Forecasting Expected Shortfall

An Extreme Value Approach

Benjamin Kjellson

Bachelor's thesis
2013:K7

LUND UNIVERSITY

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

Abstract

We compare estimates of Value at Risk and Expected Shortfall from AR(1)-GARCH(1,1)-type models (standard GARCH, GJR-GARCH, Component GARCH), to estimates produced using the Peak Over Threshold method on the residuals of these models. We find that the conditional volatility model matters less than the choice of distribution for the innovations in the loss process, for which we compare the normal and the t-distribution. The Peak Over Threshold estimates are found to improve upon the estimates of the original models, particularly in the case of normally distributed innovations.

Acknowledgements

I sincerely thank my supervisor Nader Tajvidi for his patience and wise suggestions in my writing of this thesis, and for introducing me to Extreme value theory. I am also thankful for the education that I have received at the Centre for Mathematical Sciences at Lund University, and I express my gratitude to all the inspiring and passionate lecturers I have had during my stay in Lund.

Contents

List of Tables

List of Figures

1 Introduction

The recent financial crisis illustrated, to some extent, the inadequacy of traditional risk measures such as Value at Risk. Although the current and upcoming regulatory framework for supervision and risk management of the banking sector, Basel II and Basel III respectively, still cling to this measure, financial institutions and actors still need proper risk measures for internal use. In this thesis, we consider the risk measure expected shortfall as calculated through variations of the popular AR(1)-GARCH(1,1) model, combined with the Peak Over Threshold model from Extreme value theory to better capture heavy-tail risks. We begin with a short qualitative description of risk management and what the purpose of a risk measure is, what problems are associated with financial data, and how the suggested method for estimating expected shortfall can overcome some of these problems.

1.1 Risk Management, Risk Measures, and Financial Data

Risk management, in its broadest scope, is a systematic approach to identifying, measuring, and controlling risks, whatever they may be (Jorion, 2001, p. 3). In this thesis though, we limit ourselves to financial risks on the asset level. The most frequently used risk measure for asset or portfolio risk, and indeed the measure banks and financial institutions must use according to the Basel framework, is Value at Risk (VaR).

VaR is simply a quantile of the loss distribution, telling us, for example, what our worst loss 95 days out of a hundred is expected to be. VaR is thus easy to interpret, and it further lends itself to parametric modelling and backtesting[1]. However, there are also downsides to VaR as a risk measure, particularly from a mathematical and statistical standpoint. An often-cited paper by Artzner et al. (1999) identifies a few properties that a risk measure ought to have. One of these properties is subadditivity, which for a real valued function $f\colon A \to B$ and elements $a, b \in A$ means $f(a+b) \leq f(a) + f(b)$. VaR does not have this property, as the VaR of a portfolio may sometimes be larger than the sum of the VaR for the individual assets in the portfolio (McNeil et al., 2005, p. 40).

An additional downside to VaR is that it does not tell us what the losses look like when VaR is exceeded. An alternative risk measure that solves exactly this problem is expected shortfall (ES), which is the average of the losses beyond a certain quantile of the loss distribution[2]. That is, ES does not only tell us about the probability of large losses occurring, but also informs us about the likely magnitude of these losses. ES has not only the previously mentioned subadditivity property, but also fulfills the other conditions of a "coherent" risk measure as outlined in Artzner et al. (1999). However, ES also has some downsides which may explain why it is used less often than VaR. Yamai & Yoshiba (2002) showed that the ES measure requires a much larger sample to achieve the same accuracy in backtesting than VaR, which is not strange considering the fact that ES relies on first estimating the VaR, and then adding additional estimates to that. Furthermore, VaR backtests have a much stronger theoretical underpinning than do ES backtests.

[1] Testing the measure on historical data without (intentional) look-ahead bias.

[2] Here, we consider losses to be positive real numbers. Further, we take the q-quantile to mean the value below which losses are expected to fall $100q\%$ of the time, with q often being a number in the range [0.9,1).

There are a few stylized facts of financial returns that must at least be considered in selecting a risk measure. In this paper, we will pay particular attention to three of these commonly observed phenomena. Volatility clustering is the term used to describe the fact that that large changes in asset prices tend to be followed by further large changes, and that small changes are similarly often followed by small changes (Brooks, 2008, p. 386–387). A proper risk measure ought then to take account of a sudden spike in volatility in estimating the risk for the following days, and not treat the spike as a one-off event. Another well-known fact is that financial data seem to be generated from distributions with fat tails, meaning that using a normal distribution to model returns may underestimate the frequency of large losses or gains. Therefore, the analytic simplicity of the normal distribution may need to be sacrificed for a distribution or simulation technique that better models reality, so as not to underestimate risks. A third observation about financial data is the so called leverage effect, noted e.g. by Black (1976). This effect, somewhat improperly named, describes the assymmetry in the influence of past shocks on current volatility, in the sense that a large loss in the past is associated with higher current volatility than is a equally large gain. These three stylized facts form the starting point of this paper.

1.2 Goal

This thesis aims to show how the risk measures Value at Risk and Expected shortfall can be estimated by augmenting variations of the popular AR(1)-GARCH(1,1) model with the Peak Over Threshold model from Extreme value theory. The performance of each model, with or without this augmentation, will be evaluated though backtesting the estimates on a few financial data sets. This will allow for a comparison of the different models. To be specific, the "base" models from time series analysis that we will use will all include an autoregressive component of order one for the return, and the conditional variance models that will be used are

- The GARCH(1,1) model introduced by Bollerslev (1986), which allows us to take into account volatility clustering in our ES estimates.
- The GJR-GARCH(1,1) model by Glosten et al. (1993), which can also account for the leverage effect.
- The Component GARCH(1,1) model of Lee & Engle (1999), in which the conditional variance is decomposed into two parts corresponding to transitory and permanent effects, so that both long-run and short-run movements in volatility are accounted for.

For each such combination of an AR(1) process and a GARCH(1,1) type process, the distribution of the error terms must be decided. For reasons of parsimony, we have used the normal distribution and the Student's t distribution with four degrees of freedom in estimation—the former to see if it is indeed inadequate and will lead to underestimation of risk, and the latter both of its simplicity in use, and for its fat tails. With these models, we extend the research by McNeil & Frey (2000), who covered and tested expected shortfall only for the AR(1)-GARCH(1,1) model with normal and t-distributions, and the corresponding POT models fitted to the residuals. In addition, we implement the double bootstrap algorithm by Danielsson et al. (2001) for selection of threshold in the POT models, instead of the more subjective plot-inspection approach by McNeil & Frey (2000).

1.3 Caveats

In producing this paper, some choices had to be made that were either somewhat arbitrary or just followed convention when no theoretical basis existed. These choices, and their motivation, include

- Backtesting window length: the length of 1000 observations was chosen because it was the length used in the paper by McNeil & Frey (2000). With about 250 trading days in a year, 1000 observations correspond to around 4 years of data, which admittedly may exclude extreme observations prior to calm stretches of time such as that ending in 2007.
- Data sets: we have chosen to test the models on three data sets, each from a different asset class. More data sets could of course have been included, but there is a trade-off in the amount of time it takes to run a backtest and how much information an additional data set will give us.
- Choice of forecast horizon: we limit ourselves to 1-day forecasts, while multiple-day forecasts may actually be more common in practice. This is again motivated by what the convention seems to be in the articles we have based this paper on.
- Univariate approach: we make our VAR and ES estimates on an asset-by-asset basis rather than on a portfolio of assets, which would be more realistic. For expected shortfall, this can be motivated by the subadditivity property discussed above.

1.4 Previous Research

Value at Risk and Extreme value theory is covered well in most books on risk management and VaR in particular, see for example Hull (2006), Jorion (2001), McNeil et al. (2005), and Dowd (2005). Vice versa, VaR is treated in some Extreme value theory literature, such as Embrechts et al. (1997) and Coles (2001). Expected shortfall is covered in practically every book on risk management as well, although to a much lesser extent than VaR. Instead, we look to the paper by McNeil & Frey (2000) as inspiration for this thesis; a paper that has generated many follow-up studies, such as those by Byström (2004), Gilli & Këllezi (2006), and Tolikas et al. (2007). Newer papers on expected shortfall in combination with conditional volatility models or Extreme value theory often use Bayesian methods; see for example the papers by Hoogerheide & van Dijk (2010) and Gerlach et al. (2012).

2 A Look at Some Financial Data

To gain some better intuition about the supposed fat-tailedness and dependence in financial data, we will take a look at one of the data sets used later on. We won't delve into the theory here—that is done in the next section—but rather just make a few short comments about what we can see. We will have a look at data from the S&P500 index, which is a market cap-weighted index of 500 prominent companies publicly traded in the US stock market. We look at the full data set used for our backtests later on; the models we will use later were decided beforehand, so this poses no problem in terms of a look-ahead bias. We begin by plotting the index series:

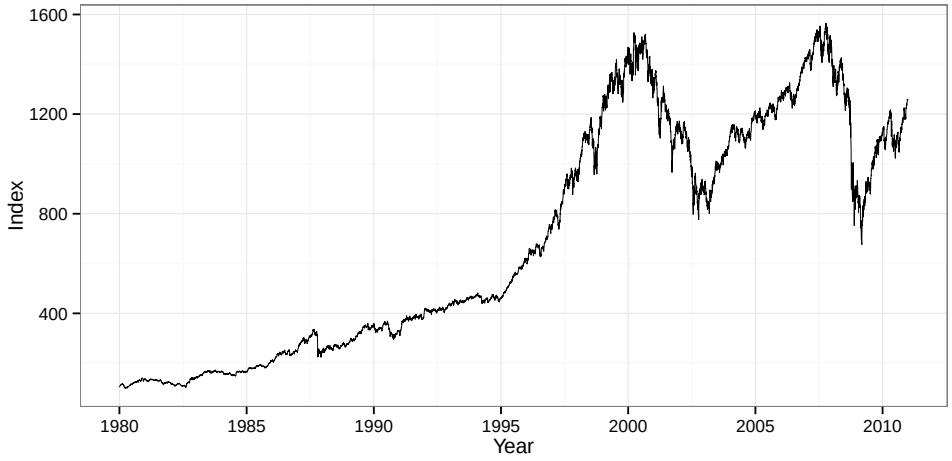


Figure 1: The S&P 500 index, with data between 1980 and 2011.

Here we see the long upward trend until the dot-com bubble and the recession in the early 2000s; the recovery in the years after, and then the financial crisis. Next, we look at the losses of the index, keeping an eye out for any trends or clusters in the volatility:

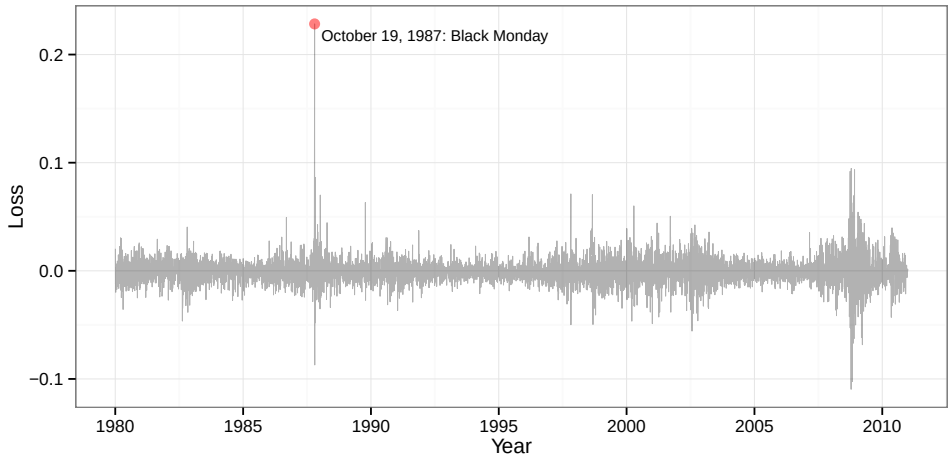


Figure 2: Losses of the S&P 500 index in the years 1980–2011.

Judging by figure 2, it seems as if there have been periods when the volatility of the losses

has been high or low for longer, indicating that it may be dependent on its past. To investigate dependence, we fit one of the models for the losses that we also use later on: an AR(1)-GARCH(1,1) process with normally distributed innovations. Looking at the conditional volatility below, it seems as if there are both short spikes in volatility, and trends that hold for longer.

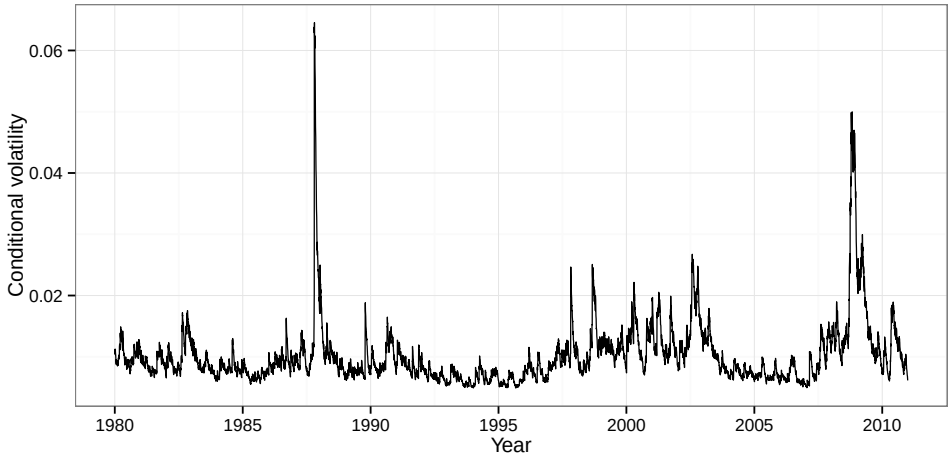
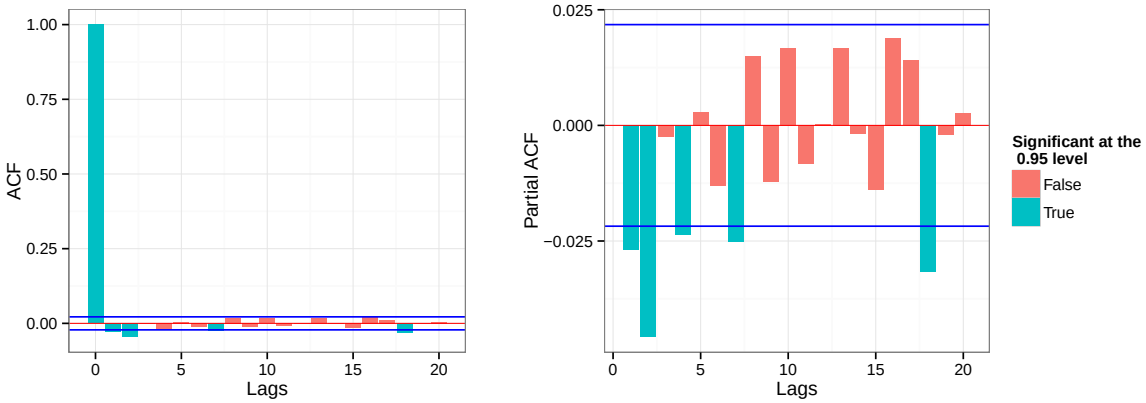


Figure 3: Conditional volatility of the S&P 500, from a fitted AR(1)-GARCH(1,1) model with normally distributed innovations.

To investigate correlations between losses at different times, we take a look at the autocorrelation and partial autocorrelation functions for the losses and for the squared losses below.



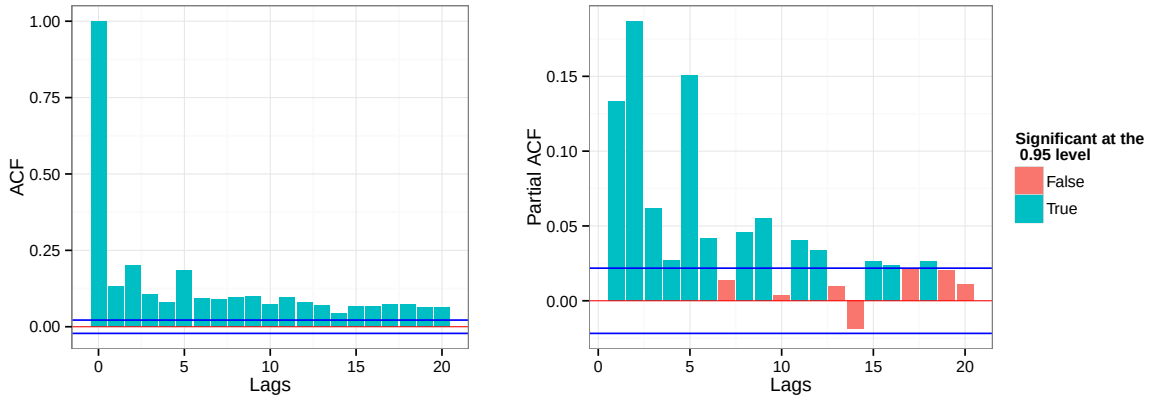
(a) ACF for S&P Losses

(b) PACF for S&P Losses

Figure 4: Autocorrelations and partial autocorrelations for the losses of the S&P 500.

The plot of the ACF indicates a nonexistent or at least low MA order for the losses, while the plot of the PACF indicates a low AR order (assuming the spike at lag 18 is spurious).

For the squared losses, seen in figure 5 below, the dependence is clearer. As the squared returns (or equivalently losses) are usually taken as an approximation of the variance, this indicates that a model in which volatility is allowed to be dependent is suitable for our purposes.

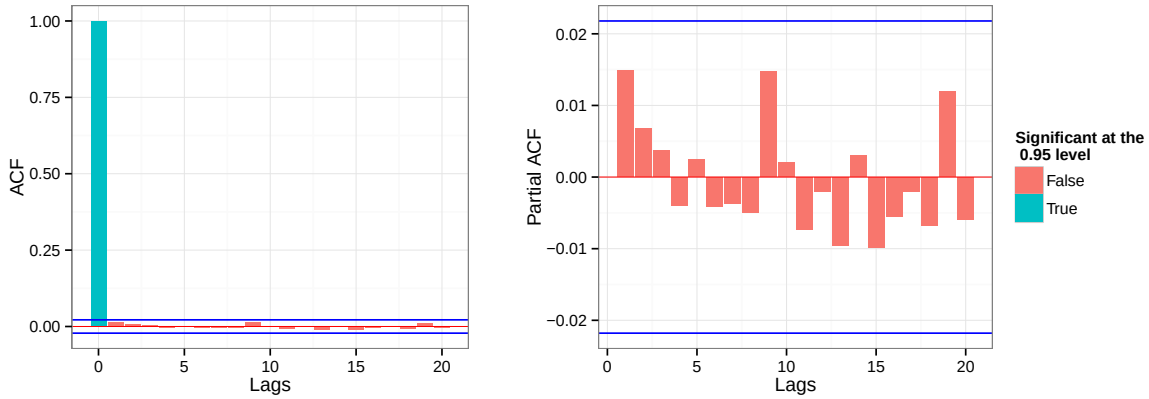


(a) ACF for S&P squared losses

(b) PACF for S&P squared losses

Figure 5: Autocorrelations and partial autocorrelations for the squared losses of the S&P 500.

It is then interesting to look at the residuals of the fitted AR(1)-GARCH(1,1) model, to see if it has successfully removed dependence in the losses and in the volatility. As the dependence in the losses as they are is already low, we focus on the squares of the residuals, and plot the ACF and PACF for these.



(a) ACF for the squared residuals

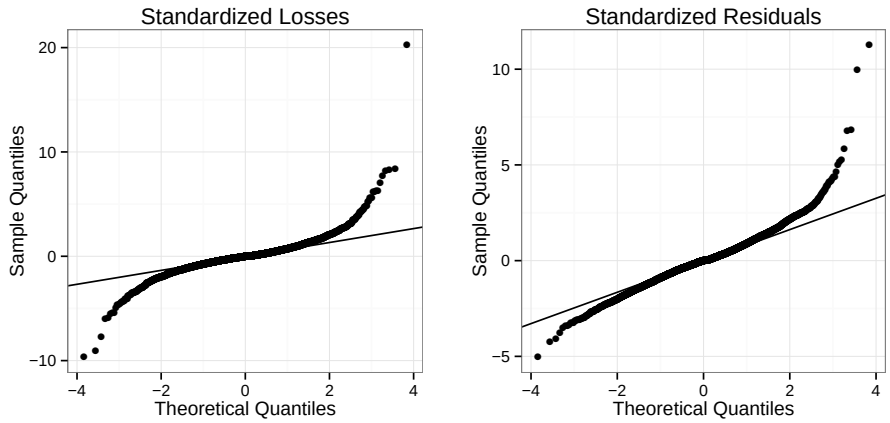
(b) PACF for the squared residuals

Figure 6: Autocorrelations and partial autocorrelations for the squared residuals of the fitted AR(1)-GARCH(1,1) model.

Clearly, the autocorrelation in the squared residuals is smaller than that of the squared losses, which indicates that fitting an AR(1)-GARCH(1,1) model to the data may be a good way of obtaining independent residuals, on which we can then apply methods from

Extreme value theory to estimate VaR and ES.

To investigate fatness of tails, we can look at Q-Q plots, which show quantiles of the empirical distribution against the theoretical distribution, here chosen to be the normal distribution. We make such plots for both the standardized losses and the standardized residuals of the AR(1)-GARCH(1,1) model, to compare.



(a) Q-Q plot of S&P 500 losses

(b) Q-Q plot of residuals

Figure 7: Q-Q plots for the standardized losses and the standardized residuals of the fitted AR(1)-GARCH(1,1) model, for the S&P 500.

As seen in the Q-Q plots above, the data shows signs of fatter tails than what would be the case had losses been normally distributed. Even though the conditional model makes the situation slightly better, the signs of fat tails are still evident. This speaks for fitting a Generalized Pareto distribution to the tails, in order to estimate value at risk and expected shortfall more accurately.

3 Theoretical Background

In this section, we present the mathematical details of the risk measures and models mentioned. The aim is to give an outline of the theory behind our estimator for expected shortfall, without delving too deep into the general theories. Before we begin, we take a moment to define the variables we will be working with, namely the return and losses on financial assets.

3.1 Asset Returns and Losses

Denoting by S_t, $t \in \mathbb{Z}$, the closing price of an asset at day t, the raw returns for day $t+1$ are defined by

$$\frac{S_{t+1} - S_t}{S_t} = \frac{S_{t+1}}{S_t} - 1 \tag{3.1}$$

Since $\log(1+x) \approx x$ for x close to zero, and since one-day returns are usually small, the one-day raw returns can be approximated by the one-day log returns, which are defined as

$$R_{t+1} = \log\left(\frac{S_{t+1}}{S_t}\right) = \log(S_{t+1}) - \log(S_t) \tag{3.2}$$

We then define the loss X_t at day t as the negative of the log-return, i.e. $X_t = -R_t$. As we are interested in future returns and future losses, we consider these to be (continuous) random variables. We treat losses as positive numbers rather than negative ones out of convenience; for example, most literature on extreme value theory deals with the upper tails of distributions, although the results apply equally well to the lower tails (through a simple transformation). We will be working with losses in this form throughout the paper, as opposed to measuring losses in money amounts (of a fictitious portfolio, or something similar).

As a basis for all models we test in this paper, we will assume that the losses $\{X_t, t \in \mathbb{Z}\}$, form a stationary time series where

$$X_t = \mu_t + \sigma_t Z_t \tag{3.3}$$

where $\{Z_t\}$ are iid (independent and identically distributed) continuous random variables with mean zero, unit variance, and come from a location-scale family, and where μ_t and σ_t are measurable with respect to the return process up to time $t-1$.

3.2 Value at Risk

The value at risk for a given confidence level $q \in (0,1)$ and time t is given by the smallest number x_q such that the loss X_{t+1} at time $t+1$ will fall below x_q with probability q:

$$\mathrm{VaR}_q^t = \inf\{x_q \in \mathbb{R} : P(X_{t+1} \leq x_q) \geq q\} = \inf\{x_q \in \mathbb{R} : P(X_{t+1} > x_q) \leq 1 - q\} \quad (3.4)$$

Thus, value at risk is a quantile of the loss distribution, and q is usually taken to be in the range [0.9,1), so that we may talk about the "95%-VaR" or "99%-VaR". Sometimes, the *coverage rate* $\alpha = 1 - q$ is used instead. Note that the above is just the one-day VaR; the concept can be extended to longer horizons of five days, ten days, or more. This can be done in several ways; a common way for shorter horizons is to simply multiply the 1-day estimate of VAR by the square root of the number of days ahead one wishes to forecast; this is the so-called square-root-of-time rule. See McNeil et al. (2005, p. 54) for more details.

With some improper functional notation, if the random variable Z is normally distributed with mean μ and variance σ^2, then for $q \in (0,1)$ the VaR of Z is given by

$$\mathrm{VaR}_q(Z) = \mu + \sigma\, \Phi^{-1}(q) \quad (3.5)$$

where Φ is the cumulative distribution function of a standard normal variable (McNeil et al., 2005, p. 39).

If instead the normalized random variable $\tilde{Z} = \frac{Z-\mu}{\sigma}$ has a standard t-distribution with $\nu > 2$ degrees of freedom, so that $\mathrm{E}(\tilde{Z}) = 0$ and $\mathrm{Var}(\tilde{Z}) = \frac{\nu}{\nu-2}$, then the VaR of Z is given by

$$\mathrm{VaR}_q(Z) = \mu + \sigma\, t_\nu^{-1}(q) \quad (3.6)$$

where t_ν denotes the distribution function of the standard Student's t-distribution (McNeil et al., 2005, p. 40).

As these two examples allude to, similar results can be obtained from any location-scale family of distributions (McNeil et al., 2005, p. 40).

For losses the type of losses we are considering (see equation (3.3)), where the mean and variance are possibly time-dependent but the innovations are not, we may write $\mathrm{VaR}_q^t(X_{t+1}) = \mu_{t+1} + \sigma_{t+1} \cdot \mathrm{VaR}_q(Z)$ for the value at risk at time t for the loss at time $t+1$.

In section 3.5 we will cover the calculation of VaR of a Generalized Pareto distribution.

3.3 Expected Shortfall

The expected shortfall at level q, in turn, is the expected value at time t of the loss in the next period, X_{t+1}, conditional on the loss exceeding VaR_q^t:

$$\text{ES}_q^t = \text{E}_t[X_{t+1}|X_{t+1} > \text{VaR}_q] \tag{3.7}$$

For this reason, expected shortfall is also referred to as Expected Tail Loss or Conditional VaR. Just as for VaR, expected shortfall can also be estimated over longer horizons than a day, and again the square-root-of-time rule can be used.

For $q \in (0, 1)$, the expected shortfall for a normally distributed random variable Z with mean μ and variance σ^2 is

$$\text{ES}_q(Z) = \mu + \sigma \frac{\phi(\Phi^{-1}(q))}{1-q} \tag{3.8}$$

where ϕ is the density of a standard normal variable (McNeil et al., 2005, p. 45).

If instead the random variable $\tilde{Z} = \frac{Z-\mu}{\sigma}$ has a standard t-distribution with $\nu > 2$ degrees of freedom, then the expected shortfall of Z is given by

$$\text{ES}_q(Z) = \mu + \sigma \frac{g_\nu\left(t_\nu^{-1}(q)\right)}{1-q} \cdot \frac{\nu + \left(t_\nu^{-1}(q)\right)^2}{\nu - 1} \tag{3.9}$$

where g_ν denotes the density of the standard Student's t-distribution (McNeil et al., 2005, p. 45–46).

Analogous to the VaR calculations, the calculation of expected shortfall will be similar for any random variable from a location-scale family. Thus, for the losses that we are considering (see equation (3.3)), where the mean and variance are possibly time-dependent but the innovations Z are not, we may write $\text{ES}_q^t(X_{t+1}) = \mu_{t+1} + \sigma_{t+1} \cdot \text{ES}_q(Z)$. This is the expected shortfall at time t for the loss at time $t+1$.

In section 3.5 we will cover the calculation of ES of a Generalized Pareto distribution.

3.4 Time Series

In this section we establish some properties of certain stochastic processes, that we will use to model the asset returns (or losses).

Consider a discrete time stochastic process $\{X_t : t \in \mathbb{Z}\}$. The process is said to be strictly or strongly stationary if the joint distribution of $(X_{t_1}, \ldots, X_{t_k})$ is the same as for $(X_{t_1+\ell}, \ldots, X_{t_k+\ell})$, for $\ell \in \mathbb{Z}$ fixed and $(t_1, \ldots, t_k)$ free to vary (but the distance between t_i and t_j, $i, j \in \{1, \ldots, k\}$ fixed). That is, the distribution is invariant to such shifts in time (Tsay, 2010, p. 30).

We say that the process is weakly or covariance stationary if instead the mean of X_t, and the covariance between X_t and $X_{t-\ell}$, are invariant with respect to a shift in time. That is, under weak stationarity, $\mathrm{E}(X_t) = \mu$ and $\mathrm{Cov}(X_t, X_{t-\ell}) = \gamma_\ell$ (independent of t), where the latter is called the lag-ℓ autocovariance. Under weak stationarity, the first two moments of X_t must be finite. Letting $\mathrm{Var}(X_t) = \gamma_0$ and noting that $\gamma_{-\ell} = \gamma_\ell$, the autocorrelation function (ACF) for the correlation between X_t and $X_{t-\ell}$ can be defined as

$$\rho_\ell = \frac{\gamma_\ell}{\gamma_0}, \tag{3.10}$$

where $-1 \leq \rho_\ell \leq 1$ and $\rho_0 = 1$ (Tsay, 2010, pp. 30-31). The ACF, together with the partial autocorrelation function (PACF), which measures correlation between the current observation and an observation ℓ periods ago conditional on the values of the intermediate lags (see Enders (2003, pp. 82–83) for a more rigorous definition), can be used to determine the order of ARMA-type models, defined below. These facts serve to give us a bit more understanding of the ACF and PACF plots in section 2.

A time series is said to be linear if it can be expressed as

$$X_t = \mu + \sum_{i=0}^{\infty} \psi_i Z_{t-i}, \tag{3.11}$$

where $\{Z_t : t \in \mathbb{Z}\}$ is a white noise process, i.e. a sequence of iid random variables with zero mean and finite variance σ_Z^2, and where ψ_i are weights with $\psi_0 = 1$. Z_t is also called the innovation at time t.

3.4.1 Autoregressive and Moving Average Models

An autoregressive model of order p, $p \in \mathbb{N}$, or AR(p)-model, is a model in which X_t can be expressed as a linear combination of past observations of the process,

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t, \tag{3.12}$$

where Z_t comes from a white noise process. For a pure AR process of order p, the partial autocorrelation function (PACF) will cut to zero after lag p, which means that a plot of the sample PACF may help in identifying the order of the process (if it is indeed an AR process).

We will use the simplest autoregressive model, the AR(1)-model, to model the mean term μ_t in equation (3.3) for the asset losses, as financial returns are known to have low serial

correlation in levels. With the notation above, an AR(1) process may be written

$$X_t = \phi_0 + \phi_1 X_{t-1} + Z_t, \tag{3.13}$$

which has the properties

$$\mathrm{E}(X_t|X_{t-1}) = \phi_0 + \phi_1 X_{t-1}, \quad \text{and} \tag{3.14}$$

$$\mathrm{Var}(X_t|X_{t-1}) = \mathrm{Var}(Z_t) = \sigma_Z^2 \tag{3.15}$$

Since we are dealing with a weakly stationary process, the unconditional expected value and variance exist as well, but in our present application (forecasting VAR and ES one day ahead), these are of lesser interest.

Though not the focus here, another type of process is where X_t is a linear combination of q lagged values of the innovations Z_t instead; this is called a moving average or MA model of order q, and similar to the role of the PACF for AR processes, plotting the sample ACF may help in determining the order of an MA process.

If X_t is a sum of both p lagged values of itself, and q lagged values of the innovations Z_t, then we have an ARMA(p, q) model. This is reminiscent of the GARCH-type models defined next.

3.4.2 Conditional Heteroscedastic Models

Financial time series have been observed to suffer from volatility clustering—higher volatility in some periods of time, low in others—and to be finite and evolve continuously over time (Tsay, 2010, p. 111). It would therefore be proper to model asset losses (returns) as having time-varying conditional[3] variance instead of a fixed conditional variance, as in the autoregressive model above. Engle (1982) introduced the ARCH (Autoregressive Conditional Heteroscedastic) model, in which the conditional variance changes over time as a function of past errors, while the unconditional variance is fixed. Bollerslev (1986) generalized these models by allowing for the variance to depend not only on past errors, but on past conditional variances as well.

The ARCH(q) model

Slightly more rigorously, if $\{Z_t\}$ is a white noise process (mean zero, variance one) and $\{\sigma_t\}$ is a strictly positive-valued stochastic process, we define the mean-adjusted process $\{X_t - \mu\} = \{\epsilon_t\}$ to be an ARCH(q) process provided it is strictly stationary and provided it satisfies the equations

$$\epsilon_t = \sigma_t Z_t \tag{3.16}$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2, \tag{3.17}$$

for all t, with $\omega > 0$ and $\alpha_i \geq 0$, and where $\sum_{i=1}^{q} \alpha_i < 1$ is a necessary (and sufficient) condition for weak stationarity (McNeil et al., 2005, p. 139).

[3] Conditional on past observations, or more generally on past information.

The GARCH(p,q) model
The Generalized ARCH (GARCH) process by Bollerslev adds a sum of lagged conditional variances to the definition (the assumptions above remaining the same): $\{\epsilon_t\}$ is a GARCH(p,q) process if it satisfies the equations

$$\epsilon_t = \sigma_t Z_t \tag{3.18}$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \tag{3.19}$$

where $\omega > 0$ and $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\sum_{i=1}^{\max(p,q)}(\alpha_i + \beta_i) < 1$. (Tsay, 2010, p. 132) We see that if the ϵ_{t-i}, $i = 1, \ldots, q$, are large in magnitude, ϵ_t will likely also be large in magnitude, which provides a reasonable explanation of volatility clusters.

A parsimonious form of the GARCH model, and the one most frequently used in the financial literature, is the GARCH(1,1) model, in which the conditional variance may be written as

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{3.20}$$

with $\alpha + \beta < 1$ guaranteeing that the process $\{\epsilon_t\}$ is stationary (Brooks, 2008, p. 394). If α is large relative to β, then volatility will react quickly to market movements and appear spiky, while if the reverse is true, then volatility will appear to be persistent, remaining at around the same level for longer (Dowd, 2005, p. 132).

In the GARCH(1,1) model, we can forecast the conditional variance one day ahead by

$$\hat{\sigma}_{t+1}^2 = \hat{\omega} + \hat{\alpha} \epsilon_t^2 + \hat{\beta} \sigma_t^2 \tag{3.21}$$

where $\hat{\omega}, \hat{\alpha}$ and $\hat{\beta}$ are usually estimated by the maximum likelihood method or similar.

The GJR-GARCH(p,q) model
Another phenomenon that has been noted in financial data is that past negative shocks, i.e. losses that exceed the mean loss (or equivalently, returns falling below the average return), will impact current volatility more so than positive shocks (the opposite situation). This phenomenon is sometimes referred to as the "leverage effect", noted by Black (1976), and has been observed particularly over short horizons, e.g. a shock yesterday affecting the volatility today. Glosten et al. (1993) introduced an extension of the GARCH model that can account for this fact, by introducing an indicator variable in the sum of the ARCH-terms (shocks) ϵ_{t-i}. We say that the mean-adjusted losses $\{\epsilon_t\}$ follow a GJR-GARCH(p,q) process if they satisfy the equations

$$\epsilon_t = \sigma_t Z_t \tag{3.22}$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} (\alpha_i + \gamma_i I_{t-i}) \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \tag{3.23}$$

where $\omega, \alpha_i, \gamma_i, \beta_i > 0$ guarantees positivity, $\sum_{i=1}^{q} \alpha_i + c \sum_{i=1}^{q} \gamma_i + \sum_{j=1}^{p} \beta_j < 1$ is necessary for stationarity (and where c is some constant depending on the distribution of ϵ_t^2; $c = 0.5$

if the distribution is symmetric), and I_{t-i} is an indicator variable such that

$$I_{t-i} = \begin{cases} 1 & \text{if } \epsilon_{t-i} > 0 \\ 0 & \text{if } \epsilon_{t-i} \leq 0 \end{cases} \tag{3.24}$$

Again, the most parsimonious case is the GJR-GARCH(1,1) process, in which

$$\epsilon_t = \sigma_t Z_t \tag{3.25}$$

$$\sigma_t^2 = \omega + (\alpha + \gamma I_{t-1})\, \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{3.26}$$

In the GJR-GARCH(1,1) model, we can forecast the conditional variance one day ahead by

$$\hat{\sigma}_{t+1}^2 = \hat{\omega} + (\hat{\alpha} + \hat{\gamma} I_t + \beta)\, \hat{\sigma}_t^2 \tag{3.27}$$

The Component GARCH(p,q) model

Another characteristic of asset returns is that the persistence of volatility is quite strong, though for both theoretical and empirical reasons, it is unlikely that the variance of an asset return series would be non-stationarity (Brooks, 2008, p. 394). However, to investigate the short- and long-run movements in asset volatility, Lee & Engle (1999) introduced the Component GARCH model, in which the conditional variance is decomposed into two parts corresponding to transitory and permanent effects. Letting q_t represent the permanent part of the conditional variance, the variance of this model may be written as

$$\sigma_t^2 = q_t + \sum_{i=1}^{q} \alpha_i \left(\epsilon_{t-i}^2 - q_{t-i}\right) + \sum_{j=1}^{p} \beta_j \left(\sigma_{t-j}^2 - q_{t-j}\right) \tag{3.28}$$

$$q_t = \omega + \rho q_{t-1} + \phi \left(\epsilon_{t-1}^2 - \sigma_{t-1}^2\right) \tag{3.29}$$

where $\rho < 1$, and from which it is seen that the intercept in the GARCH model now follows an AR(1)-type process and is time-varying. The term $\sigma_{t-j}^2 - q_{t-j}$ is here the transitory part of the conditional variance. The one-step ahead forecast may be made by plugging in the parameter estimates and the "observed" values of the day before; see the vignette of Ghalanos (2013) for more details. Again the simplest case is the Component GARCH(1,1) model with

$$\sigma_t^2 = q_t + \alpha \left(\epsilon_{t-1}^2 - q_{t-1}\right) + \beta \left(\sigma_{t-1}^2 - q_{t-1}\right) \tag{3.30}$$

$$q_t = \omega + \rho q_{t-1} + \phi \left(\epsilon_{t-1}^2 - \sigma_{t-1}^2\right) \tag{3.31}$$

Here, $0 < \alpha + \beta < \rho < 1$ and $0 < \phi < \beta$.

3.5 Extreme Value Theory

When it comes to measuring the risk of a financial asset, what matters most are the large, rare losses such as the ones that occurred on Black Monday in 1987 or in the Flash Crash of 2010. That is, the most interesting outcomes are those from the upper tail of the loss distribution. Unfortunately, high quantiles and conditional expectations far out in the tail are hard to adequately estimate using non-parametric approaches such as historical simulation, or traditional parametric approaches using the normal distribution or t-distribution. What is needed is an approach that can extract more information from the large losses we observe, and allow us to better predict large and rare losses, possibly larger than previously observed. Enter Extreme value theory.

Consider a sequence of independent random variables $Z_1, Z_2, \ldots, Z_n$ all distributed according to a distribution function F, and let $M_n = \max\{Z_1, Z_2, \ldots, Z_n\}$. It is a simple exercise to show that $P(M_n \leq z) = F^n(z)$. Since the distribution of M_n will approach a point mass at zero for z smaller than the upper end-point[4] of F as n approaches infinity, we instead want to renormalize M_n so that the limiting distribution is not degenerate. If we can find sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) \tag{3.32}$$

where G is a non-degenerate distribution function, then G belongs to the generalized extreme value (GEV) family of distributions (Coles, 2001, p. 48). These distributions have the form

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{3.33}$$

with support $\{z \in \mathbb{R} : 1 + \xi(z-\mu)/\sigma > 0\}$, and with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma > 0$, and shape parameter $\xi \in \mathbb{R}$. The case $\xi = 0$ is treated as a limit $\xi \to 0$ of G. With n large enough, $P((M_n - b_n)/a_n \leq z) \approx G(z)$, and since (3.33) is a location-scale family, $P(M_n \leq z) \approx G^*(z)$ for some other member G^* of the GEV family (Coles, 2001, p. 48).

Now to the model that we will use for the VAR and ES estimates. Suppose we have a sequence of independent random variables $Z_1, Z_2, \ldots, Z_n$ that satisfy the conditions above, so that for large enough n, $P(M_n \leq z) \approx G(z)$, where G belongs to the GEV family of distributions. Then if we denote any of the random variables Z_i by Z, then for a large enough **threshold** u, the distribution function of $Y = (Z - u)|Z > u$ is approximately

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{3.34}$$

with support $\{y \in \mathbb{R} : y > 0, (1 + \xi y)/\tilde{\sigma} > 0\}$, and with $\tilde{\sigma} = \sigma + \xi(u - \mu)$, and μ, σ, and ξ as in the GEV above (Coles, 2001, pp. 75). The case $\xi = 0$ is treated as a limit $\xi \to 0$ of H. This family of distributions is called the Generalized Pareto family, and the Generalized Pareto distribution is abbreviated GPD. The method of looking only at observations above a certain threshold and fitting a GPD to these exceedances is aptly called the Peak Over Threshold (POT) method.

[4] The point $z_+ = \inf\{z : F(z) = 1\}$.

The shape parameter ξ controls the thickness of the tail of the distribution, with $\xi > 0$ indicating a distribution with a thick tail; $\xi = 0$ a tail of "medium" thickness, and $\xi < 0$ a tail with a finite endpoint. See Chapter 4 of Coles (2001) for a deeper coverage of the properties of these distributions.

3.5.1 Quantiles and Conditional Expectations of the GPD

If exceedances of a threshold u by a variable Z can be appropriately modelled by a GPD with scale and shape parameters σ and ξ, then the q-quantile[5] (the VaR_q) z_q of Z is given by

$$z_q = u + \frac{\sigma}{\xi}\left[\left(\frac{1-q}{\zeta_u}\right)^{-\xi} - 1\right] \tag{3.35}$$

where $\zeta_u = P(Z > u)$, $z_q > u$, and $\xi \neq 0$ (Coles, 2001, p. 82). If $\xi = 0$, then $z_q = u + \sigma \log(\zeta_u/(1-q))$. It can be shown that if excesses over a threshold u_0 can be modelled well by a GPD with scale σ_{u_0} and shape ξ, then excesses of an even higher threshold u will also be modelled will by a GPD, with the same shape parameter ξ but with a scale parameter $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$. In section 4.2, we discuss how the threshold u may be chosen, but we consider it as given here.

Now, for $Z - u|Z > u \sim \text{GPD}(\sigma, \xi)$ with $\xi < 1$ and $\sigma + u\xi > 0$, we have

$$E[Z - u|Z > u] = \frac{\sigma}{1-\xi} \tag{3.36}$$

and since for $z_q > u$, $Z - z_q|Y > z_q \sim \text{GPD}(\sigma + \xi(z_q - u), \xi)$, we have

$$\text{E}[Z - z_q|Z > z_q] = \frac{\sigma + \xi(z_q - u)}{1-\xi} \quad \Leftrightarrow \tag{3.37}$$

$$\text{E}[Z|Z > z_q] = \frac{z_q}{1-\xi} + \frac{\sigma - \xi u}{1-\xi}, \tag{3.38}$$

the last expression being the ES_q for Z. Here, σ and ξ can be estimated by the maximum likelihood method, while $\zeta_u = P(X > u)$ can be estimated by $\hat{\zeta}_u = N_u/n$, where N_u is the number of points in the sample that exceed the threshold u and n is the total number of observations in the sample (Coles, 2001, p. 82).

The next step is then to combine all the results presented in this section, to produce estimates of VaR and ES.

[5] Again defined as the least point on the support of the distribution function below which q of the probability lies.

4 Methodology

4.1 Estimating Value at Risk and Expected Shortfall

Using the results of section 3.4, we can form a model for the losses X_t as

$$X_t = \mu_t + \sigma_t Z_t \tag{4.1}$$
$$\mu_t = \phi_0 + \phi_1 X_{t-1} \tag{4.2}$$

with σ_t given either by the (standard) GARCH(1,1) model in equation (3.20); the GJR-GARCH(1,1) model in equation (3.26); or the Component GARCH(1,1) model in equation (3.30). As is seen, the losses are represented by variations of the AR(1)-GARCH(1,1) model. Again, the Z_t are iid with mean zero and unit variance, and some additional conditions apply. Letting Z denote a random variable with the same distribution as any of the Z_t, we will assume Z either has a standard normal distribution, or a Student's t-distribution scaled to have variance one. Because of the properties of VaR and ES relating to location-scale families of distributions described in sections 3.2 and 3.3, the VaR and ES estimated at time t for the loss at time $t+1$ may be written as

$$\mathrm{VaR}_q^t(X_{t+1}) = \mu_{t+1} + \sigma_{t+1} \cdot \mathrm{VaR}_q^t(Z) \tag{4.3}$$
$$\mathrm{ES}_q^t(X_{t+1}) = \mu_{t+1} + \sigma_{t+1} \cdot \mathrm{ES}_q^t(Z) \tag{4.4}$$

The VaR and ES for a standard normal variable Z is again given by

$$\mathrm{VaR}_q^t(Z) = \Phi^{-1}(q) \tag{4.5}$$
$$\mathrm{ES}_q^t(Z) = \frac{\phi(\Phi^{-1}(q))}{1-q} \tag{4.6}$$

where ϕ is the density and Φ the distribution function of a standard normal variable. If Z instead has a Student's t-distribution scaled to have unit variance, then

$$\mathrm{VaR}_q^t(Z) = \sqrt{\frac{\nu-2}{\nu}} \cdot t_\nu^{-1}(q) \tag{4.7}$$
$$\mathrm{ES}_q^t(Z) = \sqrt{\frac{\nu-2}{\nu}} \cdot \frac{g_\nu\left(t_\nu^{-1}(q)\right)}{1-q} \cdot \frac{\nu + \left(t_\nu^{-1}(q)\right)^2}{\nu-1} \tag{4.8}$$

where g_ν denotes the density and t_ν the distribution function of the standard Student's t-distribution.

Lastly, if the exceedances of Z over a threshold u are distributed according to a Generalized Pareto distribution with scale parameter σ, shape parameter $\xi < 1$ and and $\sigma + u\xi > 0$, then the VaR and ES of Z are given by

$$\mathrm{VaR}_q^t(Z) = u + \frac{\sigma}{\xi}\left[\left(\frac{1-q}{\zeta_u}\right)^{-\xi} - 1\right] \tag{4.9}$$
$$\mathrm{ES}_q^t(Z) = \frac{\mathrm{VaR}_q^t(Z)}{1-\xi} + \frac{\sigma - \xi u}{1-\xi}, \tag{4.10}$$

where $\zeta_u = P(Z > u)$, $\mathrm{VaR}_q^t(Z) > u$, and $\xi \neq 0$. Parameters are estimated as described in section 3.5.1.

4.2 Threshold Choice

There is some difficulty in choosing the correct threshold in the Peak Over Threshold approach. A low threshold means more observations to use in fitting a Generalized Pareto (GP) distribution to the data, but with too low a threshold there is a risk of including observations that are not far enough into the tail for the POT method to be valid, which in turn may lead to bias in the parameter estimates. Conversely, choosing too high a threshold means fewer observations in obtaining parameter estimates, and so may lead to higher variance for these.

In section 3.5.1, we noted that if the excesses of a variable X over a threshold u_0 have a GP distribution, then the excesses of X over a threshold $u > u_0$ are also GP-distributed. This suggests a way of choosing the threshold u: plot the sample mean of the exceedances of a threshold u (with u subtracted from these exceedances) against u, for increasing values of u, and choose as a threshold the value u_0 after which the plot begins to look linear. This plot is called the **mean residual life plot**, and the procedure is described in more detail in Coles (2001, p. 79). Unfortunately, inspecting plots to determine the threshold at each step in our backtest would simply be impossible, and we would like that somehow automates the threshold selection process.

In their paper, McNeil & Frey (2000) choose to fix the number k of exceedances of the threshold, thereby always using the $(k+1)$th order statistic of the sample data as the threshold. To determine the appropriate k, they perform a Monte-Carlo simulation using random variables (t-distributed) that roughly resemble the residuals of the AR(1)-GARCH(1,1) model they fit to their data. They compare estimates of bias and MSE for different values of k, arriving at $k = 100$ as a sensible value for the number of exceedances, with 1000 observations in their backtest window. However, this approach still relies on visually inspecting plots and therefore some degree of subjectivity, and also does not adapt to changing market conditions (e.g. the long, calm period before the events in 2007, and the volatile years after).

Instead, we will use the double bootstrap methodology of Danielsson et al. (2001); what follows is a brief description of it, leaving many details out. We suppose $X_1, X_2, \ldots$ are independent (positive) random variables from a heavy-tailed distribution with distribution function F, that has associated tail index $1/\xi > 0$, where ξ is a shape parameter as in section 3.5. The aim is to minimize the asymptotic (in sample size n) MSE of the estimator $\hat{\xi} = \xi_n$ by choosing the tail fraction k, where ξ is estimated using the Hill estimator

$$\xi_n(k) = \frac{1}{k} \sum_{i=1}^{k} \left(\log X_{n,(n-i+1)} - \log X_{n,(n-k)} \right) \tag{4.11}$$

which is asymptotically normally distributed. $X_{n,(1)} \leq X_{n,(k)} \leq X_{n,(n)}$ are here the order statistics of $X_1, X_2, \ldots, X_n$. The asymptotic MSE (AMSE) of ξ_n is defined as

$$\mathrm{AMSE}(n,k) = \mathrm{Asym}\, \mathrm{E}\left(\xi_n(k) - \xi\right) \tag{4.12}$$

The AMSE is estimated by a bootstrap procedure to find the k for which it is minimized. Resamples $\mathcal{X}^*_{n_1} = \{X^*_1, \ldots, X^*_{n_1}\}$ are drawn with replacement from $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$

and for values $k_1 < n_1$ we compute

$$\xi^*_{n_1}(k_1) = \frac{1}{k_1}\sum_{i=1}^{k_1}\left(\log X^*_{n_1,(n_1-i+1)} - \log X^*_{n_1,(n_1-k)}\right) \tag{4.13}$$

The bootstrap estimate of the AMSE is then given by

$$\widehat{\mathrm{AMSE}}(n_1,k_1) = \mathrm{E}\left[\left(\xi^*_{n_1}(k_1) - \xi_n(k)\right)^2 \middle| \mathcal{X}_n\right] \tag{4.14}$$

where k has to be chosen so that $\xi_n(k)$ is consistent. Since k is unknown, $\xi_n(k)$ is replaced by

$$M_{n_1}(k_1) = \frac{1}{k_1}\sum_{i=1}^{k_1}\left(\log X_{n_1,(n_1-i+1)} - \log X_{n_1,(n_1-k)}\right) \tag{4.15}$$

which in turn is estimated in the bootstrap by

$$M^*_{n_1}(k_1) = \frac{1}{k_1}\sum_{i=1}^{k_1}\left(\log X^*_{n_1,(n_1-i+1)} - \log X^*_{n_1,(n_1-k)}\right) \tag{4.16}$$

The bootstrap estimate of the AMSE for n_1 and k_1 is then given by

$$Q(n_1,k_1) = \mathrm{E}\left[\left\{M^*_{n_1} - 2\left(\xi^*_{n_1}\right)^2\right\}^2 \middle| \mathcal{X}_n\right] \tag{4.17}$$

For a fixed n_1, k_1 is then varied to find the optimal value $\hat{k}^*_{n_1}$ which minimizes $Q(n_1,k_1)$. The same procedure is then done for a smaller value $n_2 = \lfloor n_1^2/n \rfloor$ (where $\lfloor\cdot\rfloor$ denotes the floor function) to get the optimal value $\hat{k}^*_{n_2}$ which minimizes $Q(n_2,k_2)$. Then for these fixed values of n_1 and n_2 with optimal values $\hat{k}^*_{n_1}$ and $\hat{k}^*_{n_2}$, the optimal fraction of observations in the tail (i.e. above the threshold) is given by

$$\hat{k}^{\mathrm{opt}}_n = \left\lfloor \frac{(\hat{k}^*_{n_1})^2}{\hat{k}^*_{n_2}} \cdot \left[\frac{\left(\log \hat{k}^*_{n_1}\right)^2}{2\log n_1 - \log \hat{k}^*_{n_1}}\right]^{\left(\log n_1 - \log \hat{k}^*_{n_1}\right)/\log n_1} \right\rfloor \tag{4.18}$$

Finally, to choose n_1, one simply performs the above procedure for different values of $n_1 < n$, and chooses the n_1 that minimizes

$$R(n_1) = \frac{\left(Q(n_1,\hat{k}^*_{n_1})\right)^2}{Q(n_2,\hat{k}^*_{n_2})} \tag{4.19}$$

As should be obvious, this is a very computationally intensive procedure and should preferably be done on a grid of computers. The number of bootstrap samples to use in each step should therefore be chosen wisely, and in the case of this paper, the step sizes between different values of n_1 and n_2 were chosen to be larger than one (we chose them based on the values used in the simulation study in Danielsson et al. (2001)). It would also have been impossible time-wise to re-estimate the optimal k for each dataset and model at each step during the backtesting procedure, so we chose to only re-estimate k at every 500th step instead.

4.3 Backtesting VaR and Expected Shortfall

We will use the loss data from the assets listed in section 4.4 to estimate value at risk and expected shortfall with the models under consideration. Backtesting involves fixing a window size n and stepping through the data day by day, using the past n observations to estimate the VaR and ES for the next day. Introducing some notation, we have window of raw data $(x_{t_1}, x_{t_2}, \ldots, x_{t_n})$ ordered by time, which we use to estimate our models. Estimates of VaR and ES can be made directly after fitting the AR(1)-GARCH(1,1)-type models on this data, while for the POT approach, we first create the residuals

$$(z_{t_1}, z_{t_2}, \ldots, z_{t_n}) = \left(\frac{x_{t_1} - \hat{\mu}_{t_1}}{\hat{\sigma}_{t_1}}, \frac{x_{t_2} - \hat{\mu}_{t_2}}{\hat{\sigma}_{t_2}}, \ldots, \frac{x_{t_n} - \hat{\mu}_{t_n}}{\hat{\sigma}_{t_n}} \right) \tag{4.20}$$

using the parameter estimates from each model. These should resemble Z_t in equation (3.3), i.e. have mean zero and unit variance, if the fitted model is a plausible model for the true loss process. We then fit a GPD to these residuals using a threshold $u = z_{(k+1)}$, the $(k+1)$th order statistic of these residuals, where k is chosen by the procedure described in section 4.2.

To measure the performance of a specific model, the estimates should be compared to the actual outcomes to see how well the model fared over the days for which predictions of the ES were made. That is, we want to test whether the forecasts of the model are consistent with the assumptions underlying the model choice i.e. the distribution of the losses and/or residuals in our case. This is the general description of a backtest—finding a good way to make the comparison is the tricky matter. Unfortunately, the backtesting theory and methodology is not as developed for expected shortfall as it is for value at risk. For VaR, we perform a three-staged unconditional coverage and independence test, while for ES, we implement the bootstrap backtest used by McNeil & Frey (2000), and what we may call a "V-test" as used by Embrechts et al. (2005).

4.3.1 Unconditional Coverage and Independence Test for VaR

Here we follow the presentation given in Christoffersen (2003), section 8.3.

At a given confidence level $q \in (0,1)$, we expect the actual loss X_{t+1} to exceed the estimated $\text{VaR}_q^t(X_{t+1})$ only $100(1-q)\%$ of the time. We call such exceedances VaR-breaks. Particularly with a VaR-model that adapts to recent losses and recent volatility (i.e. the AR(1)-GARCH(1,1)-type models we are considering), we also expect the VaR-breaks to be independent of each other. It follows that a way to test the performance of a VaR-model is to test if the model produces the expected number of VaR-breaks when testing the model on a set of data, and to test if the VaR-breaks are independent of each other. We therefore form a "hit sequence" of indicator variables representing the VaR-breaks as

$$I_{t+1} = \begin{cases} 1 & \text{if } X_{t+1} > \text{VaR}_q^t(X_{t+1}) \\ 0 & \text{if } X_{t+1} \leq \text{VaR}_q^t(X_{t+1}) \end{cases}, \tag{4.21}$$

and we may talk of a VaR-break as a "hit" and otherwise a "miss". With a data set of T VaR-predictions, this gives us a sequence $\{I_t\}_{t=1}^T$ (the "hit sequence"). The simplest possible null hypothesis is then that the I_t are Bernoulli variables with success probability $\alpha = 1 - q$, so that $\{I_t\}_{t=1}^T$ is a sequence of iid Bernoulli random variables. The probability

mass function of a Bernoulli(p) variable is given by

$$f(I_t; p) = p^{I_t}(1-p)^{1-I_t} \tag{4.22}$$

Unconditional Coverage Test

We first want to test if our VaR model produces as many VaR-breaks as expected at the chosen coverage rate α; we do this by a **unconditional coverage test** as described in Christoffersen (2003, pp. 185–186). We do this by comparing the theoretical sample fraction π of VaR-breaks to the promised fraction α; the null hypothesis for the unconditional coverage test is that $\pi = \alpha$. This comparison is done through a likelihood ratio test. Denoting by T_1 and T_0 the number of hits and misses in a sample of size T, the likelihood function under the null hypothesis is given by

$$L(\alpha) = \prod_{t=1}^{T} p^{I_t}(1-p)^{1-I_t} = p^{T_1}(1-p)^{T_0} \tag{4.23}$$

Next, π is estimated by $\hat{\pi} = T_1/T$, which is the maximum likelihood estimate of π. The maximized likelihood for the sample is then given by

$$L(\hat{\pi}) = \left(\frac{T_1}{T}\right)^{T_1} \cdot \left(\frac{T_0}{T}\right)^{T_0} \tag{4.24}$$

The likelihood ratio statistic is then given by

$$LR_{\text{uc}} = -2\left[L(\alpha)/L(\hat{\pi})\right] \tag{4.25}$$

and is asymptotically (in T) distributed as a χ^2 random variable with one degree of freedom, so that quantiles of the $\chi^2(1)$ distribution can be used for the test. As Christoffersen (2003, p. 186) notes, the number of observations T, and even more so the number of violations T_1 (particularly for small α), may in practice be too small for this test to be reliable. Christoffersen instead recommends doing a Monte Carlo simulation to obtain reliable p-values for this test. This is done by generating 999 samples of iid Bernoulli(α) variables and calculating the above test statistic for these samples, giving us a sequence $\{\widetilde{LR}_{\text{uc}}(i)\}_{i=1}^{999}$. A simulated p-value can then be calculated as

$$\text{p-value} = \frac{1}{1000}\left[1 + \sum_{i=1}^{999} \mathbf{1}\left\{\widetilde{LR}_{\text{uc}}(i) > LR_{\text{uc}}\right\}\right], \tag{4.26}$$

where $\mathbf{1}\{\cdot\}$ is an indicator variable which is equal to one if the condition within the curly brackets is true, and zero otherwise. Too low a p-value, and we reject the null hypothesis that the VaR model gives the correct coverage rate. Christoffersen recommends using a p-value of 0.1 on the count of type II errors being costly in practice.

Independence Test

Next, we want to test if the VaR-breaks are independent of each other—do they come in clusters or not? If they do, it would mean that the VaR-model does not adapt sufficiently and quickly enough to large losses, possibly creating a risk of bankruptcy in a very short period of time as losses pile on. Knowledge that the VaR-breaks are not independent would in practice mean that the probability of a break tomorrow given that there was a break today is larger than α. Christoffersen (2003, pp. 187–188) provides a way to test

for independence. We assume that the hit sequence $\{I_t\}_{t=1}^T$ is dependent, and that it can be described by a discrete-time Markov chain with transition probability matrix

$$\mathbf{\Pi}_1 = \begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix} = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix} \tag{4.27}$$

where π_{ij} $(i, j \in \{0, 1\})$ is the probability that $I_{t+1} = j$ conditional on $I_t = i$. For example, π_{11} is the probability that a VaR-break occurs tomorrow given that one occurred today. Under this model, only today's outcome matters for the outcome tomorrow; earlier outcomes than that say nothing about the outcome tomorrow. With T observations, the likelihood function of this process is given by

$$L(\mathbf{\Pi}_1) = (1-\pi_{01})^{T_{00}} \cdot \pi_{01}^{T_{01}} \cdot (1-\pi_{11})^{T_{10}} \cdot \pi_{11}^{T_{11}} \tag{4.28}$$

where $T_{i,j}$ is the number of days for which a j followed an i in the hit sequence, with $i, j \in \{0, 1\}$. The maximum likelihood estimates for these probabilities are then given by

$$\hat{\pi}_{01} = \frac{T_{01}}{T_{00}+T_{01}} \qquad \Rightarrow \qquad \hat{\pi}_{00} = 1-\hat{\pi}_{01} \tag{4.29}$$

$$\hat{\pi}_{11} = \frac{T_{11}}{T_{10}+T_{11}} \qquad \Rightarrow \qquad \hat{\pi}_{10} = 1-\hat{\pi}_{11}, \tag{4.30}$$

giving us a matrix of estimated transition probabilities

$$\hat{\mathbf{\Pi}}_1 = \begin{bmatrix} \hat{\pi_{00}} & \hat{\pi}_{01} \\ \hat{\pi_{10}} & \hat{\pi_{11}} \end{bmatrix} = \begin{bmatrix} \frac{T_{00}}{T_{00}+T_{01}} & \frac{T_{01}}{T_{00}+T_{01}} \\ \frac{T_{10}}{T_{10}+T_{11}} & \frac{T_{11}}{T_{10}+T_{11}} \end{bmatrix} \tag{4.31}$$

Now, if the hit sequence is dependent, then we would have $\pi_{01} \neq \pi_{11}$. If the sequence were independent, then we would instead have $\pi_{01} = \pi_{11} = \pi$. As we are most concerned about positive dependence in the sense that $\pi_{11} > \pi_{01}$, we estimate π by $\hat{\pi} = T_1/T$ as before. Thus, under independence we get a transition probability matrix

$$\hat{\mathbf{\Pi}} = \begin{bmatrix} 1-\hat{\pi} & \hat{\pi} \\ 1-\hat{\pi} & \hat{\pi} \end{bmatrix} \tag{4.32}$$

which has the same likelihood function $L(\hat{\pi})$ as in the unconditional coverage test. We may then test the independence hypothesis $\pi_{01} = \pi_{11}$ again by a likelihood ratio test with statistic

$$LR_{\text{ind}} = -2\left[L(\hat{\pi})/L(\hat{\mathbf{\Pi}}_1)\right] \tag{4.33}$$

which is asymptotically distributed as a χ^2 random variable with one degree of freedom. Again, Christoffersen (2003, p. 188) recommends using a Monte Carlo simulation to obtain an accurate p-value instead of using quantiles from the $\chi^2(1)$ distribution when testing the independence hypothesis. This is done in the same fashion as for the unconditional coverage test.

Conditional Coverage Test

Finally, the two tests above can be combined to test jointly for correct coverage and independence. Since LR_{uc} and LR_{ind} are each $\chi^2(1)$-distributed (asymptotically), their

sum should be $\chi^2(2)$-distributed, and we can create the test statistic

$$LR_{\text{cc}} = LR_{\text{uc}} + LR_{\text{ind}} \tag{4.34}$$

$$= -2\left[L(\alpha)/L(\hat{\mathbf{\Pi}}_1)\right] \tag{4.35}$$

Just as for the two previous tests, a Monte Carlo simulation is performed to get more accurate p-values.

4.3.2 Bootstrap Test for the Expected Shortfall

To backtest the ES estimates, we want to look at the difference between the next-day return X_{t+1} and the estimate of the expected shortfall at time t, $\text{ES}_q^t(X_{t+1})$, conditional on X_{t+1} exceeding the estimate of the q-quantile of X_{t+1}, i.e. the $\text{VaR}_q^t(X_{t+1})$. We introduce the notation $x_q^t := \text{VaR}_q^t(X_{t+1})$, so that $\hat{x}_q^t = \widehat{\text{VaR}}_q^t(X_{t+1})$.

As per our base model for losses (3.3) and its assumptions, i.e. that losses X_t under all five of our models can be written in the form $X_t = \mu_t + \sigma_t Z_t$, where the Z_t iid with mean zero and unit variance and come from a location-scale family, we can, with the notation and remarks of section 3.3 define errors

$$R_{t+1} = \frac{X_{t+1} - \text{ES}_q^t(X_{t+1})}{\sigma_{t+1}} \tag{4.36}$$

$$= \frac{\mu_{t+1} + \sigma_{t+1}Z_{t+1} - (\mu_{t+1} + \sigma_{t+1}\text{ES}_q^t(Z))}{\sigma_{t+1}} \tag{4.37}$$

$$= Z_{t+1} - \text{E}_t[Z|Z > z_q] \tag{4.38}$$

conditional on $X_{t+1} > x_q^t$ or equivalently $Z_{t+1} > z_q$, z_q being the q-quantile of Z. The R_t's are then iid under our model of losses and furthermore have an expected value of zero. Based on the data and our estimates of expected shortfall, we can construct the corresponding residuals on day when $x_{t+1} > \hat{x}_q^t$, i,e, on days when VaR-breaks occur. Following McNeil & Frey (2000, p. 294) we call these "exceedance residuals", denoting them by

$$\mathbf{r} = \{r_{t+1};\ \text{for } t \text{ such that } x_{t+1} > \hat{x}_q^t\}, \quad \text{where} \tag{4.39}$$

$$r_{t+1} = \frac{x_{t+1} - \widehat{\text{ES}}_q^t(X_{t+1})}{\hat{\sigma}_{t+1}} \tag{4.40}$$

and $|\mathbf{r}| = m$, where m is whatever number of VaR-breaks we get from the particular model we use.

Under the null hypothesis that we estimate μ_{t+1}, σ_{t+1}, and the expected shortfall correctly, these residuals should behave like an iid sample from a random variable with mean zero. The alternative hypothesis is that the residuals have a mean greater than zero, i.e. that the expected shortfall is systematically underestimated, which as McNeil & Frey (2000, p. 294) remarks is the more likely direction of failure. Further, it is the more dangerous way in which ES can be wrong, since it may lead to losses (as opposed to missing out on profits). The downside, however, is that the test will tend to favor models that overestimate the expected shortfall, which is undesirable in the long-run.

We may test this by a nonparametric bootstrap as outlined in chapter 16 section 4 of

Efron & Tibshirani (1993): We want to test if our original residuals $\mathbf{r}$, which are distributed according to some distribution function F, have mean $\mu_0 = 0$, as under our null hypothesis. To do this, we create a statistic

$$T = t(\mathbf{r}) = \frac{\bar{r} - \mu_0}{\bar{\sigma}/\sqrt{m}}, \quad \text{where} \tag{4.41}$$

$$\bar{r} = \frac{1}{m}\sum_{i=1}^{m} r_i, \quad \text{and} \tag{4.42}$$

$$\bar{\sigma} = \frac{1}{m-1}\sum_{i=1}^{m}(r_i - \bar{r})^2 \tag{4.43}$$

To sample data from under the null hypothesis, we translate the empirical distribution function so that it has the desired mean μ_0, by forming the shifted residuals

$$\tilde{r}_i = r_i - \bar{r} + \mu_0, \quad i = 1, 2, \ldots, m \tag{4.44}$$

From these, we sample $\tilde{r}_1^*, \tilde{r}_2^*, \ldots, \tilde{r}_m^*$ with replacement, and for each such bootstrap sample j (N_r of them in total) we compute the statistics

$$T_j^* = t(\tilde{\mathbf{r}}_j^*) = \frac{\bar{\tilde{r}}^* - \mu_0}{\bar{\tilde{\sigma}}/\sqrt{m}} \tag{4.45}$$

with notation as above.

We can then compute a p-value for our null hypothesis through

$$\text{p-value} = \frac{1 + \sum_{j=1}^{N_r} \mathbf{1}_{\{T_j^* > T\}}}{1 + N_r} \tag{4.46}$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function, which is equal to 1 if the condition within the curly brackets is true and 0 if it is false. We add 1 to both numerator and denominator in order to avoid p-values of 0. Models can then be compared based on their p-values: high p-values speak in favor of a model, while low p-values do the opposite. In our backtests, we will take $N_r = 10\,000$.

4.3.3 V-test for the Expected Shortfall

Embrechts et al. (2005, pp. 10–11) introduce a couple of methods for evaluating the performance of different ES estimates, based on the relative size of the test statistics. The first statistic V_1 simply takes the average of the difference between the actual return and the forecasted expected shortfall, for days where the actual return exceeded the VaR estimate. This should lead to a value close to zero of V_1 if the model is good, since if the model is correct the expected value of this statistic is zero. For a chosen probability q, V_1 is thus given by

$$V_1 = \frac{\sum_{t=1}^{T}\left(x_{t+1} - \widehat{\mathrm{ES}}_q^t(X_{t+1})\right)\mathbf{1}_{\{x_{t+1} > \hat{x}_q^t\}}}{\mathbf{1}_{\{x_{t+1} > \hat{x}_q^t\}}} \tag{4.47}$$

where T is the total number of estimates of the ES for a particular data set. As Embrechts

et al. (2005) note, the weakness of this measure is that it depends strongly on the VaR estimates. This could mean that we in fact take the mean over a subsample of size much different from the size $T(1-q)$ we would like, depending on how bad the VaR estimator is.

With the ES_q, we are looking for the average size of a one in a $1/(1-q)$-event. A measure which looks at these types of events is the measure V_2 defined by

$$V_2 = \frac{\sum_{t=1}^{T} \left(x_{t+1} - \widehat{\text{ES}}_q^t(X_{t+1}) \right) \mathbf{1}_{\{D_t > D_q\}}}{\mathbf{1}_{\{D_t > D_q\}}} \tag{4.48}$$

where $D_t = \left(x_{t+1} - \widehat{\text{ES}}_q^t(X_{t+1}) \right)$ and D_q is the empirical q-quantile of $\{D_t, t = 1, 2, \ldots, T\}$. We expect D_t to be negative in less than one in $1/(1-q)$ cases. A good estimator for ES would thus hopefully give us an estimate close to zero.

V_1 and V_2 can be combined into a third measure that strikes a balance between the theory-reliant V_1 measure and the more practically oriented V_2 measure. This third measure is defined as

$$V = \frac{|V_1| + |V_2|}{2} \tag{4.49}$$

and should again be close to zero if it is good.

4.4 Data

We will perform our backtests on three different data sets, representing three asset classes (stock indices, stocks, and commodity indices). The data sets are

Asset	Acronym	Source	Length	Start date	End date
S&P 500 Composite	SP500	Datastream	8088	1980-01-01	2010-12-31
Electrolux B	ELUXB	Datastream	7564	1982-01-04	2010-12-31
Continuous Commodity Index	CCI	Wikiposit	7785	1980-01-02	2010-12-31

Table 1: List of data sets tested.

Below, we present some summary statistics of these data sets.

Asset	Mean	St. dev.	Minimum	Maximum	Skewness	Ex. kurtosis
SP500	−0.03%	1.13%	−10.96%	22.83%	1.2	28.7
ELUXB	−0.05%	2.11%	−19.18%	20.80%	−0.2	6.6
CCI	−0.01%	0.72%	−5.02%	5.90%	0.3	4.7

Table 2: Summary statistics for the losses (negated returns) of the data sets tested.

4.5 Software

The code for this paper was written in R, and can be found here[6]. The packages `timeSeries` (Wuertz & Chalabi, 2010), `fGarch` (Wuertz et al., 2009) and `rugarch` were used for time series modeling; the package `ismev` (Stephenson & Heffernan, 2012) for the Extreme value theory parts; `ggplot2` (Wickham, 2009) for plotting; and packages `boot` (Canty & Ripley, 2013), and `doParallel` (Revolution Analytics, 2012) for other parts.

4.6 Notes on Implementation

Before running the backtests, we will normalize the data to have unit variance in order to avoid problems with numerical instability. This should not affect our results and the conclusions we can draw from them. In the table below, we restate some of the parameter choices we have made for our models and backtesting procedures.

Window length n	1 000
Degrees of freedom ν for t-distribution	4
Bootstrap samples N_r	10 000

Table 3: Choice of parameters.

[6] https://github.com/BenjaK/Thesis2013

5 Results

With three choices of model for the conditional volatility, two choices of distribution for the innovations, and a POT model fitted to the residuals of each combination of the former two choices, we will have twelve models to backtest on our three data sets. To shorten the notation, we introduce the following acronyms for our models:

- $\mathrm{S_n}$: AR(1)-GARCH(1,1) with normally distributed innovations
- $\mathrm{S_t}$: AR(1)-GARCH(1,1) with t-distributed innovations
- $\mathrm{G_n}$: AR(1)-GJR-GARCH(1,1) with normally distributed innovations
- $\mathrm{G_t}$: AR(1)-GJR-GARCH(1,1) with t-distributed innovations
- $\mathrm{C_n}$: AR(1)-Component-GARCH(1,1) with normally distributed innovations
- $\mathrm{C_t}$: AR(1)-Component-GARCH(1,1) with t-distributed innovations

A superscript P will indicate the POT model corresponding to one of the above models. For example, $\mathrm{G_n^P}$ will indicate the POT model fitted to the residuals of model $\mathrm{G_n}$. We will refer to the models without a superscript as the "original" models, and those with a superscript as the POT models. As it is interesting to see how the different models perform at higher and higher probabilities q, we will estimate and test VaR and ES for $q \in \{0.95, 0.975, 0.99, 0.995\}$.

A small result worth mentioning before we move on to the results from the backtests is that the double bootstrap method for threshold choice, described in section 4.2, yielded optimal values for the number of observations k in the tail that were remarkably close to the value $k = 100$ used by McNeil & Frey (2000). We found k to be entirely in the range 96–109 for all data sets.

5.1 Tests for Value at Risk

5.1.1 Tests for Value at Risk Estimates

We begin by examining the number of VaR-breaks produced by each model, for each data set and probability q. The number of breaks can be compared to the expected number of breaks.

Data set	Expected	S_n	S_n^P	S_t	S_t^P	G_n	G_n^P	G_t	G_t^P	C_n	C_n^P	C_t	C_t^P
$q = 0.95$													
SP500	354	372	366	387	376	372	388	393	389	379	369	420	383
ELUXB	328	294	372	350	367	296	380	344	373	314	370	353	367
CCI	339	367	349	345	349	373	352	349	345	366	351	371	375
$q = 0.975$													
SP500	177	241	207	195	195	241	212	186	203	249	211	212	208
ELUXB	164	172	185	154	186	173	192	154	187	189	190	157	189
CCI	170	223	190	136	184	232	186	139	187	227	188	158	202
$q = 0.99$													
SP500	71	142	79	64	74	139	82	67	84	150	84	78	88
ELUXB	66	92	70	54	70	89	67	53	73	102	71	54	66
CCI	68	102	73	29	73	107	73	34	76	104	69	36	84
$q = 0.995$													
SP500	35	96	35	25	30	91	41	29	39	101	43	27	40
ELUXB	33	65	42	31	40	65	42	29	40	71	40	30	39
CCI	34	62	41	7	40	65	43	7	43	62	41	10	47

Table 4: Expected and actual number of VaR-breaks obtained by each model.

A few general observations can be made about the results in table 4. At the two lowest confidence levels for the VaR, $q = 0.95$ and $q = 0.975$, almost all models seem to underestimate VaR, as seen by the higher number of VaR-breaks of these models compared to the expected number of breaks. It is hard to say whether a particular model stands out, or if there are any clear differences between the different GARCH-type models, or the conditional distributions, or the original models vs. the POT models.

At the higher confidence levels $q = 0.99$ and $q = 0.995$, it is immediately clear that the models for which the innovations are normally distributed tend to underestimate VaR, as seen by the high number of VaR-breaks compared to the expected number of breaks. For these models, the POT model fitted to the residuals seem to do a good job of adjusting the VaR upward, as seen by the much lower number of VaR-breaks compared to the original model (and much closer to the expected number of breaks). Conversely, the models with t-distributed innovations tend to give a low number of breaks (often too low), and the corresponding POT models adjust this number upward (often too much). That is, the models with t-distributed innovations give high estimates of VaR, and after fitting a GPD to the residuals of these models, the VaR estimate seems to be adjusted downward. It is not obvious if the original models with t-distributed innovations perform better than the corresponding POT models or vice versa.

5.1.2 Unconditional Coverage Test

Building on the previous observations, we conduct the unconditional coverage test described in section 4.3.1. From table 5 below, it is immediately clear from the low p-values that the original models with normally distributed innovations perform badly at all confidence levels. Particularly at the high confidence levels, the corresponding POT models seem to remedy the situation, as seen by the higher p-values for these models. The original models with t-distributed innovations do comparatively better, but the corresponding POT models seem to help mostly in the cases where the original model failed. In many of the cases where the original model did succeed, however, the corresponding POT model actually had a lower p-value, indicating that it performed worse comparatively. There does not seem to be any noticeable differences between the conditional volatility models alone.

Data set	S_n	S_n^P	S_t	S_t^P	G_n	G_n^P	G_t	G_t^P	C_n	C_n^P	C_t	C_t^P
$q = 0.95$												
SP500	0.349	0.515	0.072	0.230	0.323	0.077	0.037	0.062	0.175	0.388	0.001	0.127
ELUXB	0.054	0.017	0.210	0.035	0.061	0.006	0.330	0.011	0.403	0.022	0.187	0.029
CCI	0.128	0.573	0.727	0.581	0.067	0.464	0.585	0.739	0.121	0.496	0.076	0.046
$q = 0.975$												
SP500	0.001	0.017	0.181	0.164	0.001	0.010	0.491	0.057	0.001	0.011	0.014	0.025
ELUXB	0.523	0.100	0.421	0.091	0.492	0.028	0.423	0.084	0.049	0.057	0.566	0.060
CCI	0.001	0.123	0.006	0.260	0.001	0.227	0.012	0.174	0.001	0.152	0.338	0.014
$q = 0.99$												
SP500	0.001	0.328	0.409	0.691	0.001	0.192	0.615	0.113	0.001	0.123	0.365	0.063
ELUXB	0.004	0.587	0.114	0.580	0.011	0.845	0.100	0.348	0.001	0.501	0.123	0.959
CCI	0.001	0.513	0.001	0.532	0.001	0.514	0.001	0.357	0.002	0.846	0.001	0.038
$q = 0.995$												
SP500	0.001	0.936	0.052	0.314	0.001	0.348	0.232	0.481	0.001	0.214	0.155	0.470
ELUXB	0.001	0.115	0.742	0.246	0.001	0.107	0.482	0.235	0.001	0.213	0.584	0.312
CCI	0.001	0.217	0.001	0.310	0.001	0.132	0.001	0.137	0.001	0.240	0.001	0.037

Table 5: Unconditional coverage test for the VaR-breaks.

5.1.3 Independence Test

Next we test for dependence in the VaR-breaks produced by each model, as described in section 4.3.1. Unfortunately, none of the models stand out as particularly good or bad in the results in table 6, and it is not obvious how the different models could be ranked according to performance.

Data set	S_n	S_n^P	S_t	S_t^P	G_n	G_n^P	G_t	G_t^P	C_n	C_n^P	C_t	C_t^P
$q = 0.95$												
SP500	0.581	0.517	0.531	0.238	0.738	0.717	0.641	0.560	0.713	0.418	0.089	0.277
ELUXB	0.124	0.182	0.061	0.073	0.130	0.173	0.007	0.039	0.346	0.650	0.260	0.449
CCI	0.373	0.355	0.555	0.503	0.733	0.935	0.803	0.892	0.682	0.239	0.590	0.286
$q = 0.975$												
SP500	0.046	0.002	0.053	0.002	0.224	0.104	0.017	0.080	0.042	0.043	0.111	0.001
ELUXB	0.052	0.180	0.042	0.094	0.318	0.103	0.063	0.051	0.176	0.166	0.074	0.044
CCI	0.052	0.113	0.653	0.137	0.474	0.600	0.985	0.353	0.068	0.139	0.377	0.191
$q = 0.99$												
SP500	0.044	0.573	0.706	0.375	0.214	0.974	0.113	0.643	0.074	0.627	0.019	0.019
ELUXB	0.146	0.233	0.685	0.265	0.710	0.114	0.674	0.381	0.537	0.912	0.634	0.839
CCI	0.813	0.896	0.791	0.918	0.885	0.890	0.736	0.941	0.875	0.827	0.686	0.647
$q = 0.995$												
SP500	0.529	0.085	0.024	0.034	1.000	0.132	0.024	0.087	0.923	0.143	0.026	0.121
ELUXB	0.080	0.122	0.624	0.124	0.087	0.127	0.750	0.117	1.000	0.138	0.676	0.118
CCI	0.686	0.131	1.000	0.119	0.908	0.169	1.000	0.140	0.737	0.114	1.000	0.174

Table 6: Independence test for the VaR-breaks.

5.1.4 Conditional Coverage Test

The conclusions from the conditional coverage test are about the same as for the unconditional coverage test. The original models with normally distributed innovations perform badly at the higher confidence levels, while the corresponding POT models remedy the situation as indicated by the comparatively higher p-values. The performance of the models with t-distributed innovations is somewhat mixed, with both good and bad performances at all confidence levels. This seems more linked to the data set than the particular model.

Data set	S_n	S_n^P	S_t	S_t^P	G_n	G_n^P	G_t	G_t^P	C_n	C_n^P	C_t	C_t^P
$q = 0.95$												
SP500	0.566	0.641	0.216	0.239	0.599	0.244	0.159	0.186	0.421	0.510	0.001	0.207
ELUXB	0.066	0.027	0.047	0.018	0.079	0.014	0.016	0.006	0.447	0.110	0.248	0.110
CCI	0.234	0.544	0.760	0.677	0.214	0.785	0.840	0.937	0.317	0.372	0.217	0.120
$q = 0.975$												
SP500	0.001	0.002	0.070	0.001	0.001	0.015	0.034	0.054	0.001	0.008	0.024	0.002
ELUXB	0.084	0.111	0.062	0.076	0.463	0.053	0.115	0.037	0.077	0.081	0.132	0.013
CCI	0.001	0.097	0.055	0.187	0.001	0.411	0.071	0.283	0.001	0.120	0.408	0.049
$q = 0.99$												
SP500	0.001	0.403	0.757	0.567	0.001	0.509	0.331	0.161	0.001	0.183	0.111	0.017
ELUXB	0.006	0.482	0.214	0.499	0.028	0.447	0.187	0.338	0.001	0.841	0.208	0.985
CCI	0.002	0.843	0.001	0.873	0.002	0.872	0.001	0.781	0.002	0.960	0.001	0.099
$q = 0.995$												
SP500	0.001	0.343	0.036	0.130	0.001	0.216	0.081	0.312	0.001	0.185	0.063	0.303
ELUXB	0.001	0.102	0.775	0.156	0.001	0.093	0.557	0.159	0.001	0.158	0.650	0.209
CCI	0.001	0.164	0.001	0.200	0.001	0.094	0.001	0.110	0.001	0.196	0.001	0.052

Table 7: Conditional coverage test for the VaR-breaks.

We end this section for the Value at Risk test results with a plot of the $\text{VaR}_{0.99}$ from an AR(1)-GARCH(1,1) model with normally distributed innovations and from the corresponding POT model.

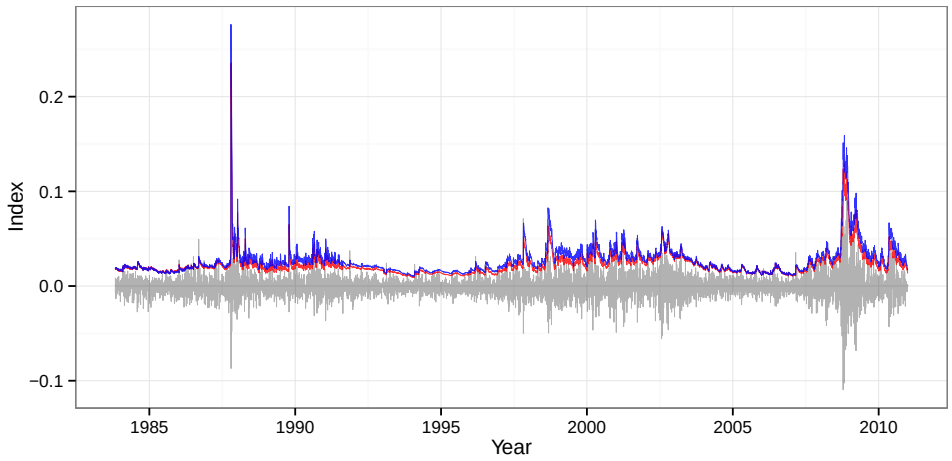


Figure 8: $\text{VaR}_{0.99}$ estimates for the S&P 500 index, from an AR(1)-GARCH(1,1) model with normally distributed innovations (red) and from its corresponding POT model (blue).

5.2 Tests for Expected Shortfall

5.2.1 Bootstrap Test

Next, we perform the bootstrap test for expected shortfall, as described in section 4.3.2. As a reminder, high p-values given by this test speak in favor of a model, while low p-values speak against a model.

Data set	S_n	S_n^P	S_t	S_t^P	G_n	G_n^P	G_t	G_t^P	C_n	C_n^P	C_t	C_t^P
$q = 0.95$												
SP500	0.000	0.038	0.695	0.152	0.000	0.040	0.764	0.169	0.000	0.012	0.690	0.068
ELUXB	0.000	0.378	0.925	0.426	0.000	0.351	0.959	0.491	0.000	0.232	0.876	0.470
CCI	0.000	0.038	1.000	0.081	0.000	0.050	1.000	0.024	0.000	0.089	1.000	0.042
$q = 0.975$												
SP500	0.000	0.229	0.717	0.179	0.000	0.145	0.619	0.172	0.000	0.183	0.718	0.169
ELUXB	0.000	0.259	0.647	0.396	0.000	0.250	0.787	0.383	0.000	0.202	0.574	0.513
CCI	0.000	0.285	1.000	0.248	0.000	0.130	1.000	0.182	0.000	0.366	1.000	0.146
$q = 0.99$												
SP500	0.000	0.038	0.382	0.041	0.000	0.020	0.381	0.047	0.000	0.040	0.589	0.131
ELUXB	0.000	0.013	0.317	0.026	0.000	0.002	0.359	0.113	0.000	0.027	0.221	0.027
CCI	0.000	0.009	1.000	0.028	0.000	0.005	1.000	0.029	0.000	0.009	1.000	0.049
$q = 0.995$												
SP500	0.000	0.005	0.085	0.003	0.000	0.005	0.171	0.011	0.000	0.024	0.109	0.028
ELUXB	0.000	0.070	0.483	0.026	0.000	0.037	0.572	0.034	0.000	0.012	0.270	0.034
CCI	0.000	0.007	0.997	0.009	0.000	0.009	0.976	0.020	0.000	0.029	0.995	0.044

Table 8: p-values for the bootstrap test of expected shortfall. A p-value of 0.000 should be interpreted as something less than 0.0005, but not zero.

Immediately clear from table 8 above is that the original models with normally distributed innovations give unsatisfactory estimates of expected shortfall; these estimates are too low. This is in line with the results of McNeil & Frey (2000, pp. 294–295), who report that "the residuals derived under an assumption of normality always fail the test with p-values in all cases much less than 0.01", and further conclude (quite strongly) that "an assumption of conditional normality is useless for the purposes of calculating expected shortfall". While the POT models corresponding to the models with normal innovations seem to improve the p-values somewhat, it is in most cases not enough, if going by a significance level of 0.05.

The models with t-distributed innovations seem to do better, with much higher p-values in all cases (all above a significance level of 0.05). However, for these models the corresponding POT models do comparatively worse, in many cases rejected where the original model was not. As noted earlier, this test tends to favor models that overestimate ES. Earlier we concluded that the original models with t-distributed innovations tend to overestimate VaR, and the corresponding POT models underestimate VaR, and we suspect that the same holds true for expected shortfall. The V_1 test statistics described in section 4.3.3 may help in resolving this matter.

Again, there is no clear winner among the different conditional volatility models (Standard/GJR/Component GARCH).

5.2.2 V-Test of Expected Shortfall

The sign of the V_1 test statistics in table 9 below can tell us whether ES is under- or overestimated on average. A negative sign indicates overestimation, and a positive sign underestimation. The closer to zero the statistic, the better the model.

As we suspected, the original models with t-distributed innovations seem to have a negative sign in most cases, so that these models tend to overestimate the expected shortfall. Conversely, the models with normal innovations exclusively have V_1 statistics with positive signs, and these are mostly larger in magnitude than the statistics of the models with t-distributed innovations. The POT models corresponding to the original models with normal innovations do better than the latter in terms of magnitude of the statistics, and mostly so in the case of t-distributed innovations. About the same conclusions can be drawn for the V_2 statistics in table 10 below.

Data set	S_n	S_n^P	S_t	S_t^P	G_n	G_n^P	G_t	G_t^P	C_n	C_n^P	C_t	C_t^P
$q = 0.95$												
SP500	0.290	0.055	-0.064	0.016	0.277	0.040	-0.085	-0.000	0.298	0.062	-0.042	0.043
ELUXB	0.254	-0.014	-0.106	-0.017	0.247	-0.016	-0.100	-0.015	0.254	-0.000	-0.097	-0.032
CCI	0.145	0.051	-0.273	0.039	0.145	0.045	-0.263	0.055	0.153	0.038	-0.237	0.044
$q = 0.975$												
SP500	0.340	0.031	-0.097	0.062	0.323	0.048	-0.075	0.036	0.340	0.041	-0.083	0.038
ELUXB	0.347	-0.021	-0.104	-0.035	0.335	-0.016	-0.124	-0.009	0.327	0.004	-0.091	-0.078
CCI	0.129	0.010	-0.450	0.024	0.127	0.046	-0.431	0.032	0.134	0.007	-0.412	0.034
$q = 0.99$												
SP500	0.445	0.282	0.108	0.309	0.439	0.288	0.096	0.244	0.417	0.259	-0.001	0.166
ELUXB	0.549	0.195	0.007	0.165	0.567	0.277	-0.006	0.126	0.496	0.152	-0.026	0.175
CCI	0.207	0.115	-0.606	0.093	0.198	0.116	-0.656	0.082	0.202	0.123	-0.659	0.086
$q = 0.995$												
SP500	0.564	0.650	0.570	0.829	0.595	0.570	0.439	0.601	0.544	0.443	0.455	0.465
ELUXB	0.661	0.177	-0.095	0.237	0.662	0.224	-0.085	0.227	0.615	0.277	0.029	0.228
CCI	0.252	0.173	-0.626	0.167	0.251	0.151	-0.818	0.134	0.261	0.119	-0.785	0.094

Table 9: p-values for the V_1 test statistics.

Data set	S_n	S_n^P	S_t	S_t^P	G_n	G_n^P	G_t	G_t^P	C_n	C_n^P	C_t	C_t^P
$q = 0.95$												
SP500	0.331	0.127	0.044	0.117	0.317	0.145	0.035	0.122	0.350	0.142	0.109	0.146
ELUXB	0.254	-0.014	-0.106	-0.017	0.247	-0.016	-0.100	-0.015	0.254	-0.000	-0.097	-0.032
CCI	0.192	0.075	-0.225	0.068	0.202	0.081	-0.205	0.083	0.200	0.069	-0.156	0.115
$q = 0.975$												
SP500	0.550	0.207	0.025	0.185	0.529	0.230	0.033	0.208	0.569	0.217	0.095	0.220
ELUXB	0.347	-0.021	-0.104	-0.035	0.335	-0.016	-0.124	-0.009	0.327	0.004	-0.091	-0.078
CCI	0.269	0.097	-0.448	0.086	0.287	0.106	-0.425	0.105	0.276	0.085	-0.364	0.142
$q = 0.99$												
SP500	1.006	0.414	0.064	0.393	0.992	0.451	0.103	0.438	1.012	0.434	0.131	0.425
ELUXB	0.549	0.195	0.007	0.165	0.567	0.277	-0.006	0.126	0.496	0.152	-0.026	0.175
CCI	0.415	0.180	-0.810	0.156	0.445	0.193	-0.789	0.181	0.423	0.159	-0.751	0.206
$q = 0.995$												
SP500	1.574	0.692	0.190	0.683	1.541	0.771	0.258	0.739	1.553	0.710	0.206	0.729
ELUXB	0.661	0.177	-0.095	0.237	0.662	0.224	-0.085	0.227	0.615	0.277	0.029	0.228
CCI	0.563	0.303	-1.165	0.267	0.598	0.299	-1.160	0.264	0.585	0.271	-1.130	0.306

Table 10: p-values for the V_2 test statistics.

Lastly, the V statistics, being the average of the absolute values of the V_1 and V_2 statistics, again indicate that the models with t-distributed innovations outperform the models with normally distributed innovations. Further, the POT models corresponding to the models with normal innovations outperform the latter as seen by the smaller magnitude of these POT models, while the case is not as clear-cut for the models with t-distributed innovations. Again, the type of conditional volatility model does not seem to matter as much as the distribution of the innovations, or whether it is the original model or the corresponding POT model.

Data set	S_n	S_n^P	S_t	S_t^P	G_n	G_n^P	G_t	G_t^P	C_n	C_n^P	C_t	C_t^P
$q = 0.95$												
SP500	0.310	0.091	0.054	0.067	0.297	0.093	0.060	0.061	0.324	0.102	0.075	0.094
ELUXB	0.226	0.054	0.068	0.055	0.221	0.062	0.069	0.057	0.246	0.057	0.054	0.058
CCI	0.169	0.063	0.249	0.054	0.174	0.063	0.234	0.069	0.176	0.053	0.197	0.079
$q = 0.975$												
SP500	0.445	0.119	0.061	0.124	0.426	0.139	0.054	0.122	0.455	0.129	0.089	0.129
ELUXB	0.369	0.078	0.090	0.079	0.359	0.083	0.106	0.067	0.381	0.079	0.079	0.092
CCI	0.199	0.053	0.449	0.055	0.207	0.076	0.428	0.069	0.205	0.046	0.388	0.088
$q = 0.99$												
SP500	0.725	0.348	0.086	0.351	0.715	0.370	0.099	0.341	0.715	0.347	0.066	0.296
ELUXB	0.696	0.252	0.057	0.219	0.705	0.304	0.065	0.201	0.685	0.229	0.055	0.212
CCI	0.311	0.147	0.708	0.125	0.321	0.155	0.722	0.132	0.313	0.141	0.705	0.146
$q = 0.995$												
SP500	1.069	0.671	0.380	0.756	1.068	0.670	0.349	0.670	1.049	0.576	0.330	0.597
ELUXB	0.991	0.336	0.096	0.358	0.997	0.378	0.115	0.329	0.981	0.405	0.029	0.357
CCI	0.408	0.238	0.896	0.217	0.424	0.225	0.989	0.199	0.423	0.195	0.957	0.200

Table 11: p-values for the V test statistics.

We end this section for the Expected shortfall test results with a plot of the $ES_{0.99}$ from an AR(1)-GARCH(1,1) model with normally distributed innovations and from the corresponding POT model.

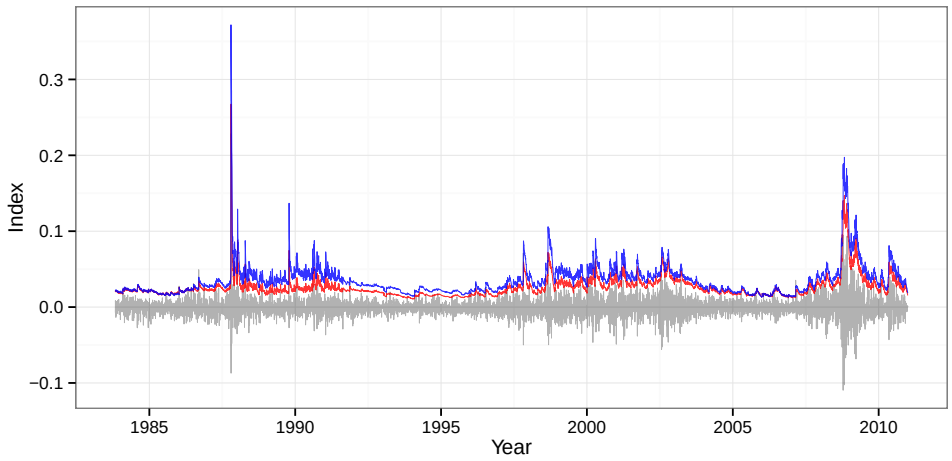


Figure 9: $ES_{0.99}$ estimates for the S&P 500 index, from an AR(1)-GARCH(1,1) model with normally distributed innovations (red) and from its corresponding POT model (blue).

6 Conclusions

The main conclusion to draw from the results presented in section 5 is that the choice of conditional volatility model—standard GARCH, GJR-GARCH, or Component GARCH in this paper—seems to matter less than the distribution of the innovations in the loss process. We found that models in which the innovations had a normal distribution tended to underestimate value at risk and expected shortfall, while t-distributed innovations gave better estimates, though sometimes too high. In applying the Peak Over Threshold method to the residuals of each model, the VaR and ES estimates were in many cases improved, though it is not clear that the POT method is needed in case of t-distributed residuals. One could perhaps estimate the degrees of freedom ν of the t-distribution instead of holding ν fixed as we have. This would lead to less uncertainty in the estimates compared to fitting a Generalized Pareto distribution to the residuals, as the model would be more parsimonious and the estimates of VaR and ES would contain fewer parameter estimates.

Our conclusion about the (lesser) importance of the choice of conditional volatility model may be due to too much similarity between these models. For further studies, it may therefore be interesting to investigate models that have a different volatility structure, or that simply have GARCH orders different from those considered here. Of course, from a practical standpoint it would be even more interesting to investigate multivariate models for value at risk and expected shortfall, both using methods from multivariate time series analysis and multivariate models from Extreme value theory, e.g. the use of copulas.

References

Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999), 'Coherent measures of risk', *Journal of Economic Perspectives* Vol. 15 (no. 4), pages 157–168.

Black, F. (1976), Studies of stock price volatility changes, *in* 'Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economic Statistics Section', pages 177–181.

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroscedasticity', *Journal of Econometrics* Vol. 31 (no. 3), pages 307–327.

Brooks, C. (2008), *Introductory Econometrics for Finance*, second edn, Cambridge University Press.

Byström, H. N. E. (2004), 'Managing extreme risks in tranquil and volatile markets using conditional extreme value theory', *International Review of Financial Analysis* Vol. 13, pages 133–152.

Canty, A. & Ripley, B. D. (2013), *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-9.

Christoffersen, P. F. (2003), *Elements of financial risk management*, Academic Press.

Coles, S. (2001), *An introduction to statistical modeling of extreme values*, Springer-Verlag London Ltd.

Danielsson, J., de Haan, L., Peng, L. & de Vries, C. (2001), 'Using a bootstrap method to choose the sample fraction in tail index estimation', *Journal of Multivariate Analysis* Vol. 76 (no. 2), pages 226–248.

Dowd, K. (2005), *Measuring market risk*, second edn, John Wiley & Sons.

Efron, B. & Tibshirani, R. J. (1993), *An Introduction to the Boostrap*, Chapman & Hall.

Embrechts, P., Kaufmann, R. & Patie, P. (2005), 'Strategic long-term financial risks: Single risk factors', *Computational Optimization and Applications* Vol. 32 (no. 1–2), pages 61–90.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997), *Modelling extremal events for insurance and finance*, Springer.

Enders, W. (2003), *Applied Econometric Time Series*, second edn, John Wiley & Sons.

Engle, R. F. (1982), 'Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation', *Econometrica* Vol. 50 (no. 4), pages 987–1007.

Gerlach, R. H., Chen, C. W. & Lin, L.-Y. (2012), Bayesian semi-parametric expected shortfall forecasting in financial markets.

Ghalanos, A. (2013), *rugarch: Univariate GARCH models*. R package version 1.0-16.

Gilli, M. & Këllezi, E. (2006), 'An application of extreme value theory for measuring financial risk', *Computational Economics* Vol. 27 (no. 2–3), pages 207–228.

Glosten, L. R., Jagannathan, R. & Runkle, D. E. (1993), 'On the relation between the expected value and the volatility of the nominal excess return on stocks', *Journal of Finance* Vol. 48 (no. 5), pages 1779–1801.

Hoogerheide, L. & van Dijk, H. K. (2010), 'Bayesian forecasting of value at risk and expected shortfall using adaptive importance sampling', *International Journal of Forecasting* Vol. 26 , pages 231–247.

Hull, J. C. (2006), *Risk Management and Financial Institutions*, Prentice Hall.

Jorion, P. (2001), *Value at Risk: The new benchmark for managing financial risk*, second edn, McGraw-Hill.

Lee, G. G. J. & Engle, R. F. (1999), A permanent and transitory component model of stock return volatility, *in* 'Cointegration Causality and Forecasting A Festschrift in Honor of Clive WJ Granger', Oxford University Press, pages pages 475–497.

McNeil, A. J. & Frey, R. (2000), 'Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach', *Journal of Empirical Finance* Vol. 7 (no. 3–4), pages 271–300.

McNeil, A. J., Frey, R. & Embrechts, P. (2005), *Quantitative risk management: Concepts, techniques, and tools*, Princeton University Press.

Revolution Analytics (2012), *doParallel: Foreach parallel adaptor for the parallel package*. R package version 1.0.1.

Stephenson, A. G. & Heffernan, J. E. (2012), *ismev: An Introduction to Statistical Modeling of Extreme Values*. Original S functions written by Janet E. Heffernan with R port and R documentation provided by Alec G. Stephenson.

Tolikas, K., Koulakiotis, A. & Brown, R. A. (2007), 'Extreme risk and value-at-risk in the German stock market', *The European Journal of Finance* Vol. 13 (no. 4), pages 373–395.

Tsay, R. (2010), *Analysis of Financial Time Series*, third edn, John Wiley & Sons.

Wickham, H. (2009), *ggplot2: elegant graphics for data analysis*, Springer New York.

Wuertz, D. & Chalabi, Y. (2010), *timeSeries: Rmetrics - Financial Time Series Objects*. R package version 2130.90.

Wuertz, D., with contribution from Michal Miklovic, Y. C., Boudt, C., Chausse, P. & others (2009), *fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling*. R package version 2110.80.

Yamai, Y. & Yoshiba, T. (2002), 'Comparative analyses of expected shortfall and value-at-risk: Their estimation error, decomposition, and optimization', *Monetary and Economic Studies* Vol. 20 (no. 1), pages 87–122.

Bachelor's Theses in Mathematical Sciences 2013:K7
ISSN 1654-6229

LUNFMS-4007-2013

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/