Univariate Extreme Value Theory, GARCH and Measures of Risk

Kaj Nyström* and Jimmy Skoglund†
Swedbank, Group Financial Risk Control
S-105 34 Stockholm, Sweden
(First draft: September 12, 2001)
September 4, 2002

Abstract

In this paper we combine ARMA-(asymmetric) GARCH and EVT methods, applying them to the estimation of extreme quantiles (i.e., beyond the 95%) of univariate portfolio risk factors. In addition to Value at Risk we consider the tail conditional expectation (or expected shortfall). A comparison is also made with methods based on conditional normality (e.g., RiskMetrics), conditional t-distribution as well as the empirical distribution function. The paper is partially self-contained as it details the ARMA-(asymmetric) GARCH model as well the GMM-estimator used. It introduces the most important theoretical results of univariate EVT, discusses risk measures on a general level as well as the notion of coherent measure of risk.

1 Introduction

The study of extreme events, like the crash of October 1987 and the Asian crisis, is at the center of risk management interest. In addition banks are required to hold a certain amount of capital against their defined market risk exposure. Under the ‘internal models’ approach capital charges are a function of banks’ own value at risk (VaR) estimates. Clearly, the accuracy of quantile based risk measures such as VaR is of concern to both risk management and regulators.

In general terms a portfolio can be thought of as a mapping \( \pi : \mathbb{R}^n \to \mathbb{R} \) where \( n \) is the dimension of the space of risk factors. Understanding this mapping is,
of course, more or less difficult depending on the complexity of the portfolio instruments. To be able to calculate/measure portfolio risk we need to define a model for our risk factors as well as relevant risk measures. In defining this model we in general try to capture, as realistically as possible, the evolution of risk factors using probability theory and stochastic processes. The model specification stage may be viewed as consisting of two steps. In the first step we specify a model for each individual risk factor, then in the second step we specify a copula function to capture dependence between the marginal distributions. Having estimated the degrees of freedom in the model, i.e., the parameters, using historical data we may consider our portfolio as a stochastic variable with support on (a subset of) $R$.

In terms of risk factor models current industry practice is to specify a conditional (on information generated up to and including time t) or unconditional normal distribution for the returns. Depending on the portfolio structure, i.e., linear or non-linear, a quantile, or possibly some other risk measure, of the profit and loss distribution may be found analytically or by simulation. It is however important to emphasize that the result of applying a risk measure to the (estimated) portfolio distribution is crucially dependent on the choice of model for the evolution of risk factors. In this regard it is not difficult to criticize current industry practice since both conditional and unconditional return distributions are characterized by the 'stylized facts' of excess kurtosis, high peakedness and are often negatively skewed. That is, neither the symmetry nor the exponentially decaying tail behavior, exhibited by the normal distribution, seems to be supported by data. Hence, using the normal distribution approximation the risk of the high quantiles may be severely underestimated. An alternative to the normal distribution may be to consider a Student-t distribution. The Student-t distribution have heavier tails than the normal, displaying polynomial decay in the tails. Hence, it may be able to capture the observed excess kurtosis although it maintains the hypothesis of symmetry. In contrast non-parametric methods (e.g., historical simulation) makes no assumptions concerning the nature of the empirical distribution function but has several drawbacks. For example it cannot be used to solve for out of sample quantiles and the variance of the highest order statistics is very high, yielding poor estimates of the tails. Furthermore, kernel based estimators are not a solution since they typically perform poorly in the smoothing of tails.

This paper applies univariate extreme value theory (EVT) to the empirical return distributions of financial series. The paper may be thought of as a first step in understanding and defining a multivariate model for the evolution of risk factors under extreme market conditions. The second step i.e., the modelling of co-dependence, is something we will return to in a forthcoming paper.

EVT and in particular the generalized pareto (GP) distribution give an asymptotic theory for the tail behavior. Based on a few assumptions the theory shifts the focus from modelling the whole distribution to the modelling of the
tail behavior and hence the symmetry hypothesis may be examined directly by estimating the left and right tail separately.

One key assumption in EVT is the assumption that extreme returns are independent and identically distributed. As there is ample evidence of volatility clustering in (extreme) returns we apply EVT to the filtered conditional residuals which form an approximately independent and identically distributed sequence. More specifically, we assume that the continuously compounded return process, \( y_t \), follows a stationary ARMA-(asymmetric) GARCH model

\[
\Phi(L)y_t = \Theta(L)\varepsilon_t
\]

where \( \Phi(L) = 1 - \sum_{i=1}^{p} \phi_i L^i \), \( \Theta(L) = 1 + \sum_{j=1}^{q} \xi_j L^j \) are the lag polynomials and with \( \varepsilon_t \) decomposed as, \( \varepsilon_t = z_t h_t \). The conditional variance process, \( h_t^2 \), is governed by the recursive equation

\[
h_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + a_2 \text{sgn}(\varepsilon_{t-1}) \varepsilon_{t-1}^2 + bh_{t-1}^2
\]

with the filtered conditional innovation, \( z_t \), being independent and identically distributed with mean zero and variance unity.

The organization of the paper is as follows. Section 2 introduces the ARMA-(asymmetric) GARCH model. Section 3 is concerned with some theoretical results of EVT as well as a simulation study, evaluating the finite-sample properties of EVT tail index estimators. In Section 4 we discuss risk measures on a general level, starting from some of the key weaknesses of VaR. Section 5 combines the ARMA-(asymmetric) GARCH and EVT methods, applying them to the estimation of extreme quantiles. A comparison is made with methods based on conditional normality (e.g., RiskMetrics), conditional t-distribution as well as the empirical distribution. In addition to Value-at-Risk we consider the tail conditional expectation (or expected shortfall) described in Section 4. In Section 6 we give a summary of the results and conclusions found in the paper. At the end the reader may also find references to the literature.

2 ARMA-GARCH models

2.1 General properties

Consider the simple ARMA\((p,q)\) model

\[
y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \xi_j \varepsilon_{t-j} + \varepsilon_t
\]

where \( \varepsilon_t \) is independent and identically distributed with mean zero and variance \( \sigma^2 \). ARMA models are traditionally employed in time series analysis to capture (linear) serial dependence. They allow conditioning of the process mean on past
realizations and is generally successful for the short-term prediction of time series.

The assumption of conditional homoskedasticity is however too restrictive for financial data where we typically observe volatility clusters, implying that a large absolute return is often followed by more large absolute returns. A GARCH model for the conditional variance process extends the simple ARMA model by assuming that

\[ \varepsilon_t = z_t h_t \]

where \(z_t\) is independent and identically distributed with mean zero and unit variance and \(z_t h_t\) are stochastically independent. It is the dynamics of \(h_t^2\), the conditional variance at time \(t\), that the GARCH model wish to capture. The GARCH model cannot only capture the volatility clustering of financial data but also to some extent excess kurtosis, since

\[ k_4 = \frac{E\varepsilon_t^4}{(E\varepsilon_t^2)^2} = v_4 \frac{Eh_t^4}{(Eh_t^2)^2} \geq v_4 \]

where \(v_4\) is the kurtosis of \(z_t\). Intuitively, the unconditional distribution is a mixture of normals, some with small variances that concentrate mass around the mean and some with large variances that put mass in the tails of the distribution.

The GARCH model is commonly used in its most simple form, the GARCH(1,1) model, in which the conditional variance is given by

\[ h_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + bh_{t-1}^2 \]
\[ = a_0 + (a_1 + b) h_{t-1}^2 + a_1 (\varepsilon_{t-1}^2 - h_{t-1}^2) \]

The term

\[ (\varepsilon_{t-1}^2 - h_{t-1}^2) = h_{t-1}^2 (z_{t-1}^2 - 1) \]

has zero mean conditional on past information and can be interpreted as the shock to volatility. The coefficient \(a_1\) therefore measures the extent to which a volatility shock in period \(j\) feeds through into the volatility in period \(j+1\), while \((a_1 + b)\) measures the rate at which this effect dies out, i.e., the discount rate.

The GARCH(1,1) model can also be written in terms of the squared errors, \(\varepsilon_t^2\). We have

\[ \varepsilon_t^2 = a_0 + (a_1 + b) \varepsilon_{t-1}^2 - b (\varepsilon_{t-1}^2 - h_{t-1}^2) + (\varepsilon_t^2 - h_t^2) \]

where as noted above \((\varepsilon_t^2 - h_t^2)\) have expectation zero conditional on past information. This form of the GARCH(1,1) model makes it clear that GARCH(1,1) is really an ARMA(1,1) for the squared returns. The parameter \(a_1\) is indeed an autoregressive parameter whereas the parameter \(b\) contributes to both the autoregressive and moving average structure. This form of the model is however not useful for estimation purposes. A standard ARMA(1,1) model has homoskedastic innovations, while here the shocks are themselves heteroskedastic.
2.2 Stationarity and persistence

In the framework of ARMA processes, the stationarity condition is characterized by the roots of the autoregressive polynomial. In the case where some roots have modulus one the process is nonstationary, called an ARIMA process, which generalize the notion of a random walk. It is interesting to examine this limiting case in the framework of the GARCH model. Consider for example the conditional variance forecast for s steps ahead in the future

$$E(h_{t+s}^2|\varepsilon_t^2) = a_0 \left[ \sum_{i=0}^{s-1} (a_1 + b)^i \right] + (a_1 + b)^s h_t^2$$

If \((a_1 + b) < 1\) a shock to the conditional variance decays exponentially whereas if \((a_1 + b) \geq 1\) the effect of a shock does not die out asymptotically. In the case where \((a_1 + b) = 1\) we have that

$$E(h_{t+s}^2|\varepsilon_t^2) = sa_0 + h_t^2$$

and hence the forecast properties corresponds to those of a random walk with drift. This analogy must however be treated with caution. A linear random walk is nonstationary in two senses. First, it has no stationary distribution, hence the process is not strictly stationary. Second, the unconditional first and second moments does not exist, hence it is not covariance stationary. For the GARCH(1,1) model \(h_t^2\) is strictly stationary although it lacks unconditional moments, hence it is not covariance stationary (which requires \(a_1 + b < 1\)). In fact if \(a_0 = 0\) the distribution of \(h_t^2\) becomes more and more concentrated around zero with fatter and fatter tails and \(h_t^2 \to 0\) almost surely whereas if \(a_0 > 0\) \(h_t^2 \to_u h_t^2\) almost surely with \(h_t^2\) strictly stationary. In this regard it is interesting to note that the case of \(a_0 = 0\) corresponds to RiskMetrics volatility model. Hence, for the RiskMetrics model

$$E(h_{t+k}^2|\varepsilon_t^2) = h_t^2$$

so the conditional variance is a martingale whereas the unconditional variance is zero!

2.3 Introducing asymmetry

The GARCH(1,1) model we have considered so far is symmetric in the sense that negative and positive shocks have the same impact on volatility. There is much stronger evidence that positive innovations to volatility are correlated with negative innovations to returns than with positive innovations to returns. One possible explanation for this observation is that negative shocks to returns tend to drive up volatility. An alternative explanation is that causality runs the other way. If expected returns increase when volatility increases (holding
expected dividends constant) then prices should fall when volatility increases. To capture this potential asymmetry we follow Glosten et al. (1993) and include an additional parameter in the GARCH(1,1) equation, yielding

\[ h_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + a_2 \text{sgn}(\varepsilon_{t-1}) \varepsilon_{t-1}^2 + bh_{t-1}^2 \]

where \( \text{sgn}(\varepsilon_t) = \text{sgn}(z_t) = 1 \) if \( z_t < 0 \) and 0 if \( z_t \geq 0 \). Hence the difference between the symmetric GARCH and the present model is that the impact of a shock (and its discounting) is captured by the term \( a_2 \text{sgn}(\varepsilon_{t-1}) \) as well. According to the discussion above we expect \( a_2 > 0 \).

### 2.4 Estimation

Under the additional assumption of conditional normality it is straightforward to set up the likelihood function for the ARMA-(asymmetric) GARCH(1,1) model and obtain estimates for the parameter vector

\[ \theta = (a_0, a_1, a_2, b, \mu, \phi', \xi') \]

where \( \phi = (\phi_1, \ldots, \phi_p)' \), \( \xi = (\xi_1, \ldots, \xi_q)' \). But as indicated earlier not even conditional normality is supported by financial return series. Nevertheless under mild moment conditions on the filtered conditional residuals (and stationarity conditions on the process) the normal likelihood may still serve as a vehicle for obtaining consistent parameter estimates although the resulting estimator is certainly not minimum variance in the class of consistent and asymptotically normal (CAN) estimators.

In our approach the efficiency of the filtering process, i.e., the construction of \( z_t \), is of paramount importance. This is so because the filtered residuals serve as an input to the EVT tail estimation. This suggests that we should search for an estimator which is more efficient under non-normality and as efficient as quasi maximum likelihood under normality. Such an estimator exists, it is based on the Generalized Method of Moments (GMM) principle. That is, it does not make specific assumptions about the distribution (of \( z_t \)) but proceeds by postulating conditional moments. We discuss this estimator only briefly here and refer to Skoglund (2001) for details.

Define the raw vector

\[ r_t = [\varepsilon_t, (\varepsilon_t^2 - h_t^2)]' \]

and the generalized vector,

\[ g_t = F_t' r_t \]

where \( F_t \) is a so-called instrumental variable function.
The GMM estimator of a parameter vector $\theta$ is then a solution to

$$\min_{\theta \in \Theta} \left[ \sum_{t=1}^{T} g_t \right] ' W_T \left[ \sum_{t=1}^{T} g_t \right]$$

with $W_T = T^{-1} \sum_{t=1}^{T} W_t$ being an appropriate weighting matrix. An efficient choice of instrumental variable function and weighting matrix corresponds to choosing

$$F_t = \Sigma_t^{-1} \frac{\partial r_t}{\partial \theta} ' \text{ and }$$

$$W_t = \Sigma_t^{-1} \frac{\partial r_t}{\partial \theta} \frac{\partial r_t}{\partial \theta} ' \Sigma_t^{-1} \frac{\partial r_t}{\partial \theta}$$

where $\Sigma_t = \text{var}(r_t)$ and $(\frac{\partial r_t}{\partial \theta})$ is the Jacobian matrix. The objective function for an operational efficient GMM estimator is then given by

$$Q_T = T^{-1} \left[ \sum_{t=1}^{T} g_t \right] ' \left( \sum_{t=1}^{T} \Lambda_t \right) \left[ \sum_{t=1}^{T} g_t \right]$$

where $\Lambda_t = g_t g_t ^{-1}$. Denoting by $v_k = E z_t^k$ the generalized moment, $g_t$ is explicitly written

$$g_t = \frac{1}{\Delta} \left[ \left( \frac{\partial h^2}{\partial \theta} \right) h^2_t \left[ v_3 h^2_t - \left( \frac{h^2_t}{h_t} - 1 \right) \right] \right]$$

with $\Delta = [(v_4 - 1) - v_3^2]$ and the derivatives $\frac{\partial h^2}{\partial \theta}$ are computed recursively as,

$$\frac{\partial h^2}{\partial \theta} = c_{t-1} + b \frac{\partial h^2}{\partial \theta}$$

where $c_t = (1, \varepsilon_t^2, \text{sgn}(\varepsilon_t) \varepsilon_t^2, h_t^2, -2 [a_1 + a_2 \varepsilon_t \text{sgn}(\varepsilon_t)] \varepsilon_t X_t)$ with

$$X_t = (1, y_{t-1}, \ldots, y_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q})$$

and

$$\frac{\partial \varepsilon_t}{\partial \theta} = \pi_t$$

with $\pi_t = (0, 0, 0, 0, -X_t)$.

Application of the GMM estimator requires an initial guess on the third and fourth moments of $z_t$. That is, we require an initial estimator of the parameter vector $\theta$. For this purpose it is convenient to use the (normal) quasi maximum likelihood estimator to obtain initial consistent estimates. In this regard we can view efficient GMM as updating the (normal) quasi-maximum likelihood estimator (adapting the initial estimator) to get more efficient estimates of the parameters and hence also the filtered residuals.

$^1$The estimator is equivalently defined by $T^{-1} \sum_{t=1}^{T} g_t$ and $Q_T$. However $Q_T$ has the advantage of being invariant to non-singular linear transformations and is the preferred choice in practice.
3 Extreme Value Theory

In this section we intend to describe the main focus and results of univariate extreme value theory. In particular we discuss

- Extreme value distributions.
- Generalized Pareto distributions.
- Application of extreme value theory to quantile estimation for high out of sample quantiles.
- Estimation of parameters in models based on extreme value theory.

3.1 Theoretical background

Let in the following $X_1, \ldots, X_n$ be $n$ observation from $n$ independent and identically distributed random variables with distribution function $F$. We are interested in understanding the distribution function $F$ with a particular focus on its upper and lower tails. From this perspective it is natural to consider the following objects:

$$M_n = \max\{X_1, \ldots, X_n\}$$
$$m_n = \min\{X_1, \ldots, X_n\}$$

Both $M_n$ and $m_n$ are random variables depending on the length $n$ of the sample and we are, in analogy with the central limit law for sums, interested in understanding the asymptotic behavior of these random variables as $n \to \infty$. Notice that $m_n = -\max\{-X_1, \ldots, -X_n\}$ and hence in the following we will only describe the theory for $M_n$, i.e., we focus on observations in the upper tail of the underlying distribution.

Before stating the main theorems of univariate extreme value theory we feel that it is appropriate to make a digression to the more well-known central limit law. Consider the sum of our observations, i.e.,

$$S_n = \sum_{r=1}^{n} X_r$$

Many classical results in probability are based on sums and results like the law of large numbers and the central limit theorem (CLT) are all based on the object $S_n$. In its most rudimentary form the CLT can be formulated as follows.

**Theorem 1** Suppose that $X_1, \ldots, X_n$ are $n$ independent and identically distributed random variables with mean $\mu$ and variance $\sigma^2$. Then

$$\lim_{n \to \infty} P\left(\frac{S_n - n\mu}{\sigma \sqrt{n}} \leq x\right) = N(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-z^2/2)dz$$
The central limit theorem is interesting for the obvious reason. It tells us that, even though we just have information about the first and second order moments of the underlying random variables, properly normalized sums behave just like a normally distributed variable. That is, if we are interested in the distribution of sums of a large number of observations from IID random variables we have to care less about a careful modelling of the underlying random variable $X$ and its distribution function $F$, as according to CLT, we know the structure of the distribution of the sum once we have information about the first and second moments of $X$.

The version of CLT stated above can to a certain extent be generalized to the case of random variables not having finite variance by considering Levy distributions. By introducing a notion of sum-stable distributions one may conclude, using the Fourier transform, that a distribution is sum-stable if and only if it is a Levy distribution. In its more general form CLT now states that if the sum $S_n$ converges, then the limit has to be sum-stable and hence a Levy distribution. Key assumptions in the theory are of course that the random variables we sum are independent and identically distributed. Now, both these assumption can be relaxed and the CLT generalized.

Generally speaking one may say that extreme value theory (EVT) gives similar results as the CLT but for the maximum of random variables, i.e., EVT gives us the structure of the asymptotic limit of the random variable $M_n$ defined above.

Before stating a few general theorems we believe that a short example could serve as a clarifying introduction. Let $F(x) = (1 - \exp(-x))\chi_{[0,\infty)}(x)$ i.e., let us consider exponentially distributed random variables. Assuming independence we have

$$P(M_n \leq x) = (1 - \exp(-x))^n$$

and letting $n \to \infty$ we see that the R.H.S. tends to zero for every positive value of $x$. Hence, without a proper normalization we can never get a non-degenerated limit. Let us therefore redo the calculation in the following manner,

$$P(M_n \leq x + \log n) = (1 - \exp(-(x + \log n)))^n$$

$$= (1 - \exp(-x)\frac{n}{n})^n \to \exp(-\exp(-x)) =: \Gamma(x)$$

Here the limit indicates that we let $n$ tend to infinity. One may in fact prove that the convergence is uniform and hence in particular for large $n$ we have

$$P(M_n \leq x) \sim \Gamma(x - \log n)$$

The function $\Gamma(x)$ is an example of an extreme value distribution. We introduce the following notation for any $\xi \in \mathbb{R}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$

$$\Gamma_{\xi,\mu,\sigma}(x) = \exp\left(-\left(1 + \frac{(x - \mu)}{\sigma}\right)_+^{-1/\xi}\right), x \in \mathbb{R}$$

9
This is the general form of the extreme value distribution. The \(1/\xi\) is referred to as the tail index as it indicates how heavy the upper tail of the underlying distribution \(F\) is. Letting \(\xi \to 0\) we see that the tail index tends to infinity and that \(\Gamma_{\xi,\mu,\sigma}(x) \to \Gamma((x - \mu)/\sigma)\). The parameters \(\mu\) and \(\sigma\) represent a translation and a scaling respectively. Obviously the distribution \(\Gamma_{\xi,\mu,\sigma}(x)\) is non zero if and only if \((1 + \xi(x-\mu)/\sigma) > 0\). As \(\sigma\) by definition is positive the subset of the real axis where this inequality is true is depending on the sign of \(\xi\). If \(\xi = 0\) the distribution spreads out along all of the real axis and in this case the distribution is often called a Gumbel distribution. If \(\xi > 0\) the distribution has a lower bound (often the distribution is called Frechet this case) and if \(\xi < 0\) the distribution has an upper bound (usually referred to the Weibull case). The first result of EVT is the following.

**Theorem 2** Suppose \(X_1, \ldots, X_n\) are \(n\) independent and identically distributed random variables with distribution function \(F\) and suppose that there are sequences \(\{a_n\}\) and \(\{b_n\}\) so that for some non-degenerated limit distribution \(G(x)\) we have,

\[
\lim_{n \to \infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = G(x), x \in \mathbb{R}
\]

Then there exist \(\xi \in \mathbb{R}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+\) such that

\[
G(x) = \Gamma_{\xi,\mu,\sigma}(x)
\]

Many of our more well-known distributions may be divided between the three classes of EVT-distributions according to the size of their tail index. For example Normal, Gamma and Lognormal distributed variables converge to the Gumbel distribution \((\xi = 0)\). Student-t, Pareto, Loggamma, Cauchy distributed variables converge to Frechet \((\xi > 0)\) and uniform distributions on \((0,1)\) as well as Beta distributed random variables converge to Weibull \((\xi < 0)\). From the statement of Theorem 2 it is natural to introduce the notion of the domain of attraction for \(\Gamma_{\xi}\), denoted \(D(\xi)\). \(D(\xi)\) is the subset of all distributions \(F\) which converges to \(\Gamma_{\xi,\cdot,\cdot}\) and it is natural to try to understand that set and state theorems which clarifies the issue in terms of the tail behavior of the underlying distribution \(F\). Using the notation of Theorem 2 we note that by independence,

\[
P\left(\frac{M_n - b_n}{a_n} \leq x\right) = (F(a_n x + b_n))^n = \exp\left(n \log (1 - (1 - F(a_n x + b_n)))\right)
\]

As for fixed \(x\) we are essentially only interested in the case where the argument \(a_n x + b_n\) is very large we may use a simple approximation of the logarithm to conclude that for large \(n\)

\[
P\left(\frac{M_n - b_n}{a_n} \leq x\right) \sim \exp\left(n(1 - F(a_n x + b_n))\right)
\]
In particular we may conclude that \( n(1 - F(a_n x + b_n)) \to \tau \) if and only if \( P((M_n - b_n)/a_n \leq x) \to \exp(-\tau) \). This is the idea behind the proof of the following theorem.

**Theorem 3** \( F \) belongs to the domain of attraction of \( \Gamma_\xi \) with normalizing constants \( \{a_n\} \) and \( \{b_n\} \) if and only if for every \( x \in \mathbb{R} \)

\[
\lim_{n \to \infty} n(1 - F(a_n x + b_n)) = \Gamma_\xi(x)
\]

Hence, whether or not a given distribution function \( F \) is in the domain of attraction of a certain EVT-distribution depends on the tail behavior of \( F \) in the way described in the Theorem above. For every standard EVT-distribution it is possible to characterize its domain of attraction more precisely. Such theorems can be found in the literature and we have chosen to include one such theorem valid in the Frechet case. This is the case that is a priori thought of as being the most interesting one from the point of view of financial time-series and the result can be stated clearly and simply.

**Theorem 4** Suppose \( X_1, \ldots, X_n \) are \( n \) independent and identically distributed random variables with distribution function \( F \) and suppose that

\[
\lim_{t \to \infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\xi}, x \in \mathbb{R}_+, \xi > 0
\]

Then for \( x > 0 \)

\[
\lim_{n \to \infty} P \left( \frac{M_n - b_n}{a_n} \leq x \right) = \Gamma_{\xi,0,1}(x)
\]

where \( b_n = 0 \) and \( a_n = F^{-1}(1 - 1/n) \).

The converse is also true in the sense that if the last limit holds then the tail behaves as stated. Define

\[
L(x) = (1 - F(x))x^{-1/\xi}
\]

The theorem states that \( F \) is in the domain of attraction of \( \Gamma_{\xi,0,1}(x) \) for some positive \( \xi \) if and only if

\[
\frac{1 - F(tx)}{1 - F(t)} = \frac{L(tx)}{L(t)} x^{-1/\xi}
\]

for some slowly varying function \( L \), i.e., for some function fulfilling

\[
\lim_{t \to \infty} \frac{L(tx)}{L(t)} = 1
\]

for all positive \( x \). An example of a slowly varying function is \( \log(1 + x) \). The function \( 2 + \sin(x) \) is not slowly varying. Theorem 4 also throws some light on how to choose the normalization parameters \( a_n \) and \( b_n \) in this particular case.
We now intend to make a smooth transition to Generalized Pareto distributions. Let $x_F$ be the end of the upper tail. $x_F$ is of course quite often $\{+\infty\}$. The following characterization of the domain $D(\xi)$ may be proven.

**Theorem 5** For $\xi \in R$, $F \in D(\xi)$ if and only if there exists a positive and measurable function $a(\cdot)$ such that for all $x \in R$ such that $(1+x\xi) > 0$

$$\lim_{u \to x_F} \frac{1 - F(u + xa(u))}{1 - F(u)} = (1 + \xi x)^{-1/\xi}$$

The condition in the theorem may be reformulated as

$$\lim_{u \to x_F} P\left(\frac{X-u}{a(u)}|X>u\right) = (1 + \xi x)^{-1/\xi}$$

where $X$ is a random variable having $F$ as its distribution function. Hence, that $F \in D(\xi)$ is equivalent to a condition for scaled excesses over a threshold $u$.

Based on Theorem 5 we will in the following concentrate on the conditional distribution function $F_u$. Define in the following manner for $x > u$,

$$F_u(x) = P(X \leq x|X>u)$$

We also introduce the following notation for any $\xi \in R, \beta \in R_+$:

$$GP_{\xi,\beta}(x) = 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi}_+, x \in R$$

Here, GP stands for Generalized Pareto. Basically Theorem 5 states a $1-1$ correspondence between EVT-distributions and GP distributions. The formal connection can be stated

$$1 - GP_{\xi,\beta}(x) = -\ln \Gamma_{\xi,0,\beta}(x)$$

Using the continuity of the GP-distributions and letting $\beta(u) = a(u)$ we may conclude using Theorem 5 that $F \in D(\xi)$ if and only if for some function $\beta : R_+ \to R_+$,

$$\lim_{u \to x_F} \sup_{u < x < x_F} |F_u(x) - GP_{\xi,\beta(u)}(x-u)| = 0$$

which states that excesses over a threshold is asymptotically (large threshold) described by the GP distributions and that there is a 1-1 correspondence between the statement that $F$ is in the domain of attraction of $\Gamma_{\xi,\cdot},$ and that $F$ has a tail behavior described by the GP-distribution with index $\xi$. 

12
### 3.2 Applying Extreme Value Theory

In the following we focus on:

- Application of extreme value theory and in particular GP-distributions to quantile estimation for high out of sample quantiles.
- Estimation of parameters in models based on extreme value theory.

The crucial question in any applied situation is how to make use of Generalized Pareto distributions in an estimate of the distribution $F$. By definition

$$F_u(x) = P(X \leq x | X > u) = \frac{F(x) - F(u)}{1 - F(u)}$$

Hence,

$$1 - F(x) = (1 - F(u))(1 - F_u(x - u))$$

We now want to make use of extreme value theory in an estimate of the tail for large $x$. In order to do so we have to make a choice for the threshold $u$, i.e., we have to decide where the tail “is to start”. Consider our original sample of $n$ points sorted according to their size, $X_{1,n} \leq \ldots \leq X_{n,n}$. Suppose that we let the upper tail be defined by an integer $k << n$ hence considering $X_{k,n} \leq \ldots \leq X_{1,n}$ to be the observations in the tail of the distribution. This implies that we choose $u = X_{k+1,n}$ to be our threshold. A natural estimator for $1 - F(u)$ is $k/n$. Using the Generalized Pareto distribution for the tail we get, for $x > u$, the following estimator of the distribution function $F(x)$

$$\tilde{F}(x) = 1 - \frac{k}{n} \left(1 + \tilde{\xi} \left(\frac{x - X_{k+1,n}}{\tilde{\beta}}\right)\right)^{-1/\tilde{\xi}}$$

where $\tilde{\xi}$ and $\tilde{\beta}$ are estimates of the parameters in the GP-distribution. Given this as an estimate for the upper tail of the distribution function we solve the equation $q = F(x_q)$ for $q > 1 - k/n$ i.e., for high quantiles. Using the formula above we get

$$\tilde{x}_q = \tilde{x}_{q,k} = X_{k+1,n} + \tilde{\beta} \left(\left(\frac{1 - q}{k/n}\right)^{-\tilde{\xi}} - 1\right)$$

Let us point out that for $x \leq u$ we may choose the empirical distribution as an estimate for $F(x)$. It is obvious that all the estimates stated above are depending on the size of the sample, $n$, and on the threshold $u$ implicitly through $k$. There is furthermore a trade-off when choosing the size of the quotient $k/n$. If this quotient is too small we will have to few observations in the tail, giving rise to large variance in our estimators for the parameters in the GP-distribution. If the quotient is very large the basic model assumption, i.e., the fact that from the
point of view of the asymptotic theory $k(n)/n$ should tend to 0 may be violated. Hence, there is a delicate trade off when choosing $k$. What one therefore have to do is to construct estimators for the parameters $\xi$ and $\beta$ based on the data and a choice of $k$ and then understand the stability of the estimators with respect to the choice of $k$. In the next section we discuss different estimators as well the issue of parameter stability with respect to the parameter $k$.

### 3.3 Monte-Carlo study

In the literature there are several estimators for the parameters in the GP-distributions. In this section we consider two of these, the maximum likelihood estimator and the Hill estimator, and compare their efficiency. The first estimator, i.e., the maximum likelihood estimator, is based on the assumption that the tail under consideration exactly follows a GP-distribution. By differentiating we may find the density of the GP-distribution, hence we can write down a likelihood function. We can then find estimators of the parameters $\hat{\xi}_k$ and $\hat{\beta}_k$ using the standard maximum likelihood approach. This estimator is referred to as the ML-estimator. Provided that $\xi > -1/2$ the ML-estimator of the parameters is consistent and asymptotically normal as the number of data points tends to infinity.

When estimating $\hat{\xi}_k$ one could, assuming a priori that $\hat{\xi}_k > 0$, use the semiparametric result described in Theorem 4 combined with a maximum-likelihood method. In this case we assume that the tail is described as in Theorem 4 and redoing the maximum likelihood approach we get a maximum likelihood estimator of the parameter $\hat{\xi}_k > 0$. This estimator is referred to as the Hill estimator, see Danielson and de Vries (1997). One may also prove that the Hill estimator is consistent and asymptotically normal of $\hat{\xi}_k > 0$.

As emphasized several times already the key issue when applying EVT is the choice of threshold. This is because since we have no natural estimator of the threshold we in fact have to assume, arbitrarily of course, that the tail of the underlying distribution begins at the threshold $u$. Given the choice of $u$, $k < n$ sample points will exceed this threshold. In practice we will however choose a fraction $k/n$ of the sample, hence implicitly choosing a threshold $u$. In a step of estimation the ML and Hill estimators described above, are fitted to the excesses over that threshold. Hence the data used is

$$\{X_{1,n} - X_{(k+1,n)}, \ldots, X_{k,n} - X_{(k+1,n)}\}$$

To evaluate the finite-sample properties (i.e., sensitivity to threshold and sample size) of the ML and Hill based quantile estimators we conduct a Monte-Carlo experiment. For $r = 10,000$ replications samples of size $n = 1000$ and $n = 5000$ is generated from the $t$-distribution with 5 degrees of freedom. Denoting by $\tilde{x}_q$ a quantile estimator and by $x_q$ the true value we compare the relative percent bias
\[ RPS = 100 \times \left[ \frac{1}{r} \sum_{i=1}^{r} \left( \frac{\hat{x}_q - x_q}{x_q} \right) \right] \]

and relative percent standard deviation (RPSD)

\[ RPSD = 100 \times \left[ \sqrt{\frac{1}{r} \sum_{i=1}^{r} \left( \frac{\hat{x}_q - x_q}{x_q} \right)^2} \right] \]

of the Hill and ML implied quantile estimators. This is done for a grid of quantiles and threshold values, specified by choices of \( k/n \). The fraction \( k/n \) is chosen from the interval \([0.05, 0.12]\). The choice of grid for the quantiles range from the 95\% quantile to the 99.5\% quantile but since the results for intervening quantiles can be obtained essentially by linear interpolation we only present results for the 95\%, 99\% and 99.5\% quantiles. Inclusion of the 99\% quantile reflects an interest in the Bank of International Settlements (BIS) capital adequacy directive where an estimate of the 99\% quantile is required.

For the case of \( n = 1000 \) Table 1 show the RPS of the ML and Hill quantile estimators for the quantiles 95\%, 99\% and 99.5\% as \( k/n \in [0.05, 0.12] \). The ML quantile estimator has very small or no RPS at all quantile levels. Moreover it is not sensitive to the choice of threshold. In contrast the Hill estimator has a large RPS for all the quantiles and is very sensitive to the choice of threshold. Looking at the RPSD for \( n = 1000 \) (Table 2) we find that the ML quantile estimator outperforms the Hill estimator again. The RPSD of the ML estimator is invariant to the choice of threshold and never above that of the Hill estimator.

Corresponding figures for the case of \( n = 5000 \) (Table 3 and 4 respectively) show that the Hill quantile estimators large negative RPS at the 95\% level and large positive RPS at the 99% and 99.5% level decreases only slowly with sample size. The RPS is also, as in the case of \( n = 1000 \), very sensitive to the choice of threshold. In addition RPSD of the Hill estimator has not decreased by much whereas RPSD for the ML estimator has decreased by a factor of approximately 2.5 at all quantiles.

The result of the simulation shows that ML is the preferred estimator at all quantiles of interest i.e., \{95, ..., 99.5\}. It always performs better in terms of RPS and RPSD than the Hill estimator. In addition it has the useful property of being almost invariant to the choice of threshold. This is in sharp contrast to the Hill estimator which is very sensitive to the threshold and one may doubt the usefulness of the Hill estimator in empirical applications since the results obtained may be highly dependent on the (arbitrary) choice of threshold. There are also theoretical reasons to prefer the ML estimator. The ML estimator is applicable to light-tailed data as well whereas the Hill estimator is designed specifically for the heavy-tailed case.
Table 1 RPS for n=1000

<table>
<thead>
<tr>
<th>k/n</th>
<th>95% quantile</th>
<th>99% quantile</th>
<th>99.5% quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML Hill</td>
<td>ML Hill</td>
<td>ML Hill</td>
</tr>
<tr>
<td>0.05</td>
<td>-0.307</td>
<td>-0.307</td>
<td>-0.004</td>
</tr>
<tr>
<td>0.06</td>
<td>-0.239</td>
<td>-1.598</td>
<td>0.281</td>
</tr>
<tr>
<td>0.07</td>
<td>-0.153</td>
<td>-2.545</td>
<td>0.337</td>
</tr>
<tr>
<td>0.08</td>
<td>-0.177</td>
<td>-3.341</td>
<td>0.357</td>
</tr>
<tr>
<td>0.09</td>
<td>-0.032</td>
<td>-3.881</td>
<td>0.622</td>
</tr>
<tr>
<td>0.1</td>
<td>0.061</td>
<td>-4.292</td>
<td>0.724</td>
</tr>
<tr>
<td>0.11</td>
<td>-0.033</td>
<td>-4.710</td>
<td>0.692</td>
</tr>
<tr>
<td>0.12</td>
<td>0.016</td>
<td>-4.913</td>
<td>0.754</td>
</tr>
</tbody>
</table>

Table 2 RPSD for n=1000

<table>
<thead>
<tr>
<th>k/n</th>
<th>95% quantile</th>
<th>99% quantile</th>
<th>99.5% quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML Hill</td>
<td>ML Hill</td>
<td>ML Hill</td>
</tr>
<tr>
<td>0.05</td>
<td>5.438</td>
<td>5.438</td>
<td>7.419</td>
</tr>
<tr>
<td>0.06</td>
<td>4.986</td>
<td>5.164</td>
<td>7.468</td>
</tr>
<tr>
<td>0.07</td>
<td>4.893</td>
<td>5.312</td>
<td>7.184</td>
</tr>
<tr>
<td>0.08</td>
<td>5.022</td>
<td>5.731</td>
<td>7.252</td>
</tr>
<tr>
<td>0.09</td>
<td>5.087</td>
<td>6.012</td>
<td>7.282</td>
</tr>
<tr>
<td>0.1</td>
<td>4.882</td>
<td>6.101</td>
<td>7.067</td>
</tr>
<tr>
<td>0.11</td>
<td>4.933</td>
<td>6.427</td>
<td>7.101</td>
</tr>
<tr>
<td>0.12</td>
<td>4.863</td>
<td>6.519</td>
<td>7.081</td>
</tr>
</tbody>
</table>

Table 3 RPS for n=5000

<table>
<thead>
<tr>
<th>k/n</th>
<th>95% quantile</th>
<th>99% quantile</th>
<th>99.5% quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML Hill</td>
<td>ML Hill</td>
<td>ML Hill</td>
</tr>
<tr>
<td>0.05</td>
<td>-0.035</td>
<td>-0.035</td>
<td>0.322</td>
</tr>
<tr>
<td>0.06</td>
<td>-0.155</td>
<td>-1.311</td>
<td>0.512</td>
</tr>
<tr>
<td>0.07</td>
<td>-0.182</td>
<td>-2.293</td>
<td>0.711</td>
</tr>
<tr>
<td>0.08</td>
<td>-0.239</td>
<td>-3.114</td>
<td>0.714</td>
</tr>
<tr>
<td>0.09</td>
<td>-0.271</td>
<td>-3.761</td>
<td>0.896</td>
</tr>
<tr>
<td>0.1</td>
<td>-0.289</td>
<td>-4.271</td>
<td>0.917</td>
</tr>
<tr>
<td>0.11</td>
<td>-0.229</td>
<td>-4.607</td>
<td>1.119</td>
</tr>
<tr>
<td>0.12</td>
<td>-0.122</td>
<td>-4.806</td>
<td>1.187</td>
</tr>
</tbody>
</table>
### Table 4 RPSD for n=5000

<table>
<thead>
<tr>
<th>k/n</th>
<th>95% quantile</th>
<th>99% quantile</th>
<th>99.5% quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML</td>
<td>Hill</td>
<td>ML</td>
</tr>
<tr>
<td>0.05</td>
<td>2.340</td>
<td>2.340</td>
<td>3.298</td>
</tr>
<tr>
<td></td>
<td>4.095</td>
<td>5.685</td>
<td></td>
</tr>
<tr>
<td>0.06</td>
<td>2.253</td>
<td>2.576</td>
<td>3.340</td>
</tr>
<tr>
<td></td>
<td>4.226</td>
<td>7.376</td>
<td></td>
</tr>
<tr>
<td>0.07</td>
<td>2.193</td>
<td>3.116</td>
<td>3.330</td>
</tr>
<tr>
<td></td>
<td>4.241</td>
<td>9.499</td>
<td></td>
</tr>
<tr>
<td>0.08</td>
<td>2.199</td>
<td>3.723</td>
<td>3.229</td>
</tr>
<tr>
<td></td>
<td>4.098</td>
<td>11.67</td>
<td></td>
</tr>
<tr>
<td>0.09</td>
<td>2.221</td>
<td>4.258</td>
<td>3.332</td>
</tr>
<tr>
<td></td>
<td>4.198</td>
<td>14.54</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>2.218</td>
<td>4.713</td>
<td>3.332</td>
</tr>
<tr>
<td></td>
<td>4.229</td>
<td>17.41</td>
<td></td>
</tr>
<tr>
<td>0.11</td>
<td>2.172</td>
<td>4.997</td>
<td>3.375</td>
</tr>
<tr>
<td></td>
<td>4.269</td>
<td>20.93</td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td>2.212</td>
<td>5.196</td>
<td>3.422</td>
</tr>
<tr>
<td></td>
<td>4.249</td>
<td>24.61</td>
<td></td>
</tr>
</tbody>
</table>

### 3.4 General remarks

We here just shortly want to emphasize a few features of EVT.

- EVT techniques make it possible to concentrate on the behavior of the extreme observations. It can be used to estimate out of sample quantiles. Hence, being a theory of extrapolation.

- The extreme value method does not assume a particular model for returns and therefore the model risk is considerably reduced.

- The parameter estimates of the limit distributions depend on the number of extreme observations used. The choice of a threshold should be large enough to satisfy the conditions to permit its application \((u \text{ tends towards infinity})\), while at the same time leaving sufficient observations for the estimation. Different methods of making this choice can be used, but in general estimation risk can be an issue.

- An important assumption in the theory is that observations are independent and identically distributed. Deviations from the assumption of independent and identically distributed random variables may result from trends, periodicity, autocorrelation and clustering. By one or more methods one have to try to filter out some residuals to which the estimation methods can be applied. Clustering effects which may be the result of stochastic or time-dependent volatility may be handled using a two-step approach based on GARCH and EVT as described in this paper.

### 4 Risk measures

A key issue in risk management is the definition of risk as well as the definition of relevant risk measures. The most commonly used risk measure, Value-at-Risk,
is defined as follows at the \((1 - \alpha) \times 100\%\) confidence level and time horizon one day:

\[
VaR_\alpha(X) = -\inf \{x | P(X \leq x) > \alpha\}
\]

Note that in this definition we have defined VaR to be a positive number. VaR has become something of a industry standard but as a measure of risk VaR is quite often criticized. The most relevant criticism stems from the following two issues.

- **VaR** gives no information about the conditional probability

\[
P(X \leq x | X < -VaR_\alpha(X))
\]

for \(x < -VaR_\alpha(X)\) i.e., it does give us no information about the probability of losses larger than VaR.

- **VAR** is by its very nature not a subadditive measure of risk, i.e., in general it is not true that \(VaR_\alpha(X + Y) \leq VaR_\alpha(X) + VaR_\alpha(Y)\) and hence VaR is not able to properly detect diversification effects.

The first of these two issues concerns the tail of the distribution function of the risk factor under consideration and hence finding the “correct” model for the tail one may completely understand the VaR-number for higher order quantiles. In that case the first issue evaporates. The conclusion is therefore that the first issue is not too much of a problem if we are able to correctly model the tails of distributions. The second issue, i.e., the lack of subadditivity, is a more serious problem and is on the theoretical level clearly a drawback for VaR as a risk measure. One should note though that an analytic VaR-number like Delta-Normal VaR is in fact subadditive. This can easily be seen from the inequality

\[
x^2 + 2\rho xy + y^2 \leq (x + y)^2
\]

for \(\rho \in [-1, 1]\).

Both from a theoretical and practical point of view it is of importance to try to define and perhaps axiomatize what to demand from an ideal measure of risk as well as how to construct appropriate risk measures. A theoretical solution to this problem have been suggested by Artzner et al. (1999) by introducing the concept of coherent measures of risk. They have created a unified framework for the analysis, construction and implementation of measures of risk. In order to formulate their axioms we need some notation. Let \(\Omega\) denote the states of the world and let \(\Upsilon\) be the set of all random variables \(X : \Omega \rightarrow R\). We may think of \(\Omega\) as the set of all potential scenarios for the world and \(\Upsilon\) as all possible risks. They define a risk measure, i.e., a function \(m : \Upsilon \rightarrow R\) to be coherent if it satisfies the following conditions for all \(X, Y \in \Upsilon, t \in R_+, r \in R\),

- \(m(X + Y) \leq m(X) + m(Y)\) (subadditivity)
• \(m(tX) = tm(X)\) (homogeneity)
• \(m(X) \geq M(Y)\) if \(X \leq Y\) (monotonicity)
• \(m(X + r) = m(X) - r\) (risk-free condition)

They also prove the following theorem stating that all coherent measures of risk are obtained by means of “generalized scenarios”.

**Theorem 6** A risk measure \(m\) is coherent if and only if there exists a family \(\Phi\) of probability measures on \(\Omega\) such that

\[
m(X) = \sup \{E_Q(X) | Q \in \Phi\}
\]

for all \(X \in \mathcal{Y}\) where \(E_Q\) denotes expectation w.r.t \(Q\).

For a set \(\Phi\) containing only one probability measure \(Q\) the mean \(E_Q[X]\) obviously defines a coherent risk measure. On the other hand, the more scenarios or rather the more probability measures used in a set \(\Phi\), the larger risk measure obtained and hence the more conservatively we measure risk. The theorem reduces the issue of constructing coherent risk measures to the construction of appropriate probability measures. In particular for the purpose of constructing risk measures for stress tests it is natural to consider

\[
m(X) = \max \{X(w_1), \ldots, X(w_k)\}
\]

with \(w_1, \ldots, w_k\) being a finite set of scenarios, i.e., a finite set of states of the world. That this is a coherent risk measure is easily seen by choosing \(\Phi = \{\delta_{w_1}, \ldots, \delta_{w_k}\}\) where \(\delta_w\) is the idealized probability measure having all its mass concentrated in \(w\). A relevant quantity in this context is

\[
m(X) = E(X | X \leq -VaR_\alpha(X))
\]

where as always \(\alpha > 0\) corresponds to a suitable confidence level. This quantity is usually referred to as expected short-fall. Choosing \(\Phi = \{P(\cdot | A)\}\) where \(A\) ranges over all events with probability \(\geq \alpha\) we may conclude, using Theorem 6, that expected shortfall is a coherent measure of risk. It is interesting to note that if the random variable \(X\) has a polynomial decaying Pareto type tail and if we plot the expected shortfall as a function of \(VaR_\alpha(X)\) a straight line will appear.

Although it is easy to agree with the framework suggested by Artzner et al. there is a theoretical criticism towards the axiom stating that a coherent risk measure has to be subadditive. Of course subadditivity is a natural condition but the following argument can be found in the literature. Suppose that we have two independent risks or random variables \(X\) and \(Y\) both having polynomial decaying tails of the type described by the Generalized Pareto distribution. Hence
we assume that \( P(X > x) = c_1 x^{-\alpha_1} \) and that \( P(Y > y) = c_2 y^{-\alpha_2} \). Let in the following for simplicity \( c_1 = c_2 = c \) and assume that \( \alpha_1 = \alpha_2 = \alpha \). If \( \alpha > 1 \) the distributions have finite first order moments. Assuming independence of \( X \) and \( Y \) one may prove that for large \( x \),

\[
P(X + Y > x) \sim P(X > x) + P(Y > x)
\]

Letting the \( p\%-\)quantile be our risk measure we have \( m(X) = \text{VaR}_p(X) \) and \( m(X) \) solves the equations \( 1 - p = P(X > m(X)) = c(m(X))^{-\alpha} \). Hence,

\[
m(X) = (c/(1 - p))^{1/\alpha}
\]

and we get

\[
m(X + Y) \sim (2c/(1 - p))^{1/\alpha} = 2^{1/\alpha}(c/(1 - p))^{1/\alpha}
\]

\[
m(X) + m(Y) = 2(c/(1 - p))^{1/\alpha}
\]

If \( \alpha < 1 \) then obviously \( 2^{1/\alpha} > 2 \) and the measure \( m \) is not subadditive. This argument shows that for really heavy tailed distributions the condition of subadditivity may be irrelevant. Still this is not the case for the financial risk factors we consider as none of them will show tails resulting in a lack of finite first order moments.

> From our perspective the following conclusions can be drawn.

- Once a careful analysis have detected the “correct” tail behavior of a risk factor using Generalized Pareto distributions one may plot the VaR-number for all high quantiles hence understanding its variation. From this perspective it is of less interest whether or not VaR is a coherent measure of risk. That issue becomes more relevant when the complexity increases in terms of larger portfolios.

- Expected shortfall is a natural construction and coherent risk measure. Once the distribution of our risk factor is properly understood expected shortfall must be considered as the first risk measure beyond VaR to be seriously considered.

- The canonical way of constructing coherent risk measures is through the specification of a set of generalized scenarios, states of the world. The problem is hence reduced to the construction of probability measures on the states of the world. This way of looking at the construction of risk measures fits very well into a program for stress testing.

## 5 Application

In the following we describe the empirical study we have carried out and summarize the conclusions that we think can be drawn. We have applied the methods
to several Equity, FX (Foreign Exchange) and IR (Interest Rate) risk factors. To conserve space and also because the obtained results are very similar for the different time series we only present results for the risk factors

- Equity. ABB (1993/01/20-2001/09/06), S&P 500 (1921/02/01-1991/10/01)
- FX. USD/SEK (1993/01/20-2001/09/06)

where the dates within brackets after each risk factor gives the length of the time series used. Given each time series we transform the data to log returns, \(y_t\).

The analysis then proceeds in two steps. First we use a combination of ARMA and GARCH in order to capture dependence between consecutive returns. In this step we found it sufficient to use the following ARMA model for returns.

\[
y_t = \mu + \sum_{i=1}^{5} \phi_i y_{t-i} + \epsilon_t
\]

We also have an asymmetric GARCH process for \(\epsilon_t = z_t h_t\) where

\[
h_t^2 = a_0 + a_1 \epsilon_{t-1}^2 + a_2 \text{sgn}(\epsilon_{t-1}) \epsilon_{t-1}^2 + b h_{t-1}^2
\]

This means that in this preliminary (filtering) step there are in total 10 parameters to be estimated. This vector of parameters is denoted

\[
\theta = (\mu, \phi_1, \phi_2, \phi_3, \phi_4, \phi_5, a_0, a_1, a_2, b)
\]

which is estimated using GMM and with the (normal) quasi-maximum likelihood estimator providing initial consistent estimates. Using the model just described we filter the original process, \(y_t\), hence obtaining a residual process, \(z_t\)

\[
z_t = (y_t - \mu - \sum_{i=1}^{5} \phi_i y_{t-i}) / h_t
\]

Now, consecutive realizations of the process \(z_t\) should in the best of worlds be close to being independent and identically distributed. Looking at the autocorrelation of \(z_t^k\) and \(z_{t-j}^k\), \(k = 1, 2\), and also more formally considering Box-Pierce tests one may conclude that this is indeed the case.

In the second step we focus entirely on the residual process \(z_t\) and we try to model the tails of the distribution \(F\) from which we assume the data \(z_t\) to be drawn. First we construct the empirical distribution function (EDF). Then we consider three different models for \(F\), two symmetric and one asymmetric. The symmetric models are the normal distribution (Normal) and the Student-t distribution. In the case of the t-distribution the degrees of freedom parameter \(\nu\) is also estimated. These two models are compared with an asymmetric model for the tails based on extreme value theory and in this case we differentiate between
the upper an lower tails. Based on the analysis in Section 3 of the paper we fit a GP-distribution (using ML estimation) to the upper and lower tails of the empirical distribution using uniform thresholds for all the time-series in the sense that the upper tail is defined to be the largest 10% of the realizations of the process $z_t$ and the lower tail is defined to be the smallest 10%.

First we focus on the calculation of daily risk measures. We proceed as follows. Suppose that we have estimated the parameter vector $\theta$ in the ARMA-GARCH model using data up to time $T$. This also implies that we have trimmed a model for the residual process $z_t$ using data up to time $T$. We denote by $Z$ a random variable having the constructed distribution as it distribution function. Then,

$$y_{T+1} = \mu + \sum_{i=1}^{5} \phi_i y_{T+1-i} + h_{T+1} Z$$

where the volatility factor $h_{T+1}$ is a deterministic forecast obtained from the GARCH process. Hence, $y_{T+1}$ is a random variable with randomness quantified by the random variable $Z$. In Figure 1 the reader finds quantile plots of the upper and lower tails of the distribution of $y_{T+1}$ for the time series we consider here (ABB and USD/SEK). The plots contain the EDF as well as plots of the quantile of $y_{T+1}$ based on each of the three models for $Z$. The forecasted volatility is part of the plots and the effect of a higher or lower volatility forecast $h_{T+1}$ would be reflected in a shift of the plots upwards respectively downwards. In particular if we today at time $T$ experience extremely high volatility this would be fed through the GARCH to a much higher forecast volatility $h_{T+1}$ than if we today experienced moderate volatility.

If we apply a risk measure $m$ to $y_{T+1}$ and assuming that the risk measure is homogenous and fulfills a risk-free condition (see Section 4) we get

$$m(y_{T+1}) = \mu + \sum_{i=1}^{5} \phi_i y_{T+1-i} + h_{T+1} m(Z)$$

Hence, the risk in $y_{T+1}$ is determined by the risk in $Z$ and we again see that a higher forecasted volatility give higher risk. We have also considered expected shortfall as a daily risk measure at two levels of confidence, 95% and 99%, and we have only calculated the measures for the lower tail, see Table 5.

The following general conclusions for daily risk measurement can be drawn from our empirical studies.

- In general the Normal model under-estimates the lower tail but over-estimates the upper tail. The t-distribution tend to over-estimate both the upper and lower tails.
- In general the tail index, i.e., $1/\xi$ tends to be smaller for the lower than for the upper tail implying that the lower tail is in general fatter than the upper tail.
Figure 1 Quantile plots for daily returns of ABB and USD/SEK

1a) Left tail (ABB)

1b) Right tail (ABB)

1c) Left tail (USD/SEK)

1d) Right tail (USD/SEK)
Table 5: Expected shortfall for daily returns of ABB and USD/SEK

<table>
<thead>
<tr>
<th>Density</th>
<th>ABB ES(95%)</th>
<th>ABB ES(99%)</th>
<th>USD/SEK ES(95%)</th>
<th>USD/SEK ES(99%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NID</td>
<td>8.76</td>
<td>11.19</td>
<td>1.40</td>
<td>1.81</td>
</tr>
<tr>
<td>EDF</td>
<td>9.64</td>
<td>15.33</td>
<td>1.51</td>
<td>2.17</td>
</tr>
<tr>
<td>EVT</td>
<td>11.41</td>
<td>17.18</td>
<td>1.46</td>
<td>2.10</td>
</tr>
</tbody>
</table>

- Focusing on the lower tail, at the 95% confidence level VaR is rather insensitive to the choice of model for the residuals. For the same level of confidence the expected shortfall shows a higher model dependence. The difference in relative terms between the EVT-model and the Normal-model is in general about 30% with the EVT-model giving a higher risk.

- For higher quantiles like the 99% confidence level both risk measures show a significant model dependence. For VaR the relative difference between the EVT-model and the Normal-model is often as much as 30%. For expected shortfall the difference can often be as much as 50 – 100%.

- Concerning differences between risk factors the EVT-model tend to yield more value-added in the case of stocks than in the case of FX and IR.

In general we can conclude that for quantiles beyond say the 97 – 98% level of confidence refined methods like the EVT-model give a significant contribution.

It is also natural to consider s-days (for s > 1) ahead risk measurement and in this case we need to consider the distribution of \( \sum_{i=1}^{s} y_{T+i} \) which is in general unknown (e.g., even for normal Z with GARCH variance). However, we can of course simulate cumulative paths of \( y_{T+s} \) (using the assumed distribution of Z) and hence thereby obtain an estimate of the distribution. We refer the interested reader to McNeil and Frey (2001) for an application of this method for s = 10 i.e., 10-days ahead risk measurement as demanded by Basel.

Instead, we consider risk measures focusing on the cumulative tail probabilities of daily maxima and minima losses within a given time period. Such measures are interesting for the obvious reason and we exemplify with the S&P 500 equity index and a window of 20 days (monthly) yielding approximately 850 observations. Figure 2 contains the quantile plots in this case. The reader notices that in contrast to daily risk measurement we have big differences between models even at low levels of probability. Furthermore, and not very surprising, the conditional distribution of returns is essentially the same as the unconditional i.e., the risk measure is largely time-independent. Hence, the filtering step is largely unnecessary.
6 Summary and conclusions

When considering financial time-series the modelling and forecasting of volatility is a key issue. It is obvious that if volatility fluctuates in a forecastable manner a good forecast of volatility can improve the management of risk. In any model of time-series one could potentially work under the following two paradigms.

- Volatility is not forecastable.
- Volatility is forecastable.

Deciding, in a particular situation, which of the complementary paradigms that is the relevant one has strong methodological implication. If volatility is forecastable an accurate modelling must be based on a dynamic, conditional description of the extreme movements in financial returns. If volatility is not forecastable then a direct, unconditional modelling of the extreme returns ought to be the correct approach.

In a particular situation the time-scale over which returns are considered is of most importance in a decision about whether one should consider volatility as forecastable or not. For one day returns volatility tend to be forecastable as the analysis in this paper shows. If we increase the time-scale from one day returns to say 20 days returns, a 20 day volatility probably has less (but still potentially
significant) influence on the next 20 days volatility. In our case we have so far considered one day returns and one conclusion is that the GARCH process does indeed explain much of the dependence structure in the time-series we have considered. In particular after filtering we have seen that, for quantiles less than say 97 – 98% it is more important over a short time scale to properly model stochastic volatility than it is to fit more or less sophisticated distributions to the filtered residuals. This implies that for 95% VaR the standard model based on normally distributed variables combined with the GARCH model for volatility is good enough. Still as our studies show, for quantiles higher than 97 – 98% the use of EVT does indeed give a substantial contribution and the generalized Pareto distributions are more able than the normal distribution to accurately model the empirically observed tail.

As discussed above, for larger time-scales, the importance of the GARCH process will decrease but also by considering the return over the larger time-scale as a sum of daily returns one realize that a modelling of the tails of the returns using Extreme value theory will become even less important. However, something fundamentally different will occur if we divide the data into blocks of say 10 – 20 days and take the maxima of the daily returns within each block. The stochastic variables constructed in this way will show limited dependence and in particular their dependence would go to zero if the block size was allowed to go to infinity in suitable fashion.
References

Most of the references below are referred to in the bulk of the paper. Here we just want to make a few remarks on the literature. Apart from the standard Embrechts, Mikosch and Klüppelberg reference for the extreme value theory we want to make the reader aware of the book of Resnick as a good reference on regularly varying functions and random variables. Concerning many aspects of estimation principles and econometric models for time-series analysis we refer the reader to the book of Hamilton.


