Backtesting Beyond VaR

W. Härdle and G. Stahl
Humboldt-Universität zu Berlin and
Bundesaufsichtsamt für das Kreditwesen, Berlin

December 6, 1999

Abstract

VaR models are related to statistical forecast systems. Within that framework different forecast tasks including Value-at-Risk and shortfall are discussed and motivated. A backtesting method based on the shortfall is developed and applied to VaR forecasts of a real portfolio. The analysis shows that backtesting based on shortfall is very sensitive with respect to the underlying assumptions.

1 Forecast tasks and VaR Models

With the implementation of Value-at-Risk (VaR) models a new chapter of risk management was opened. Their ultimate goal is to quantify the uncertainty about the amount that may be lost or gained on a portfolio over a given period of time. Most generally, the uncertainty is expressed by a forecast distribution P_{t+1} for period $t+1$ associated with the random variable L_{t+1}, denoting the portfolio's profits and losses (P&L).
In practice, the prediction P_{t+1} is conditioned on an information set at time t and, typically calculated through a plug-in approach, see Dawid (1984). In this case, P_{t+1} is output of a statistical forecast system, here the VaR model, consisting of a (parametric) family of distributions, denoted by $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ together with a prediction rule. Assumed that P_{t+1} belongs to this parametrized family $\mathcal{P}$ the estimates $\hat{\theta}_t$ are calculated by the prediction rule on the basis of a forward rolling data history $\mathcal{H}_t$ of fixed length n (typically $n = 250$ trading days) for all t, i.e.

$$P_{t+1}(\cdot) = P_{\hat{\theta}_t}(\cdot \mid \mathcal{H}_t).$$

One example for $\mathcal{P}$ also pursued in this paper is the RiskMetrics (1996) delta normal framework, i.e., the portfolios considered are assumed to consist of linear (or linearised) instruments and the common distribution of the underlyings' returns $Y \in I\!R^d$, i.e., the log price changes $Y_{t+1} = logX_{t+1} - logX_t$, is a (conditional) multinormal distribution, $N_d(0, \Sigma_t)$, where Σ_t (resp. and σ_t^2) denotes a conditional variance, i.e., $\mathcal{H}_t$ measurable function.
Consider for simplicity a position of λ_t shares in a single asset (i.e., d = 1) whose market value is x_t. The conditional distribution of L_{t+1} for this position with exposure $w_t = \lambda_t\ x_t$ is (approximately)

$$\begin{aligned} \mathcal{L}(L_{t+1} \mid \mathcal{H}_t) &= \mathcal{L}(\lambda_t(X_{t+1} - x_t) \mid \mathcal{H}_t) = \mathcal{L}\Big(w_t \frac{X_{t+1} - x_t}{x_t} \mid \mathcal{H}_t\Big) \\ &\approx \mathcal{L}(w_t Y_{t+1} \mid \mathcal{H}_t) = N(0, w_t^2\ \sigma_t^2), \end{aligned}$$

where the approximation refers to

$$lnX_{t+1} - lnx_t = \frac{X_{t+1} - x_t}{x_t} + o(X_{t+1} - x_t).$$

The generalization to a portfolio of (linear) assets is straightforward. Let w_t denote a $d-$dimensional exposure vector, i.e., $w_t = (\lambda_t^1 x_t^1, \cdots, \lambda_t^d x_t^d)$. Hence, the distribution of the random variable $w_t^T Y_{t+1}$ belongs to the family

$$\mathcal{P}_{t+1} = \{N(0, \sigma_t^2) : \sigma_t^2 \in [0, \infty)\}, \tag{1}$$

where $\sigma_t^2 = w_t^T \Sigma_t w_t$.
The aim of the VaR analysis is to estimate $\theta = \sigma_t$ and thereby to establish a prediction rule. For L_{t+1} we adopt therefore the following framework:

$$L_{t+1} = \sigma_t\ Z_{t+1} \tag{2}$$

$$Z_{t+1} \overset{iid}{\sim} N(0, 1) \tag{3}$$

$$\sigma_t^2 = w_t^T \Sigma_t w_t. \tag{4}$$

For a given $(n \times d)$ data matrix $\mathcal{X}_t = \{y_i\}_{i=t-n+1,\cdots,t.}$ of realisations of the underlying vector of returns with dimension d, two estimators for Σ_t will be considered. The first is a naive estimator, based on a rectangular moving average (RMA)

$$\hat{\Sigma}_t = \frac{1}{n} \mathcal{X}_t^T \mathcal{X}_t. \tag{5}$$

This definition of $\hat{\Sigma}_t$ makes sense since the expectation of Y_t is assumed zero. The second, also recommended by Taylor (1986) to forecast volatility, is

built by an exponential weighting scheme (EMA) applied to the data matrix $\tilde{\mathcal{X}}_t = \{\ diag(\lambda^d, \lambda^{d-1}, \cdots, \lambda, 1)^{1/2} y_i\}_{i=t-n+1,\cdots,t}$:

$$\hat{\Sigma}_t = (1-\lambda)\tilde{\mathcal{X}}_t^T \tilde{\mathcal{X}}_t \tag{6}$$

These estimates are plugged-into (4) and (2), yielding two prediction rules for

$$P_{t+1} \in \mathcal{P} = \{N(0, \sigma_t^2) \mid \sigma_t^2 \in [0, \infty)\}.$$

By their very nature VaR models contribute to several aspects of risk management. Hence, a series of parameters of interest - all derived from P_{t+1} - arise in natural ways. The particular choice is motivated by specific forecast tasks, e.g., driven by external (e.g., regulatory issues) or internal requirements or needs (e.g., VaR-limits, optimisation issues).
A very important part of risk management is the implementation of a systematic process for limiting risk. In the light of that task, it is at hand that forecast intervals defined by the $\widehat{VaR}_t$,

$$\widehat{VaR}_t = F_{t+1}^{-1}(\alpha) := inf\{x \mid F_{t+1}(x) \geq \alpha\},$$

where F_{t+1} denotes the cdf of P_{t+1}, are substantial.
If the main focus is to evaluate the forecast quality of the prediction rule associated to a VaR model, transformations of F_t should be considered, see Dawid (1984), Sellier-Moiseiwitsch (1993) and Crnkovic and Drachman (1996). For a given sequence of prediction-realisation pairs (P_t, l_t) - where l_t denotes a realisation of L_t - the prediction rules works fine if the sample $u = \{u_t\}_{t=1}^k = \{F_t(l_t)\}_{t=1}^k$ looks like an *iid* random sample from $U[0,1]$. A satisfactory forecast quality is often interpreted as an adequate VaR model. The focus of this paper is to consider the expected shortfall of L_{t+1}, as the parameter of interest and to derive backtesting methods related to this parameter - this will be done in the next section. The expected shortfall - also called tail VaR - is defined by

$$\begin{aligned} E(L_{t+1} \mid L_{t+1} > VaR_t) &= E(L_{t+1} \mid L_{t+1} > z_\alpha\ \sigma_t) && (7) \\ &= \sigma_t\ E(L_{t+1}/\sigma_t \mid L_{t+1}/\sigma_t > z_\alpha) && (8) \end{aligned}$$

where z_α is a α-quantile of a standard normal distribution. The motivation to consider this parameter is threefold. Firstly, McAllister and Mingo (1996) worked out the advantage of (7) compared to VaR if these parameters are plugged-into the denominator of a risk performance measures, e.g. a Sharpe-ratio or a RAROC (risk-adjusted return - that constitutes the numerator - on capital) numbers which are used to benchmark divisional performance, see Matten (1996) and CorporateMetrics (1999), - the economic motivation.

Secondly, Artzner et al. (1997) and **?** pointed out that (7) can be used as an approximation for the worst conditional expectation which is a coherent risk measure, a conceptual consideration. Thirdly, Leadbetter (1995) emphasized in the context of environmental regulation the need for incorporating the height of exceedances violating regulatory thresholds and critized those methods solely based on counts, neglecting the heights - statistical arguments. The paper is organised as follows. In the next section we present our approach on backtesting using the expected shortfall risk. In section 3 we apply this methodology to real data and visualise the difference betweeen RMA and EMA based VaRs. Section 4 presents the conclusions of this work.

2 Backtesting based on the expected shortfall

As pointed out by Baille and Bollerslev (1992), the accuracy of predictive distributions is critically dependent upon the knowledge of the correct (conditional) distribution of the innovations Z_t in (2). For given past returns $\mathcal{H}_t = \{y_t, y_{t-1}, \cdots, y_{t-n}\}$, σ_t in (4) can be estimated either by (5) or (6) and then $\mathcal{L}(L_{t+1} \mid \mathcal{H}_t) = N(0, \hat{\sigma}_t^2)$. Hence,

$$\mathcal{L}(L_{t+1}/\hat{\sigma}_t \mid \mathcal{H}_t) = N(0,1).$$

This motivates to standardize the observations l_t by the predicted STD, $\hat{\sigma}_t$,

$$\frac{l_{t+1}}{\hat{\sigma}_t}$$

and to interpret these as realisations of (2).

$$Z_{t+1} = \frac{L_{t+1}}{\sigma_t} \sim N(0,1) \tag{9}$$

For a fixed u we get for Z_{t+1} in (2)

$$\vartheta = E(Z_{t+1} \mid Z_{t+1} > u) = \frac{\varphi(u)}{1-\Phi(u)} \tag{10}$$

$$\varsigma^2 = Var(Z_{t+1} \mid Z_{t+1} > u) = 1 + u \cdot \vartheta - \vartheta^2 \tag{11}$$

where φ, Φ denotes the density, resp. the cdf of a standard normal distributed random variable.
For a given series of standardized forecast distributions and realisations,

$(F_{t+1}(\cdot/\hat{\sigma}_t), l_{t+1}/\hat{\sigma}_t)$ we consider (9) as parameter of interest. For fixed u, ϑ is estimated by

$$\hat{\vartheta} = \frac{\sum_{t=0}^{n} z_{t+1}\ I(z_{t+1} > u)}{\sum_{t=0}^{n} I(z_{t+1} > u)} \tag{12}$$

where z_{t+1} denotes the realisations of the variable (2). Inference about the statistical significance of $\hat{\vartheta} - \vartheta$ will be based on the following asymptotic relationship:

$$\sqrt{N(u)}\ \Big(\frac{\hat{\vartheta} - \vartheta}{\hat{\varsigma}}\Big) \xrightarrow{\mathcal{L}} N(0,1) \tag{13}$$

where $N(u)$ is the (random) number of exceedances over u and $\hat{\vartheta}$ is plugged-into (11) yielding an estimate $\hat{\varsigma}$ for ς. The convergence in (13) follows from an appropriate version of the CLT for a random number of summands in conjunction with Slutzky's Lemma, see Leadbetter (1995) for details. Under sufficient conditions and properly specified null hypothesis it is straight forward to prove the complete consistency and an asymptotic $\alpha-$level for a test based on (13), see Witting and Müller-Funk (1995), pp. 236.
Though these asymptotic results are straight forward they should be applied with care. Firstly, because the truncated variables involved have a shape close to an exponential distribution, hence, $\hat{\vartheta}$ will be also skewed for moderate sample sizes, implying that the convergence in (13) will be rather slow. Secondly, in the light of the skewness, outliers might occur. In such a case, they will have a strong impact on an inference based on (13) because the means in the nominator and in the denominator as well are not robust. The circumstance that the truncated variables' shape is close to an exponential distribution motivates classical tests for an exponential distribution as an alternative to (13).

3 Backtesting in Action

The Data The prediction-realisation (P_t, l_t) pairs to be analysed are stemming from a real bond portfolio of a German bank that was hold fixed over the two years 94 and 95,i.e., $w_t \equiv w$. For that particular (quasi) linear portfolio the assumptions met by (2) - (4) are reasonable and common practice in the line of RiskMetrics.
The VaR forecasts are based on a history $\mathcal{H}_t$ of 250 trading days and were calculated by two prediction rules for a 99%-level of significance. The first rule applies a RMA, the second is based an EMA with decay factor $\lambda = 0.94$ as proposed by RiskMetrics to calculate an estimate of $\hat{\Sigma}_t$ different from (5). Remembering the bond crisis in 1994, it is of particular interest to see how

these different forecast rules perform under that kind of stress. Their comparison will also highlight those difficulties to be faced with the expected shortfall if it would be applied e.g. in a RAROC framework.

Exploratory Statistics The following analysis is based on two distinctivce features in order to judge the difference of the quality of prediction rules by elementary exploratory means: calibration and resolution, see Murphy and Winkler (1987), Dawid (1984) and Sellier-Moiseiwitsch (1993). The exploratory tools are timeplots of prediction- realisation pairs (Fig. 1) and indicator variables (Fig. 4) for the exceedances to analyse the resolution and Q-Q-plots of the variable

$$\frac{L_{t+1}}{VaR_t} = \frac{L_{t+1}}{2.33\sigma_t} \tag{14}$$

to analyse the calibration (Fig 2, 3). A further motivation to consider variable (14) instead of (2) is that their realisations greater than one are just the exceedances of the VaR forcasts. Of course these realisations are of particular interest. If the predictions are perfect, the Q-Q-plot is a straight line and the range of the Y-coordinate of the observations should fill out the interval $[-1, 1]$. Hence, the Q-Q-plot for (14) visualises not only the calibration but also the height of exceedances. A comparison of Figure 2 with Figure 3 shows clearly that EMA predictions are better calibrated than RMA ones. The second feature, resolution, refers to the *iid* assumption, see Murphy and Winkler (1987). Clusters in the timeplots of exceedances, Figure 4,

$$(t, I(l_{t+1} > \widehat{VaR}_t)_{t=1}^{260}$$

indicate a serial correlation of exceedances. Again EMA outperforms RMA. ¿From Figure 1, we conclude that in 94 (95) 9 (4) exceedances were recorded for the EMA and 13 (3) for the RMA. Evidently, the window-length of 250 days causes an underestimation of risk for RMA if the market moves from a tranquile regime to a volatile one, and overestimates vice versa. On the other hand the exponential weighting scheme adapts changes of that kind much quickier.

The poor forecast performance, especially for the upper tail is evident. The asymmetry and outliers are caused by the market trend. For a particular day the VaR forecast is exceeded by almost 400 %. If the model (2) - (4) would be correct, the variable (14) has a STD of 0.41. The STD calculated from the data is 0.62. Hence, in terms of volatility the RMA underestimates risk on the average of about 50%.

The plot for EMA, Figure 3, shows the same characteristics as those in Figure 2 but the EMA yields a better calibration than RMA. The STD from

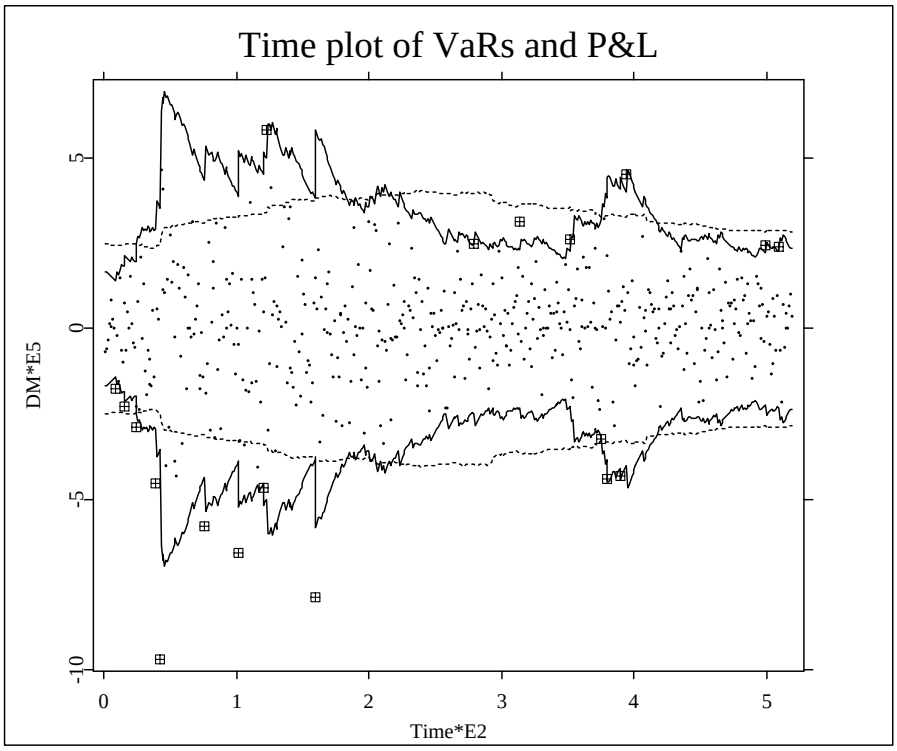

Figure 1: The dots show the observed change of the portfolio values, l_t. The dashed lines show the predicted VaRs based on RMA (99% and 1%). The solid lines show the same for EMA.

the data yields 0.5. Hence, an underestimation on the average of 25%. This indicates clearly that EMA gives a better calibration then RMA. Q-Q-plots for 95 are omitted. The two models give similar results, though even in that case the EMA is slightly better.

Inference The exploratory analysis has shown notable differences between the acurracy of RMA and EMA for the year 94. In this paragraph their statistical significance will be investigated. The inference will be based on the observations

$$\frac{l_{t+1}}{\hat{\sigma}_t}$$

and the underlying model (2) - (4). The threshold u is set to the 80%-quantile of L_{t+1}/σ_t yielding $\vartheta = 1.4$, by (10). Now, based on (13) an asymptotic

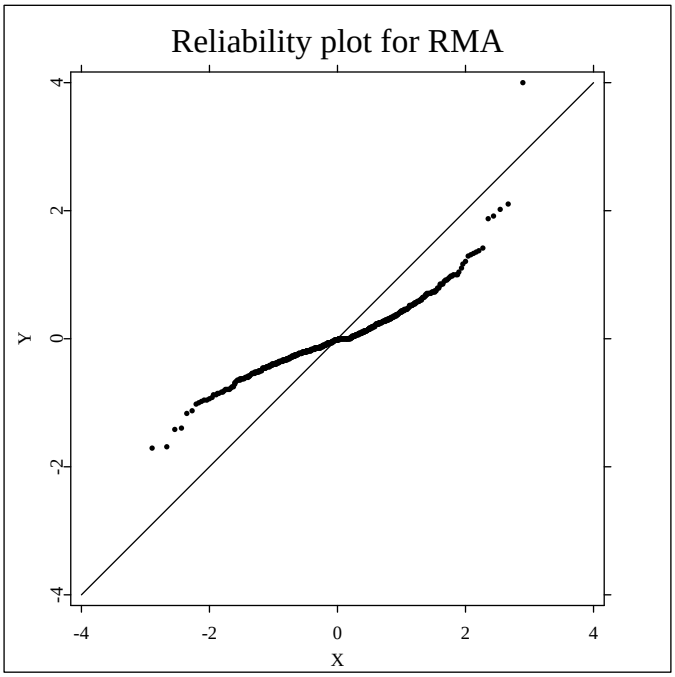

Figure 2: Q-Q plot of $l_{t+1}/\widehat{VaR}_t$ for RMA in 94.

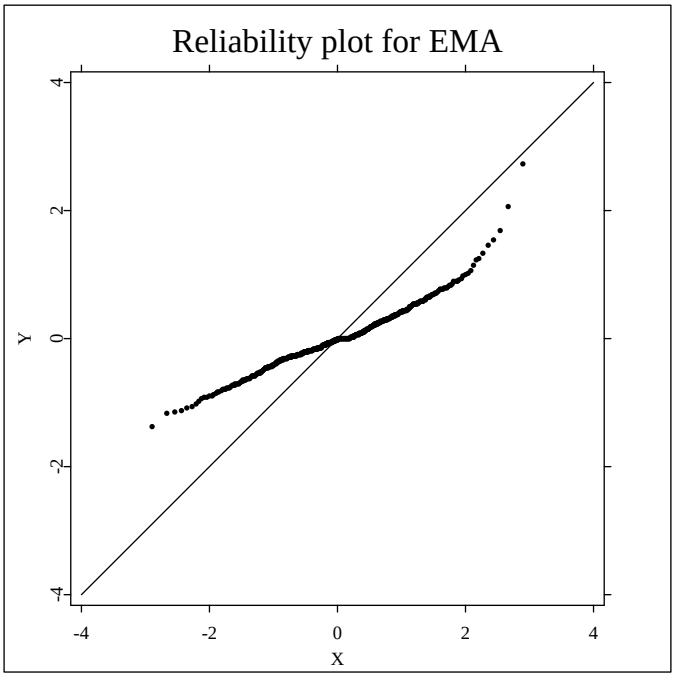

Figure 3: Q-Q plot of $l_{t+1}/\widehat{VaR}_t$ for EMA in 94.

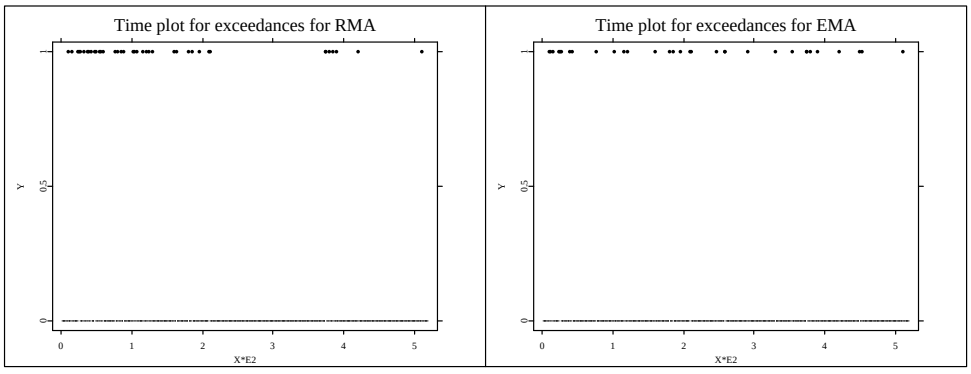

Figure 4: Timeplots of the exceedances over VaR of 80% level for RMA (left) and EMA. The better resolution of EMA is evident.

significance test for the hypothesis

$$H_0 \ : \vartheta \stackrel{(<)}{=} 1.4 \tag{15}$$

will be used. This setting - especially (2) - seems reasonable for RMA and the given sample of size $n = 250$.
As mentioned by Skouras and Dawid (1996) plug-in forecasting systems have the disadvantage that the uncertainty of the estimator for σ_t is not incorporated in the predictive distribution P_{t+1}. This applies especially to Z_{t+1} if the EMA is used. In that case a $t(n)$-distribution is indicated. A reasonable choice - motivated by generalized degrees of freedom - is

$$Z_{t+1} = \frac{L_{t+1}}{\sigma_t} \sim t(20). \tag{16}$$

Though the particular thresholds $u_N = 0.854$ - for the normal distribution - and $u_t = 0.86$ - for the $t(20)$ distribution differ only slightly (0.5 %), the associated means ϑ change about 5 % and the STD ς even about 18%. Parallel to (15) the hypothesis

$$H_0 \ : \vartheta \stackrel{(<)}{=} 1.47 \tag{17}$$

will be tested.
Tables 1 to 4 summarise the empirical results.

Method	$\vartheta = 1.4$	$\varsigma = 0.46$	$\frac{\sqrt{N(u)}(\hat{\vartheta}-\vartheta)}{\hat{\varsigma}}$	significance	nobs
EMA	$\hat{\vartheta} = 1.72$	$\hat{\varsigma} = 1.01$	2.44	0.75%	61
RMA	$\hat{\vartheta} = 1.94$	$\hat{\varsigma} = 1.3$	3.42	0.03%	68

Table 1: $H_0 \ : \vartheta \stackrel{(<)}{=} 1.4$

Method	$\vartheta = 1.47$	$\varsigma = 0.546$	$\frac{\sqrt{N(u)}(\hat{\vartheta}-\vartheta)}{\hat{\varsigma}}$	significance	nobs
EMA	$\hat{\vartheta} = 1.72$	$\hat{\varsigma} = 1.01$	2.01	2.3%	61
RMA	$\hat{\vartheta} = 1.94$	$\hat{\varsigma} = 1.3$	3.04	0.14%	68

Table 2: $H_0 \ : \vartheta \stackrel{(<)}{=} 1.47$

Firstly from tables 1 and 2, the observed exceedances over threshold u indicate again that the EMA is superior than the RMA. For a sample of 260

prediction-realisation pairs 52 exceedances are to be expected (STD 6.45). For the EMA 61 (61 - 52 $\approx$ 1.5 STD) exceedances were observed and 68 (68 - 52 $\approx$ 2.5 STD) for the RMA.
A comparison of table 1 with 2 shows that random errors strongly influence the significance of the test. Recalling the impressive outliers in the Q-Q-plots it is worthwile to exclude these from the data and re-run the test. The results are given in tables 3 and 4. Again, a serious change in the level

Method	$\vartheta = 1.4$	$\varsigma = 0.46$	$\frac{\sqrt{N(u)}(\hat{\vartheta}-\vartheta)}{\hat{\varsigma}}$	significance	nobs
EMA	$\hat{\vartheta} = 1.645$	$\hat{\varsigma} = 0.82$	2.31	1%	60
RMA	$\hat{\vartheta} = 1.83$	$\hat{\varsigma} = 0.93$	3.78	0.00%	67

Table 3: $H_0 : \vartheta \overset{(<)}{=} 1.4$ - largest outlier excluded

Method	$\vartheta = 1.47$	$\varsigma = 0.546$	$\frac{\sqrt{N(u)}(\hat{\vartheta}-\vartheta)}{\hat{\varsigma}}$	significance	nobs
EMA	$\hat{\vartheta} = 1.645$	$\hat{\varsigma} = 0.82$	1.65	5%	60
RMA	$\hat{\vartheta} = 1.83$	$\hat{\varsigma} = 0.93$	3.1	0.15%	67

Table 4: $H_0 : \vartheta \overset{(<)}{=} 1.47$ - largest outlier excluded

of significance for the RMA is observed indicating the non robustness of the test. These results show furthermore that inference about the tails of a distribution is subtle. In addition the *iid* assumption - cluster of exceedances - might also be violated. One possible source for that is the overlap of the $\mathcal{H}_t$. Hence, the estimates may correlate. Techniques like moving blocks and resampling methods see **?** and **?** are good remedies.
To overcome the problems related to the slow convergence of (13) an exponential distribution may be fitted to the data and then, again a classical test will be applied. The following table reports the significance levels based on a one-sided Kolmogoroff-Smirnov test. Again, the results emphasize the impact of random errors. The number in brackets refers to that case, where the largest outlier is deleted.

4 Conclusions

VaR models were introduced as specific statistical forecast systems. The backtesting procedure was formulated in terms of measuring forecast qual-

Method	$\sigma = 0.46$	$\sigma = 0.546$
EMA	0.25%	10% (14%)
RMA	$< 0.1\%$	$< 0.1\%$

Table 5: Kolmogoroff-Smirnov Test

ity. The empirical results highlight the better calibration and resolution of VaR forecasts based on (exponentially weights) EMA compared to (uniformly weights) RMA. However, more interesting is the impressive difference in amount (50%). A surprising result is the strong dependence of inferences based on expected shortfall from the underlying distribution. Hence, if expected shortfall will be used in practice in order to calculate performance measures like RAROC the inferences resp. the estimates should be robustified.

Acknowledgements: The authors would like to express their warmest thanks to Zdeněk Hlávka for his help by providing the graphics in XploRe. They also wish to thank for the support by the Sonderforschungsbereich 373. Last but not least the second author disclaims that the views expressed herein should not be construed as being endorsed by the Bundesaufsichtsamt.

References

Artzner, P., Dealban, F., Eber, F.-J. & Heath, D. (1997) Thinking Coherently, *RISK MAGAZINE*.

Baille, R. T. & T. Bollerslev (1992) Prediction in Dynamic Models with Time-Dependent Conditional Variances. *Econometrica*, **50**: 91–114.

Crnkovic, C. & J. Drachman (1996) A Universal Tool to Discriminate Among Risk Measurement Techniques, *RISK MAGAZINE*.

Dawid, A. P. (1984) The prequential approach. *J. R. Statist. Soc.*, A, **147**: 278–292.

Härdle, W. & Klinke, S. & Müller, M. (1999) XloRe Learning Guide. www.xplore-stat.de, Springer Verlag, Heidelberg.

Jaschke, S. & Küchler, U. (1999) Coherent Risk Measures, Valuation Bounds, and $(\nu, p)-$Portfolio Optimazation. *Discussion Paper*, **No 64**, Sonderforschungsbereich 373 of the Humboldt Universiät zu Berlin

Leadbetter, M. R. (1995) On high level exceedance modeling and tail inference. *Journal of Planning and Inference*, **45**: 247–260.

Matten, C. (1996) *Managing Bank Capital*. John Wiley & Sons: Chicheseter.

McAllister, P. H. & J.J. Mingo (1996) Bank Capital requirements for securitzed loan portfolios. *Journal of Banking and Finance*, **20**: 1381–1405.

Murphy, A. H. & R. L. Winkler (1987) A General Framework for Forecast Verification. *Monthly Weather Review*, **115**: 1330–1338.

RiskMetrics (1996) Technical Dokument, 4th Ed.

CorporateMetrics (1999) Technical Dokument, 1st. Ed.

Sellier-Moiseiwitsch, F. (1993) Sequential Probability Forecasts and the Probability Integral Transform. *Int. Stat. Rev.*, **61**: 395–408.

Skouras, K. and A. P. Dawid (1996) On efficient Probability Forecasting Systems. *Research Report* **No. 159**, Dep. of Statistical Science, University College London.

Taylor, S. J. (1986) *Modelling Financial Time Series*. Wiley, Chichester.

Witting H. and U. Müller-Funk (1995) *Mathematische Statistik II*. Teubner, Stuttgart.