in this section assume that the forecasted distributions $(P_t)_{t\in\mathcal{T}_T}$ of the profit/loss are given. It seems natural to penalize banks with a plus factor for using inappropriate model families, but not for just having to estimate a correctly specified model (assuming that they use their data efficiently). In order to do so, we derive in this section backtest procedures that take estimation risk into account.

Again, we use the standardization procedure described in Section III. We assume given a random estimation sample $(y_t)_{t\in\mathcal{T}_e}$, $\mathcal{T}_e = \{-N+1, ..., 0\}$, and a random testing sample $(y_t)_{t\in\mathcal{T}_T}$ $\mathcal{T}_T = \{1, ..., T\}$ with $y_t \sim Q$ ($Q = G$ under the null). We then have

$$\sqrt{n}\left(\varrho(Q_n) - \varrho(Q)\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\psi^2(Q)\right), \qquad n = T, N$$

where $\psi(\cdot)$ is the influence function of $\varrho(\cdot)$. This yields (still under the null)

$$\begin{aligned}\sqrt{T}\left(\varrho(Q_T) - \varrho(Q_N)\right) &= \sqrt{T}\left(\varrho(Q_T) - \varrho(Q)\right) - \sqrt{\frac{T}{N}}\sqrt{N}\left(\varrho(Q_N) - \varrho(Q)\right) \\ &\xrightarrow{d} \mathcal{N}\left(0, (1+c)\,\mathbb{E}\psi^2(G)\right), \end{aligned} \tag{29}$$

when $\frac{T}{N} \to c$ as $N \to \infty$ and $T \to \infty$.

If the estimation period would grow with time, c would tend to zero. In practice, one usually specifies a finite fixed estimation period (for example, 2 years) and computes

the risk measure based on this estimation period. This is a so-called rolling window estimation procedure, which can be approximated in our setting by taking $c = \frac{T}{N}$ in (29).

For the examples in IV.B we can derive the critical values for the yellow and red zones in the same way by replacing V by $(1+c)V$. With the incorporation of estimation risk in the backtesting procedure we introduce an additional degree of freedom for the regulator, namely the choice of c (or N, since T could already be chosen by the regulator).

V. Simulation results

In this section we compare the finite sample behavior of the backtest procedures. First, we determine the actual size of the tests for the exceedances ratio, value-at-risk, and expected shortfall. For simplicity, we take $F_t = \mathcal{N}(0,1)$, the standard normal distribution, for $t \in \mathcal{T}_T$. To check the performance of the tests for size, we take $P_t = F_t$, $t \in \mathcal{T}_T$, and set the significance level $\alpha = 0.05$. We verify the performance of the tests given in the examples in Section IV.B using $G = \Phi$, the standard normal distribution function.[17] The tests are compared to the censored LR test of Berkowitz (2001), which we denote as the Berkowitz tail test. Table II shows the results of the performance of the size of the tests. We see that the size for the three tests (Exceedances, value-at-risk, and expected shortfall) seem reasonable for the common sample size of 250. The Berkowitz tail test seems to converge a bit faster.

Next, we investigate the power of the different tests. In practice, financial time series often exhibit excess kurtosis with respect to the normal distribution and have longer left tails. We consider three alternatives that replicate (parts of) this behavior. First, we use the student $t-$distribution with 5 degrees of freedom, that is, $F_t = t_5$, . This distribution has heavier tails than the normal distribution, but is still symmetric.

[17] Using $G = U[0,1]$ results in very poor results for smaller sample sizes. The reason is that by transforming the data to uniform random numbers the symmetry in the test is lost due to the non-linear shape of F.

Table II
Simulation results for size of tests

This table presents the Type I errors (in percentages) if $F_t = P_t = \mathcal{N}(0,1)$ for $t \in T_T$ for $T = 125$, 250, 500, and 1000. The argument H_0 denotes that the variance used is $\mathbb{E}\psi_t^2(G)$ and H_1 denotes that the variance used is $V = \frac{1}{T}\sum_{t=1}^{T}\left(\psi_t(Q) - \frac{1}{T}\sum_{t=1}^{T}\psi_t(Q)\right)^2$. $\text{Tail}_{0.025}$ denotes Berkowitz tail test. The number of simulations equals 10,000.

T	Exceedances	$\text{VaR}_{0.01}(H_0)$	$\text{VaR}_{0.01}(H_1)$	$\text{ES}_{0.025}(H_0)$	$\text{ES}_{0.025}(H_1)$	$\text{Tail}_{0.025}$
125	3.75	2.75	1.81	2.64	3.24	3.05
250	4.17	4.81	2.87	5.14	4.64	5.42
500	6.63	2.91	2.27	9.38	8.10	5.16
1000	4.51	3.87	2.98	4.34	2.63	5.33

Second, we use two alternatives from the Normal Inverse Gaussian (NIG) family.[18] The NIG distribution allows one to control both the level of excess kurtosis and the skewness. We consider two cases: a symmetric case with a moderately high kurtosis, $\beta = 0$, $\alpha = \sqrt{\beta^2+1}$, $\delta = 1/\left(1+\beta^2\right)$, $\mu = 0$ and a case where the distribution is very skewed to the left and has a large kurtosis, $\beta = -0.25$, $\alpha = \sqrt{\beta^2+1}$, $\delta = 1/\left(1+\beta^2\right)$, $\mu = 0$. Third, we take a GARCH(1,1)-process,[19] with parameter values $\omega = 0.05$, $\gamma_1 = 0.25$, and $\gamma_2 = 0.7$ to allow for a time-dependent distribution under the alternative hypothesis. For the time-independent cases we present the results for VaR and ES with the test statistic estimated under the null as well as under the alternative (see footnote 16). Table III contains the results. We see that in case of a time-independent alternative

[18]The density of the $NIG(\alpha,\beta,\mu,\delta)$ is given by

$$f_{NIG}(x) = \frac{\alpha \exp\left(\delta\sqrt{\alpha^2-\beta^2} - \beta\mu\right)}{\pi} q\left(\frac{x-\mu}{\delta}\right)^{-1} K_1\left\{\delta\alpha q\left(\frac{x-\mu}{\delta}\right)\right\} \exp\left\{\beta(x-\mu)\right\},$$

with $q(x) = \sqrt{1+x^2}$ and $K_1(x)$ the modified Bessel function of the third kind. See, for example, Barndorff-Nielsen (1996).

[19]The GARCH(1,1) model (see Bollerslev (1986)) is given by the following return and volatility equations:

$$\begin{aligned} r_t &= \sqrt{h_t}\epsilon_t \\ h_t &= \omega + \gamma_1 r_{t-1}^2 + \gamma_2 h_{t-1} \end{aligned}$$

Table III: **Simulation results for power of tests**

This table presents the Type II errors (in percentages) if $F_t = t_5$, $F_t = NIG(\alpha, 0, \delta, \mu)$, $F_t = NIG(\alpha, -0.25, \delta, \mu)$, and $F_t = N\left(0, \sigma_t^2\right)$ (GARCH(1,1)) ; $\alpha = \sqrt{\beta^2 + 1}$, $\delta = 1/\left(1 + \beta^2\right)$, $\mu = 0$. σ_t^2 follows the volatility equation of a GARCH(1,1) model with $\omega = 0.05$, $\gamma_1 = 0.25$, and $\gamma_2 = 0.7$. $P_t = N(0, 1)$ for $t \in \mathcal{T}_T$ for $T = 125$, 250, 500, and 1000. The number of simulations equals $10,000$.

T	Exceedances	$\text{VaR}_{0.01}(H_0)$	$\text{VaR}_{0.01}(H_1)$	$\text{ES}_{0.025}(H_0)$	$\text{ES}_{0.025}(H_1)$	$\text{Tail}_{0.025}$
t_5						
125	11.72	22.44	10.41	26.77	6.73	20.51
250	17.64	35.98	14.98	45.65	14.22	42.43
500	32.86	38.57	17.54	69.86	35.93	63.13
1000	42.89	57.60	32.68	82.39	52.12	87.91
$NIG(\alpha, 0, \delta, \mu)$						
125	16.08	25.08	14.22	30.27	0.00	22.84
250	25.53	44.73	22.93	52.51	22.72	45.29
500	47.06	51.17	29.25	78.51	51.11	69.90
1000	63.32	74.38	53.43	90.13	71.44	91.41
$NIG(\alpha, -0.25, \delta, \mu)$						
125	33.94	45.81	31.03	54.26	21.41	41.52
250	52.97	71.94	47.48	81.00	48.41	72.54
500	83.40	85.53	67.25	97.15	87.42	92.96
1000	95.97	97.93	91.87	99.76	98.39	99.71
GARCH(1,1)						
125	11.08	11.60		13.66		17.63
250	14.45	20.49		24.02		19.23
500	24.17	20.10		40.66		25.78
1000	27.34	29.63		43.37		39.93

for both the value-at-risk and the expected shortfall the tests with variance evaluated under the null hypothesis have (far) more power. The difference with the test using the estimated variance under the alternative narrows when the sample size increases. The test for expected shortfall performs best in detecting the misspecification, also when the alternative is GARCH$(1,1)$ for $T \geq 250$; the number of exceedances test has less power than the value-at-risk test and the expected shortfall test. The Berkowitz tail test also performs well and, therefore, seems a worthwhile auxiliary test, but, in general, trails the test for expected shortfall. Especially for the shorter sample sizes the test for expected shortfall performs better with only GARCH$(1,1)$ for $T = 125$ as an exception.

Finally, we take estimation risk into account. In Table IV the results are shown for an equal estimation and testing period. It gives the expected result that the longer the samples the better the power of the tests. However, the performance of the test for value-at-risk with the variance evaluated under the alternative (in the time-independent cases) is quite bad. In Table V we fixed the testing period to 1 year (250 days) and varied the estimation period. As expected the results improve for longer estimation periods. Again, the performance of the test for value-at-risk with the variance evaluated under the (time-independent) alternative is quite bad.

Concluding, we find that the performances of the tests with the variance evaluated under (a time-independent) H_0 have far more power than the tests with the variance evaluated under H_1 for sample sizes realistic for financial data. Furthermore, we find that the performance for the size of the tests of the 2.5% expected shortfall is about equal to the 1% value-at-risk. However, the power of the 2.5% expected shortfall test is much better than that of the 1% value-at-risk.

VI. Multiplication factors

In this section we propose a method to compute multiplication factors for capital requirements determination. Our starting point is the test statistic (11). If the test statistic results in rejection of the null hypothesis, then we might conclude that $\varrho\,(G)$ is taken too low. The question then is by which multiplication factor $\varrho\,(G)$ at least should be

Table IV: **Simulation results for power of tests in case of estimation risk**

This table presents the Type II errors (in percentages) if $F_t = t_5$, $F_t = NIG(\alpha, 0, \delta, \mu)$, $F_t = NIG(\alpha, -0.25, \delta, \mu)$, and $F_t = N\left(0, \sigma_t^2\right)$ (GARCH(1,1)) ; $\alpha = \sqrt{\beta^2 + 1}$, $\delta = 1/\left(1 + \beta^2\right)$, $\mu = 0$. σ_t^2 follows the volatility equation of a GARCH(1,1) model with $\omega = 0.05$, $\gamma_1 = 0.25$, and $\gamma_2 = 0.7$. $P_t = N(0, 1)$ for $t \in \mathcal{T}_T$ and $\mathcal{T}_T$ for $T = 125$, 250, 500, and 1000. The number of simulations equals $10,000$.

$N = T$	Exceedances	$\text{VaR}_{0.01}(H_0)$	$\text{VaR}_{0.01}(H_1)$	$\text{ES}_{0.025}(H_0)$	$\text{ES}_{0.025}(H_1)$	$\text{Tail}_{0.025}$
t_5						
125	18.40	15.49	0.34	22.91	4.87	15.93
250	13.51	22.84	0.38	37.81	6.69	27.49
500	19.25	21.30	0.27	59.23	15.85	47.79
1000	28.50	30.91	1.40	72.24	23.92	74.94
$NIG(\alpha, 0, \delta, \mu)$						
125	21.85	17.07	0.23	24.79	6.66	15.84
250	18.11	25.00	0.38	41.62	10.68	26.15
500	27.89	26.99	0.63	66.16	25.62	48.37
1000	45.44	41.88	3.66	80.08	40.71	76.80
$NIG(\alpha, -0.25, \delta, \mu)$						
125	38.36	31.03	0.85	45.10	12.81	33.55
250	41.01	47.08	1.95	69.98	24.90	54.97
500	61.90	57.61	4.67	91.94	58.48	81.22
1000	86.61	81.47	20.17	98.71	85.74	97.86
GARCH(1,1)						
125	18.79	12.10		13.13		7.83
250	13.31	13.22		19.78		10.28
500	16.28	11.37		31.81		14.46
1000	20.30	13.81		32.28		21.09

Table V: **Simulation results for power of tests in case of estimation risk**

This table presents the Type II errors (in percentages) if $F_t = t_5$, $F_t = NIG(\alpha, 0, \delta, \mu)$, $F_t = NIG(\alpha, -0.25, \delta, \mu)$, and $F_t = N\left(0, \sigma_t^2\right)$ (GARCH(1,1)) ; $\alpha = \sqrt{\beta^2 + 1}$, $\delta = 1/\left(1 + \beta^2\right)$, $\mu = 0$. σ_t^2 follows the volatility equation of a GARCH(1,1) model with $\omega = 0.05$, $\gamma_1 = 0.25$, and $\gamma_2 = 0.7$. $P_t = N(0, 1)$ for $t \in \mathcal{T}_T$ and $\mathcal{T}_T$ for $T = 125$, 250, 500, and 1000. The number of simulations equals $10,000$.

(N, T)	Exceedances	$\text{VaR}_{0.01}(H_0)$	$\text{VaR}_{0.01}(H_1)$	$\text{ES}_{0.025}(H_0)$	$\text{ES}_{0.025}(H_1)$	$\text{Tail}_{0.025}$
t_5						
$(125, 250)$	17.58	16.71	0.02	33.06	4.46	43.28
$(250, 250)$	13.56	22.91	0.33	37.53	6.51	55.80
$(500, 250)$	21.46	28.14	0.92	42.29	9.18	63.21
$(1000, 250)$	20.17	31.37	1.34	44.09	12.26	68.02
$NIG(\alpha, 0, \delta, \mu)$						
$(125, 250)$	18.31	13.33	0.10	33.01	5.03	18.60
$(250, 250)$	18.00	25.00	0.36	42.75	10.19	27.00
$(500, 250)$	29.45	34.51	1.30	47.71	16.56	34.48
$(1000, 250)$	29.61	40.17	2.35	50.37	20.21	39.54
$NIG(\alpha, -0.25, \delta, \mu)$						
$(125, 250)$	41.32	30.37	0.52	62.26	13.57	9.52
$(250, 250)$	41.17	47.10	1.83	70.38	24.84	27.49
$(500, 250)$	55.13	57.11	5.31	74.50	35.27	57.55
$(1000, 250)$	54.40	62.11	7.87	76.06	41.90	85.70
GARCH(1,1)						
125	18.98	12.82		19.07		8.72
250	13.36	13.60		19.82		9.98
500	17.98	15.84		22.37		13.38
1000	16.27	17.47		23.04		15.23

increased, such that the test statistic does no longer result in rejection of the null. Let $\varrho^*(Q_T)$ the realized value of $\varrho(Q)$. Then the minimum multiplication factor, mf, for which the null hypothesis would not be rejected follows from setting (11) equal to k_α, the critical value of the test at the significance level α

$$\sqrt{T}\frac{(\varrho(Q_T) - \text{mf}(s_T^*)\varrho(G))}{\sqrt{V}} = z_\alpha, \tag{30}$$

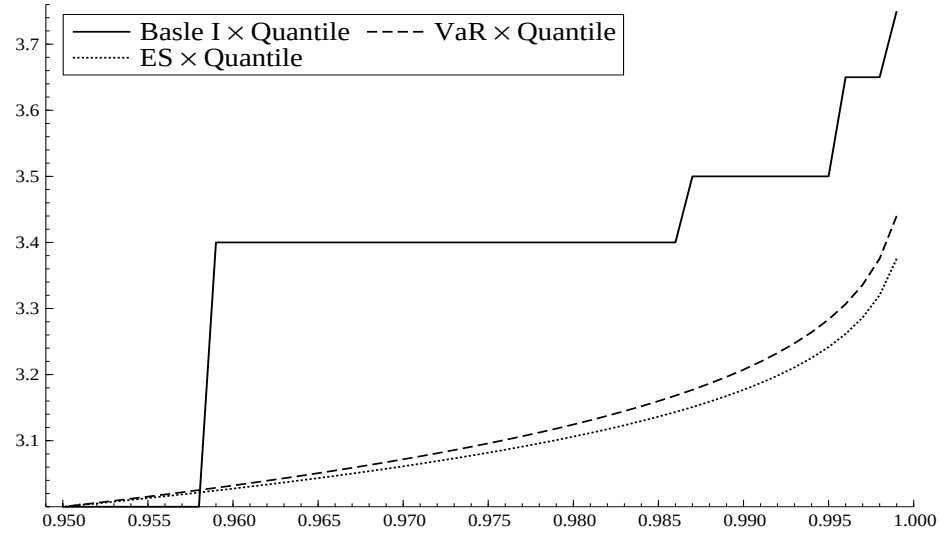
where s_T^* denotes the realized value of the test statistic. More generally, we may want to use a basis multiplication factor (bmf) and we may want to cap the multiplication factor at some upper value (limit). Using the fact that $\rho(Q_T) = \rho(G) + \sqrt{\frac{Vs_T^*}{T}}$ our proposal for the multiplication factor becomes

$$\text{mf}(s_T^*) = \min\left\{\left(\text{bmf}\cdot\max\left\{1, 1 + \frac{\sqrt{\frac{Vs_T^*}{T}} - \sqrt{\frac{Vk_\alpha}{T}}}{\varrho(G)}\right\}\right), \text{limit}\right\}, \tag{31}$$

We show the results for our proposed multiplication factor applied to value-at-risk, and expected shortfall in Figure 1, where we use $G = \Phi$, $\alpha = 0.05$, bmf $= 3$, and limit $= 4$. As the variance in (29) is larger than without estimation risk, the basis multiplication factor should be taken higher is one takes estimation risk into account. This is probably also one of the reasons that the multiplication factor of the BIS is rather high. For reasons of comparison with the BIS scheme, we use here a bmf of 3 and a limit of 4. See **?** for suggestions on setting the bmf for markets depending on the reliability with which the market can be modeled. On the horizontal axis we plot the quantiles of the distribution of the test statistic in (11) under the null hypothesis and on the vertical axis the resulting multiplication factors. As a benchmark we also plot the multiplication factors when using the current Basel procedure (now as a function of the quantiles of the corresponding test under the null). We see that the multiplication factors according to our proposal seem to compare favorably with those according to the Basel procedure. Moreover, the multiplication factors for expected shortfall are slightly lower than for value-at-risk. This has to do with the result that expected shortfall is more accurately estimated under the null than value-at-risk, i.e., the variance V in case

Figure 1. Multiplication factors

This figure shows the multiplication factors on the vertical axis against the quantiles of the test statistic on the horizontal axis. We used $G = \Phi$, $\alpha = 0.05$, and a basic multiplication factor bmf= 3.

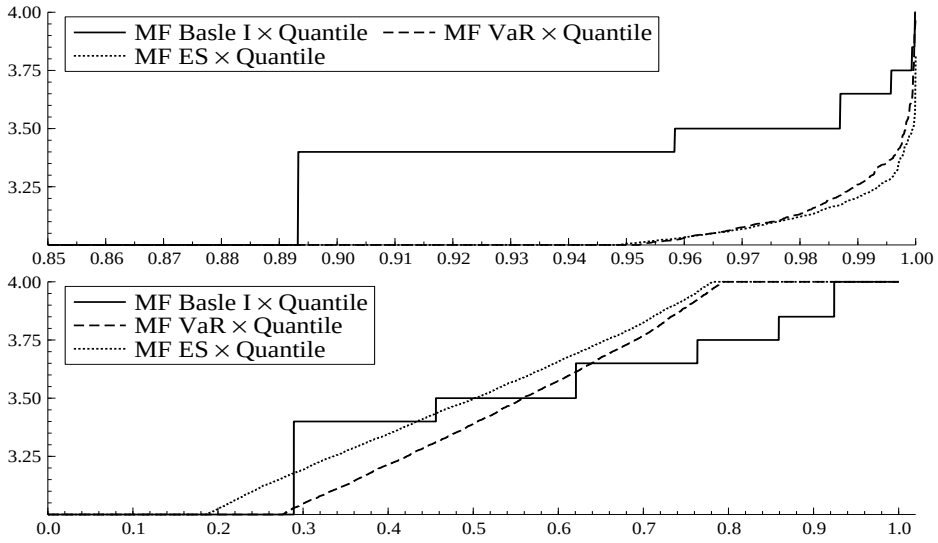


of expected shortfall is smaller than in case of value-at-risk.

In Figure 2 we report the results of applying the multiplication factors from (31) to value-at-risk and expected shortfall, using again the outcomes of the Basel procedure as a benchmark. We consider two cases: first, we look at the case where the model is correct, $P_t = F_t = \mathcal{N}\left(\mu, \sigma^2\right)$; second, the case of a seriously misspecified model, $P_t = \mathcal{N}\left(\mu, \sigma^2\right)$ and $F_t = NIG(\alpha, -0.25, \delta, \mu)$ with α, δ, μ as before, being the case where the distribution is very skewed to the left and has a large kurtosis.

The results of the correctly specified case reflect the outcomes presented in the previous figure: expected shortfall, having the lowest multiplication factors, performs best. Notice that the multiplication factor scheme from the current Basel Accord results in (too) large multiplication factors. In the second case of a misspecified model we see that the test using expected shortfall results in higher factors in more cases (due to the higher power) than the test using value-at-risk. For both expected shortfall and value-at-risk the punishment depends smoothly on the outcome of the test. The multiplication

Figure 2. Multiplication factors (size, power)
This figure shows the simulated cdf of the multiplication factors. In the upper panel the case of $F_t = \mathcal{N}\left(\mu, \sigma^2\right)$ is shown. In the lower panel we have the case where $F_t = NIG(\alpha, -0.25, \delta, \mu)$. In both panels $P_t = \mathcal{N}\left(\mu, \sigma^2\right)$. The number of days equals 250 and the number of simulations equals $10,000$.



factors according to the current Basel Accord more or less correspond to those of value-at-risk and expected shortfall, but in a heavily non-smooth way.

Concluding, in the case that the bank uses a correctly specified model, we find that the capital requirement scheme using expected shortfall leads to the least severe punishments. On the basis of the current Basel Accord banks would be punished more often and then also severely. Furthermore, in case of a misspecified model, we find that the capital requirement scheme using expected shortfall rejects the misspecified models most often, the multiplication factor depends smoothly on the size of the misspecification found and the variance in the multiplication factors is low.

VII. Conclusions

In this paper we suggested a backtest framework for a large and relevant group of risk measurement methods using the functional delta method. We showed that, for a large

group of risk measurement methods containing all currently used risk measurement methods, the backtest procedure can readily be found after computing the appropriate influence function of the risk measurement method. The influence functions for value-at-risk and expected shortfall are provided. Since this general framework is based on asymptotic results, we investigated whether the procedure is appropriate for realistic finite samples sizes. The results indicate that this is indeed the case, and that, contrary to common belief, expected shortfall is not harder to backtest than value-at-risk if we adjust the level of expected shortfall. Furthermore, the power of the test for expected shortfall is considerably higher than that of value-at-risk. Since the probability of detecting a misspecified model is higher for a given value of the test statistic, this allows the regulator to set lower multiplication factors. We suggested a scheme for determining multiplication factors. This scheme results in less severe penalties for the backtest based on expected shortfall compared to backtests based on value-at-risk, and the current Basel Accord backtesting scheme in case the test incorrectly rejects the model. In case of a misspecified model the multiplication factors are on average about the same for all tests. However, the multiplication factors based on the expected shortfall test are smooth and have low variance.

Thus, the prospects for setting up viable capital determination schemes based on expected shortfall seem promising.

References

Acerbi, C. and Tasche, D.: 2002, On the coherence of expected shortfall, *Journal of Banking and Finance* **26**, 1487–1503.

Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D.: 1997, Thinking coherently, *Risk* **10**, 68–71.

Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D.: 1999, Coherent measures of risk, *Mathematical Finance* **9**, 203–228.

Barndorff-Nielsen, O. E.: 1996, Normal inverse gaussian distributions and stochastic volatility modelling, *Scandinavian Journal of Statistics* **24**, 1–13.

Basel Committee on Banking Supervision: 1996a, *Amendment to the Capital Accord to Incorporate Market Risks,* Bank for International Settlements, Basel.

Basel Committee on Banking Supervision: 1996b, *Supervisory Framework for the Use of "Backtesting" in Conjunction with the Internal Models Approach to Market Risk Capital Requirements,* Bank for International Settlements, Basel.

Berkowitz, J.: 2001, Testing density forecasts, with applications to risk management, *Journal of Business and Economic Statistics* **19**, 465–474.

Berkowitz, J. and O'Brien, J.: 2002, How accurate are value-at-risk models at commercial banks?, *Journal of Finance* **57**, 1093–1111.

Bollerslev, T.: 1986, Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**, 307–327.

Christoffersen, P., Hahn, J. and Inoue, A.: 2001, Testing and comparing value-at-risk measures, *Journal of Empirical Finance* **8**, 325–342.

Delbaen, F.: 2000, Coherent risk measures on general probability spaces, *Working paper ETH* pp. 1–35.

Diebold, F. X., Gunther, T. A. and Tay, A. S.: 1998, Evaluating density forecasts, *Internation Economic Review* **39**, 863–883.

Duffie, D. and Pan, J.: 1997, An overview of value at risk, *Journal of Derivatives* **4**, 7–49.

Jorion, P.: 2000, *Value at Risk: The New Benchmark for Managing Financial Risk*, 2 edn, McGraw-Hill, New York.

Kerkhof, J., Melenberg, B. and Schumacher, H.: 2002, Model risk and regulatory capital, *CentER discussion paper 2002-27* pp. 1–56.

McNeil, A. and Frey, R.: 2000, Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach, *Journal of Empirical Finance* **7**, 271–300.

Risk Magazine: 1996, Value at risk, *Risk Magazine Special Supplement* pp. 68–71.

RiskMetrics: 1996, *Technical Document*, 4 edn, JP Morgan.

Tasche, D.: 2002, Expected shortfall and beyond, *Journal of Banking and Finance* **26**, 1519–1533.

Van der Vaart, A. W.: 1998, *Asymptotic Statistics*, Cambridge University Press.

Van der Vaart, A. W. and Wellner, J. A.: 1996, *Weak Convergence and Empirical Processes*, Springer-Verlag, New York.

West, K. D.: 1996, Asymptotic inference about predictive ability, *Econometrica* **64**, 1067–1084.

Yamai, Y. and Yoshiba, T.: 2002, On the validity of value-at-risk: Comparative analyses with expected shortfall, *Monetary and Economic Studies* **20**, 57–86.