Value–at–Risk Prediction: A Comparison of Alternative Strategies[*]

Keith Kuester[a]   Stefan Mittnik[b c d †]   Marc S. Paolella[e]

[a]*Faculty of Economics and Business Administration, University of Frankfurt, Germany*
[b]*Department of Statistics, University of Munich, Germany*
[c]*Center for Financial Studies, Frankfurt, Germany*
[d]*Ifo Institute for Economic Research, Munich, Germany*
[e]*Institute of Statistics and Econometrics, University of Kiel, Germany*

January 2005

Abstract

Given the growing need for managing financial risk, risk prediction plays an increasing role in banking and finance. In this study, we compare the out-of-sample performance of existing methods and some new models for predicting Value-at-Risk. Using more than 30 years of the daily return data on the NASDAQ Composite Index, we find that most approaches perform inadequately, although several models are acceptable under current regulatory assessment rules for model adequacy. A hybrid method, combining a heavy-tailed GARCH filter with an extreme value theory-based approach, performs best overall.

Keywords—Empirical finance; Extreme value theory; Fat-tails; GARCH; Quantile regression.

1   Introduction

The market crash in October 1987, recent crises in emerging markets, and disasterous losses resulting from trading activities of institutions—such as Orange County, Long-Term Capital Management Fund and Metallgesellschaft—have increased the regulatory demand for reliable quantitative risk management tools. The *Value-at-Risk* (VaR) concept has emerged as the most prominent measure of downside market risk. It places an upper bound on losses in the sense that these will exceed the VaR–threshhold with only a small target probability, λ, typically chosen between 1% and 5%. More specifically, conditional on the information given up to time t, the VaR for period $t+h$ of one unit of investment is the negative λ-quantile of the conditional return distribution, i.e.,

$$\mathrm{VaR}_{t+h}^{\lambda} := -Q_{\lambda}(r_{t+h} \mid \mathcal{F}_t) = -\inf_{x}\{x \in \mathbb{R} : \mathrm{P}(r_{t+h} \leq x \mid \mathcal{F}_t) \geq \lambda\}, \quad 0 < \lambda < 1, \qquad (1)$$

where $Q_{\lambda}(\cdot)$ denotes the quantile function; r_t is the return on an asset or portfolio in period t; and $\mathcal{F}_t$ represents the information available at date t.[1]

Regardless of some of its criticisms,[2] regulatory requirements are heavily geared towards VaR.[3] In light of this practical relevance of the VaR concept, the need for reliable VaR estimation and prediction strategies arises. The purpose of this paper is to compare alternative approaches for VaR prediction, introduce some new models, and to provide some guidance for choosing an appropriate strategy.

For implementing VaR–based measures, one seeks a precise quantile estimate relatively far out in the left tail of the return distribution for some specified future date. Existing approaches for obtaining such an estimate may be classified as follows: *historical simulation* simply utilizes empirical quantiles based on the available past data; *fully parametric* models describe the entire distribution of returns including possible volatility dynamics; *extreme value theory* (in

---

[1] We subsequently suppress superscript λ for simplicity.

[2] The appropriateness of VaR as a risk-measure has been questioned. Evaluating prospects by VaR for varying $\lambda \in (0,1)$ is equivalent to checking for first-order stochastic dominance (as, for example, implied by the results in Bawa, 1978) and, thus, does not use the concept of risk aversion to rank prospects. But, by taking just one target probability level, λ, in contrast to using first–order stochastic dominance, any investment will be ranked. Thus, for a specific λ, VaR is a risk-measure (*cf.* Pedersen and Satchell, 1998).
According to the definition in Artzner *et al.* (1999), VaR fails to be a coherent risk measure. As such, VaR–based allocation of risks among agents can lead to Pareto-inferior allocations if agents are risk averse. Further, VaR can fail to appropriately account for portfolio–risk diversification (Artzner *et al.,* 1999).

[3] Besides risk reporting to senior management and shareholders, VaR is applied for allocating financial resources and risk–adjusted performance evaluation (*cf.* Jorion, 1997, Ch. 1). Furthermore, with the advent of the *Internal Model Approach* (Basle Committee, 1995 and 1996a), banks in the main financial jurisdictions may use their in-house VaR models for calculation of regulatory market–risk capital requirements.

short, EVT) parametrically models only the tails of the return distribution; and, finally, *quantile regression* directly models a specific quantile rather than the whole return distribution.

Below, we provide out-of-sample performance comparisons of models arising from these alternative strategies. The assessment is based on daily return data on the NASDAQ Composite Index, which is a typical representative of a portfolio of volatile financial assets.[4] We show that, with only a few exceptions, all of the methods perform less than desirable from a statistical viewpoint, although this may go undetected by the current regulatory assessment rules for model adequacy. In addition, we advance methodological concepts to VaR modeling by extending the EVT framework as applied in McNeil and Frey (2000) and by introducing a new specification for VaR quantile regressions. Both measures lead to significant improvements in out–of–sample VaR forecasts.

The remainder of the paper is organized as follows. Section 2 briefly summarizes different statistical approaches to VaR estimation. Section 3 examines methods for testing the adequacy of VaR forecasts. In Section 4, we describe the data and discuss the empirical results of the alternative forecasting methods. The final section provides concluding remarks; some technical details are given in the Appendix.

2 Statistical Approaches to Value-at-Risk

In practice, VaR prediction is hampered by the fact that financial returns exhibit "non–standard" statistical properties. Specifically, they are not independently and identically distributed (iid) and, moreover, they are not normally distributed. This is reflected by three widely reported stylized facts: (i) volatility clustering, indicated by high autocorrelation of absolute and squared returns; (ii) substantial kurtosis, i.e., the density of the unconditional return distribution is more peaked around the center and possesses much fatter tails than the normal density; and (iii) mild skewness of the returns, possibly of a time-varying nature. As a consequence, "standard" methods, based on the assumption of iid-ness and normality, tend not to suffice, which has led to various alternative strategies for VaR prediction. The most prominent of these are outlined in the following subsections.

---

[4] For an empirical study of actual commercial banks' VaR models based on their profit and loss data, see Berkowitz and O'Brian (2002).

2.1 Historical Simulation

The arguably simplest way to estimate VaR is to use the sample quantile estimate based on historic return data. Somewhat misleadingly, this approach is referred to as *historical simulation*. Specifically, the VaR estimate for $t + 1$ is given by the empirical λ–quantile, $\hat{Q}_\lambda(\cdot)$, of a moving window of w observations up to date t, i.e., $\widehat{\text{VaR}}_{t+1} = -\hat{Q}_\lambda(r_t, r_{t-1} \ldots, r_{t-w+1})$. For example, with a moving window of length, say, $w = 1,000$ observations, the 5% VaR estimate is simply the negative of the 50th sample order statistic. Besides ignoring the oftentimes blatant non iid–nature of the data, predictions extending beyond the extreme returns observed during the past w observations are not possible with this method. Also, the resulting VaR estimates can exhibit predictable jumps when large negative returns enter or drop out of the start of the window.

2.2 Fully Parametric Models

Fully parametric models are based on the assumption that returns belong to a location-scale family of probability distributions of the form

$$r_t = \mu_t + \epsilon_t = \mu_t + \sigma_t z_t, \tag{2}$$

where location μ_t and scale σ_t are $\mathcal{F}_{t-1}$–measurable parameters and $z_t \overset{\text{iid}}{\sim} f_Z(\cdot)$, where f_Z is a zero–location, unit–scale probability density. The h-period–ahead VaR forecast based on information up to time t is $\widehat{\text{VaR}}_{t+h} = -(\hat{\mu}_{t+h} + \hat{\sigma}_{t+h} Q_\lambda(z))$, where $Q_\lambda(z)$ is the λ–quantile implied by f_Z. Approaches differ with respect to specification of the conditional location, μ_{t+h}, the conditional scale, σ_{t+h}, and the density, f_Z.

Unconditional parametric models set $\mu_t \equiv \mu$ and $\sigma_t \equiv \sigma$, thereby assuming that the returns are iid with density $\sigma^{-1} f_Z\left(\sigma^{-1}(r_t - \mu)\right)$. Conditionally homoskedastic parametric models allow for a time-varying conditional mean, possibly captured by an ARMA(p, q) process, i.e.,

$$\mu_t = a_0 + \sum_{i=1}^{p} a_i\, r_{t-i} + \sum_{j=1}^{q} b_j\, \epsilon_{t-j}, \tag{3}$$

with $\sigma_t \equiv \sigma$, $t = 1, \ldots, T$. In light of the observed volatility clustering, this model class will be of marginal use. Instead, conditionally heteroskedastic parametric models, which allow the scale parameter to be a function of past information, are frequently used. Arguably the most popular formulation is the GARCH(r, s) model

$$\sigma_t^2 = c_0 + \sum_{i=1}^{r} c_i \epsilon_{t-i}^2 + \sum_{j=1}^{s} d_j \sigma_{t-j}^2, \tag{4}$$

3

introduced by Bollerslev (1986).

In the empirical analysis below, we utilize three different assumptions for the innovation distribution, f_Z, in (2): the normal; the Student's t with $\nu \in \mathbb{R}_+$ degrees of freedom (in short, t distribution); and the generalized asymmetric t, (in short, skewed t) with density

$$f(z; d, \nu, \theta) = C\left(1 + \frac{(-z\theta)^d}{\nu}\right)^{-\frac{\nu+1}{d}} \mathcal{I}(z < 0) + C\left(1 + \frac{(z/\theta)^d}{\nu}\right)^{-\frac{\nu+1}{d}} \mathcal{I}(z \geq 0), \qquad (5)$$

where $d, v, \theta \in \mathbb{R}_+$, $\mathcal{I}(\cdot)$ is the indicator function, $C = [(\theta + \theta^{-1}) d^{-1} \nu^{1/d} B(d^{-1}, \nu)]^{-1}$, and $B(\cdot, \cdot)$ denotes the beta function. The rth raw integer moment, $0 \leq r < \nu d$, for the skewed t is

$$\frac{(-1)^r \theta^{-(r+1)} + \theta^{r+1}}{\theta^{-1} + \theta} \frac{B\left(\frac{r+1}{d}, \nu - \frac{r}{d}\right)}{B\left(\frac{1}{d}, \nu\right)} \nu^{\frac{r}{d}},$$

from which, for example, variance, skewness and kurtosis can be computed if they exist. The cumulative distribution function (cdf) of the skewed t (as required for VaR calculation) is given by

$$F(z) = \begin{cases} \dfrac{I_L\left(\nu, \frac{1}{d}\right)}{1 + \theta^2}, & \text{if } z \leq 0, \\ \dfrac{I_U\left(\frac{1}{d}, \nu\right)}{1 + \theta^{-2}} + \left(1 + \theta^2\right)^{-1}, & \text{if } z > 0, \end{cases}$$

where $L = \nu / \left[\nu + (-z\theta)^d\right]$, $U = (z/\theta)^d / \left[\nu + (z/\theta)^d\right]$, and

$$I_x(a, b) = \frac{B_x(a, b)}{B(a, b)} = \frac{1}{B(a, b)} \int_0^x t^{a-1}(1 - t)^{b-1} dt \quad (a, b > 0)$$

is the incomplete beta ratio.

2.3 Extreme Value Theory in Risk Management

Extreme value theory (EVT) is concerned with the distribution of the smallest and largest order statistics. Whereas the models discussed in the previous subsection specify the entire return distribution, the EVT approach focuses only the tails of the return distribution. We follow convention and restrict attention to the right tail, an implication of which is that, if the left tail of the data is of interest (as is more often the case in a financial risk context), then the EVT analysis should be applied to the absolute value of the negative returns.

To briefly review the concepts which will be required in the subsequent analysis,[5] let $\{X_t\}_{t=1}^T$ be a sequence of iid random variables, and $M_T = \max(X_1, X_2, \ldots, X_T)$. If there exist norming

---

[5] For a comprehensive and accessible overview of the subject, the interested reader is referred to Embrechts *et al.* (1997), on which the following discussion is mainly based.

constants $c_T > 0$ and $d_T \in \mathbb{R}$ and some non-degenerate distribution function H such that

$$\frac{M_T - d_T}{c_T} \xrightarrow{d} H, \tag{6}$$

then, for $1 + \xi x > 0$,

$$H_\xi(x) = \begin{cases} \exp\left\{-(1+\xi x)^{-\frac{1}{\xi}}\right\}, & \text{if } \xi \neq 0, \\ \exp\left\{-\exp\{-x\}\right\}, & \text{if } \xi = 0, \end{cases}$$

where H_ξ is called the generalized extreme value distribution (GEV) (*cf.* Embrechts *et al.*, 1997, pp. 121, 152). In other words, there is one, and only one, possible family of limiting distributions for sample maxima. A random variable X, with distribution function F, is said to belong to the maximum domain of attraction of an extreme value distribution (in short, $X \in \mathrm{MDA}(H_\xi)$), if there exist constants $c_T > 0$ and $d_T \in \mathbb{R}$, such that (6) holds. Parameter ξ is crucial because it governs the tail behavior of $F(x)$. Distributions $F \in \mathrm{MDA}(H_\xi)$ are heavy tailed for $\xi > 0$, in which case their tails satisfy the power law

$$\bar{F}(x) = 1 - F(x) = x^{-\alpha}L(x), \tag{7}$$

for some $\alpha > 0$, with α referred to as the *tail index*, and slowly varying function $L(x)$. For these distributions, $\mathrm{E}|X|^\delta < \infty$ for any $\delta < \alpha$. Distributions in this class include, *inter alia*, the Pareto distribution and the stable Paretian distributions with tail index $\alpha = 1/\xi < 2$. For $\xi = 0$, $\bar{F}(x)$ decreases at an exponential rate as $x \to \infty$, as is, for example, the case for the normal and lognormal distributions, while distributions with $\xi < 0$ have a finite right end point, $x_F < \infty$, like the uniform and beta distributions. Indeed, $\mathrm{MDA}(H_\xi)$ includes essentially all the common continuous distributions occurring in applied statistics.

Consider now the distribution function of excesses, $Y = X - u$, of the iid random variable X over a high, fixed threshold $u < x_F$, that is,

$$F_u(y) = \mathrm{P}(X - u \leq y \mid X > u), \quad y \geq 0.$$

For excesses over thresholds, a key result, due to Pickands (1975), is that, for every fixed $\xi \in \mathbb{R}$, $F \in \mathrm{MDA}(H_\xi)$ if and only if

$$\lim_{u\uparrow x_F} \sup_{0 \leq y < x_F - u} \left| F_u(y) - G_{\xi,\beta(u)}(y) \right| = 0, \tag{8}$$

for some positive, measurable function β (*cf.* Embrechts *et al.*, 1997, p. 165). Thus, the generalized Pareto distribution (in short, GPD)

$$G_{\xi,\beta}(y) = \begin{cases} 1 - (1 + \frac{\xi y}{\beta})^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0, \\ 1 - e^{-\frac{y}{\beta}}, & \text{if } \xi = 0, \end{cases} \tag{9}$$

with support

$$y \geq 0, \qquad \text{if} \quad \xi \geq 0,$$
$$0 \leq y \leq -\frac{\beta}{\xi}, \quad \text{if} \quad \xi < 0,$$

and scale parameter β, arises naturally as the limit distribution of scaled excesses of iid random variables over high thresholds.

Two main strands of the current literature exist. The first (see, for example, Danielsson and de Vries, 2000) assumes fat-tailed data and uses (7) to approximate

$$\bar{F}(x) = \mathrm{P}(X > x) \approx K x^{-\alpha}, \quad x > u,$$

for some threshold $u = X_{k+1,T}$, given by the $(k+1)^{\mathrm{st}}$ descending order statistic. The most common estimator of the tail index is the Hill (1975) estimator:

$$\widehat{\alpha}^{(H)} = \widehat{\alpha}_{k,T}^{(H)} = \left(\frac{1}{k} \sum_{j=1}^{k} \log X_{j,T} - \log X_{k,T} \right)^{-1}, \tag{10}$$

where $k = k(T) \to \infty$ and $T/k(T) \to \infty$ as $T \to \infty$. For iid data, the estimator is consistent and asymptotically normal, assuming that k grows at a suitable rate relative to the sample size, T. In finite samples, however, the properties of $\widehat{\alpha}^{(H)}$ heavily depend on the choice of parameter k, which specifies the beginning of the tail area.[6]

Below, we focus on the second and more general strand of literature, which makes use of the limit result for peaks over thresholds (POT) in (8) and is not confined to fat-tailed data. Suppose that the X_t are iid with distribution function $F \in \mathrm{MDA}(H_\xi)$. Then, for a chosen threshold $u = X_{k+1,T}$ (so that there are k exceedances), define

$$F_u(y) = \mathrm{P}(X - u \leq y \mid X > u) = \frac{F(u+y) - F(u)}{1 - F(u)}, \quad y \geq 0,$$

which can be rewritten as

$$\bar{F}(u+y) = \bar{F}(u)\bar{F}_u(y). \tag{11}$$

In (11), $\bar{F}(u)$ can be estimated by its empirical counterpart, $\bar{F}_T(u) = k/T$, with $F_T(u)$ being the empirical distribution function of X. For a high enough threshold, (8) and (9) suggest utilizing

$$\bar{F}_u(y) \approx 1 - G_{\xi,\beta(u)}(y) \tag{12}$$

---

[6] Finite sample properties of the Hill as well as other related tail index estimators under various distributions are given by Embrechts *et al.* (1997); Mittnik *et al.* (1998); Danielsson and de Vries (1998); and the references therein.

so that, given estimates $\widehat{\xi}$ and $\widehat{\beta}$, $1 - G_{\widehat{\xi},\widehat{\beta}}(y)$ provides an estimate for $\bar{F}_u(y)$.[7] Thus, the tail probability for $X > u$ can be estimated by

$$\widehat{F(x)} = \frac{k}{T}\left(1 + \widehat{\xi}\,\frac{x - u}{\widehat{\beta}}\right)^{-1/\widehat{\xi}}. \tag{13}$$

A quantile estimator, $F(x_p) > 1 - k/T$, is obtained by inverting (13), i.e.,

$$\widehat{x}_{p,k} = X_{k+1,T} + \frac{\widehat{\beta}}{\widehat{\xi}}\left(\left(\frac{1-p}{k/T}\right)^{-\widehat{\xi}} - 1\right), \tag{14}$$

recalling $u = X_{k+1,T}$. Here, the choice of k suffers from similar problems as for the Hill estimator. Choosing u too high leads to very few exceedances and, thus, a high variance for the estimator, while low threshold values induce bias, as approximation (12) works well only in the tail.[8]

Although EVT is a natural candidate for VaR modeling, in light of the aforementioned stylized facts, EVT's iid assumption is inappropriate for most asset return data. While extensions which relax the independence assumption exist (*cf.* Embrechts *et al.,* 1997, pp. 209ff), one may alternatively apply the EVT analysis to appropriately filtered data. Diebold *et al.* (1998) propose fitting a time–varying volatility model to the data and estimating the tail of the filtered or standardized residuals, $z_t = (r_t - \mu_t)/\sigma_t$, by an EVT model, yielding an estimate for the standardized quantile, $Q_\lambda(z)$, as defined by (1), and, thus, for the Value–at–Risk

$$\mathrm{VaR}_t = -(\mu_t + \sigma_t Q_\lambda(z)). \tag{15}$$

With a correct model specification of the location and scale dynamics and use of consistent parameter estimates, the filtered model residuals will be approximately iid, as assumed in EVT modeling. Using the ideas advanced in Diebold *et al.* (1998), McNeil and Frey (2000) applied the POT technique to the residuals of a Gaussian AR(1)–GARCH(1, 1) model fitted to the sign–switched returns (i.e., multiplied by -1).

A natural generalization which we consider here is to use a parametric model as in Section 2.2, i.e., a location-scale model $r_t = \mu_t + \sigma_t z_t$, with z_t being a standardized zero–location, unit–scale random variable, to filter the returns and produce a set of approximate iid residuals, whose left tail we then approximate by an EVT model. The Gaussian AR(1)–GARCH(1, 1) filter is just a special case which, while indeed capable of removing the majority of volatility clustering

---

[7] The estimates for ξ and β can be obtained via maximum likelihood; see the Appendix for details.

[8] However, simulation evidence in McNeil and Frey (2000) for Student's t data suggests that the mean squared error for quantile estimates based on POT is far less sensitive to the choice of the threshold than for the Hill estimator.

and rendering the data approximately iid, will almost always be a (drastically) misspecified model for financial return data of daily or higher frequency. Much of the misspecification can be accommodated for by using a fat–tailed and asymmetric distribution such as the skewed t for f_Z, given in (5). As such, its use would be expected to result in more accurate AR and GARCH parameter estimates, as well as filtered (estimated) σ_t values, which, in turn, lead to improved scale forecasts, $\hat{\sigma}_{t+h}$.

On the other hand, Bollerslev and Wooldridge (1992) show that, if the conditional mean and volatility dynamics are properly specified, then the conditional mean and volatility are consistently estimated by pseudo–maximum likelihood—that is, ML–estimation under normality assumptions, even when innovations are not normally distributed.[9] Because the proper specification of the volatility dynamics is clearly an unattainable goal for actual return series, it is far from obvious which specification will be optimal, and the decision should be rather based on out–of–sample VaR forecasting performance. The results below for the NASDAQ returns indicate that the Gaussian AR(1)–GARCH(1,1) filter can be significantly improved upon by replacing the Gaussian assumption with the skewed–t. This holds for both the fully parametric and the EVT approaches.

Finally, irrespective of the filter used to render the data iid, when adopting the EVT approach, the choice of the shortfall probability, λ, should be compatible with the tail area for which the tail approximation starts to work well. While Diebold *et al.* (1998) favor only the extreme quantile values, say, one tenth of a percent, McNeil and Frey (2000) assume a reasonable fit already at the 5% level.[10]

2.4   Quantile Regression Approach

The determination of VaR naturally lends itself to the concept of quantile regression. To estimate *conditional* quantiles, the time series of the specified quantile is explicitly modeled using any information deemed to be relevant. No distributional assumptions for the time series behavior of returns is needed. The basic idea is to model the conditional λ-quantile, $Q_\lambda(r_t \mid X_t) = -\text{VaR}_t$, as some function of the information $X_t \in \mathcal{F}_{t-1}$, i.e.,

$$\text{VaR}_t \equiv -g_\lambda(X_t; \beta_\lambda), \tag{16}$$

---

[9] Their proofs rest on the existence of conditional variances. This seemingly innocuous assumption would, for example, rule out use of the asymmetric stable Paretian distribution, which has been demonstrated in Mittnik and Paolella (2003) to be particularly effective in Value–at–Risk prediction.

[10] Some details on the implementation of the GPD–ML are given in the appendix.

where $g(\cdot, \cdot)$ and parameter vector β explicitly depend on λ. A good choice of relevant information and of the functional form should yield a close approximation to the population quantile (*cf.* Chernozhukov and Umantsev, 2001). Koenker and Basset (1978) generalize the common linear regression framework by shifting the focus from the conditional *mean* to conditional *quantiles*. As shown, for example, in Koenker and Portnoy (1997), the *unconditional* sample λ–quantile, $\lambda \in (0,1)$, can be found as the solution to

$$\min_{\beta \in \mathbb{R}} \left\{ \sum_{r_t \geq \beta} \lambda |r_t - \beta| + \sum_{r_t < \beta} (1 - \lambda)|r_t - \beta| \right\}. \tag{17}$$

Extending this to the classical linear regression framework, Koenker and Bassett (1978) define the λth *regression quantile* estimator by

$$\widehat{\beta}(\lambda) = \operatorname*{argmin}_{\beta \in \mathbb{R}^k} \left\{ \sum_{r_t \geq x_t'\beta} \lambda |r_t - x_t'\beta| + \sum_{r_t < x_t'\beta} (1 - \lambda) |r_t - x_t'\beta| \right\},$$

where the x_t are non-random vectors.[11] The key assumption in the linear quantile regression model is that $r_t = x_t'\beta_\lambda + u_{t,\lambda}$. Note that the distribution of the error term is left unspecified. The only assumption made is that the *conditional quantile function* is given by $Q_\lambda(r_t \mid x_t) = x_t'\beta_\lambda$ and, thus, $Q_\lambda(u_{t,\lambda} \mid x_t) = 0$.[12]

One natural extension of the objective function for the general, possibly non-linear case of (16), proposed by Engle and Manganelli (2002), is

$$\min_{\beta \in \mathbb{R}^k} \left\{ \sum_{r_t \geq -\text{VaR}_t} \lambda |r_t + \text{VaR}_t| + \sum_{r_t < -\text{VaR}_t} (1 - \lambda) |r_t + \text{VaR}_t| \right\} \tag{18}$$

with, according to (16), $\text{VaR}_t \equiv -g(x_t; \beta_\lambda)$ or, in the linear case, $\text{VaR}_t \equiv -x_t'\beta_\lambda$.[13]

Chernozhukov and Umantsev (2001) use quantile regressions to model VaR—without, however, examining the model performance in terms of the sequence of VaR violations, as is done below. Taylor (1999) deals with the estimation of multiperiod VaR in the context of exchange rates, specifying (16) as linear functions of (transforms of) volatility estimates and the return

---

[11] Their proof of consistency and joint asymptotic normality relies on iid error terms with a continuous distribution function and fixed regressors. Non-linear absolute ARCH-power specifications are the subject of Koenker and Zhao (1996). For more general asymptotic results, see Chernozhukov and Umantsev (2001) and references therein.

[12] The well–known least absolute deviation (LAD) estimator for the regression median arises as the special case $\lambda = 0.5$. It yields more efficient estimates for the population mean than least squares in the presence of fat-tailed error distributions (see Bassett and Koenker, 1978, for the iid–error case with fixed regressors).

[13] Consistency and asymptotic normality of the non-linear regression quantiles for the time series case are established in Engle and Manganelli (2002).

horizon. As is common in the VaR literature, Taylor (1999) judges the *efficiency* of VaR estimates only on the basis of *unconditional* coverage (to be defined in Section 3 below).

Because our focus is exclusively on one–step forecasting performance, we more closely examine the conditional VaR approach formulated in Engle and Manganelli (2002), which is amenable to our maintained assumption that return data contain sufficient information for forecasting. In their specification of (16), they link VaR to the *conditional* standard deviation of the returns such that an increase in the latter leads to a more dispersed return distribution and, thus, *ceteris paribus*, to a higher VaR. Their conditional autoregressive VaR (in short, CAViaR) specifications include VaR_{t-1} as an explanatory variable in x_t, to adapt to serial dependence in volatility and mean. A function of r_{t-1} is also included to link the conditional quantile to return innovations.

As mentioned above, no explicit distributional assumptions need to be made, guarding against this source of model misspecification. Although many specifications of (16) are conceivable, we first adopt those put forth in Engle and Manganelli (2002). The *baseline* CAViaR model is given by

$$\text{VaR}_t = \text{VaR}_{t-1} + \beta \left[\mathcal{I} \left(r_{t-1} \leq -\text{VaR}_{t-1} \right) - \lambda \right].$$

As typically $\lambda \leq 0.05$ for risk management purposes, we have an asymmetric response: VaR_t will jump upward when a violation occurs and will slowly decrease otherwise—provided that the *a priori* conjecture $\beta > 0$ holds. In the baseline model the adaptive process "learns" nothing from the actual *size* of returns (except whether or not the returns are in line with VaR), as is the case with the *symmetric absolute value* CAViaR specification,

$$\text{VaR}_t = \beta_0 + \beta_1 \text{VaR}_{t-1} + \beta_2 |r_{t-1}|.$$

It allows the autoregressive parameter, β_1, to be different from one, and introduces a direct response of the quantile to the return process, treating the effect of extreme returns on VaR—and implicitly, on volatility—symmetrically. The symmetric assumption is relaxed in the *asymmetric slope* CAViaR specification,

$$\text{VaR}_t = \beta_0 + \beta_1 \text{VaR}_{t-1} + \beta_2 \max[r_{t-1}, 0] + \beta_3 \max[-r_{t-1}, 0],$$

which allows the VaR prediction to respond asymmetrically to positive and negative returns and so can accommodate the leverage effect. The *indirect* GARCH(1,1) CAViaR process,

$$\text{VaR}_t = (\beta_0 + \beta_1 \text{VaR}_{t-1}^2 + \beta_2 r_{t-1}^2)^{1/2}, \tag{19}$$

proposed by Engle and Manganelli (2002) would be appropriate if the data were generated by a location-scale model (2), with a GARCH(1, 1) process for the conditional scale, σ_t, and with zero location parameter, μ_t.

Autocorrelation in financial returns is often nonnegligible (see, e.g., Danielsson and Morimoto, 2000). This property can be incorporated by extending the existing CAViaR framework by allowing the returns to have a time–varying mean of the form

$$\mu_t = \mathrm{E}\left(r_t \mid \mathcal{F}_{t-1}\right), \tag{20}$$

which may, for example, be captured by a regression, ARMA or ARMAX model. An indirect GARCH specification of orders r and s with conditional mean μ_t can then be written as

$$\left(\mathrm{VaR}_t - \mu_t\right)^2 = c_0 + \sum_{i=1}^{r} c_i \left(r_{t-i} + \mu_{t-i}\right)^2 + \sum_{j=1}^{s} d_j \left(\mathrm{VaR}_{t-j} - \mu_{t-j}\right)^2. \tag{21}$$

The indirect conditional–mean GARCH(r,s) CAViaR model (20) and (21) reduces to that in Engle and Manganelli (2002) for $r = s = 1$ and $\mu_t = 0$. In the application below, it will be demonstrated that this more general CAViaR specification leads to a significant improvement in performance. There, we specify an AR(1) model, $r_t = ar_{t-1} + \epsilon_t$ for (20), and GARCH orders $r = s = 1$, leading to the *indirect* AR(1)–GARCH(1,1) CAViaR[14]

$$\mathrm{VaR}_t = ar_{t-1} + \left(\beta_0 + \beta_1 \left(\mathrm{VaR}_{t-1} - ar_{t-2}\right)^2 + \beta_2 \left(r_{t-1} - ar_{t-2}\right)^2\right)^{1/2}, \quad \beta_1, \beta_2, \beta_3 \in \mathbb{R}_+. \tag{22}$$

3   Comparing and Testing the Fit of Value–at–Risk Models

To assess the predictive performance of the models under consideration, we follow Christoffersen's (1998) framework, which is designed for evaluating the accuracy of out–of–sample interval forecasts. Defining $H_t = \mathcal{I}\left(r_t < -\mathrm{VaR}_t\right)$, Christoffersen (1998) terms the sequence of VaR forecasts *efficient* with respect to $\mathcal{F}_{t-1}$ if

$$\mathrm{E}[H_t \mid \mathcal{F}_{t-1}] = \lambda, \tag{23}$$

which, by applying iterated expectations, implies that H_t is uncorrelated with any function of a variable in the information set available at $t-1$. If (23) holds, then VaR violations will occur

---

[14] To see this, when the process generating the returns is a GARCH$(1,1)$ with the AR(1) mean equation $r_t = ar_{t-1} + \epsilon_t$ with $a_0 \equiv 0$. Then, $\sigma_t^2 = c_0 + c_1(r_{t-1} - \mu_{t-1})^2 + d_1\sigma_{t-1}^2$, where $c_0, c_1, d_1 > 0$. Substitution of $\mathrm{VaR}_{t-1} = -\mu_{t-1} - \sigma_{t-1}z_\lambda$ yields

$$\sigma_t^2 = c_0 + c_1(r_{t-1} - \mu_{t-1})^2 + d_1 S_{t-1}, \quad S_{t-1} = \left(\frac{\mathrm{VaR}_{t-1} + \mu_{t-1}}{z_\lambda}\right)^2,$$

and, with $\mu_t = a_1 r_{t-1}$ and $\mathrm{VaR}_t = -ar_{t-1} - (c_0 + c_1(r_{t-1} - ar_{t-2})^2 + d_1 S_{t-1}z_\lambda)^{1/2}$. Taking z_λ (which is a constant for iid innovations) into the root and noting that $z_\lambda < 0$ for small λ (so $z_\lambda = -\sqrt{z_\lambda^2}$), we obtain the desired CAViaR expression (22) after appropriately relabelling the parameters.

with the correct conditional and unconditional probability, and neither the forecast for VaR_t nor that for H_t could be improved.

Although a general test of (23) is desirable, we follow Christoffersen (1998) in using intermediate statistics for testing specific implications of the general hypothesis, so that particular inadequacies of a model can be revealed. By specifying $\mathcal{F}_{t-1}$ to include at least $\{H_1, H_2, \ldots, H_{t-1}\}$, it is straightforward to show (Christoffersen, 1998, Lemma 1) that efficiency implies

$$H_t \mid \mathcal{F}_{t-1} \overset{\text{iid}}{\sim} \text{Ber}(\lambda), \quad t = 1, 2, \ldots, T, \tag{24}$$

where $\text{Ber}(\cdot)$ denotes the Bernoulli distribution. Below, property (24) is referred to as *correct conditional coverage*.

3.1 Test of Unconditional Coverage

By taking iterated expectations, (23) implies correct *unconditional* coverage of the interval forecasts. We test for the correct number of violations by

$$H_0 : \text{E}[H_t] = \lambda \quad \text{vs.} \quad H_A : \text{E}[H_t] \neq \lambda.$$

Under the null, (24) implies the likelihood–ratio test statistic

$$\text{LR}_{\text{uc}} = 2 \left[\mathcal{L}(\widehat{\lambda}; H_1, H_2, \ldots, H_T) - \mathcal{L}(\lambda; H_1, H_2, \ldots, H_T) \right] \overset{\text{asy}}{\sim} \chi^2_1, \tag{25}$$

where $\mathcal{L}(\cdot)$ denotes the log binomial likelihood. The MLE, $\widehat{\lambda}$, is the ratio of the number of violations, n_1, to the total number of observations, $n_0 + n_1 = T$, i.e., $\widehat{\lambda} = n_1/(n_0 + n_1)$.

3.2 Test of Independence

VaR forecasts which do not take temporal volatility dependence into account may well be correct on average, but will produce violation clusters (*cf.* Christoffersen, 1998), a phenomenon which is ignored when considering unconditional coverage.

Several tests for independence have been proposed in the literature, including the runs tests and the Ljung-Box-Test (1978). More recently, a test based on the time between exceedances was proposed in Danielsson and Morimoto (2000). Under the null, a violation today has no influence on the probability of a violation tomorrow. Christoffersen (1998) models $\{H_t\}$ as a binary first-order Markov chain with transition probability matrix

$$\Pi = \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{bmatrix}, \quad \pi_{ij} = \text{P}(H_t = j \mid H_{t-1} = i),$$

as the alternative hypothesis of dependence.[15] The approximate joint likelihood, conditional on the first observation, is

$$\mathcal{L}(\Pi; H_2, H_3, \ldots, H_T \mid H_1) = (1 - \pi_{01})^{n_{00}} \, \pi_{01}^{n_{01}} \, (1 - \pi_{11})^{n_{10}} \, \pi_{11}^{n_{11}}, \tag{26}$$

where n_{ij} represents the number of transitions from state i to state j, i.e.,

$$n_{ij} = \sum_{t=2}^{T} \mathcal{I}(H_t = i \mid H_{t-1} = j);$$

and the ML estimators under the alternative hypothesis are

$$\widehat{\pi}_{01} = \frac{n_{01}}{n_{00} + n_{01}} \quad \text{and} \quad \widehat{\pi}_{11} = \frac{n_{11}}{n_{10} + n_{11}}.$$

Under the null of independence, we have $\pi_{01} = \pi_{11} \equiv \pi_0$, from which the conditional binomial joint likelihood under the null follows as

$$\mathcal{L}(\pi_0; H_2, \ldots, H_T \mid H_1) = (1 - \pi_0)^{n_{00} + n_{10}} \, \pi_0^{n_{01} + n_{11}}. \tag{27}$$

The MLE, $\widehat{\pi}_0$, is analogous to that in the unconditional coverage test; and the LR-test is given by

$$\mathrm{LR}_{\mathrm{ind}} = 2 \left[\mathcal{L}(\widehat{\Pi}; H_2, \ldots, H_T \mid H_1) - \mathcal{L}(\widehat{\pi}_0; H_2, \ldots, H_T \mid H_1) \right] \overset{\mathrm{asy}}{\sim} \chi_1^2. \tag{28}$$

3.3 Conditional Coverage

Because $\widehat{\pi}_0$ is unconstrained, the test in (28) does not take correct coverage into account. Christoffersen (1998) suggests combining (25) and (28) to

$$\mathrm{LR}_{\mathrm{cc}} = 2 \left[\mathcal{L}(\widehat{\Pi}; H_2, \ldots, H_T \mid H_1) - \mathcal{L}(\lambda; H_2, \ldots, H_T \mid H_1) \right] \overset{\mathrm{asy}}{\sim} \chi_2^2, \tag{29}$$

in order to test correct conditional coverage (24). By conditioning on the first observation in (25), we have

$$\mathrm{LR}_{\mathrm{cc}} = \mathrm{LR}_{\mathrm{uc}} + \mathrm{LR}_{\mathrm{ind}},$$

which provides a means to check in which regard the violation series $\{H_t\}$ fails the correct conditional coverage property (24).[16]

---

[15] The runs test is uniformly most powerful against this alternative (see, e.g., Lehmann, 1986). We opt for the framework used here, because it can be easily integrated into a test of the more general hypothesis (24).

[16] As with all asymptotically motivated inferential procedures, the actual size of the tests for finite samples can deviate from their nominal sizes. Lopez (1997) examines the size of unconditional and conditional coverage

3.4 Dynamic Quantile Test

Condition (23) is stronger than correct conditional coverage; it suggests that *any* $\omega_{t-1,i} \in \mathcal{F}_{t-1}$ be uncorrelated with H_t. In particular, Engle and Manganelli (2002) remark that conditioning violations on the VaR for the period itself is essential. To illustrate this point, they let $\{\text{VaR}_t\}_{t=1}^T$ be a sequence of iid random variables such that

$$\text{VaR}_t = \begin{cases} K, & \text{with probability } 1 - \lambda, \\ -K, & \text{with probability } \lambda. \end{cases}$$

Then, for K very large and conditioning also on VaR_t, the violation sequence exhibits correct conditional coverage, as tested by (29), but, conditional on VaR_t, the probability of a violation is either almost 0 or almost 1. None of the above tests has power against this form of *inefficiency*.

To operationalize (23), one can, similar to Christoffersen (1998) and Engle and Manganelli (2002), regresses H_t on a judicious choice of explanatory variables in $\mathcal{F}_{t-1}$, i.e.,

$$H_t = \lambda_0 + \sum_{i=1}^p \beta_i H_{t-i} + \beta_{p+1} \widehat{\text{VaR}}_t + \sum_{j=1}^n \beta_{p+j+1} f(\omega_{t-1,j}) + u_t,$$

where, under the null hypothesis, $\lambda_0 = \lambda$ and $\beta_i = 0$, $i = 1, \ldots, p+n+1$. In vector notation, we have

$$\mathbf{H} - \lambda \boldsymbol{\iota} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad u_t = \begin{cases} -\lambda, & \text{with probability } 1 - \lambda, \\ 1 - \lambda, & \text{with probability } \lambda, \end{cases} \tag{30}$$

where $\beta_0 = \lambda_0 - \lambda$ and $\boldsymbol{\iota}$ is a conformable vector of ones. Under the null (23), the regressors should have no explanatory power, i.e., $H_0 : \boldsymbol{\beta} = \mathbf{0}$. Because the regressors are not correlated with the dependent variables under the null, invoking a suitable CLT yields[17]

$$\widehat{\boldsymbol{\beta}}_{\text{LS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{H} - \lambda \boldsymbol{\iota}) \stackrel{\text{asy}}{\sim} \text{N}\left(0, (\mathbf{X}'\mathbf{X})^{-1}\lambda(1 - \lambda)\right),$$

from which Engle and Manganelli (2002) deduce the test statistic

$$\text{DQ} = \frac{\widehat{\boldsymbol{\beta}}_{\text{LS}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{LS}}}{\lambda(1 - \lambda)} \stackrel{\text{asy}}{\sim} \chi^2_{p+n+2}. \tag{31}$$

---

tests via simulation, as well as their power against various model misspecifications. For a sample size of 500 observations, he finds both tests to be adequately sized. Even for such a small sample, power appears to be reasonable. For the LR_{uc} test, for example, he reports that, for λ values of 0.05 or smaller, if the true data generating process is conditionally heteroskedastic, then power is well above 60% for wrong distributional assumptions for the innovations. In general, the tests have only moderate power when volatility dynamics are closely matched but power increases under incorrect innovation distributions, especially further out in the tails.

[17] Note that, for the asymptotics under the null, it is irrelevant whether $H_t - \lambda$ is regressed on lags of H_t or lags of $H_t - \lambda$, as proposed by Engle and Manganelli (2002).

In the empirical application below, we use two specifications of the Dynamic Quantile (henceforth DQ) test: For the first, denoted by $\mathrm{DQ_{Hit}}$, regressor matrix $\mathbf{X}$ contains a constant and four lagged Hits, $H_{t-1}, \ldots, H_{t-4}$; the second, $\mathrm{DQ_{VaR}}$, uses, in addition, the contemporaneous VaR estimate.

4 Empirical Analysis

We examine the VaR forecasting performance for a portfolio which is long in the NASDAQ Composite Index.[18] The index itself is a market–value weighted portfolio of over 5,000 stocks listed on the NASDAQ stock market. The data comprise daily closing levels, p_t, of the index from its inception on February 8, 1971 up to June 22, 2001, yielding a total of $T = 7,681$ observations of percentage log–returns, $r_t := 100(\ln p_t - \ln p_{t-1})$. Table 1 presents some relevant summary statistics. The sample skewness $\hat{\mu}_3/\hat{\mu}_2^{3/2} = -0.466$ indicates considerable asymmetry which, taken together with the sample kurtosis $\hat{\mu}_4/\hat{\mu}_2^2 = 17.3$, indicates a substantial violation of normality.[19]

> Table 1 somewhere here.

For all models considered, we allow the corresponding parameters to change over time: Using moving windows of size $w = 1000$ (corresponding to roughly four years of trading data) we update the model parameters for each moving window with increments of one trading day. This leaves us with 6,681 one–step–ahead VaR forecasts to study the predictive performance of the models. To save space, we refer to the AR(1)–GARCH(1, 1) specification simply as "GARCH".

In addition to the GARCH–EVT approach of Diebold *et al.* (1998) and McNeil and Frey (2000), which relies on the normal assumption for the GARCH filter, we consider an alternative conditional EVT implementation which specifies the skewed t instead of the normal distribution in order to better account for conditional asymmetry and heavy–tailedness. These are referred to in Table 3 and the subsequent discussion as N–EVT and ST–EVT, respectively.

The results under the alternative modeling assumptions are reported in Tables 2, 3 and 5. With a few exceptions, all models tend to underestimate the frequency of extreme returns.

---

[18] The data and further information may be obtained from http://www.marketdata.nasdaq.com, maintained by the Economic Research Department of The National Association of Securities Dealers, Inc.

[19] The fact that heavy–tailed distributions may not possess low–order moments implies that usual significance tests for skewness and kurtosis are most likely unreliable and are not worth reporting. See Loretan and Phillips (1994), Adler *et al.* (1998), Paolella (2001), and the references therein for further discussion.

Although the performance varies substantially across the modeling approaches as well as the distributional assumptions, some clear patterns emerge. We first discuss the performance of the unconditional models, then the GARCH–based, and, finally, CAViaR models.

As the unconditional models do not account for volatility clustering, none of them is able to produce iid VaR violations, causing us to strongly reject independence of the H_t sequences for all unconditional models (see Table 2).[20] At the 1% λ–level, the naive historical simulation performs well with respect to violation frequencies, along with the skewed t and the unconditional EVT. The superior performance of the skewed t distribution relative to the normal and t is due to the fact that it allows for skewness, which is clearly present in the *unconditional* return data (see Table 1).

Table 2 somewhere here.

Current regulatory performance assessment focuses on the unconditional coverage property, leaving other implications of *efficiency* unexamined. The three–zone framework suggested by the Basle Committee (1996b) deems a VaR model acceptable (Green Zone) if the number of violations of 1% VaR remains below the binomial(0.01) 95% quantile. A model will be disputable (Yellow Zone) up to the 99.99% quantile and deemed seriously flawed (Red Zone) whenever more violations occur. Translated to our sample size, a model passes regulatory performance assessment if at most 79 violations (1.18%) occur and will be disputable when between 80 and 98 (1.47%) violations occur. The results reported in Table 2 suggest that, as far as the unconditional coverage is concerned, the normal and Student's t unconditional models are seriously flawed (Red), while historical simulation, unconditional EVT and the unconditional skewed t models are disputable (Yellow), though the latter are still inadequate in terms of clustering of violations. As none of the unconditional models is acceptable in any of the testing categories at both the 2.5% and 5% target probability, a conditional approach should be preferred.

Introducing GARCH volatility dynamics almost uniformly improves VaR prediction performance: The great majority of models do better than the unconditional approach across all λ–levels and all distributions considered—as was to be expected from the apparent volatility clustering in the return series. Regarding the percentage of violations (see Table 3), among the distributional assumption for the fully parametric models, the skewed–t is by far the best for all three λ–levels, followed by the t, though the t was occasionally outperformed by the normal

---

[20] This finding contradicts Danielsson and de Vries (2000) and is, instead, more in line with the observations of Danielsson and Morimoto (2000), who—for their data sets—still observe considerable (though decreasing) dependence in extreme returns above (and including) the 1% λ–level.

assumption. Both the t and normal performed quite poorly. While this is not surprising for the normal, some empirical studies, such as McNeil and Frey (2000) and Danielsson and Morimoto (2000), show that normal–GARCH might have some merit for larger values of λ. Our findings indicate that, at least for this data set, the 5% quantile is still not large enough for the normal assumption to be adequate. This is made clear in Figure 1, which graphically displays the accuracy of the violation percentages. The normal assumption yields quite accurate results for $\lambda = 0.1$, in stark comparison to the Student's t assumption, which tends to perform worse as λ increases and yields the worst result among all the conditional models.

Table 3 somewhere here.

Figure 1 somewhere here.

The N–EVT and ST–EVT outperformed the fully parametric conditional models, with preference given to the latter. In particular, from Table 3, the N–EVT had 0.97%, 2.50% and 5.33% violations for $\lambda = 0.01, 0.025$ and 0.05, respectively, while ST–EVT had 0.97%, 2.47% and 5.06%. This is also made clear from Figure 1: The two EVT formulations are always the best two performers among all conditional models and for all values of $0 < \lambda < 0.1$. Moreover, the ST–EVT delivers virtually exact results for all $0 < \lambda < 0.1$, while N–EVT is competitive for $0 < \lambda < 0.025$ and then weakens considerably as λ increases towards 0.1.

We now turn to the information in the sequence of violations, as reflected in the p–values of the LR and DQ test statistics in Table 3. Entries in bold–face are greater than 0.01, signifying that the hypothesis of independence cannot be rejected at the 1% significance level. Four models exhibit all p–values greater than 0.01, thus hinting at *efficient* VaR forecasts: the fully parametric GARCH with skewed t innovations for $\lambda = 0.01$; the N–EVT for $\lambda = 0.01$ and 0.05; and the ST–EVT for $\lambda = 0.01$. From this list and the other tabulated p–values of the LR and DQ statistics, it is clear overall that, with respect to independence, the AR(1)-GARCH(1, 1) filter coupled with a skewed, fat–tailed distributional assumption or the EVT model results in VaR violations which contain virtually no information about the probability of a future violation. Note that the violation frequencies for the AR(1)-GARCH(1, 1) filter with either the normal or t innovation assumption are considerably worse than for the S&P and DAX series reported in McNeil and Frey (2000). However, none of their tests includes the high–volatility regime following the Asian and Russian Crises, as well as the beginning of the recent slump in the market.

Table 4 provides summary information for each model by averaging over all values of λ in $(0, 0.1)$. Specifically, we computed the mean absolute deviation (MAD) and mean squared error

(MSE), based on the mean of (approximately) 668 probability deviations shown in Figure 1. Both summary measures indicate the same result: The ST–EVT has (with considerable margin) the lowest deviation.

Summarizing the results for the fully parametric and EVT models: Major improvements in VaR predictions are achieved in all aspects when accounting for the volatility dynamics. VaR violations are reasonably independent when using either the fully parametric GARCH model with skewed t innovations, or an EVT model based on a GARCH filter with either normal or skewed t innovations, the latter being preferred overall.[21]

Table 4 somewhere here.

Next, we turn to the results of the CAViaR models, which deliver mixed results (see Table 5). Only the very simple adaptive CAViaR specification performs adequately at all λ–levels with regard to *unconditional* coverage. Because the adaptive CAViaR model increases VaR once a violation occurs and decreases it slightly otherwise, it is not surprising that it cannot produce cluster–free violations. This also agrees with the results reported in Engle and Manganelli (2002).

Table 5 somewhere here.

The symmetric absolute–value specification, which adapts VaR to the size of returns, is only acceptable with respect to unconditional coverage at the higher λ–levels. Otherwise, it blatantly fails the independence tests, rendering its performance as inadequate as that of the adaptive CAViaR. While the asymmetric slope specification is a generalization of the symmetric absolute–value CAViaR, it exhibits quite different characteristics: It passes the independence test at all λ–levels, it fails all the other tests at all λ levels.

None of the traditional CAViaR models passes any DQ–test at any λ–level. The indirect GARCH(1, 1) specification performs well at the 1% and 5% λ–levels, with respect to correct unconditional coverage only. This was expected, given the first–order autocorrelation in the data. Engle and Manganelli (2002) report a weaker CAViaR performance for index data than for individual stock returns. The fact that our NASDAQ sample comprises two additional highly

---

[21] The performance of other GARCH formulations and distributional assumptions was also investigated. In particular, the A-PARCH model proposed by Ding, Granger and Engle (1993), which nests (4), and use of the asymmetric stable Paretian distribution. The more general A-PARCH structure actually led to a decrease in forecast quality compared to (4), as did the popular absolute value GARCH model advocated by Nelson and Foster (1994) and which is also subsumed in the A-PARCH model. The stable Paretian distributional assumption, while better than the normal and t, was inferior to use of the skewed t, the latter also being more straightforward and faster to estimate. Of course, these findings pertain only to the NASDAQ returns.

volatile years, which presumably deteriorate overall performance, may help to reconcile the poor out-of-sample performance of the established CAViaR models with the more positive findings of Engle and Manganelli (2002).

The need to incorporate an explicit autoregressive link to returns is seen when looking at the indirect CAViaR AR(1)–GARCH(1, 1) model proposed here. It passes altogether 12 out of the 15 tests, whereas the second best CAViaR specification passes only three out of the 15. While resulting in too many VaR violations for the 2.5% level, the model exhibits correct unconditional coverage at the 1% and 5% λ–levels. While all other CAViaR models fail to pass any DQ–test, the proposed specification passes it for the 2.5 and 5% VaR–levels. Hence, in the sense that it is less prone to violations clustering, the new model improves considerably upon the previous CAViaR specifications.

5   Conclusions

The predictive performance of several recently advanced and some new Value–at–Risk models has been examined. The majority of these suffer from excessive VaR violations, implying an underestimation of market risk. Employing more informative tests than is common in the literature, we find that regulatory forecasting assessment is flawed. Most notably, all of the unconditional models produce clustered VaR violations, yet some may still pass as acceptable when considering only the (unconditional) violation frequencies.

Conditional VaR models lead to much more volatile VaR predictions than unconditional models and may arguably cause problems in allocating capital for trading purposes (see, e.g., Danielsson and Morimoto, 2000). However, our results show that only conditionally heteroskedastic models may yield acceptable forecasts. For the fully parametric models, a major improvement in terms of violation frequencies is achieved when accounting for scale dynamics. In addition, taking heteroskedasticity into account yields reasonably unclustered VaR violations. Considerable improvement upon normality is achieved by using innovation distributions which allow for skewness and fat tails. The conditional skewed t and the two EVT approaches, N–EVT and ST–EVT, perform best, with the EVT models being favored in most cases. Among the latter, preference goes to the ST–EVT, given that it yields the best overall conditional tail fit for $\lambda \in (0, 0.1)$.

Finally, none of the CAViaR models perform well overall, though the proposed indirect AR(1)–GARCH(1, 1) model is most promising in the CAViaR class in the sense that it passes most of the tests.

Appendix

We implemented the EVT approach unconditionally, using the raw return data, and conditionally (i.e., GARCH–filtered) assuming normal and and skewed t innovations. In simulations of size 1000, for t-residuals, McNeil and Frey (2000) found that $k \approx 100$ minimizes the MSE of the resulting quantile estimates $\hat{z}'_{1-\lambda,k}$ and that the results are rather insensitive to the choice of k for a wide range of k. While this choice may not be adequate for other innovation distributions, there exist no automatic mechanisms capable of choosing k consistently. We therefore followed their choice. Differentiating (9) and assuming excesses are iid GPD, the log-likelihood function is

$$
\mathcal{L}(\xi, \beta; y_1, \ldots, y_k) = \begin{cases} -k\log\beta - \left(\dfrac{1}{\xi} + 1\right)\sum\limits_{j=1}^{k}\log\left(1 + \dfrac{\xi}{\beta}y_j\right), & \text{if} \quad \xi \neq 0, \\[2ex] -k\log\beta - \sum\limits_{j=1}^{k}\dfrac{y_j}{\beta}, & \text{if} \quad \xi = 0, \end{cases}
$$

with support $y_j \geq 0$, if $\xi \geq 0$, and with $0 \leq y_j \leq -\beta/\xi$, if $\xi < 0$. Smith (1987) showed that ML estimation works well for $\xi > -1/2$, a constraint which was never violated in our empirical analysis. Moreover, for iid data, we have

$$
\sqrt{n}\left(\hat{\xi} - \xi, \frac{\hat{\beta}_n}{\beta} - 1\right) \xrightarrow{d} N(0, M^{-1}), \quad n \to \infty, \tag{32}
$$

where

$$
M^{-1} = (1+\xi)\begin{bmatrix} 1+\xi & 1 \\ 1 & 2 \end{bmatrix}, \tag{33}
$$

and the usual consistency and efficiency properties of MLE apply (Smith, 1987).

The support constraint for $\xi < 0$ is not easily implemented in standard constrained optimization. We chose to penalize the likelihood function proportionally to the amount of violation, in order to "guide" the estimates back into regions covered by the GPD. We used the Nelder-Mead simplex algorithm, as implemented in the Matlab 6 routine "fminsearch", employing the default options.

Our implementation of the CAViaR approach was essentially the same as that in Engle and Manganelli (2002). Here, we discuss some issues concerning the implementation. Due to the multitude of local minima in objective function (18), the baseline CAViaR model was estimated using a genetic algorithm (for details on the algorithm, see Engle and Manganelli, 2002). The initial population consisted of 500 members chosen from the interval $[0, 1]$. A population size of 20 was maintained in each of a maximum of 500 generations. The programs used for calculation in Engle and Manganelli (2002) were unconstrained, which, for the NASDAQ data we considered,

resulted occasionally in negative estimates for parameter $\widehat{\beta}$. In this case, each violation lowers the VaR measure. If towards the end of the in-sample period the VaR measure is sufficiently low, the number of violations tends to accelerate, rapidly driving down the out-of-sample VaR forecasts to levels well below zero (toward $-\infty$ for long forecast horizons). This drawback may be overcome by decreasing the "genetic fitness" if this happens. We "punish" the fit by adding a value of 3 (about $1/10$ of the fitness function for the 1% λ–level and correspondingly less for the higher levels) whenever VaR and/or the parameter estimate is negative. This proved to be successful in guiding the process back into more suitable regions. To estimate the remaining specifications, we used the Matlab 6 functions "fminunc" and "fminsearch", as described in Engle and Manganelli (2002).

The indirect GARCH model needs to be constrained to prevent the term in parentheses in (19) from becoming negative. This was achieved by setting the objective function to 10^{10} whenever this occurred. We also penalized in case a parameter estimate became a complex number, as was also done for the symmetric absolute value and the asymmetric slope CAViaR. The restrictions for the indirect AR(1)–GARCH(1, 1) are analogous to those of the GARCH(1, 1) counterpart.

References

Adler, R. J., Feldman, R. E., and Gallagher, C. (1998). Analysing Stable Time Series. In Adler, R. J., Feldman, R. E., and Taqqu, M. S., editors, *A Practical Guide to Heavy Tails*, pages 133–158. Birkhäuser, Boston.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent Measures of Risk. *Mathematical Finance*, 9(3):203–228.

Basle Committee on Banking Supervision (1995). An Internal Model-Based Approach to Market Risk Capital Requirements. http://www.bis.org.

Basle Committee on Banking Supervision (1996a). Overview of the Amendment to the Capital Accord to Incorporate Market Risks. http://www.bis.org.

Basle Committee on Banking Supervision (1996b). Supervisory Framework for the Use of "Backtesting" in Conjunction with the Internal Models Approach to Market Risk Capital Requirements. http://www.bis.org.

Bassett, G. and Koenker, R. (1978). Asymptotic Theory of the Least Absolute Error Regression. *Journal of the American Statistical Association*, 73:618–622.

Bawa, V. S. (1978). Safety-First, Stochastic Dominance, and Optimal Portfolio Choice. *Journal of Financial and Quantitative Analysis*, 13(2):255–271.

Berkowitz, J. and O'Brien, J. (2002). How Accurate Are Value–at–Risk Models at Commercial Banks? *Journal of Finance*, LVII:1093–1111.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31:307–327.

Bollerslev, T. and Wooldridge, J. M. (1992). Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances. *Econometric Reviews*, 11(2):143–172.

Chernozhukov, V. and Umantsev, L. (2001). Conditional Value-at-Risk: Aspects of Modeling and Estimation. *Empirical Economics*, 26(1):271–292.

Christoffersen, P. (1998). Evaluating Interval Forecasts. *International Economic Review*, 39(4):841–862.

Danielsson, J. and de Vries, C. G. (1998). Beyond the Sample: Extreme Quantile and Probability Estimation. Discussion paper, no. 98-016, Tinbergen Institute.

Danielsson, J. and de Vries, C. G. (2000). Value-at-Risk and Extreme Returns. *Annales d'Economie et de Statistique*, 60:239–270.

Danielsson, J. and Morimoto, Y. (2000). Forecasting Extreme Financial Risk: A Critical Analysis of Practical Methods for the Japanese Market. *Monetary and Economic Studies*, 18(2):25–48.

Diebold, F. X., Schuermann, T., and Stroughair, J. D. (1998). Pitfalls and Opportunities in the Use of Extreme Value Theory in Risk Management. Working paper no. 98-10, The Wharton School, University of Pennsylvania.

Ding, Z., Granger, C. W., and Engle, R. F. (1993). A Long Memory Property of Stock Market Returns and a New Model. *Journal of Empirical Finance*, 1:83–106.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.

Engle, R. F. and Manganelli, S. (2002). CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. Manuscript, University of California, San Diego.

Fisher, R. A. and Tippet, L. (1928). Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. *Proc. Cambridge Philos. Soc.*, 24:180–190.

Hill, B. M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. *The Annals of Statistics*, 3:1163–1174.

Jorion, P. (1997). *Value at Risk: The New Benchmark for Controlling Market Risk*. McGraw-Hill, New York.

Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46(1):33–50.

Koenker, R. and Portnoy, S. (1997). Quantile Regression. Working paper no. 97-0100, University of Illinois at Urbana-Champaign.

Koenker, R. and Zhao, Q. (1996). Conditional Quantile Estimation and Inference for ARCH Models. *Econometric Theory*, 12:793–813.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses, 2nd ed.* John Wiley & Sons, New York.

Ljung, G. and Box, G. (1979). On a Measure of Lack of Fit in Time Series Models. *Biometrika*, 66:265–270.

Lopez, J. A. (1997). Regulatory Evaluation of Value-at-Risk Models. Staff reports no. 33, November 1997, Federal Reserve Bank of New York.

Loretan, M. and Phillips, P. (1994). Testing the Covariance Stationarity of Heavy-tailed Time Series. *Journal of Empirical Finance*, 1:211–248.

McNeil, A. J. and Frey, R. (2000). Estimation of Tail-related Risk Measures for Heteroscedastic Financial Time Series: An Extreme Value Approach. *Journal of Empirical Finance*, 7(3/4):271–300.

Mittnik, S. and Paolella, M. S. (2003). Prediction of Financial Downside-Risk with Heavy-Tailed Conditional Distributions. In *Handbook of Heavy Tailed Distributions in Finance*. North-Holland.

Mittnik, S., Paolella, M. S., and Rachev, S. T. (1998). A Tail Estimator for the Index of the Stable Paretian Distribution. *Communications in Statistics–Theory and Methods*, 27(5):1239–1262.

Nelson, D. B. and Foster, D. P. (1994). Asymptotic Filtering Theory for Univariate ARCH Models. *Econometrica*, 62(1):1–41.

Paolella, M. S. (2001). Testing the Stable Paretian Assumption. *Mathematical and Computer Modelling*, 34:1095–1112.

Pedersen, C. S. and Satchell, S. E. (1998). An Extended Family of Financial-Risk Measures. *The Geneva Papers on Risk and Insurance Theory*, 23(2):89–117.

Pickands III, J. (1975). Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3(1):119–131.

Smith, R. L. (1987). Estimating Tails of Probability Distributions. *The Annals of Statistics*, 15(3):1174–1207.

Taylor, J. W. (1999). A Quantile Regression Approach to Estimating the Distribution of Multi-period Returns. *The Journal of Derivatives*, 7(1):64–78.

Table 1: Summary Statistics for NASDAQ Returns

Sample Size	Mean	Std. Dev.	Skewness	Kurtosis	Min	Max
7681	0.0392	1.1330	-0.4656	17.302	-12.048	13.255

Table 2: VaR Prediction Performance: Unconditional Models[a]

Model	100λ	% Viol.	LR_{uc}	LR_{ind}	LR_{cc}	DQ_{Hit}	DQ_{VaR}
Historical Simulation	1	1.30	**0.02**	0.00	0.00	0.00	0.00
	2.5	3.26	0.00	0.00	0.00	0.00	0.00
	5	6.00	0.00	0.00	0.00	0.00	0.00
Normal	1	2.80	0.00	0.00	0.00	0.00	0.00
	2.5	4.27	0.00	0.00	0.00	0.00	0.00
	5	6.18	0.00	0.00	0.00	0.00	0.00
Student's t	1	2.10	0.00	0.00	0.00	0.00	0.00
	2.5	4.52	0.00	0.00	0.00	0.00	0.00
	5	7.74	0.00	0.00	0.00	0.00	0.00
Skewed t	1	1.30	**0.02**	0.00	0.00	0.00	0.00
	2.5	3.46	0.00	0.00	0.00	0.00	0.00
	5	6.17	0.00	0.00	0.00	0.00	0.00
EVT	1	1.29	**0.02**	0.00	0.00	0.00	0.00
	2.5	3.40	0.00	0.00	0.00	0.00	0.00
	5	6.03	0.00	0.00	0.00	0.00	0.00

[a] λ is the target probability. Entries in the last 5 columns are the significance levels (p-values) of the respective tests. Bold–faced entries are not significant at the 1% level. For DQ_{Hit}, $H_t - \lambda$ is regressed onto a constant and 4 lagged violations, for DQ_{VaR}, in addition the contemporaneous VaR estimate. The unconditional EVT model does not "pre-filter" with an estimated ARMA–GARCH structure and is instead directly applied to the (negative) return data.

Table 3: VaR Prediction Performance: AR(1)-GARCH$(1,1)$[a]

Model	100λ	% Viol.	LR_{uc}	LR_{ind}	LR_{cc}	DQ_{Hit}	DQ_{VaR}
Normal	1.0	2.23	0.00	**0.03**	0.00	0.00	0.00
	2.5	3.92	0.00	**0.04**	0.00	0.00	0.00
	5.0	6.21	0.00	**0.21**	0.00	0.00	0.00
Student's t	1.0	1.81	0.00	0.01	0.00	0.00	0.00
	2.5	4.04	0.00	**0.02**	0.00	0.00	0.00
	5.0	6.89	0.00	**0.06**	0.00	0.00	0.00
Skewed t	1.0	1.20	**0.12**	**0.35**	**0.19**	**0.16**	**0.04**
	2.5	2.72	**0.25**	0.00	0.00	0.00	0.00
	5.0	5.12	**0.65**	**0.03**	**0.08**	**0.07**	0.00
N–EVT	1.0	0.97	**0.82**	**0.16**	**0.37**	**0.12**	**0.08**
	2.5	2.50	**1.00**	0.01	**0.02**	0.00	0.00
	5.0	5.33	**0.22**	**0.07**	**0.09**	**0.05**	**0.08**
ST–EVT	1.0	0.97	**0.82**	**0.17**	**0.37**	**0.09**	**0.02**
	2.5	2.47	**0.87**	0.00	0.00	0.00	0.00
	5.0	5.06	**0.82**	**0.02**	**0.06**	**0.08**	0.00

[a] "N–EVT" refers to the use of AR(1)–GARCH$(1,1)$ with Gaussian innovations as the filter used with the conditional EVT model; "ST–EVT" is similar, but uses density (5) instead. See also the footnote in Table 2.

Table 4: Overall Deviation Measures[a]

Model	MAE	MSE
Normal	0.970	1.147
Student's t	1.474	2.504
Skewed t	0.146	0.031
N–EVT	0.217	0.065
ST–EVT	0.050	0.004

[a]The mean of the absolute (MAE) and squared (MSE) deviation of (100 times) the empirical and theoretical tail probability, computed over the first 10% of the sorted out–of–sample cdf values for the GARCH–based models.

Table 5: VaR Prediction Performance: CAViaR[a]

| Model	100λ	% Viol.	LR_{uc}	LR_{ind}	LR_{cc}	DQ_{Hit}	DQ_{VaR}
Adaptive	1	1.14	**0.27**	0.00	0.00	0.00	0.00
	2.5	2.80	**0.12**	0.00	0.00	0.00	0.00
	5	5.10	**0.70**	0.00	0.00	0.00	0.00
Symmetric Abs. Value	1	1.33	0.01	0.01	0.00	0.00	0.00
	2.5	2.83	**0.09**	0.00	0.00	0.00	0.00
	5	5.45	**0.10**	0.00	0.00	0.00	0.00
Aymmetric Slope	1	1.60	0.00	**0.83**	0.00	0.00	0.00
	2.5	3.35	0.00	**0.01**	0.00	0.00	0.00
	5	6.02	0.00	**0.11**	0.00	0.00	0.00
Indirect GARCH(1,1)	1	1.32	**0.01**	0.00	0.00	0.00	0.00
	2.5	3.08	0.00	0.00	0.00	0.00	0.00
	5	5.54	**0.05**	0.00	0.00	0.00	0.00
Indirect AR(1)– GARCH(1,1)	1	1.32	**0.01**	**0.15**	**0.02**	0.00	0.00
	2.5	3.04	0.00	**0.74**	**0.02**	**0.03**	**0.01**
	5	5.55	**0.04**	**0.89**	**0.12**	**0.01**	**0.02**

[a] See the footnote in Table 2.
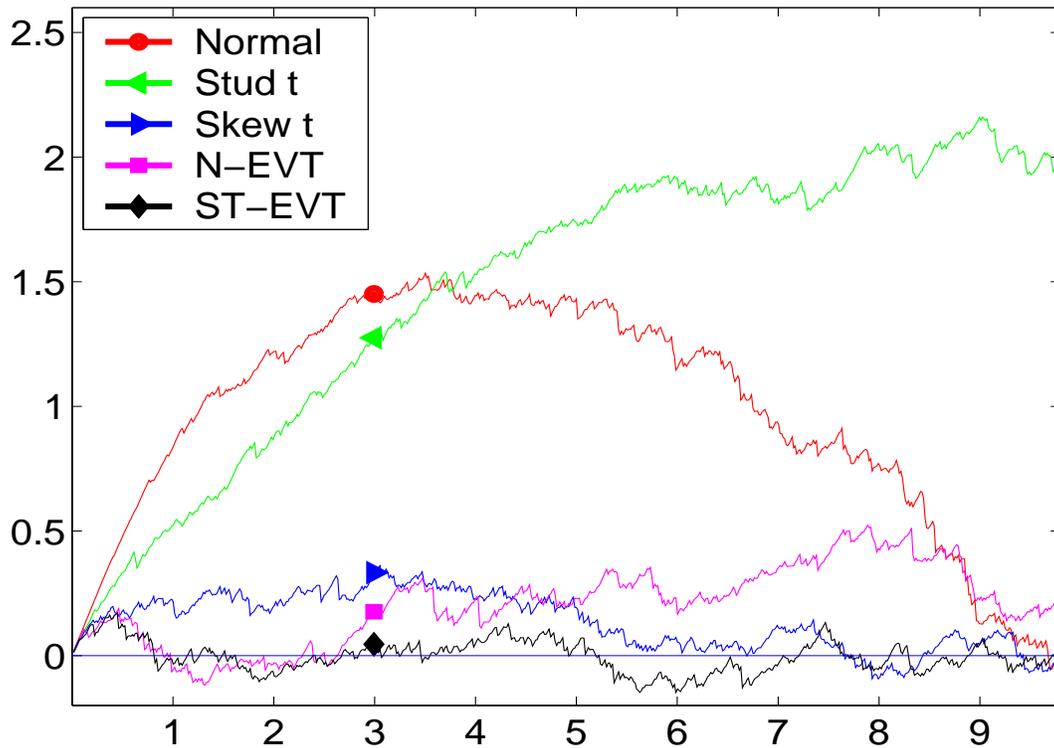
Figure 1: Deviation probability plot for the five GARCH–based conditional models. Plotted values are $100(F_U(u) - \hat{F}(u))$ versus $100u$, where F_U is the cdf of a uniform random variable; $\hat{F}$ refers to the empirical cdf formed from evaluating the 6,681 one–step, out–of–sample distribution forecasts at the true, observed return; and u is a grid of values between 0 and 0.1.