BACKTESTING USING A GENERALISATION OF THE TRAFFIC-LIGHT-APPROACH

*Gerhard Stahl**

Abstract: In an important regulatory innovation the Basle Commitee on Banking Supervision has allowed for banks to use their own internal models - so-called Value-at-Risk (VaR) models - for setting capital requirements to back their trading activities. In order to validate VaR models a quality check has to be run. This backtesting exercise is also a cornerstone in the regulatory framework for internal models.

This article reviews both current backtesting methods as proposed by regulators and those of practitioners as well and furthermore employs some refinements of current methods.

Key words: *Backtesting, goodness-of-fit tests, exploratory methods*

Received: May 2, 1997
Revised and accepted: September 8, 1997

1. Introduction

Value-at-Risk (VaR) models have been accepted by banking regulators as tools for setting capital requirements for market risk exposure. The basic idea behind the use of these models for internal decision making process and the setting of capital requirement purposes is to calculate an upper limit for price movements in the underlying instruments (i.e. the market risk) and derive from this figure an upper limit for the risk of loss inherent to a position, which could be realised within a given prespecified probability, called level of "confidence", for a fixed period of time (holding period). An in-depth description of the theory of VaR-models is given in [1] and [2]. Many banks that have adopted an internal model-based approach to market risk measurement routinely compare daily profits and losses with model generated risk measures to gauge the quality and accuracy of their risk measurement systems. This process known as backtesting is not only a useful exercise per se, but also a cornerstone of the BASLE COMMITTEE ON BANKING SUPERVISION (BIS) framework for internal models, as described in [3]. As a technique for evaluating the quality of a firm's risk management model, backtesting continues to evolve. This paper consists of 5 sections. After some preliminary and introductory remarks in section 1 and 2 section 3 is devoted to review [4], [5],

*Gerhard Stahl
Suppervisory Office, German Banking, Gardeschutzenweg 71-101, D-12203 Berlin, Germany

[6], [7], mirroring the state of current research on backtesting. The second part is separated by section 4 and section 5. In section 4 we review and define some nonparametric tests, which give rise to generalisations of the BIS' backtesting rules. In section 5 we present some tools of exploratory statistics order to complement and strengthen the jugdement based on classical inferential backtesting procedures of the preceding section.

2. Notation and Motivation

Consider the simplest case of a linear portfolio, this means, that the potential change in the market value of the portfolio is,

$$\pi_t = \sum_{i=1}^{N} w_{i,t} r_{i,t} = w_t r_t,$$

where $w_t = (w_{1,t}, \cdots, w_{N,t})'$ denotes a vector of portfolio weights, i.e. the exposure of the portfolio to risk factor i and $r_t = (r_{1,t}, \cdots, r_{N,t})$ denotes a vector of risk factors or assets; $w_t, r_t \in \mathbb{R}$. As usual, the components of r_t are relative (or logarithmic) price changes of the underlying instruments. Subscript t denotes a discrete (daily) time index and x' the transpose of vector x.

In the context of VaR models some, may be even all, of the following assumptions are usually encountered:

- A1: The portfolio consists of linear or linearised instruments.
- A2: It is assumed, that r_t is a stationary Gaussian process with mean vector zero, e.g. $r_t \sim N(O, \Sigma_t)$.
- A3: The realizations of r_t are *iid*.
- A4: The random variables π_t are independent.

Assumption A1 implies that the total risk of the portfolio is just

$$\sigma_t^2 = w_t' \Sigma_t w_t. \tag{1}$$

As an immediate consequence of A2, π_t is a univariate normally distributed random variable

$$\pi_t \sim N(0, \sigma_t^2).$$

We denote with h_t and H_t the density and the cumulative distribution function (cdf) of any - not necessarily normally distributed - random variable π_t. The VaR of π_t , denoted by v_t , is defined as the α-quantile of the random variable π_t, i.e the root of

$$\int_{-\infty}^{v_t} h_t(-x)dx = \alpha,$$

where α is a given level of confidence, the BIS demands a 99% -level. Of course, h_t is in general not known and has to be estimated from data. Together with the

specific VaR models there are a lot of natural estimates, in the following denoted by $\hat{h}_t$ and $\hat{H}_t$ respectively. Given some estimate $\hat{h}_t$ or $\hat{H}_t$, we may calculate $\hat{v}_t$, the estimate of v_t. In the framework of the BIS' rules the estimates have to be based on $T \geq 250$ observations, where the latest observation is sampled at time $t-1$. The observed trading loss at time t is denoted by l_t, i.e. $l_t = \pi_t(\omega)$.

We want to stress, that any kind of VaR calculations depends on assumptions closeley related to ours. From a general point of view A1 reads as an assumption, that assures that the observed P&L are realisations of the P&L forecast, the so-called clean backtesting. If nonlinear instruments are included, the details relating used approximations of price functions and the observed prices deserve deeper considerations. We mention furthermore, that A1 still covers a number of important instruments like $\mathbb{R}$ swaps, bonds, FRNs, FX-forwards, FRAs, currency swaps and options with small gamma. Hence, linear portfolios might serve as a proxy, if the non-linear instruments do not contribute too much risk for portfolio at hand.

If A2 is read in a flexible way, it covers not only the unconditional gaussian framework for the innovations of a (multivariate) random walk model, which fails to fit most financial market data. This is especially true, if the level of time aggregation is beyond two weeks, see [8] for examples in FX market. Readers, that are familiar with conditional models to capture the phenomena of leptokurtosis and heteroscedasticity, see [9], might switch to conditional normal distributions.

Besides these two equation models an approach based on elliptically symmetric distributions, c.f. [11], is also worth to be considered. These models generalise the normal distribution to a class of transformation models, keeping many calculation rules valid. An application of these models is given in [1], where mixtures of normal distributions are fitted to the innovations.

Assumption 4 is severe in respect to backtesting purposes. The independence of the π_t is clearly an important statistical premise, but usually violated in practice. This may be explained by the weights w_t that depend in general on the estimated covariance matrix through some optimizing process. In this case, the estimates $\hat{v}_t$ tend to be biased. As a simulation study in [10] has shown, this bias may be of considerable amount. A second source also responsible for correlation of $\hat{v}_t$, is the persistence in the estimates $\hat{\Sigma}_t$.

In addition to independence the backtesting is sharpened by day by day changing portfolio weights w_t. From (1) we conclude that the weights w_t act as a (linear) transformation on $N(O, \Sigma_t)$. This has the important implication, that the distributions H_t of π_t are not time invariant, i.e.

$$H_t \neq H_s \quad \text{for} \quad t \neq s \tag{2}$$

Aside from the assumptions mentioned above, practitioners neglect estimation errors by assuming implicitly

$$\hat{\Sigma}_t = \Sigma_t \quad \text{resp.} \quad \hat{H}_t = H_t. \tag{3}$$

Such identifications might be justified by the means of asymptotic theorems as the strong law of large numbers, but the investigations of [10] and [13] indicate, that this may be very optimistic in the field of VaR models. To summarise so far, backtesting is complicated by the following obstacles:

1. the variables π_t are in general not identically distributed,

2. the variables π_t and the estimators $\hat{v}_t$ are not independent, they are autocorrelated,

3. the involved distributions of π_t and r_t are estimated.

In the following we refer to these difficulties as the problem of aggregation, independence and estimation error.

3. Previous Research on Backtesting

The fundamental critique of [5] on various backtesting methods hits parametric and nonparametric VaR-models as well. It's emphasis is put on parametric VaR-models and on such backtesting methods which are based on various likelihood ratio (LR) statistics that use observations from a Bernoullian process, where A4 is assumed.

For given T and α the probability to observe a proportion $x = \#(l_t > \hat{v}_t)$, where $\#(l_t > \hat{v}_t)$ is read, the number of l_t 's greater than v_t, is

$$P(X = x) = \binom{T}{x} a^x (1-\alpha)^{T-x}.$$

By the Lemma of Neyman and Pearson [13] the statistic

$$\log \lambda(x) = 2 \log \left((x/T)^x (1 - x/T)^{T-x} \right) - \log \left(\alpha_0^x (1-\alpha_0)^{T-x} \right)$$

defines a uniformly most powerful test for testing

$$H_0 : \alpha \geq \alpha_0 \quad \text{against} \quad H_1 : \alpha < \alpha_0.$$

The investigation of the LR-tests' power is the focus of [5]. We quote only one, brit typical example, where $H_0 : \alpha_0 = 0.99$ is tested against $H_1 : \alpha = 0.98$ for $T = 255, 510, 1000$. The resulting type II errors are: 0.749, 0.557 and 0.218. This seems to be a serious blow against elementary backtesting strategies. In the light of such verdicts, there is a need for methods that take all available information into account. A very important innovation towards this goal is [4], where a generalisation of the binomial test is proposed and A4 is checked by the BDS-test [16]. This generalisation employs a test, that is based on the whole forecast distribution and not just one quantile, as in the binomial case. The transformed sequence $H_t(\pi_t)$ delivers the probabilistic framework, and solves at the same time the aggregation problem. It is easily seen, that under A4,

$$\Pi_t \quad iid \quad U, \tag{4}$$

where U denotes the uniform distribution on $[0, 1]$ and $\Pi_t := H_t(\pi_t)$. Now, the diagonal in the unit square may be used as a yardstick to judge the VaR model's acurracy.

If a VaR model works perfectly, i.e. true and estimated distributions of π_t coincide, the ordered pairs $\hat{\Pi}_t, t$ are scattered around this diagonal, where $\hat{\Pi}_t$ denotes a realisation of Π_t. Assume, the VaR model works acurrately, then the deviations

from the diagonal are purely random. Hence, large deviations from this ideal line are improbable events indicating an imprecise VaR forecast. To measure the deviation, an appropriate distance function δ on the space of all cdfs has to be defined, which can be used as a test statistic of a goodness-of-fit test.

Common choices for δ are variants of the Kolmogorov-Smirnov's (KS) supremum criterion D_T and integral criteria which are connected with the names of von Mises and Cramér. Let us first consider the supremum criterion of KS. Integral criteria will be considered in section 4.

Suppose, the random variables $X_1, \cdots, X_t$ are iid F, then the empirical cumulative distribution function (ecdf) is defined as

$$F_{(T)}(x) = \#(X_j \leq x)/T, \qquad -\infty < x < \infty.$$

The Kolmogorov-Smirnov (KS) statistics

$$\begin{aligned} D_T &:= \sup_x \left|F_{(T)}(x) - F(x)\right| \\ D_T^+ &:= \sup_x \left(F_{(T)}(x) - F(x)\right) \\ D_T^- &:= \sup_x \left(F(x) - F_{(T)}(x)\right) \end{aligned}$$

and Kuiper's statistic

$$K_T := D_T^+ + D_T^-$$

are well-known distance measures. Kuiper's statistic in [4] is preferred to the statistics of KS-type, in order to test the hypothesis

$$H_0 : F(x) = x \quad \text{against} \quad H_1 : F(x) \neq x, \tag{5}$$

where $F(x)$ denotes the cdf of Π_t and $f(x) \equiv x$ is the cdf of U.

A weight function $w(x)$ may be used to capture the relevance of the distributions' tails, [4] advocate

$$w(x) = -0.5 \ln(x(1-x)).$$

In contrast to the methods in [4] and [5], which are based on inferences from testing hypothesises, the methodologies in [6] and [7] switch to an estimation framework, under A4.

The approaches in [6] follow decision theoretic and Bayesian thoughts as outlined in [15]. These concepts intend to incorporate all information of both objective and subjective nature, to draw an adequate picture of the risk in order to make the best decision. Such procedures incorporate the information given by the data as well as other sources of non experimental information such as (asymmetrical) loss functions and priors on the parameter space Θ. The main ideas in [7] are inspired by the work of [17], which is devoted to the problem of evaluating weather forecasts. The point is to specify a (regulatory) loss function Q and an event of interest - in other words a parameter of interest - in order to measure the VaR's

accuracy. The accuracy of the VaR forecasts is gauged by how well they minimize this loss function. The selected quadratic loss function is defined by

$$Q = \frac{1}{T}\sum_{t=1}^{T} 2(P_t - R_t)^2. \tag{6}$$

R_t denotes an indicator function of the regulator's event and P_t denotes the forecast of the event of interest, that depends on the estimated distribution of P_t.

4. Backtesting Based on Testing Stochastic Dominance

As shown in the preceding sections the backtesting problem is complex. The simultaneous presence of the problem of aggregation, independence and estimation error, makes it difficult, to strictly statisfy the assumptions of the statistical procedures involved. The various measures of accuracy applied to VaR estimates in [10] highlight, that backtesting methods solely based exclusively on the 99% quantile fail, to give reliable results. In this aspect we agree with the criticism in [4] and [5]. Therefore, we pursue such inferential methods with a sound statistical basis and statistics that allow for a meaningful interpretation within the field of risk management. Such statistics could be analysed by means of exploratory and confirmatory techniques, [18] and [19].

One of risk management's cornerstones is historical volatility. This statistical measure expresses the riskiness of a portfolio by a single figure, usually the standard deviation. The standard deviation is a special case of the more general concept of a random variable's dispersion around the mean. To introduce the latter, we consider two symmetric variables say Y and X with mean μ and η respectively. In the context of VaR models we may assume $\mu = \eta = 0$, and say that Y more dispersed than X, if

$$P(|X| > t) \leq P(|Y| > t) \quad \text{for all} \quad t. \tag{7}$$

A shorthand notation for (7) is $|X| <_{st} |Y|$. If furthermore

$$X <_{st} Y \tag{8}$$

is fulfilled, we say, that X is stochastically dominated by Y; [20] gives a good survey on this topic. Obviously, the concepts of more dispersed and stochastic dominance coincide, if X and Y are symmetric about zero. In terms of the involved cdfs, relation (8) is equivalent to (9)

$$F_X(t) \geq F_Y(t) \quad \text{for all} \quad t \in \mathbb{R}. \tag{9}$$

To tie volatility with dispersion, we introduce a class of general dispersion measures in some detail. Following [21] and [22], we define the functionals

$$\tau(F) = \left\{ \int_0^1 \left[F^{-1}_{|X-\mu|}(x) \right]^\gamma d\Lambda(x) \right\}^{\frac{1}{\gamma}}, \tag{10}$$

where F is assumed to be symmetric about μ, F, denotes the distribution of $|X-\mu|$, Λ is any probability distribution on $(0,1)$ and γ any positive number. We focus

on two important cases of (10): the standard deviation ($\gamma = 2, \Lambda = U(0,1)$) and the α-quantile (Λ assign probability 1 to α, γ is arbitrary). Obviously, volatility and value-at-risk are measures of dispersion. As outlined in [21], (7) and (10) are related by

$$\tau(F) \leq \tau(G) \quad \text{whenever} \quad F_X(t) \geq G_Y(t) \quad \text{for all} \quad t \in \mathbb{R}. \tag{11}$$

Now, in the light of (11), the concept of stochastic dominance provides a non-parametric framework, to interpret Y as more volatile, and hence more riskier, than X. This relation can be used, to compare VaR-forecast distributions.

Kuiper's test statistic is symmetric and rejects those models which deviate too much from the diagonal line, even those, which are based on too conservative forecasts. From a regulator's but also from a risk manager's point of view cautious and conservative VaR estimates have an intrinsic value. Let us therefore consider a statistician whose risk estimate of π_t is based on a conservative statistical model G_t that dominates the true distribution H_t.

$$G_t(x) \leq H_t(x) \tag{12}$$

Let C_t denote the random variable $G_t(\pi_t)$, and F_t its cdf. Then (12) implies

$$C_t <_{st} \Pi_t, \tag{13}$$

equivalent to

$$F_t(x) \geq x.$$

As mentioned earlier, the Π_t are *iid* U. Whereas, the random variables C_t are not necessarily identically distributed. However their mixture $M := 1/T \sum_{t=1}^{T} F_t$ still satisfies

$$M(x) \geq x \quad \text{for all} \quad x \in (0,1)$$

i.e. the mixture of cdfs dominating the uniform's cdf, also dominates x. Under the assumption that the random variables C_t are independent, we consider one-sided goodness-of-fit tests to measure the agreement of the data with the composite null hypothesis, that the VaR forecasts are conservative estimates:

$$H_0 : M(x) \geq x \tag{14}$$

against the composite alternative

$$H_1 : M(x) < x \quad \text{for some} \quad x \in (0,1).$$

To run a goodness-of-fit test, we have to define a statistic δ that rejects H_0 if the distance of the ecdf $F_{(T)}$ to H_0 is too large. Let us first consider the KS-type statistic

$$\delta(F, F_{(T)}) = D_T^-.$$

We have to solve two problems, in order to run a goodness-of-fit test based on D_T^- for (14). Firstly, we have to determine a null distribution $\hat{F}_0$ for data under H_0. This is necessary, because H_0 in (14) is composite. Secondly we have to determine the distribution of D_T^- under $\hat{F}_0$. This is achieved with the help of the bootstrap method [24] applied to non parametric tests [25]. We estimate $\hat{F}_0$ by

$$\hat{F}_0(x) := \max(x, F_{(T)}(x)) \quad x \in (0,1). \tag{15}$$

This estimate $\hat{F}_0 \in H_0$ is well justified. It is a nonparametric maximum likelihood estimate in the sense of [26]. Now, the bootstrap machinery works, in order to test (14).

We follow [25] and [27] by bootstrapping D_T^- under H_0, i.e. we calculate

$$S_T^* = \sup_x (F_{(T)}^* - \hat{F}_0) \tag{16}$$

by resampling conditionally on $\hat{F}_0$, where $F_{(T)}^*$ denotes the ecdf of sample size T from $\hat{F}_0$. Define C_T^* as the cdf of S_T^*

$$C_T^*(s) = P(S_T^* \leq s | \hat{F}_0)$$

and

$$q_T^* = \inf \{s : C_T^*(s) \geq 1 - \alpha\},$$

then the bootstrap rejection region for H_0 is just the area below the graph of

$$\hat{F}_0 - q_T^*.$$

In other words the null hypothesis is rejected, whenever there is an $x \in (0,1)$, such that

$$F_{(T)}(x) > \hat{F}_0(x) - q_T^*.$$

For appropriate choices of α (e.g.$\alpha = 0.01, 0.05$) we get a generalisation of the traffic light approach outlined in [3]. When the KS-test is applied with the focus on the tails, we encounter its weakness: the constant vertical widths. This makes the band unnecessarily broad in the tails [29]. We can either apply a weighting function $w(x)$ within the class of KS statistics or switch to statistics, that are based on the integral criterion. We continue with the first strategy and consider the second there after. Natural choices for $w(x)$, that yield Rényi's type statistics, are

$$R_T^- = \sup_{F(x)>b} \left(\frac{F(x) - F_{(T)}(x)}{F_{(x)}} \right)$$

or

$$R_T^D = \sup_{F(x)>b} \left(\frac{F(x) - F_{(T)}(x)}{F(x)(1 - F(x))} \right),$$

see [30], [31]. Usually Rényi type statistics yield narrower confidence bands especially in the tails of F, which was our crucial aim. Furthermore R_T^- admits the interpretation of the maximal relative error estimating $F(x)$ by F_T over a certain range. R_T^D is a variance-weighted version of D_T^-. The bootstrap applies in the same manner as outlined above. More details on Rényi's and related statistics are found in [28] and [32].

We focused so far on supremum tests, which exploit only the extreme deviation between F_T and F.

Hence, these tests have the disadvantage to be sensitive only with respect to the greatest difference. The family of statistics (for details see [28] and [33]) based on the integral criterion are a smooth measure of discrepancy. As a typical example we consider the von Mises statistic

$$MS = T \int w(x)[F_{(T)}(x) - F(x)]^\gamma \quad dx, \tag{17}$$

where $w(x)$ is a suitable weight function and $\gamma \in \mathbb{R}^+$, usually $\gamma \in \{1, 2\}$. As mentioned in [31], this two-sided statistic (17) is not appropriate for our one-sided setting. Therefore, we propose a modification of (17) by

$$\bar{MS} = T \int w(x)(\max[0, F_{(T)}(x) - F(x)])^\gamma \quad dx \tag{18}$$

in order to test

$$H_0 : M(x) \geq x$$

against

$$H_1 : M(x) < x \quad \text{for all} \quad x \quad \text{on some interval} \quad I \subset (0, 1).$$

Well-known specifications of (17) are: Moses' test ($\gamma = 1, \quad w(x) \equiv 1$) and the von Mises test ($\gamma = 2, \quad w(x) \equiv 1$).

In this context we want to mention Fisher's test [30], [31]

$$Fi_T = \int F_{(T)}(x)/F(x)dF(x),$$

which is closely related to R_T^-. We want to point out, that Fi_T contains no parameter to be determined by the statistician, in contrary to the parameter b in R_T^+. Of course, Fi_T may be modified analogously to (18).

The test procedure is carried out within the same bootstrap framework as outlined before. It is to be expected that a test based on MS admits a power superior to those of KS-type.

We conclude this section with some general remarks on the procedures we have looked at so far. The bootstrap tests above are closely related to so-called Monte Carlo tests, see [34], [35] and [36]. The latter refer usually, but not exclusively to Monte Carlo simulations in the framework of parametric tests. Bootstrap-tests are recommended [24], if the alternative hypothesis is composite. We see two concrete advantages applying the bootstrap. Firstly, it is a problem to determine distributions of test statistics, if an arbitrary weight function $w(x)$ is involved.

Secondly, certain difficulties with estimated parameters [37] of the null hypothesis may be circumvented.

We further want to stress, that we have not adressed the important problem of *iid* assumption. It seems reasonable that exceedences of VaR limits are positively correlated. Under the assumption of positive correlation, it is shown in [38], that edf-tests and χ^2-tests reject the true null hypothesis too often, by confounding a positive dependence with lack of fit. These results apply to Pearson, Kolmogorov-Smirnov and Cramér-von Mises type tests. We see two strategies, to overcome the difficulties caused by dependence. Firstly, to find a model, describing the dependence. Secondly, to rule out the dependence with methods of moving blocks, see [39], [40], [41], [42]. Our next paper on backtesting will be devoted to these topics.

5. Exploratory Analysis and Backtesting

In addition to our remarks with respect to the assumptions of VaR calculations we are confronted with another kind of problem: the immense dimension of the portfolio. In practice, we encountered implemented VaR models including between 150 and 4000 risk factors. Though it is expected, that such high dimensional models still capture important features and structures of the data, the nominal levels of significance of calculated forecast intervals should be interpreted with care.

Taking these critical points together, we see a strong indication for applying exploratory methods [18] in addition to the inference procedures considered in section 3 and 4. Exploratory methods take the data as such without further probalistic assumptions.

The problem, to evaluate weather forecasts [45] is closely related to the backtesting problem of VaR models. These problems coincide - at least in principle -, if the target variable is continuous, but differ in the following detail.

Meteorologists make probability statements about forecasted events, such as: with a probability of 10 %, tomorrows temperature is between 25°C and 26°C. Hence, they deal with two distributions, the forecast distribution and the distribution of the observations. Applying appropriate grids to the variable's domain leads to χ^2-tests, an important tool in the area of weather forecast evaluation [45], [46]. If the forecast distributions of VaR were standardised, these methods could be also applied, the analysis of interplay between the two distributions would give valuable insights [45]. Surprisingly, meteorologists plot so-called reliability diagrams. In such plots, the observed relative frequency of a hit within each forecast probability category is plotted against forecasted probability, with the 45° diagonal line representing perfect reliability [47]. But this is nothing else but an exploratory interpretation of what we have done in the last section. Further graphical applications are P-P and Q-Q plots [43].

It is recommended to apply the bootstrap as a confirmatory mean in an exploratory framework. For example [44] suggests to use bootstrapping to complement the edf, because bootstrapped estimates of bias an standard deviations of quantiles can help the decision maker locate the order statistics about which he is most uncertain.

We close our tentative remarks on exploratory backtesting methods by proposing a score function, which mirrors the average amount of forecast exceedances. The statistics in [4] and [5] and those introduced in section 4 are based on observed percentiles given by the aggregation-step (4),

$$\hat{H}_t(l_t).$$

These statistics lost an important information given in the difference resp. the ratio

$$\hat{v}_t - l_t \quad \text{resp.} \quad v_t/l_t$$

about the amount of how much the observed loss surpassed the VaR forecast.

With $\hat{\mathbf{p}}_T$ we denote a sample of T observed percentiles $\hat{p}_t = \hat{F}_t(l_t)$. With $\hat{\mathbf{p}}_T = (p_{(1)}, \cdots, p_{(T)})$ we denote the associated order statistic, the basis of the edf-tests. Now, we calculate

$$\tilde{v}_t := \hat{H}_t^{-1}(\hat{p}_{(t)},$$

the VaR of level $\hat{p}_{(t)}$ at time t. If the ratio

$$\lambda_t := \hat{v}_t/l_t$$

is greater than 1, the VaR estimate is conservative at time t, else it fails to be conservative. We propose the score function:

$$S = \sum_{i=1}^{T} q^{T-i} \ln(\lambda_i)$$

where q defines a geometric weighting scheme. A value $S > 0$ indicates conservativism of the VaR model and a value $S < 0$ indicates the contrary.

6. Summary

We gave a tour d' hôrizon of actual applied and possible future backtesting methods. We touched classical estimation and testing inferential procedures as well as exploratory methods. We take the point of view that there exists so far no definite method that hits all purposes. But the whole spectrum of methods should give a distinctive conclusion about the VaR model's forecasting quality. Positive results of a fourty year experience in meteorology for related questions should encourage us to look optimistically to future backtesting of VaR models. We will apply our methods to the real VaR data, which were calculated by means of exponentially weighted and equally weighted observations.

Acknowledgement

Gerhard Stahl is a statistician with the Bundesaufsichtsamt für das Kreditwesen (Federal Banking Supervisory Office), Berlin. This article is an extended elaboration of the author's talk given at the Research Workshop on "Internal Models" held

at the Deutsche Bundesbank, Frankfurt, 14-16 October 1996. It is a great pleasure for the author to express his warmest thanks to Professor Huschens, University of Dresden, for his current support, very stimulating and patient discussions on various topics of VaR models. Of course, the author is solely responsible for all remaining errors. The views expressed in this paper are the author's very private opinions. These should not be quoted as those of the Bundesaufsichtsamt.

References

[1] Longerstaey J.: RiskMetrics. Technical Dokument, 3rd, 4th Ed., 1995, 1996.

[2] Jorion P.: Value At Risk. IRWIN, Chicago, 1997.

[3] Basle Committee on Banking Supervision: Proposal to Issue a Supplement to the Basle Capital Accord to Cover Market Risks. Basle, December, 1996.

[4] Crnkovic C., Drachman J.: A Universal Tool to Discriminate Among Risk Measurement Techniques. RISK, 1996.

[5] Kupiec P.: Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives, December 1995.

[6] Stahl G.: Backtesting or Backestimating. Proceedings of European SAS user conference 1996, Hamburg.

[7] Lopez J.A.: Regulatory Evaluation of Value-at-Risk Models. Preprint, Federal Reserve Bank New York, 1996.

[8] Baille R., McMahan P.: The foreign exchange market. Cambridge University Press, 1989.

[9] Bollerslev T., Chou R., Kroner K.: ARCH Modelling in Finance: A Review of the Theory and Empirical Evidence. Journal of Econometrics, **52**, 1992, 5-59.

[10] Davé R., Stahl G.: On the accuracy of VaR estimates. Forthcoming Proceedings of the 6-th Econometric Workshop of the University of Karlsruhe, Germany, 1997.

[11] Fang K.T., Kotz S., Ng K.W.: Symmetric Multivariate an Related Distributions. Chapman Hall, London, 1990.

[12] Jamshidian F., Zhu Y.: Scenario Simulation: Theory and Methodology. Finance and Stochastics, **1**, 1997, 43-67.

[13] Huschens S.: Confidence Intervals of Value at Risk Estimates. Forthcoming Proceedings of the 6-th Econometric Workshop of the University of Karlsruhe, Germany, 1997.

[14] Lehmann E.L.: Testing Statistical Hypotheses. Wiley, New York, 1986.

[15] Berger J.O.: Statistical Decision Theory and Bayesian Analysis. 2nd Ed. Springer New York, 1985.

[16] Brock W.A., Hsieh D.A., LeBaron B.: Nonlinear Dynamics, Chaos and Instability: Statistical Theory and Economic Evidence. MIT Press, Cambride, Massachusetts, 1993.

[17] Brier G.W.: Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Reviev, **75**, 1950, 1-3.

[18] Tukey J.W.: The Future of Data Analysis. Annals of Mathematical Statistics, **30**, 1962, 1-67.

[19] Tukey J.W.: We need both exploratory and confirmatory. The American Statistician, **34**, 1980, 23-25.

[20] Levy H.: Stochastic Dominance and Expected Utility: Survey and Analysis. Management Science, **38**, 1992, 555-593.

[21] Bickel P.J., Lehmann E.L.: Descriptive Statistics For Nonparametric Models. III Dispersion. Annals of Statistics, 4, 1976, 1139-1158.

[22] Bickel P.J., Lehmann E.L.: Descriptive Statistics For Nonparametric Models. IV Spread. In Contributions to Statistics, Jaroslaw Hajek Memorial Volume, J. Jureckova, ed. Reidel, Dodrecht, 1979, 33-40.

[23] Lehmann E.L.: Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, Oakland, 1975.

[24] Efron B., Tibshirani R. J.: An Introduction to the Bootstrap. Chapman Hall, London, 1993.

[25] Romano R.P.: A Bootstrap Revival of Some Nonparametric Distance Tests. Journal of the American Statistical Association, **83**, 1988, 698-708.

[26] Scholz F.W.: Towards a Unified Definition of Maximum Likelihood. The Canadian Journal of Statistics, **8**, 1980, 193-203.

[27] Bickel P.J., Krieger A.M.: Confidence Bands for a Distribution Function Using the Bootstrap. Journal of the American Statistical Association, **84**, 1989, 95-100.

[28] Durbin J.: Distribution Theory for Tests based on the Sample Distribution Function. SIAM, Philadelphia, 1973.

[29] Cheng R.C.H., Iles T.C.: Confidence Bands for Cumulative Distribution Functions of Continuous Random Variables. Technometrics, **25**, 1983, 77-86.

[30] Sahler W.: A survey of distribution-free statistics based on distances between distribution functions. Metrika, **13**, 1968, 149-169.

[31] Chapman D.G.: A comparative study of several one-sided goodness-of-fit tests. Annals of Mathematical Statistics, **29**, 1958, 655-674.

[32] Shorack G.R., Wellner J.A.: Empirical Processes with Applications to Statistics. Wiley, New York, 1986.

[33] Stephens M.A.: Tests Based on EDF Statistics. In Goodness-of-Fit Techniques, edited by D'Agostono R.B. and M.A. Stephens. Marcel Dekker, New York, 1986.

[34] Hope A.C.A.: A simple Monte Carlo Test Procedure. J. R. Statist. Soc., B, **30**, 1968, 582-598.

[35] Diggle P.J., Gratton R. J.: Monte Carlo Methods of Inference for Implicit Statistical Methods. J. R. Statist. Soc., B, **46**, 1984, 193-227.

[36] Hall P., Titterington D.M.: The Effect of Simulation Order on Level Accuracy and Power of Monte Carlo Tests. J. R. Statist. Soc., B, **51**, 1989, 459-467.

[37] Stephens M.A.: Asymptotic Results for Goodness-of-Fit Statistics with Unknown Parameters. Annals of Statistics, **4**, 1976, 357-369.

[38] Gleser L.J., Moore D.S.: The Effect of Dependence on Chi-Squared and Empiric Distribution Tests of Fit. Annals of Statistics, **11**, 1984, 1100-1108.

[39] Künsch H.: The Jackknife and the Bootstrap for Generally Stationary Observations. Annals of Statistics, **17**, 1989, 1217-1241.

[40] Beran J., Ghosh S.: Goodness-of Fit tests and Long Range Dependence. In Directions in Robust Statistics and Diagnostics, 21-33. Ed. by Stahel W. and S. Weisberg. Springer, New York, 1991.

[41] Liu R., Singh K.: Moving Blocks Jackknife and Bootstrap Capture Weak Dependence. In Exploring the Limits of Bootstrap. Ed. by LePage R. and L. Billard. Wiley New York, 1992.

[42] Carlstein A.: Resampling Techniques For Stationary Time Series: Some Recent Developments, 75-85. In New Directions in Time Series, Ed. Brillinger D. et al. Springer, New York, 1993.

[43] D'Agostiono R.B.: Graphical Analysis. In Goodness-of-Fit Techniques, edited by D'Agostono R.B. and M.A. Stephens. Marcel Dekker, New York, 1986.

[44] Nelson R.D., Pope R.D.: Bootstrapped Insights into Empircal Applications of Stochastic Dominance. Management Science, **37**, 1991, 1182-1194.

[45] Murphy A.H., Winkler R.L.: A Gerneral Framework for Forecast Verification. Monthly Weather Review, **115**, 1987, 1330-1338.

[46] Panofsky H.A., Brier G.W.: Some Applications of Statistics to Meteorology. Penn. St. University, 1968.

[47] Doswell C.A., Flueck J.A.: Forecasting and Verifying in a Field Research Project: DOPLIGHT '87. Weather and Forecasting, **4**, 1989, 97-109.