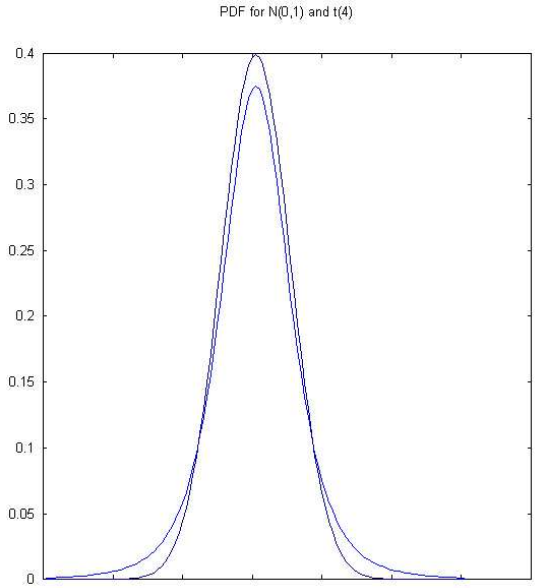

Tails and outliers in financial time series

Fredrik Strandberg

27th December 2001

Contents

Abstract

This paper is about the modelling of tails for real world financial data, with special attention paid to aberrant values, "outliers", their modelling and detection within univariate and multivariate frameworks. Problems encountered in practice are missing values, inhomogeneous series, non-synchronous series, small samples and time-varying volatility.

A major part of the work was done at HypoVereinsbank Risk Control in Munich in the autumn of 2001, with the goal of constructing a robust and automatic data quality control system, applicable for all kinds of financial instruments. The system was finally implemented in the programming language Formula Engine for the financial market data base system Asset Control.

My thanks to my supervisor Jan Grandell, KTH Stockholm, Timo Teräsvirta, Stockholm School of business, Stefan Lundbergh, Skandia Asset management, Thomas Mikosch, University of Copenhagen, Claudia Klüppelberg, Munich University of Technology, Michael Auer and Gabriela Pop, HypoVereinsbank Risk Control, Munich.

1 Introduction

Over the last few years, financial mathematics has attracted a dramatically increasing amount of interest. The benchmark for this science is the Black-Scholes derivative pricing formula that appeared 1973 and brought the Brownian motion, the Girsanov transform and the theory of stochastic integration into the spotlight. Parallel to the derivative pricing theory, econometricians have given their attention to time series modelling. Time series are discrete stochastic processes, and the connection in the limit to continuous time and stochastic integration is not trivial. However, the aims of time series modelling are more in the direction of finding a reasonable model, that is mathematically tractable and can be understood and used by practitioners.

There exists a large amount of proposed financial time series models, more or less complicated. Since there also exists a large variety of financial data, it not possible to tell which model is the generally best one.

1.1 Goals and problems

The central banks in several countries have agreed on following the BIS (Bank for International Settings) Basel Committee agreement. This includes, among many other things, the obligation of estimating the covariance by the standard sample covariance.

Typically, a group of chosen series (such as important stocks or indices), "riskfactors", are used to derive risk measures, such as Expected Shortfall or Value-at-Risk. If a matrix X is constructed, where x_{ij} is observation nr $i, i = 1, ..., n$ of return series j, the covariance is then estimated by

$$\Sigma^* = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X})$$

If a multivariate normal distribution is assumed, which is very common, the joint distribution is then completely specified.

The aim of this work has been to investigate possible solutions of the problems encountered in the real world, by estimating Σ. Practical problems to deal with are:

- Missing values due to technical failures or holidays.
- Outliers. There are sometimes values occurring that should not be used in the estimation of Σ.
- Non-synchronous series due to geographical and political differences. For example, the markets in Stockholm, New York and Tokyo all trade at different times of the day. There are also locally defined holidays in each country, such as Christmas or the first of may in Sweden.

These problems are all connected. Non-synchronousity gives rise to missing values and if it is decided that a value should not be used (an outlier), it also has to be replaced by another value.

Since an outlier is typically defined as present when the distance between the realisation and what it "should" be, a prediction, is too large, the whole concept practically boils down to the problem of prediction in time series.

The aim of this paper is to give an overview of possible techniques used to deal with these problems. First, however, it is necessary to have an understanding of the typical behaviour of financial time series, and how they can be modelled.

The paper is organised as follows: First we summarise some empirical properties of financial time series and consider how this behaviour can be modelled. Extra attention is paid the GARCH model.

Next, we consider the univariate case and consider how some outlier types can be modelled and included in a time series model. Then, the multivariate case is considered. We investigate joint distributions, conditional marginal distributions and a multivariate distance measure. We also give an application of the multivariate techniques in the univariate case, and we generalise some outlier definitions.

Then, we give a review of tail modelling with Extreme Value Theory (EVT). The POT method and the Hill estimator are described, and we give an application where the EVT is combined with a GARCH model.

Finally, a short review of robust estimation techniques is given and we show how a missing value can be optimally interpolated within a given univariate time series model. In appendix A, we investigate the problems of fitting a GARCH model and propose a solution.

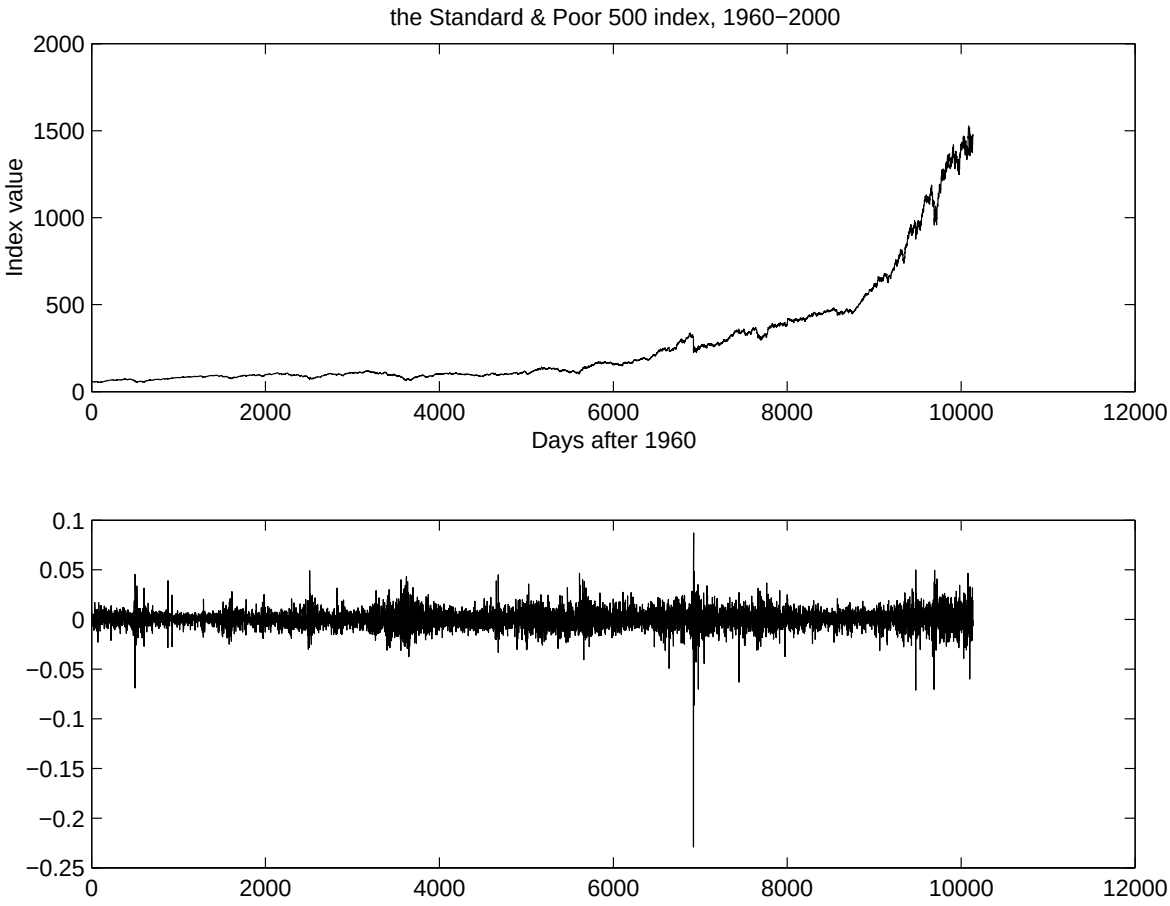

Figure 1: The SP500 index, 1960-2000. **Upper**: The index series. **Lower**: The log-return series

2 Empirical properties

Due to the broad spectrum of financial data series types, one should also be careful of saying too much about common properties. However, if one focuses on share prices, stock indices or foreign exchange rates, they all behave very much the same way after the extremely common transformation

$$X_t = log(P_t/P_{t-1})$$

The series $\{X_t\}$ is referred to as the *(log) return* series. Also commonly used is the *relative return* series

$$\frac{P_t - P_{t-1}}{P_{t-1}}$$

The latter is perhaps more intuitive. By a Taylor series argument, the relative returns are close to the log-returns.

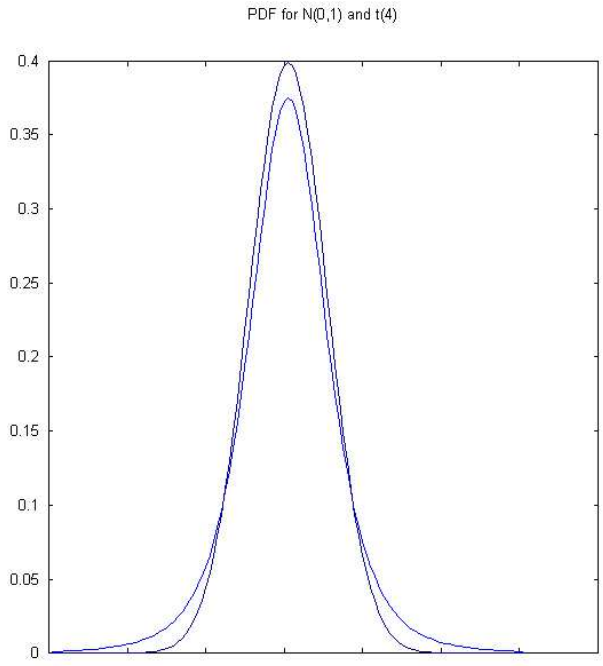

Figure 2: The PDF for the standard normal distribution and the $t(4)$-distribution

The return series is a mean-reverting martingale sequence. The mean μ_t is sometimes modelled as an AR(p)-process, but in practice the constant zero assumption is good and most often used. The unconditional sample variance is usually of order 10^{-4} and smaller.

Financial time series typically possess the following further properties:

2.1 Leptokurtosis

Leptokurtosis is a property of a probability distribution that gives a higher peak, a thinner midrange, and fatter tails than a normal distribution. The most common model of stock price movements is the geometrical Brownian motion (used in the Black-Scholes model)which would imply the log-returns (from now on referred to as "returns") to be an IID normal series. However, an inspection of practically any series will show that this is not quite correct. A histogram will show that the distribution is roughly symmetric but *sharply peaked around zero and with heavy tails on both sides.* This can also be clearly seen in a QQ-plot against the normal distribution.

A QQ-plot, for Quantile-Quantile plot, is a plot of the empirical quantiles against the theoretical quantiles of a given distribution. If the values come from this distribution, the plot will be an approximately straight line $y(x) = x$

A heavy tailed distribution such as the student t-distribution of perhaps 3,4 or 5 degrees of freedom is usually more appropriate.

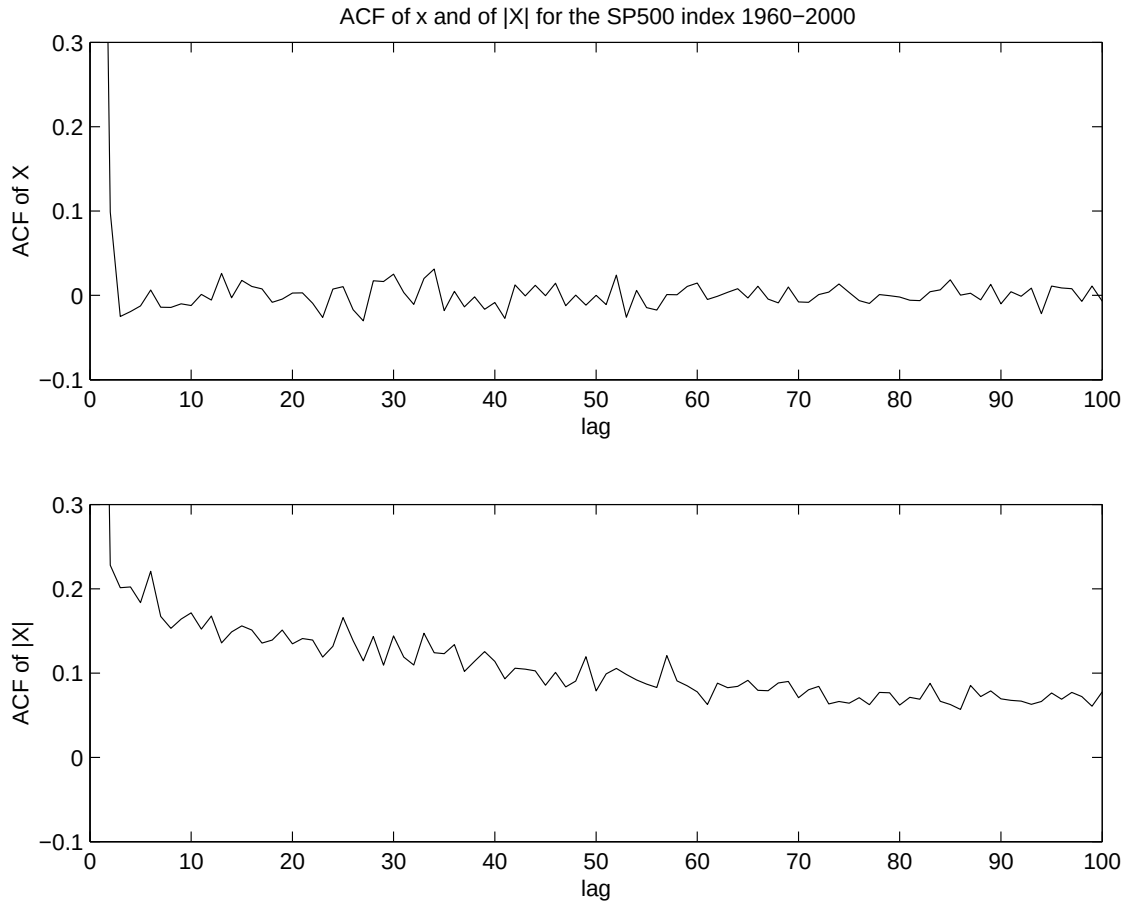

Figure 3: A typical feature of financial time series: The ACF is negligible at almost all lags, but the ACF of the absolute values are different from zero for a large range of lags

2.2 Autocorrelation

This is another important contradiction to the Black-Scholes normal white noise (WN) assumption. A typical feature of financial time series is the interesting structure of the autocorrelation function: The sample autocorrelations are negligible at almost all lags, i.e. the WN assumption seems plausible, *but*, the sample autocorrelations of the *squares* (or absolute values) are different from zero for a large range of lags. See figure 3, where the autocorrelation function (ACF) of the SP500 index (figure 2) is plotted for the return series and for the squared return series. There are different opinions of how the latter property is to be interpreted, the standard idea is to interpret the slow decay as a long memory or *long range dependence*. Mikosch (2000) however, claims that this might not be correct, first since it is empirically shown that often the fourth moment does not exist, implying the ACF to produce meaningless results, and secondly due to non-stationarity effects.

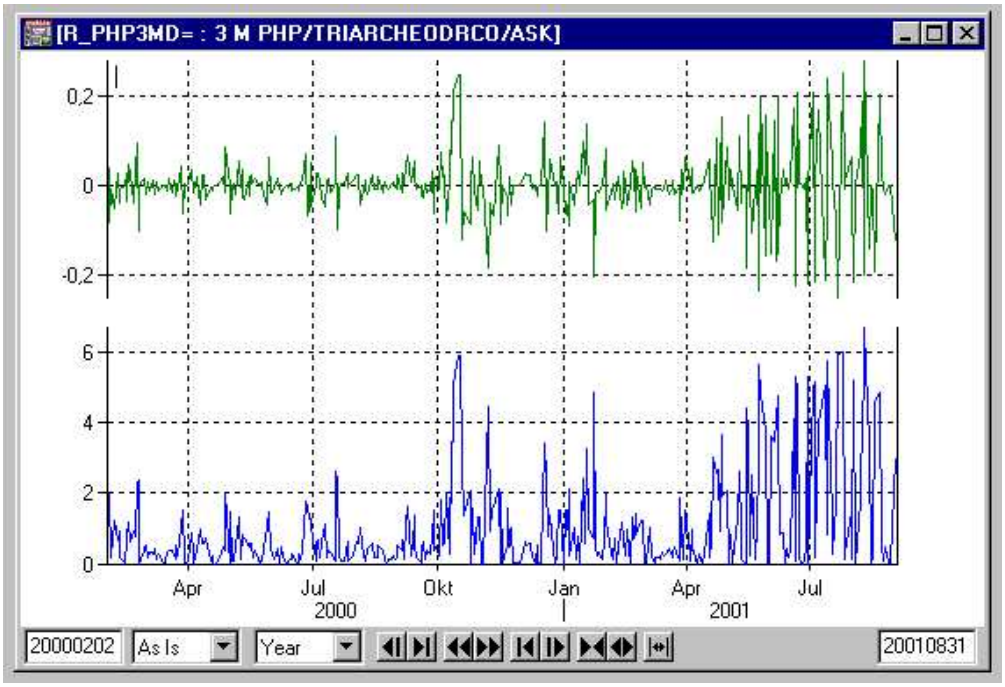

Figure 4: Volatility clustering

2.3 Volatility clustering

Large and small values often occur in clusters. Periods of low variance seem often to be followed by periods of high variance. Recall that the BS formula assumes the variance to be constant in time, which is a questionable assumption. The latter property has given rise to many different models for the conditional variance $\sigma_t^2 = Var(X_t|x_{t-1}, x_{t-2})$.

For the large (and small) returns, there is also dependence in the tails. The analysis of the tail data is called Extreme Value Theory (EVT) and treated in section 10.

3 Stochastic volatility models

Most models for financial returns X_t used in practice are of the form

$$X_t = \mu_t + \sigma_t Z_t,$$

The standard assumptions for the noise Z_t is that $Z_t \in IID, E(Z) = 0$ and $Var(Z) = 1$, with Z_t independent of σ_t and Z is typically assumed to be normal or t-distributed. Hence, conditional upon $X_{t-1}, X_{t-2}, ..., X_t$ has variance σ_t^2.

Sometimes an AR(p) process is assumed for the mean, but usually a constant zero mean $\mu_t = 0$ is used.

The standard deviation process $\{\sigma_t\}$ is called *volatility* by econometricians and often assumed to be a function of past values X_{t-1}, X_{t-2}, and $\sigma_{t-1}, \sigma_{t-2},$. This gives together with the assumed distribution of Z_t an easy distributional forecast - a very central question for every bank is to estimate the quantiles of X_t, the so called Value-at-Risk (VaR).

3.1 The ARCH family

ARCH stands for AutoRegressive Conditional Heteroskedasticity. The first model ARCH(p) for the conditional volatility was proposed by Engle 1982

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2$$

Recall that usually X_t is close to white noise, but that there is significant dependence in the squared returns X_t^2.

However, the $ARCH(p)$ model does not fit data very well, unless the order p is chosen very large. Therefore, some improvements have been suggested. Since the rationale is that σ_t^2 is a time-changing weighted average of the past squared observations, it is quite natural to define σ_t^2 not only as a weighted average of past X_j^2's but also of past σ_j^2's. This leads to the Generalised ARCH model, GARCH, introduced by Bollerslev 1986. The volatility process, "the stochastic volatility", is

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 := \alpha_0 + \alpha(B) X_t^2 + \beta(B) \sigma_t^2$$

where the $\alpha_i's$ and $\beta_j's$ are non-negative parameters, B is the backshift operator and $\alpha(B), \beta(B)$ are the corresponding polynomials with coefficients α_i, β_j.

4 GARCH

Since their interception, GARCH processes have gained a very fast acceptance in the financial literature, and is now one the most frequently used models. Several different ARCH-type models have been proposed, see for instance Kaiser (1996), but the "normal" GARCH(p,q) model is by far the most used in practice. One reason for the popularity is probably the existence of software packages, such as the GARCH modules for MATLAB or Splus. Another reason might be the connection to ARMA processes: It is straightforward that the equation for σ_t^2 can be rewritten as an ARMA equation with noise $v_t := X_t^2 - \sigma_t^2$:

$$(1 - \phi(B))X_t^2 = \alpha_0 + (1 - \beta(B))v_t, \qquad \phi(B) := \alpha(B) + \beta(B)$$

If $\{X_t\}$ is strictly stationary and $EX^2 < \infty$, $\{v_t\}$ constitutes a strictly stationary martingale difference sequence. However, even though the GARCH process might be rewritten as an ARMA process for the squares, $\{X_t\}$ is a non-linear process and not all theoretical properties are known.

4.1 Stationarity

For the GARCH(p,q) model with IID innovations Z_t such that $EZ = 0$ and $EZ^2 = 1$, $\{X_t\}$ is strictly stationary with finite variance if

$$\alpha_0 > 0 \text{ and } \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$$

A source of confusion can be the borrowed terminology from ARMA models. In analogy with ARMA and "integrated ARMA" (ARIMA) Engle and Bollerslev coined the name "integrated GARCH(p,q)" (the IGARCH(p,q) process) for the situation when

$$\alpha_0 > 0 \text{ and } \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j = 1$$

However, the situation is *not* the same in the ARCH case as for the ARMA case: In the IGARCH case, it has actually been shown that the process is strictly stationary but with infinite second moment.

4.2 Persistence in variance

Definition 4.1 (Persistence in variance)
A process $\{X_t\}$ is said to be persistent in variance if

$$\lim_{t\to\infty} sup|E(X_t^2|X_0, X-1, ...) - E(X_t^2|X_1, X_0, ...)| > 0, \;\; P \; a.s$$

This means that a process $\{X_t\}$ is persistent in variance if the differences between the forecasts of the conditional variances at times 0 and 1 will never disappear, or if the shocks to the conditional variance persist indefinitely.

For the GARCH(1,1) model, it can easily be shown from the ARMA representation that this process is non-persistent in variance iff $\alpha_1 + \beta_1 < 1$ i.e. when X_t has finite variance. However, the stationary IGARCH(1,1) model is persistent in variance. This is important to notice, since one *very* often used model in practice is the IGARCH(0.06,0.94) model, often referred to as the RiskMetrics Exponential Moving Average (EWMA) Model.

4.3 Long-Range Dependence Effect

Consider the ACF $\rho_X(h)$.

Definition 4.2 (Long Range Dependence)
$\{X_t\}$ is said to exhibit long-range dependence (LRD) if

$$\sum_{h=0}^{\infty} |\rho_X(h)| = \infty$$

The property shown in figure 3 is an example of this effect. However, the function $\rho_X(h)$ for a GARCH(p,q) model can be found analytically, and it is found that it is exponentially decaying. See Mikosch and Starica (2000). Hence, a GARCH model is not designed to model the LRD effect. But if the GARCH model cannot capture this behaviour, then how can it be explained?

Before answering, we consider another interesting stylised fact for the GARCH models:

4.4 IGARCH effect

It has been observed empirically that the estimated parameters $\alpha_1, ..., \alpha_p$ and $\beta_1, ..., \beta_q$ sum up to values that often are very close to one. We call this IGARCH effect. This motivated the introduction of the IGARCH model as a possible generating process for log-returns. As mentioned, the IGARCH has a strictly stationary solution, but with infinite variance. To see this, just take expectations of 4.1 and note that $EX^2 = E\sigma^2$:

$$E\sigma^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i EX^2 + \sum_{j=1}^{q} \beta_j E\sigma^2 = \Big(\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j\Big) E\sigma^2 + \alpha_0$$

Since $\alpha_0 > 0$ is necessary for strict stationarity, $E\sigma^2 = \infty$ follows. Under some mild assumptions (See Mikosch and Starica (2000)), such as the existence of a density with infinite support, it can be shown that

$$P(X > x) = const \cdot x^{-2}, x \to \infty$$

To conclude: *If* the IGARCH model was the generating process of the log-returns, in particular, with the variance infinite, the sample ACF of X_t, $|X_t|$ and X_t^2 would estimate nothing meaningful and the observed LRD-effect has nothing to do with LRD.

In the case where $\alpha_1 + \beta_1 \approx 1$ but $\alpha_1 + \beta_1 \neq 1$, the ACF are well-defined and the corresponding sample ACF have, possibly, a meaning. Miksosch and Starica (2000) investigated this case and concluded that X has a power low behaviour in the tails, such that there exists a $\kappa > 2$ such that

$$P(X > x) = const \cdot x^{-\kappa}, x \to \infty$$

However, as has been shown empirically by very many authors, a direct semi-parametric estimate of this so-called tail index will typically give a value of $\kappa \in (3, 5)$. The deviation of κ from 2 can *not* (only) be explained by poor estimates. The main reason is because the GARCH process is not a suitable model for the data!

It is found that the IGARCH effect, the value of $\alpha_1 + \beta_1$ is typically increasing towards 1 with the sample size. This supports the claim that

> *The IGARCH effect is due a bad fit of the GARCH model, and the bad fit of the model is due to non-stationarity, which is more likely for a larger sample size*

By considering a sample that consists of two subsamples from different GARCH(p,q) processes, Mikosch and Starica (2000) investigated the so-called Whittle estimate of the parameters in the GARCH(1,1) case and found analytical evidence that the IGARCH effect will build up and $\alpha_1 + \beta_1$ tend to one when $\{X_t)\}$ is non-stationary.

A closer analysis of the sample ACF and the periodogram (the natural estimator of the spectral density for a stationary process) for such a sample will also show that non-stationarity can *also* responsible for the observed LRD-effect in the sample ACF, and this LRD-effect must therefore not have anything to do with LRD!

It can also be noted that there seems to be no economic reason for such a LRD effect in the returns. It should also be noted that there do exist models that can capture the LRD behaviour in the absolute returns and their squares.

To summarise

- Standard GARCH models cannot explain LRD
- GARCH models with constant parameters give a good fit only for short time horizons. Either, the parameters must be modelled as time-dependent as well, or the model must be re-estimated frequently.
- The observed LRD-effect and the IGARCH-effect might be both due to non-stationarity
- Sometimes the fourth moment does not exist. In that case, the ACF will produce meaningless results.

Mikosch and Starica (2000) constructed a test-statistic for checking the model goodness-of-fit to the data and detecting regime changes.

4.5 The GARCH(1,1) model

Of special interest in practice is the very simplest GARCH(1,1) model, normally written as

$$X_t = \sigma_t Z_t, \quad Z \in \text{IID}N(0,1)$$
$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$
$$\alpha_0 > 0, \alpha_1, \beta_1 \geq 0$$

(We throughout assume a constant zero mean). For reasonable sample sizes, this model is considered to capture the properties of financial time series acceptably well, despite its simplicity. Another important aspect is that the GARCH(1,1) can be shown to be stationary iff $\alpha_1 + \beta_1 < 1$. For the rest of this paper, the GARCH(1,1) model will be the only GARCH model considered.

Stationarity gives that

$$E(X_t^2) = E(X_{t-1}^2) = \sigma^2$$

and the independence assumption gives that

$$E(X_t^2) = E(\sigma_t^2 Z_t^2) = E(\sigma_t^2)E(Z_t^2) = E(\sigma_t^2)$$

Since

$$E(X_t^2) = E(X_{t-1}^2) = E(\sigma_t^2) = E(\sigma_{t-1}^2) = \sigma^2$$

we have

$$\sigma^2 = \alpha_0 + \alpha_1 \sigma^2 + \beta_1 \sigma^2$$

Hence, the unconditional variance (sometimes called "long-run" variance) is

$$Var(X) = \sigma^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}$$

Now we see that the model can be interpreted as a weighted sum

$$\sigma_t^2 = \gamma \sigma^2 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

where $\gamma + \alpha_1 + \beta_1 = 1$

If the unconditional variance σ^2 is estimated with the usual sample variance, the model is determined only by α_1 and β_1, which should to be estimated by maximum likelihood.

Unfortunately, it turns out that the GARCH(1,1) model is very troublesome to ML-estimate, which hampers its use in practice. Several software packages such as MATLAB, Splus or Xplore have functions for estimating GARCH parameters, and they all seem to work badly. The fundamental problem is numerical: For a fixed σ, the likelihood surface (α_1, β_1) is globally very flat, but locally not very smooth, making it hard for any automatic routine to find the maximum, and all standard optimizing programs diverge and return a boundary value, such as $\alpha_1 = 0, \beta_1 = 1$.

In this work, a lot of effort has been made in order to find a safe way of estimating the parameters. It turns out that the global flatness problem can be solved with a simple

coordinate transformation, but the local problem of a non-smooth surface is fundamentally due to a too small sample. See appendix A for details.

The risk of obtaining nonsense parameters rises the question of how to measure the accuracy of the model. Some alternatives are described in the next chapter.

4.6 Accuracy of the GARCH(1,1)

Adequacy tests for conditional variance are not as common as for many other models. For GARCH, some tests have been proposed though:

- **LM test:** Bollerslev (1986) suggested a "score" or Lagrange multiplicator (LM) test for testing a GARCH model against a higher-order GARCH model.
- **Remaining ARCH effect test:** Li and Mak (1994) derived a so-called portmanteau type test for remaining ARCH effect in the residuals. Lundbergh and Teräsvirta (2000) constructed an asymptotically equivalent version, using a LM test.
- **Remaining autocorrelation test:** Hull (2000) proposes to test for remaining autocorrelation in the residuals.
- **Asymmetry test:** Engle and Ng (1993) considered testing the GARCH specification against asymmetry using the so-called sign-bias and size-bias tests.
- **Parameter constancy test:** Chu (1995) derived a test of parameter constancy against a single structural break. This test has a nonstandard asymptotic null distribution. Lin and Yang (1999) derived another test against a single structural break, based on empirical distribution functions.
- **A structural change test:** Mikosch and Starica (2000) constructed a moving window test statistic in the spectral domain.

Here, we will only consider the case of constant parameters. Considering time-varying parameters and structural changes in samples smaller than, say, 500-1000 observations is pointless. Instead, the tests on the standardized residuals for remaining autocorrelation and remaining ARCH-effect (the robust test) were implemented.

4.6.1 Testing for autocorrelation

In the "standard" literature by Hull, a very simple and intuitive test is proposed: Consider the model residuals $Z_t = X_t/\sigma_t$. If the model is good, there should be no dependence in the residuals. Therefore, a standard Ljung-Box test was used to check for autocorrelation. The Ljung-Box test statistic (LB) is

$$LB(X) = \sum_{t=0}^{|h|} \frac{\rho_X^2(t+1)}{n-t-1}$$

where n is the length of the time series and h is the "lag". $h = 15$ was used. LB is χ^2 distributed with h degrees of freedom, but recall that *the ACF might not produce reliable results, since the fourth moment sometimes does not exist!*. (See Mikosch and Starica (2000)).

Instead of considering the value of the autocorrelation, a simple measure may be to just consider the *reduction* of autocorrelation as Goodness-of-Fit value:

$$GoF = \frac{LB(X^2)}{LB(Z^2)}$$

4.6.2 Testing for ARCH effect

The "ARCH-in-GARCH" test by Lundbergh and Teräsvirta (2000) is a parametric so called Lagrange multiplier (LM) test. The null hypothesis of no remaining ARCH effect is tested in the standardized residuals. It is a very easy test to perform; easy to implement and without any heavy computations at all. It can also be modified to a be robust against the normality assumption, with a minimal loss of efficiency! It is also asymptotically equivalent to the portmanteau test by Li and Mak (1994).

Let η denote the parameter vector, for our GARCH(1,1) $\eta = (\alpha_0, \alpha_1, \beta_1)$. η should if possible be estimated by quasi maximum likelihood (QML), but any other consistent estimate may be used. (Quasi means under the assumption of normal distribution). A necessary assumption for the test is that the QML estimator η^* is consistent and asymptotically normal. This assumption is very difficult to verify in practice for a GARCH(p,q) model, but the GARCH(1,1) is an exception; Here it can be shown that the stationarity condition $\alpha_1 + \beta_1 < 1$ is enough.

Now, for $X_t = \sigma_t Z_t$, the innovations Z_t are in fact

$$Z_t = z_t \sqrt{g_t}$$

where $\{z_t\}$ is a standard normal IID sequence and

$$g_t = 1 + \phi' \mathbf{v}_t$$

where

$$\mathbf{v}_t = (Z_{t_1}^2, ..., Z_{t_m}^2)'$$

and

$$\phi = (\phi_1, ..., \phi_m)'$$

The "extended" GARCH model then becomes

$$X_t = Z_t \sqrt{\sigma_t g_t}$$

which could be called an "ARCH nested in GARCH" model.

The hypothesis to test is $H_0 : \phi = \mathbf{0}$ against $\phi \neq \mathbf{0}$. Under this hypothesis, $g_t = 1$ and the model collapses into a GARCH(p,q) model.

Now, let X_t^* and σ_t^* denote X_t and σ_t estimated under H_0, and let $Z^{*2}_t = X^{*2}_t/\sigma^{*2}_t$. Let

$$\mathbf{y}_t^* = \frac{1}{\sigma_t^*}\frac{\partial \sigma_t^*}{\partial \eta'}$$

where $\frac{\partial \sigma_t^*}{\partial \eta'}$ denotes $\frac{\partial \sigma_t^*}{\partial \eta'}$ estimated under H_0.

For $p = q = 1$ we get

$$\mathbf{y}_t^* = \frac{1}{\alpha_0^* + \alpha_1^* X_t^2 + \beta^* \sigma^{2*}}(1, X^{*2}_{t-1}, \sigma_t^*)$$

Finally, let

$$\mathbf{v}_t^* = (Z^{*2}_{t_1}, ..., Z^{*2}_{t_m})'$$

The robust version of the test is then carried out as follows:

1. Obtain the QML estimates for η under H_0, compute $(X^*)_t^2/\sigma^{*2}_t - 1, \mathbf{v}_t^*$ and $\mathbf{y}_t^*, t = 1, ..., T$
2. Regress $\mathbf{v}_t^*$ on $\mathbf{y}_t^*$ and compute the $m \times 1$ residual vectors $\mathbf{r}_t$, t=1,...,T
3. Regress 1 on $(X^*)_t^2/(\sigma^*)_t^2 - 1)\mathbf{r}_t$ and compute the residual sum of squares SSR from that regression. The test statistic

$$LM = T - SSR$$

has an asymptotic χ^2 distribution with m degrees of freedom under H_0.

For details, the non-robust version and the proof, see Lundbergh and Teräsvirta (2000).

5 Outliers

There is no such thing as just an "outlier". An outlier is always an outlier with *respect to something*, usually a model, telling what the value "should" be. For a k-dimensional time series $\{x_t\}$,

$$x_t = (x_t^{(1)}, ..., x_t^{(k)})$$

we divided in this work "outliers" into 3 categories:

- **Multivariate outliers**.

  A multivariate outlier is occurring when all or many of the component series are jointly much out of their expected position (at some time point), i.e. if the multi-dimensional random variable that the multidimensional time series at a fixed time point constitutes is much out of it expected position, measured in some suitable way.

  In practice, a multivariate outlier in a financial time series is simply a day when something important happened, affecting all or many instruments. For example, 11 sep 2001. (The attacks on World Trade Center).

- **Marginal outliers** A marginal outlier is a component $x_t^{(i)}$ of the variable x_t, that is much out of its position as predicted by all other components of x_t at a fixed time point t. It could be moving in the unexpected direction for some unknown reason, or it could be a technical error. It also covers the case when it did not move, even though the rest of the market moved.

- **Univariate outliers** A univariate outlier is a value X_t in a univariate (component) time series $\{X_t\}$, which is much out of its position as predicted by all other values of $\{X_t\}$.

More explicit definitions will be given later. This classification was made up in this work, and there is no such standard classification appearing in the literature.

We do *not* define an outlier as a value not belonging to the distribution. That is a very hard statistical problem, in fact, it would often be impossible to tell weather or not a value is an "outlier" in that case.

We already know that financial time series are typically heavy tailed, and we concentrate on finding the values that "look" strange, leaving the question open how to interpret the empirical meaning.

In fact, often in the real world, the *definition* of an outlier is a value "looking strange", a value that seems not to fit into the picture. Banks and financial institutes typically remove values "not representative for the market" in the correlation/variance estimation. For them, it is of great importance to get estimates consistent with the empirically expected "economically reasonable" values

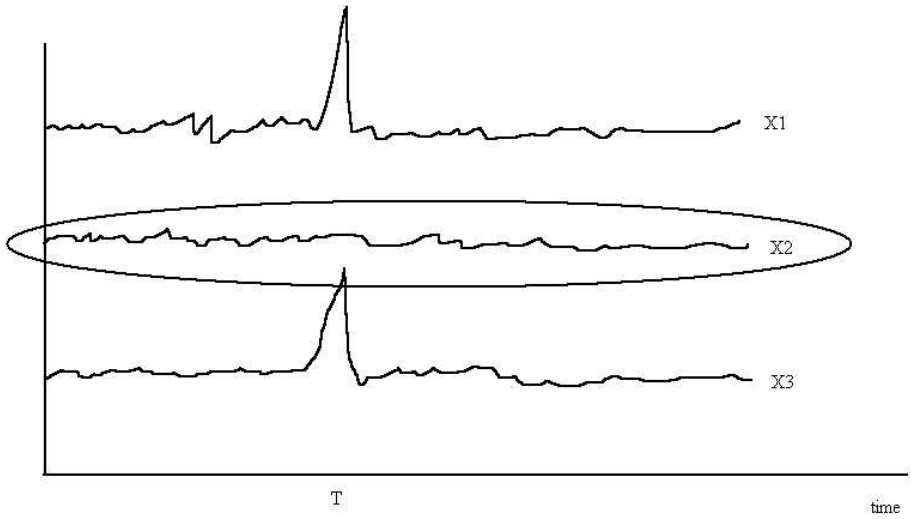

Figure 5: A marginal outlier, but not a univariate outlier. If the series X2 is only compared to itself, the point T will not be found suspect

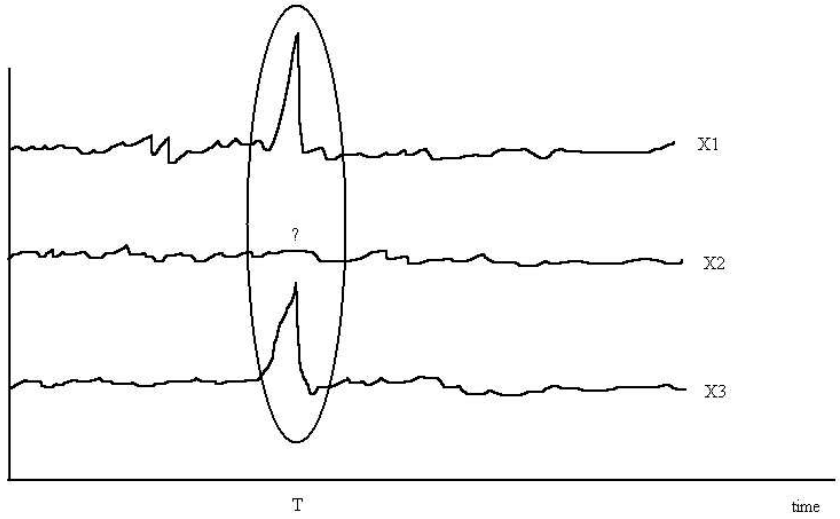

Figure 6: A marginal outlier, but not a univariate outlier. If X2 is compared to X1 and X3, the X2(T) will be found suspect. (Notice: 2 assumptions are made. 1. High correlations. 2. The movements of X1 and X3 are "true".)

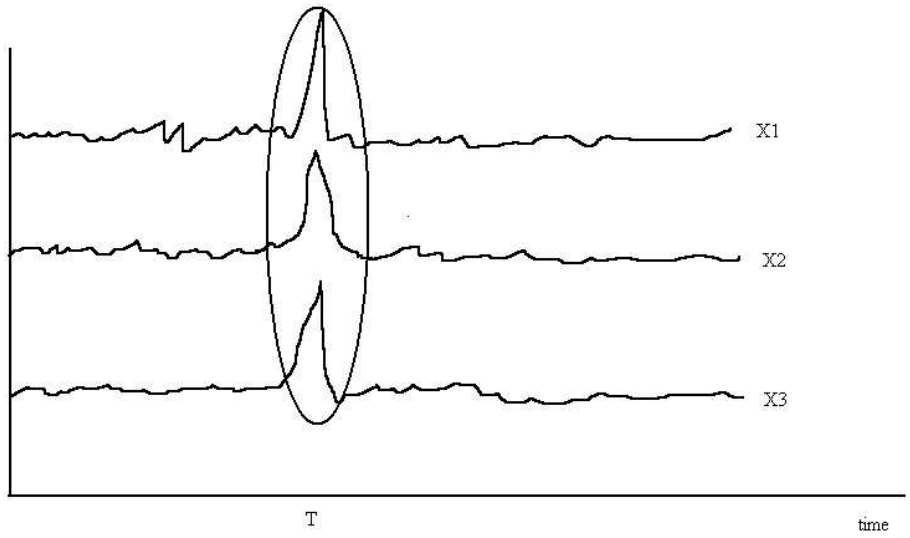

Figure 7: A multivariate outlier. No time dynamics is taken into account.

5.1 The subjectivity

The outlier problem is of a very subjective nature. This is mainly because it always includes an implicit assumption that the model that it is compared to is a correct model, and that the values that it is compared to are correct.

A statistical treatment of the general case is difficult. Therefore, we throughout assume the number of outliers to be small, typically less than 1 %. Since our fuzzy definition of an outlier is a value that does not fit into the picture, this is a natural thing to do - almost a consequence of the definition.

Still, it should be kept in mind that statistics *is* a fuzzy subject. The very common quantiles 95 % and 99% are being widely used to determine weather or not something is "significant". Just like in the outlier case, it depends on the situation to determine what "significant" is supposed to mean.

5.2 Outlier removal versus robust statistics

One should also be aware of the difference of focusing on the outliers themselves, or on the data. The latter approach gives rise to the robust statistics, which tries to treat the data in such a way that the outlier effects are minimized, typically using ordering statistics (and weight functions) as a basic workhorse, instead of summation. Generally, one could say that robust estimators usually give (much) better estimates in the presence of (big) outliers, but worse estimates when no outliers are present. See Lucas (1999) for a detailed review of robust methods in empirical finance.

Due to the BIS, we *must* estimate the covariance matrix as the usual sample covariance, i.e. in the non-robust way, using summations. Therefore, we must find the outliers and replace them with better estimates.

5.3 Problems

For outlier detection in practice, there are mainly two problems:

- **Masking effect**. Masking effect means that an outlier hides another outlier from being found.
- **Swamping effect**. Swamping effect means that an outlier causes a true value to look like an outlier.

In $\Re^N$, these phenomena often become severe, and it is crucial to have clear definitions. These effects take different forms in the univariate and the multivariate framework, as well as for different models. In the univariate case, a big outlier will typically induce a masking effect. In the multivariate case, there will typically be a swamping effect in the cross-component direction.

5.4 Effects of outliers

The most naive method of detecting/defining outliers (not necessary in dependent data), is to simply use a quantile in the tail as limit, like $x \geq 3\sigma$. However, usually σ is estimated as $\sqrt{\frac{1}{n-1}\Sigma(x_i - \bar{x})^2}$. This is a very unrobust and sensitive estimator, due to the squares. Hence, a big outlier would cause σ to be too big, implying a masking effect.

This illustrates the fundamental problem with big outliers: The model will not be correctly estimated. As for the variance, least squares estimates are very sensitive as well. Also Maximum Likelihood estimates are sensitive. This shows as that we need robust estimates for the model in which the outliers are to be detected.

For ARMA models, the Expectation Maximization (EM) algorithm gives robust approximate ML estimates, see Tolvi (1999). The EM algorithm is described in section 8.5. Estimates using the ACF or PACF are typically biased.

For the GARCH(1,1) model, which is to be estimated with ML, outliers make this already troublesome task even harder. It typically affects the α_0 parameter. See Franses and van Dijk (2000) for detailed simulation results. For a review of time series model selection in the presence of outliers, see Tolvi (2001).

6 Univariate outliers

This paper concerns outliers in dependent data, but for the sake of completeness a short review of the univariate IID case is given:

6.1 Outliers in univariate IID data

In general, every method derives some kind of test statistic. Under the null hypothesis of no outliers, its distribution (found analytically or empirically) is used to derive a critical value for a given significance level (i.e. a quantile). A few examples are given next.

6.1.1 The score test

The most common way of defining an outlier is to simply derive a confidence interval for the standardized variable $Z = (X - E(X))/\sigma$. The "method" of defining outliers in this situation as, say, $\mu \pm 3\sigma$, is often referred to as the "Z-score" test.

The catch with this "method", is that the moments are unknown. Using the standard estimators for the mean and the variance, implies a masking effect. If a big outlier X_ω is present, μ will be biased of order X_ω/n and σ^2 of order X_ω^2/n.

Hence, this is not a very good method. Though, in the case of one outlier, a value being deviating large, this outlier will probably be found with this test, but possible others will not. A way of reducing this effect is to replace X_ω with an estimate and re-estimate the moments.

6.1.2 Rosner's test:

A modification of the Z-score method is to use so-called trimming, which means that one or more of the biggest values are discarded. Ordering the sample s.t

$$\{X_1, ..., X_n\} \rightarrow \{Y_1, ..., Y_n\}$$

$$\{Y_1 \leq ... \leq Y_n\}$$

This very simple test for multiple outliers, both high and low, removes the observation most far away from the estimated mean and uses the test statistic

$$R_{i+1} = \left| \frac{Y_i - \bar{X}}{S_X} \right|$$

where S_y is the sample standard deviation. The test statistic is then compared to a critical value, computed under the null hypothesis of no outliers (assuming normal distribution).

6.1.3 Dixon's test

For small samples, about 3-25 observations, another trivial test is a so-called nearest-neighbor test. A τ-statistic is computed for the highest and the lowest value as

Sample size n	τ_{high}	τ_{low}
3-7	$\frac{Y_n - Y_{n-1}}{Y_n - Y_1}$	$\frac{Y_2 - Y_1}{Y_n - Y_1}$
8-10	$\frac{Y_n - Y_{n-1}}{Y_n - Y_2}$	$\frac{Y_2 - Y_1}{Y_{n-1} - Y_1}$
11-13	$\frac{Y_n - Y_{n-2}}{Y_n - Y_2}$	$\frac{Y_3 - Y_1}{Y_{n-1} - Y_1}$
14-25	$\frac{Y_n - Y_{n-2}}{Y_n - Y_3}$	$\frac{Y_3 - Y_1}{Y_{n-2} - Y_1}$

As for the other tests, τ is compared to a critical value, computed under the null. None of these tests use robust estimators. This is treated in section 11.

6.1.4 Grubb's test

For environmental data, this trivial test was recommended by an American statistical organization. Such data are often log-normally distributed (just as financial data). Similar as in Dixon's test, the ordered data are considered and a τ-statistic for the smallest and largest value is computed:

$$\tau_{high} = \frac{Y_n - \bar{Y}}{S}$$

$$\tau_{low} = \frac{\bar{Y} - Y_1}{S}$$

where S is the sample standard deviation. Other tests are built on the idea of considering the effect of leaving one observation out.

6.2 Outliers in univariate time series

A lot of work, more or less useful, on the subject outliers in independent data has been done. For dependent data however, less has been done. The fundamental article on this subject was written by Fox 1972. This was practically the first article considering outliers in dependent data. He proposed a classification of time series outliers to type I and type II outliers, based on an autoregressive model. These two types have later been renamed as additive outliers (AO) and innovational outliers (IO). These two types are the most common appearing in time series analysis. Two further types are also often mentioned are level shifts (LS) and temporary changes (TC). The definitions of these four types, and their modelling in an ARMA(p,q) time series, are described next.

Denote an uncontaminated (underlying) ARMA process

$$\Phi(B)X_t = \theta(B)Z_t$$

where $\Phi(B)$ and $\theta(B)$ are lag polynomials with all roots outside the unit circle, B is the time backshift operator and $\{Z_t\}$ is an IID $N(0,1)$ sequence. See Grandell (1999) or Brockwell and Davis (1996) for the basics and details of ARMA processes.

To describe a time series subject to the influence of a nonrepetitive event, the following model is considered:

$$X_t^* = X_t + \omega\alpha(B)1_{t=\tau}$$

Here X_t^* is the observed time series, τ is the (unknown) time for the outlier, $1_{t=\tau}$ is an indicator function for the occurrence of the outlier impact, ω is the size of the outlier and the series $\alpha(B)$ is the *dynamic pattern* of the outlier.

When this model for the outlier contaminated series is defined, a detection procedure can be developed. At this stage, the presence of exactly one outlier is assumed.

6.2.1 Additive outliers

An additive outlier can be viewed as an observation which is the genuine data point plus or minus some value. This latter value can be nonzero because of a recording error or by misinterpreting sudden news flashes, which in turn can cause returns on stock markets to take unexpectedly large absolute values.

In other words, in the case of an AO, the data point is aberrant because of a factor outside the intrinsic economic environment that generates the time series data. Given a time series x_t, it is clear that AO's cannot be predicted using the historical information $\mathcal{F}_{t-1}$. The AO only affects one value, and after the disturbance the series returns back to its path as if nothing had happened. An AO can therefore be described by

$$X_t^* = X_t + \omega 1_{t=\tau}$$

and we have

$$\alpha(B) = 1$$

6.2.2 Innovative outliers

An IO is an outlier occurring in the noise process. For the ARMA(p,q) process it is modelled as having a lag structure analogous to the ARMA form of the noise. The model is

$$X_t^* = X_t + \omega\frac{\theta(B)}{\Phi(B)}1_{t=\tau}$$

or

$$\Phi(B)X_t^* = \theta(B)(Z_t + \omega 1_{t=\tau})$$

Hence

$$\alpha(B) = \frac{\theta(B)}{\Phi(B)}$$

6.2.3 Level shifts

Consider an AR(p) model

$$\phi_p(B)X_t = \phi + Z_t$$

The extension to include level shifts is by an extra term to the mean:

$$\phi_p(B)X_t = \phi + \omega 1_{t \geq \tau} + Z_t$$

This means that the mean shifts at $t = \tau$ from

$$\frac{\phi}{1 - \phi_1 - \phi_2 - ... - \phi_p}$$

to

$$\frac{\phi + \omega}{1 - \phi_1 - \phi_2 - ... - \phi_p}$$

Hence

$$\alpha(B) = \frac{1}{1 - B}$$

6.2.4 Temporary changes

A generalization of the LS, that has a permanent effect from time τ, is the temporary change (TC). This outlier works similar to the LS, but with an exponentially decreasing effect. This is modelled by multiplying a factor $\delta \in [0, 1]$ and we have

$$\phi_p(B)X_t = \phi + \omega 1_{t \geq \tau} Z_t$$

and

$$\alpha(B) = \frac{1}{1 - \delta B}$$

Notice, that for $\delta = 1$, the TC becomes a LS, and for $\delta = 0$ it becomes an AO. In some models, the TC may be close to an IO. Chen and Liu (1993) recommend a value of $\delta \approx 0.7$. Also notice that only the IO is dependent of the model.

These definitions have been adopted in most studies of outliers in time series. There are also other types defined in the literature though, such as Variance Changes (VC) and Reallocation outliers (RO). It is, of course, always up to the situation to decide if a type makes sense or not. In this work, only the described four types were considered.

In each case, the "outlier" is merely a model change at an unknown time point, and probably not everyone would call this an "outlier", but recall that an outlier is always an outlier with respect to a model. Notice, that it does not tell whether or not the outlier is an outlier with respect to the reality, i. e. if the value is "wrong" or if the model is "wrong".

Also notice that the effects of differentiation must be remembered, since we usually work with the differentiated log-prices. For example, a LS in the prices P_t becomes an AO in the returns X_t.

7 Univariate outlier detection

The standard way of identifying a univariate outlier, in dependent or independent data, is to simply define it as a value exceeding a chosen quantile of an assumed or estimated distribution. For example, if it was true that $X \in N(\mu, \sigma^2)$, then the standardized variable

$$Z = \frac{X - \mu}{\sigma} \in [-\lambda_\alpha, \lambda_\alpha]$$

with probability 1-α, where $\lambda_\alpha = \phi^{-1}(1 - \frac{\alpha}{2})$
($\phi(x)$ is the CDF for the standard normal distribution).

As mentioned, the problem with this score test is that σ is usually estimated as $\sqrt{\frac{1}{n-1}\Sigma(x_i - \bar{x})^2}$. A big outlier X_ω would, due to the squares, overestimate σ, implying Z to become too small. Probably X_ω will be found, but if there are other smaller outliers as well, they might not be found. This is a typical example of the masking effect. To cope with this, robust estimates can be used.

7.1 Moving median

A simple but effective and robust outlier filter, often used in signal preprocessing, is a moving median filter. Such a filter was constructed and used to "pre-filter" the price series $\{P_t\}$ *before* differentiating the logarithms to get the return series $\{X_t\}$.

A symmetrical moving window of size $2n + 1$ is run over the series P_t. For every point (except the boundaries), the mean or "trend" $m_t$is estimated by the median over the window. If there are not more than n outliers within a window, this gives a robust trend estimate.

Next, the trend deviation series $d_t = |P_t - m_t|$ can be used as outlier detector. This series is mean-reverting and with the same volatility clustering structure as the return series. Hence, d_t can be normalized with its unconditional standard deviation or some other scatter measure, (measured in some robust way). It is more robust to use d_t than the "noise" or contaminated series X_t for normalization.

It turned out that the very robust MAD-estimation (described in section 11) of $\sigma_d = std(d_t)$ gave in general too bad estimates. An empirical but in general well working scatter estimate was therefore constructed: Trimming on 1 % (discarding the 1 % absolute largest values) and measuring in 1-norm (absolute values) instead of 2-norm (squares), with assumed zero mean, gave better (= more consistent) results. For every found outlier, σ_d was then re-estimated. Compensating for the downbias caused by the trimming (under the null hypothesis of no outliers) with an empirically found linear compensation function (found by linear regression) of the number of values disregarded further improved the results. To reduce the effects of non-stationarity, different parts of the series were analyzed separately, and compared with each other as well as with the filtration of the whole series.

By "consistent", we mean a measure that has a consequent behaviour for a wide range of series, with or without outliers. It is *not* important to estimate the variance or the standard deviation, any consistent scattering measure will work.

7.2 The Joint Estimation method

A fundamental problem is that outliers are outliers w.r.t a model, and that, usually, the model parameters (such as the variance) must be estimated and may be biased by existing outliers. As explained, for the purpose of outlier detection, this has the consequences of masking and swamping, and one must choose either the approach of removing the outliers and then estimate the model or using robust estimates.

A more sophisticated method of identifying outliers with respect to a univariate time series model, is the so called Joint Estimation (JE) method by Cheng and Liu (1993). It is an iterative method for finding AO's, IO's, LS's and TC's in ARMA models, estimating the model parameters and the outlier effects *jointly*, using a linear regression method. The contaminated ARMA model and the outlier definitions are denoted as in section 6.2.

7.2.1 Residuals computation

The intuition of the method is to filter the observed process with the underlying process to get the residuals. Define the $\pi(B)$ polynomial as

$$\pi(B) = \frac{\{\Phi(B)\}}{\{\theta(B)\}} = 1 - \pi_1 B - \pi_2 B - ...$$

These weights will become smaller and smaller and eventually essentially zero, since all roots of $\phi(B)$ are outside the unit circle. The estimated residuals can be expressed as

$$e_t = \pi(B) Y_t^*, t = 1, 2, ...$$

For the 4 defined outlier types, we have

$$\begin{aligned} AO: &\quad e_t = \omega\pi(B)1_{t=\tau} + Z_t \\ IO: &\quad e_t = \omega 1_{t=\tau} + Z_t \\ LS: &\quad e_t = \omega\tfrac{\pi(B)}{1-B}1_{t=\tau} + Z_t \\ TC: &\quad e_t = \omega\tfrac{\pi(B)}{1-\delta B}1_{t=\tau} + Z_t \end{aligned}$$

We define for each type an "indicator series" $\{y_t, t = 1, ..., n\}$ such that

$$y_t = \begin{cases} 0, & 1 \leq t < \tau \\ 1, & t = \tau \end{cases}$$

for all four types.

For $t = \tau + k, k \geq 1$, define

$$y_{\tau+k} = \begin{cases} 0, & \text{for IO} \\ -\pi_k, & \text{for AO} \\ 1 - \sum_{i=1}^{k} \pi_i & \text{for TC} \\ \delta^k - \sum_{i=1}^{k-1} \delta^{k-i}\pi_i - \pi_k, & \text{for LS} \end{cases}$$

7.2.2 Outlier effect estimation

Now the observed residuals can be expressed as a simple linear regression on y_t: (we assume mean zero and that the noise $Z_t \in$ IID $N(0, \sigma_Z^2)$)

$$e_t = \omega y_t + Z_t$$

(Notice that the time τ is in reality unknown, but here modelled as known.)

Now we can use standard methods and *estimate* the outlier $\omega(\tau)$ with the ordinary least squares estimate

$$\omega^*(\tau) = \frac{(e_t, y_t)}{(y_t, y_t)}$$

where $(\cdot, \cdot)$ denotes inner product. For the basics and details of linear regression, see for instance Sundberg (1997).

For AO,LS and TC we then have

$$\omega^*_{AO}(\tau) = \omega^*_{LS}(\tau) = \omega^*_{TC}(\tau) = \frac{\sum_{t=1}^n e_t y_t}{\sum_{t=1}^n y_t^2}$$

and for IO

$$\omega^*_{IO}(\tau) = e_\tau$$

Notice that for the last observation (if $\tau = n$)

$$\omega^*_{AO}(\tau) = \omega^*_{LS}(\tau) = \omega^*_{TC}(\tau) = \omega^*_{IO}(\tau) = e_n$$

Hence, it is impossible to empirically distinguish the type of an outlier at the very end of the series, which is also intuitive, since we are looking for outlier *effects*, characteristic patterns.

Now for *detecting* the outliers, the idea is to standardize the estimates of the outlier effects with their standard deviations, and then to maximize this test statistics over all times. From standard theory of linear regressions, see for instance Sundberg (1997), we know that the distribution of $\omega^*(\tau)$ is

$$\omega^*(\tau) \in N\Big(\omega(\tau), \frac{\sigma_Z^2}{\sum_{t=1}^n y_t^2}\Big)$$

(recall that we are assuming mean zero and Gaussian IID noise)

Now, we need to estimate the residual standard deviation σ_Z. Some robust method for estimating σ_Z may be used. Cheng and Liu (1993) propose to use interquantile trimming or the median absolute deviation, MAD. Franses and van Dijk (2000), however, conclude that the standard sample deviation should be used.

Whatever estimate is used for σ_Z, we obtain for AO,LS and TC

$$\omega^*(\tau)/\sigma_{\omega^*(\tau)} = \frac{\omega^*(\tau)}{\sigma^*_Z}\sqrt{\sum_{t=1}^n Z_t^2}$$

and for IO

$$\omega^*(\tau)/\sigma_{\omega^*(\tau)} = \frac{\omega^*(\tau)}{\sigma^*_Z}$$

7.2.3 Detection criterion

We now take the maximum of $\omega^*(\tau)/\sigma_{\omega^*(\tau)}$ over all times as a test statistic for outliers:

$$t_{max} = \max_{1 \leq \tau \leq n} |\omega^*(\tau)|/\sigma_{\omega^*(\tau)}$$

For a given location, and under our model assumptions, t_{max} is approximately normal distributed. However, the sampling behaviour is also associated with the sample size n, the type of outlier, the pattern of weights π for the fitted model and the estimate of σ_a.

Cheng and Liu (1993) perform some Monte Carlo simulations for all the different cases, and estimate the standard quantiles. In summary, it can be noticed that the quantiles are, usually, increasing functions of n, and that the IO test statistic is quite homogeneous w.r.t different models. The latter fact is not surprising, since the IO test statistic is simply the maximum of the standardized residuals. The user should consider which model, sample size etc that seems most appropriate for the current situation. Details and tables of critical values can be found in Cheng and Liu (1993). The ambitious user will of course perform his own Monte Carlo simulations for t_{max}.

Given an empirical distribution of t_{max} and a critical value C, the method can be used in an iterative manner to find multiple outliers. Basically:

1. Estimate a model (ML)
2. Compute t_{max}. If $t_{max} < C$, no outliers are found.
3. If $t_{max} > C$, remove the outlier. Re-estimate the model and go to step 1.

7.3 The Joint Estimation method adapted to GARCH

The JE method can be adapted to the GARCH(1,1) model. This has been done by Franses and van Dijk (2000), see this paper for more details. In this work, this method was implemented and tested on a variety of financial series.

In summary, it could be said that the method works remarkably well, but only under "good" conditions, that is, a good estimated model. Unfortunately, we have seen that it is troublesome to estimate a GARCH(1,1) for a sample size of $n \approx 250$. (see appendix A). Another thing is the fundamental outlier problem; The already troublesome estimation may be (heavily) biased, especially the term α_0, by an outlier.

7.3.1 A GARCH-Jump model

First, we assume the existence of exactly one outlier of magnitude ω by time τ. Next, we *extend* our observed GARCH(1,1) model to include this jump by an extra term:

$$X_t^* = \sigma_t Z_t + 1_{t=\tau}\omega$$

The assumption that Z is standard normally distributed is not crucial in the sense that our outlier detection method also is applicable for other distributions. Notice that the contamination does not affect the conditional volatility by time τ.

Next, we rewrite the process. Recursively, σ_t^2 can be rewritten as

$$\sum_{i=1}^{t-1} \beta^{i-1}(\alpha_0 + \alpha_1 (X^*)_{t-i}^2) + \beta^{t-1}\sigma_1$$

We only assume that σ_1 is known. Now, for $t > \tau$, the estimated conditional variance will be affected by the outlier with an extra exponentially decreasing term, that simplifies to

$$\beta^{j-1}\alpha_1(\omega^2 + 2\omega X_\tau^*)$$

for $t = \tau + j, j = 1, 2, ...$ Defining

$$v_t^* = (X^*)_t^2 - \sigma_t^2$$

and

$$v_t = X_t^2 - \sigma_t^2$$

where $\{X_t\}$ is the true, uncontaminated return series. (This transforms the GARCH-outlier model to an ARMA(1,1) model). It follows that

$$v_t^* = \begin{cases} v_t, & t < \tau \\ v_\tau + \omega^2 + 2\omega X_\tau, & t = \tau \\ v_{\tau+j} - \beta_1^{j-1}\alpha_1(\omega^2 + 2\omega X_\tau), & t = \tau + j, j = 1, 2, ... \end{cases}$$

7.3.2 Outlier effect estimation

Define

$$y_t = \begin{cases} 0, & t < \tau \\ 1, & t = \tau \\ -\beta_1^{j-1}\alpha_1, & t = \tau + j, j = 1, 2, ... \end{cases}$$

to get a one-dimensional linear regression

$$v_t^* = \gamma y_t + v_t, t = 1, 2, ...$$

where

$$\gamma = f(\omega) = 2\omega X_\tau^* - \omega^2$$

Using ordinary least squares (OLS) estimation for the parameter γ and solving for ω gives the contamination magnitude estimate

$$\omega(\tau) = X_\tau - sgn(X_\tau)\sqrt{(X^*)_\tau^2 - \gamma(\tau)}$$

where γ is the OLS estimate

$$\frac{(y, v)^2}{(y, y)^2}$$

Recall that this is all conditional upon knowing the time point τ. The standard deviation σ_v of the "noise" v_t in the regression is estimated with the usual sample standard deviation.

Finally, we standardize $\omega(\tau)$ to obtain the test statistic

$$t_\omega = \frac{\omega_\tau}{\sigma_v(\frac{\partial f}{\partial \omega}(\sum_{t=\tau}^{n} y_t^2)\frac{\partial f}{\partial \omega})^{-1/2}}$$

where $\frac{\partial f}{\partial \omega}$ simplifies to

$$\frac{\partial f}{\partial \omega} = 2X_\tau^*$$

7.3.3 Critical values

The basic idea of the method is that a conditional normal distribution for the returns can not capture all excess kurtosis. When the properties of return series are examined more closely, it turns out that the excess kurtosis may be caused almost entirely by only a few extreme observations. Therefore, instead of using a fat-tailed distribution, such as the student t-distribution, we extend the normal model with jump terms for those shocks. For normally distributed innovations, it can be shown that the kurtosis κ_X is

$$\kappa_X = \frac{3[1-(\alpha_1+\beta_1)^2]}{1-(\alpha_1+\beta_1)^2-2\alpha_1^2}$$

Note that $\kappa_X = 0$ if $\alpha_1 + \beta_1 = 1$, as for the non-stationary RiskMetrics EWMA model (IGARCH(0.06,0.94)).

Next, the distribution of t_{max} is estimated. It can not be evaluated analytically, a Monte-Carlo method must be used. t_{max} is a function of α_1 and β_1 and, unfortunately, the sample size n. κ_X is always greater than 3, and typically smaller than 4. For this region, it can be seen that there is an almost linear relationship between the quantiles of t_{max} and κ_X. It turns out that the multiple regression

$$PC(\alpha, n) = b_0 + b_1\alpha_1 + b_2\beta_1 + b_3\kappa_X + \eta_t$$

provides an almost perfect fit. Thus, we have an explicit function for the critical value.

In the implementation, values of b_0, b_1, b_2, b_3 were tabulated for different α_1, β_1 and n. Linear interpolation was used in-between. The quantiles of t_{max} were taken from 500 realizations, for 16 combinations of α_1, β_1 and n. Not only the value of t_{max} changes with n, so does the computational time: For $n = 1000$, the 16 combinations took about 500 hours to compute, using the MATLAB-compatible GNU-program Octave on a 500 MHz Pentium III

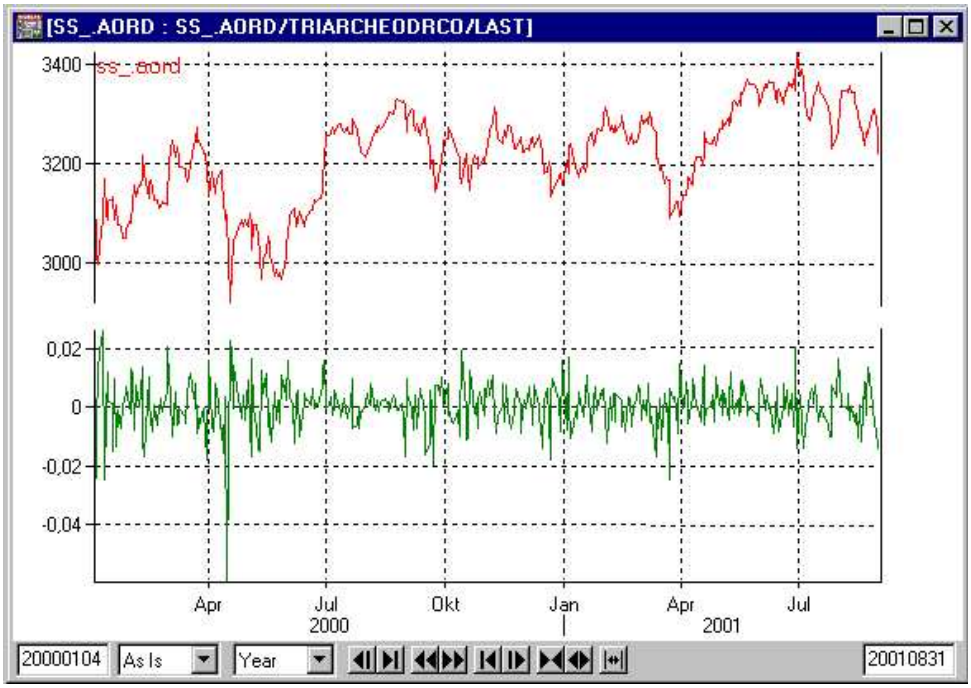

Figure 8: The AORD index. **Upper graph:** The index series. **Lower graph:** The log-differentiated series.

7.3.4 Example

The method was tested on a variety of financial indices. The residual standard deviation σ_a was estimated with the sample standard deviation, the GARCH parameters (α_1, β_1) were ML-estimated as described in appendix A, and α_0 was estimated by $\frac{\sigma^2}{1-\alpha_1-\beta_1}$, where σ^2 is the unconditional sample variance $Var(X)$.

In the Australian All Ordinaries Index (AORD), a capitalization-weighted index of common stocks listed on the Australian Stock Exchange, an outlier was found, see figure 8 and 9.

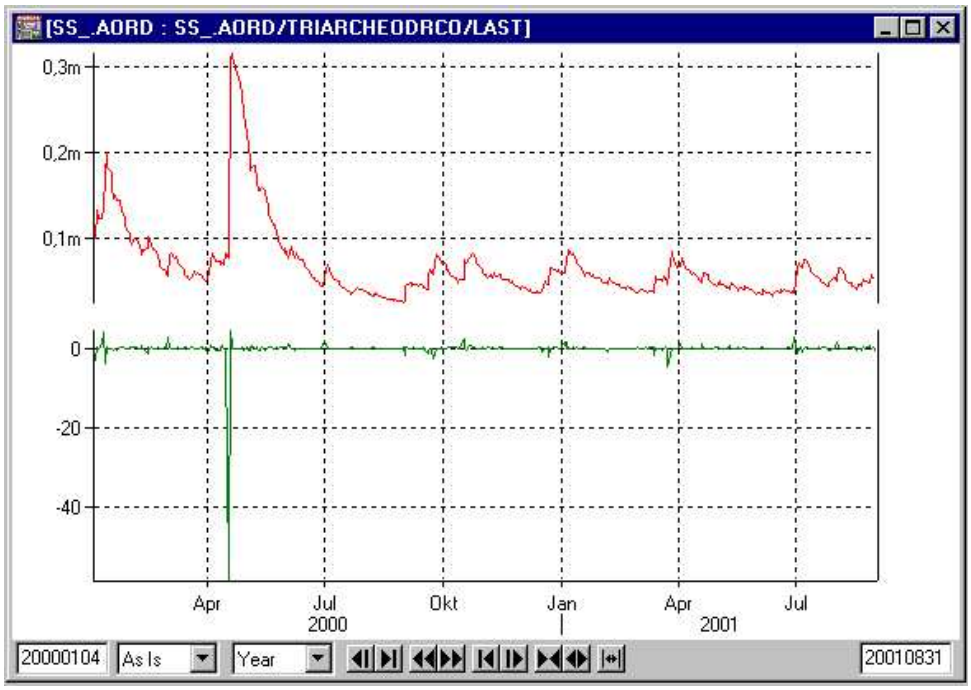

Figure 9: The AORD index. **Upper graph:** The conditional volatility σ_t for the estimated GARCH(1,1) model. **Lower graph:** The JE-GARCH outlier detection test statistic $t_{max}(t)$ (plotted without taking the absolute value).

8 Marginal outliers

Since marginal outliers are defined through a cross-section comparison, we need an appropriate joint distribution for $X(t)$, the time series at a fixed time:

8.1 Elliptical symmetric distributions

Most of the work in multivariate statistics has been done on the multivariate normal distribution (MND). A reason for this is its nature; The central limit theorem ensures that all distributions will approach the Gaussian as the number of observations tends to infinity. Another reason is its pleasant analytical properties. Of special interest to us is the fact that all marginal distributions, also conditional, are also Gaussian.

The Gaussian distribution and the t-distribution both belong to a class of distributions called *elliptical distributions:*

Definition 8.1 (Elliptical symmetric distributions)
The stochastic vector $X \in \Re^m$ is said to be elliptically distributed with parameters $\mu(m \times 1)$ and $\mathbf{V}(m \times m)$ if the distribution is of the form

$$\mathbf{f_X}(\mathbf{x}) = c_m(det\mathbf{V})^{-1/2}h[(\mathbf{x}-\mu)^T\mathbf{V}^{-1}(\mathbf{x}-\mu)]$$

where c_m is a constant not depending on $\mathbf{x}$, $h(\cdot)$ is an arbitrary function and $\mathbf{V}$ is positive definite

When $\mathbf{X}$ has an elliptical distribution, this is written $X \in E_m(\mu, \mathbf{V})$ and

Property 8.1

$$E(\mathbf{X}) = \mu$$

$$Cov(\mathbf{X}) = \alpha\mathbf{V}$$

for some constant α given by the characteristic function.

Notice that when writing $X \in E_m(\mu, \mathbf{V})$, this does not mean any *particular* elliptic distribution; For every choice of $(\mu, \mathbf{V})$, $E_m(\mu, \mathbf{V})$ describes a whole family of distributions.

The Gaussian distribution is elliptic with $\alpha = 1$ and $h(x) = e^x$. In fact, many of the tractable properties of the Gaussian distribution also holds for elliptical distributions. Of special interest to us is the following:

Property 8.2 (Partition property)
If $\mathbf{X}, \mu$ and $\mathbf{V}$ are partitioned as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{pmatrix}$$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\mathbf{V} = \begin{pmatrix} \mathbf{V_{11}} & \mathbf{V_{12}} \\ \mathbf{V_{21}} & \mathbf{V_{22}} \end{pmatrix}$$

with $\mathbf{X_1}$ *and* $\mu_1 k \times 1$ *and* $\mathbf{V} k \times k$*, then the marginal distribution* $\mathbf{X_1} \in E_k(\mu_1, \mathbf{V_1 1})$ *and, if they exist,*

$$E(\mathbf{X_1}|\mathbf{X_2}) = \mu_1 + \mathbf{V_{12}}\mathbf{V_{22}^{-1}}(\mathbf{X_2} - \mu_2))$$
$$Cov(\mathbf{X_1}|\mathbf{X_2}) = \xi(\mathbf{X_2})(\mathbf{V_{11}} - \mathbf{V_{12}}\mathbf{V_{22}^{-1}}\mathbf{V}_{21})$$

for some function $\xi(\cdot)$*. The distribution of* $\mathbf{X_1}|\mathbf{X_2})$ *is k-variate elliptic.*

For the Gaussian distribution, $\xi(\mathbf{X_2}) = 1$.

8.2 The multivariate t-distribution

As explained, a t-distribution with ν degrees of freedom is usually a better model than the normal, where usually $\nu in(3, 5)$. The univariate t-distribution includes the Gaussian distribution as a special case, as $\nu \to \infty$. There are several possible generalizations to the multivariate case, and the most common is

Definition 8.2 (The multivariate t-distribution)
The stochastic vector $X \in \Re^m$ *is said to be multivariate t-distributed with parameters* $\mu(m \times 1)$, $\mathbf{V}(m \times m)$ *and* ν *if its density is of the form*

$$\mathbf{f_X}(\mathbf{x}) = \frac{\Gamma((nu+m)/2)}{\Gamma(\nu(2)(\phi\nu)^{m/2}}(det\mathbf{V})^{-1/2}[1 + \frac{1}{\nu}(\mathbf{x}-\mu)^T\mathbf{V}^{-1}(\mathbf{x}-\mu)]^{-(\nu+m)/2}$$

This expression is valid for all $\nu > 0$. We denote a multivariate t-distributed variable $\mathbf{X} \in T_m(\mu, \mathbf{V}, \nu)$ and we have

Property 8.3

$$E(\mathbf{X}) = \mu, \qquad \nu \geq 2$$

$$Cov(\mathbf{X}) = \tfrac{\nu}{\nu-2}\mathbf{V}, \quad \nu \geq 3$$

What we are interested in is the conditional marginal distribution. For the partitioned vector, the following holds:

Property 8.4 (Partition property)

For $\mathbf{X} \in T_m(\mu, \mathbf{V}, \nu)$,

$$\mathbf{X} = \begin{pmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{pmatrix}$$

the marginal distribution of $\mathbf{X_1}$ *is*

$$\mathbf{X_1} \in T_k(\mu_1, \mathbf{V_{11}}, \nu)$$

and the conditional distribution is k-variate elliptic, but not k-variate t-distributed. Also,

$$E(\mathbf{X_1}|\mathbf{X_2}) = \mu_\mathbf{1} + \mathbf{V_{12}}\mathbf{V_{22}^{-1}}(\mathbf{X_2} - \mu_\mathbf{2})), \nu \geq 2$$

and

$$Cov(\mathbf{X_1}|\mathbf{X_2}) = \frac{(\nu + \mathbf{X_2} - \mu_\mathbf{2})^T V_{22}^{-1}(\nu + \mathbf{X_2} - \mu_\mathbf{2})}{\nu + m - (k+2)}(\mathbf{V_{11}} - \mathbf{V_{12}}\mathbf{V_{22}^{-1}}\mathbf{V}_{21}), \nu \geq 3$$

In the limit $\nu \to \infty$, the t-distribution converges in distribution to the Gaussian distribution. For $\nu = 1$ it coincides with the Cauchy distribution.

8.3 Marginal outlier identification

If we believe that some elliptical distribution is a good model for the data, considering the conditional marginal distribution is then a logical way of identifying a marginal outlier, i.e. an observation being much out of its cross-dimensional expected position. As we have seen, a multivariate t-distribution with an estimated or ad-hoc assumed degree of freedom, is usually - especially for the tail - a better model than the normal.

Unfortunately, the *conditional* marginal distribution is not t-distributed, we only know the first two moments and that it is k-variate elliptic. Since the normal distribution does have normal conditional marginal distribution, we start with a multivariate normal model $X \in N(0, \Sigma)$ anyway.

We fix a time point t (and hence drop the time index t in the notation) and consider the random variable $W_i \in \Re^n$

$$W_i = (X_i|Y_i = y)$$

where

$$Y_i = X \backslash X_i, Y_i \in \Re^{n-1}$$

The partition of Σ for the case $k = n - 1$ is generalized as follows. (We keep the notations, i.e. denote the relevant component i with 1, and the rest $\{1, 2, ..., n\} \backslash i$ with 2.)

- $\Sigma_{12} \in \Re^{n-1} : \Sigma_i \backslash \Sigma_{ii}$
  (i.e. row (column) i with element i removed)
- $\Sigma_{21} = \Sigma_{12}^T$
- $\Sigma_{12} \in \Re^{n-1} \times \Re^{n-1}$
  (i.e. Σ with row i and column i removed)
- $\Sigma_{11} (= Var(X_i))$

Now we put $m = n$ and $k = m-1$ in formula 8.4 and 8.4. Further, we put $\mu_1 = \mu_\mathbf{2} = 0$, i.e. we assume that that we really have a mean-reverting process with zero mean. We obtain

$$W \in N(\mu, \sigma^2)$$

where

$$\mu = \Sigma_{xy}\Sigma_{yy}^{-1}y$$
$$\sigma = \sigma_x^2\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}$$

We standardize W (we now drop the index i)

$$Z = \frac{W-\mu}{\sigma}$$

and can obtain an approximate (we ignore that the variance was estimated) 2-sided confidence interval on level α

$$I = [\mu - \lambda_\alpha\sigma, \mu +_\alpha \sigma]$$

where λ_α is the quantile

$$\Phi^{-1}(1-\frac{\alpha}{2})$$

However, as described, the normal assumption is usually not very good for describing the tails. A $T_n(\nu)$ distribution is usually a better alternative, but the price is that ν has to be estimated.

Since the final result is the variable Z, we inspect Z closer by plotting the empirical quantiles of the SEK/USD FX series against the theoretical quantiles of the normal distribution (figure 10) and the t-distribution (figure 11)with 4 degrees of freedom.

We see that the $t(4)$ distribution is a lot better. If we would choose the critical values λ_α as the quantile $\Phi^{-1}(1-\frac{\alpha}{2})$, "outliers" would be found even on an extremely high level, which is bad. Even though the "significance" level might be fuzzy in a probability context, we need a reasonable scaling between a high and a low level.

The degree of freedom ν for a t-distribution can be estimated by maximum likelihood or by some quantile based method, such as the Hill estimator or some modification of it. In the implemented system,ν was not estimated for every time series, due to computational time but also to *consistency*, and an ad-hoc chosen $t(4)$-distribution was used, which in general gave good results.

Definition 8.3 (Marginal outliers)
We call a value $x_i \in \Re, x_i \subset X \in \Re^n$ a marginal outlier on significance level $\alpha \in (0,1)$ if

$$\Big|\frac{W_i - E(W_i)}{\sqrt{VarW_i}}\Big| \geq F_{t(4)}^{-1}\Big(1-\frac{\alpha}{2}\Big)$$

where

$$W_i = (X_i|Y_i = y), W_i \in \Re$$
$$Y_i = X\backslash X_i, Y_i \in \Re^{n-1}$$

and $F_{t(4)}$ is the scalar CDF of the $t(4)$-distribution.

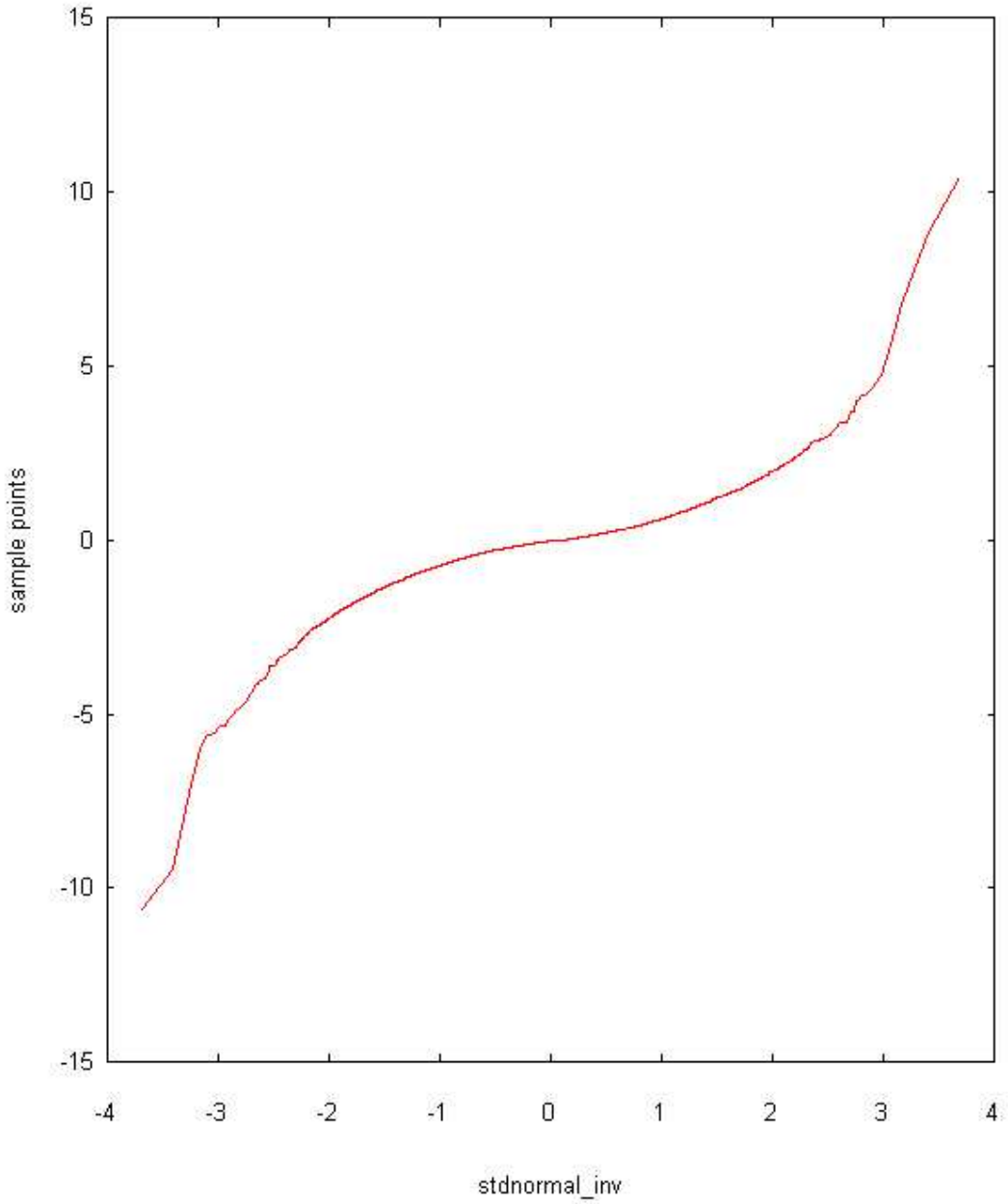

Figure 10: QQ-plot of Z against the normal distribution. The tales are too thin.

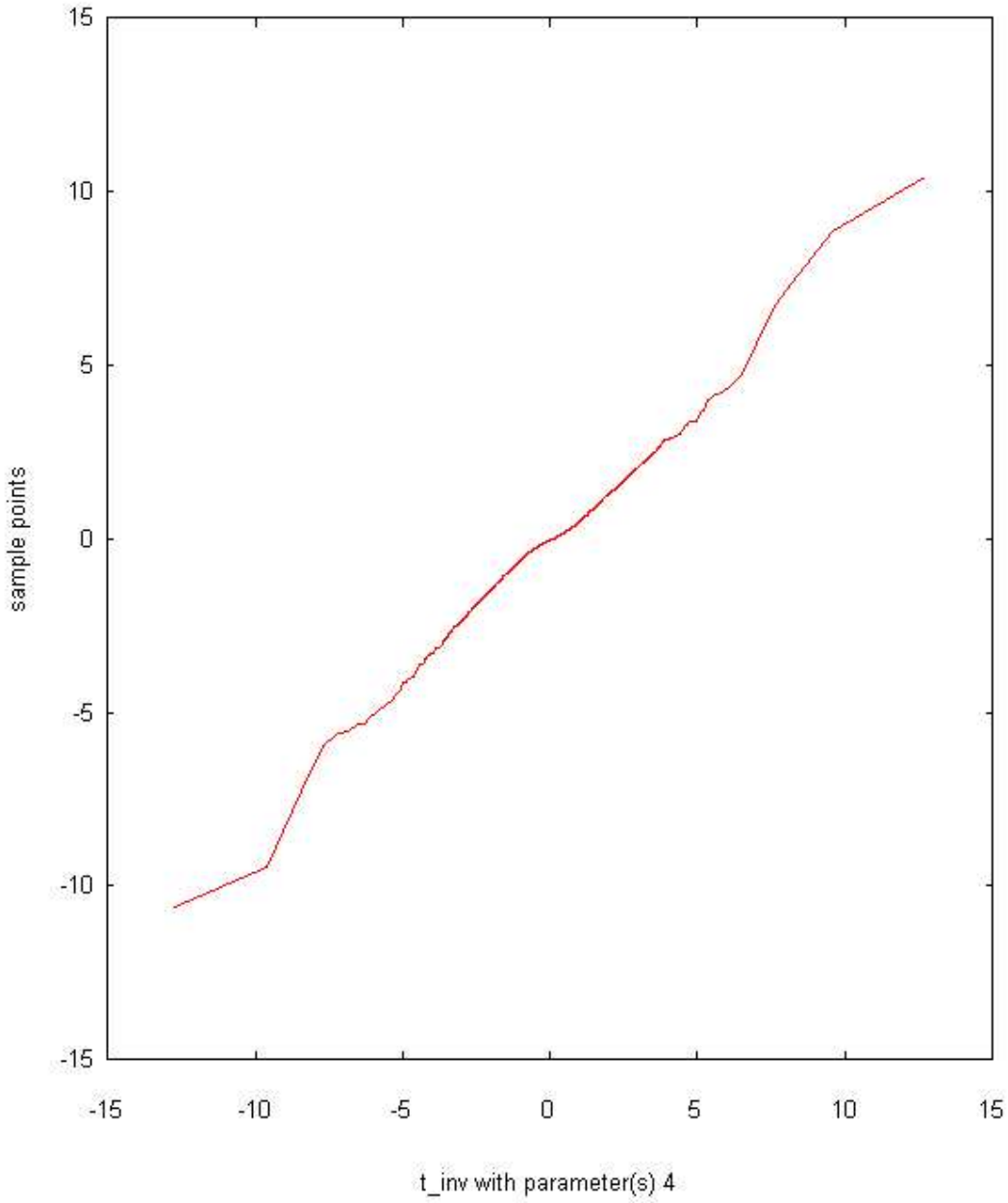

Figure 11: QQ-plot of Z against the $t(4)$ distribution. This model seems appropriate.

8.4 Missing values and low correlations

Since badly correlated instruments are more harmful than helpful for the estimation, we only used a subdimension $0 \leq m \leq n$. For a given component i, the m highest correlated instruments were compared and the corresponding covariance matrix computed. In addition, we required $Cov(X_i, X_j) > \epsilon \qquad \forall j$ for some suitable ϵ and also the existence of X_j. If any of these criteria failed, Z was set to zero. This simply means that

- If there is no value present, it can of course not be an outlier
- If there are no correlated instruments available, no comparison should be done

In practice, the subdimension m was kept low, typically $m \in (3, 5)$ depending on the total dimension n and on the nature of the instruments.

8.5 The EM-algorithm

A similar in spirit, but somewhat more sophisticated estimator in the cross-sectional direction is the Expectation-Maximization (EM) algorithm.

The EM-algorithm is an iterative procedure for computing the maximum likelihood estimator when only a subset of the data is available. The idea is simply to replace the unobserved likelihood with the *expected* likelihood, *conditional* on the observed subset.

Since its formal exposition 1977 the EM-algorithm has been one of the most successful methods of estimation when the data in study is incomplete, but it has also a wide applicability in other areas as well. For example, it can be used as a robust estimator in some model estimations.

Following the notations in Brockwell and Davis (1996), the "complete" data vector $\mathbf{W}$ is made up of observed data $\mathbf{Y}$ and unobserved data $\mathbf{X}$. For example, in a state-space model, a model that consists of two coupled equations, one for the observed observations and one for the underlying unobserved, $\mathbf{X}$ could be the latter.

Each iteration of the EM-algorithm consists of 2 steps, in most literature called the E-step and the M-step:

$$\begin{array}{ll} \text{E-step:} & \text{Calculate } Q(\theta|\theta_i) = E_{\theta_i}[L(\theta, \mathbf{X}, \mathbf{Y})|\mathbf{Y}] \\ \text{M-step:} & \text{Maximize } Q(\theta|\theta_i) \text{ w.r.t } \theta \end{array}$$

where $E_{\theta_i}[\cdot|\mathbf{Y}]$ denotes integration over the conditional density

$$f(\mathbf{x}|\mathbf{y}, \theta_i) = \frac{f(\mathbf{x}, \mathbf{y}, \theta_i)}{f(\mathbf{y}, \theta_i)}$$

and

$$L(\theta, \mathbf{X}, \mathbf{Y}) = lnf(\mathbf{x}, \mathbf{y}, \theta)$$

(It is often practical to maximize the logarithm of the density instead, which gives the same result since $ln(\cdot)$ is monotone increasing).

θ_{i+1} is then set equal to the max of $Q(\theta|\theta_i)$ in the M-step, and so the process iterates.

Now, the power of the algorithm lies in the fact that *when a global maximum exists, the EM-estimate converges to the ML-estimate!* More specific: If the sequence $\{\theta_i\}$ has a limit

$$\lim_{i \to \infty} \theta_i = \hat{\theta}$$

then $\hat{\theta}$ must be a solution of

$$L(\theta, \mathbf{Y}) = 0$$

This follows from the fact that it can be shown that $L(\theta_i, \mathbf{Y})$ is non-decreasing in i

Now, consider the missing value problem. Suppose the complete data set consists of the observations

$$Y_1, Y_2, ..., Y_n$$

of which r are observed and $n - r$ are missing. Similar to the conditional expectation of a scalar marginal distribution described, the case $r = n - 1$ is of special interest, since it can be used as an outlier detector by comparing the realisation with the estimate. However, we consider the general case: Denote the observed data

$$\mathbf{Y} = (Y_{i_1}, Y_{i_2}, ..., Y_{i_r})$$

and the unobserved

$$\mathbf{X} = (X_{j_1}, X_{j_2}, ..., X_{i_{n-r}})$$

Assume now that the (partitioned) random variable

$$\mathbf{W} = (\mathbf{X}', \mathbf{Y}') \in N(\mathbf{0}, \Sigma(\theta))$$

Then the log-likelihood for the complete data is given by

$$L(\theta, \mathbf{W}) = -ln(2\pi)n/2 - 0.5ln|\Sigma| - 0.5\mathbf{W}'\Sigma\mathbf{W}$$

($|\cdot|$ denotes the determinant). Given $\theta = \theta_i$, we want to compute the conditional expectation

$$E_{\theta_i}[L(\theta, \mathbf{W})]$$

As in the previous section, we consider the partition

$$\Sigma = \Sigma(\theta) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

and our familiar "marginal outlier detector"

$$E[\mathbf{X}|\mathbf{Y}] = \{\mu = \mathbf{0}\} = \mathbf{0} + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Y}$$

combined with

$$\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

It can be shown that

$$E_{\theta_i}[(\mathbf{X}', \mathbf{Y}')\Sigma^{-1}(\theta)(\mathbf{X}', \mathbf{Y}')|\mathbf{Y}] = \text{trace}[\Sigma_{11|2}(\theta_i)\Sigma_{11|2}^{-1}(\theta)] + (\hat{\mathbf{X}}', \mathbf{Y}')\Sigma^{-1}(\theta)(\hat{\mathbf{X}}', \mathbf{Y}')$$

and we get the E-step estimate

$$Q(\theta|\theta_i) = L[\theta, (\hat{\mathbf{X}}', \mathbf{Y}')] - 0.5\text{trace}[\Sigma_{11|2}(\theta_i)\Sigma_{11|2}^{-1}(\theta)]$$

We see that the first term on the RHS is the log-likelihood based on the complete data, but with **X** replaced by its best estimate **X'**, namely, the estimate from the previous iteration. If the increments $\theta_{i+1} - \theta_i$ are small, the second term is nearly constant and can be ignored.

As described, the multivariate normal distribution is a reasonable, but not perfect model for financial returns. However, it has been shown that the "pseudo"- or "quasi" ML estimates are asymptotically consistent.

9 Multivariate outliers

Also within the multivariate framework, but now in the time direction instead of the cross-sample direction, are the multivariate outliers. We need a measure of the expected *joint* position, and will define an outlier as a value being jointly much out of its expected position.

In the implemented system, no time dynamics was taken into account, i.e. the time axis only constitutes a set of realizations, with no respect to time dependence. The 4 common types additive outliers (AO), innovative outliers (IO), level shifts (LS) and temporary changes (TC), however, have been generalized recently by Tsay et al (1999) to multivariate ARIMA time series models.

9.1 Multivariate outliers in uncorrelated data

For multivariate outliers, we follow Gather et al (1999) and use the concept of so called α-outliers. For a $N(\mu, \Sigma)$ model, an α-outlier is just defined as an element of

$$out(\alpha, \mu, \Sigma) := \{X \in \Re^n; (x-\mu)^T \Sigma^{-1} (x-\mu) > \chi^2_{1-\alpha}(n)\}$$

which we call the α-outlier region w.r.t $N(\mu, \Sigma)$. For a sample size of N, one can also speak of an α_N-outlier region $out(\alpha_N, \mu, \Sigma)$ according to the condition

$$P(\bigcap_{i=1}^{N} \{x_i \in \Re^n \backslash out(\alpha_N, \mu, \Sigma)\} = 1-\alpha)$$

for $x_i \in N(\mu, \Sigma), i = 1, ..., N$ and $\alpha \in (0, 1)$. For an IID sample this leads to

$$\alpha_N = 1 - (1-\alpha)^{1/N}$$

Notice that the α_N-outlier region depends on the unknown parameters μ and Σ. Therefore, the outlier region is typically unknown and has to be estimated from the data and estimating an α-outlier region is equivalent to identifying all α-outliers in the set.

For this reason, α_N-outliers are now defined as follows:

Definition 9.1 (α_N-outliers)
For a multivariate data set $\{x\} = \{x_1, ..., x_N\}, x_i \in \Re^n$, assume that more than half, say $\frac{N}{2} < m \leq N$ of the observations come IID (!) from an n-variate normal distribution. Then, for $\alpha_N \in (0, 1)$ an α_N-outlier w.r.t $N(\mu, \Sigma)$is defined as an observation $x \in OR(\{x\}, \alpha_N)$ where

$$OR(\{x\}, \alpha_N) := \{z \in \Re^n; (z-v)^T S^{-1} (z-v) \geq c\}$$

where $S = S(\{x\}) \in Re^{n \times n}$ is symmetric and positive definite, $v = v(\{x\}) \in \Re^n$ and $c = c(n, N, \alpha_N), c \in Re, c \geq 0$.

The normalization constant c can be chosen analogously to 9.1 as

$$P(x_i \in) OR(\{x\}, \alpha_N)) = 1 - \alpha$$

Taking $v = \{\bar{x}\}$ and $S = S_N$, the sample covariance matrix, leads to the classical *Mahalanobis distance* (MD) as outlier identifier

$$MD : \Re^N \to \Re$$

$$MD := \sqrt{(x - \mu_x)^T \Sigma^{-1} (x - \mu_x)}$$

For all time series relevant to us, we can assume that $\mu_x = 0$, giving

$$MD = \sqrt{x^T \Sigma^{-1} x}$$

The intuition of the Mahalanobis distance is that it is a measure of distances between expectations in distributions, normalized by the deviation, giving a comparable measure without dimension. It is a better measure than the Euclidean measure $||x|| = \sqrt{x^T x}$, since it also takes the direction into account, not just the distance from the "center". In one dimension, MD boils down to $MD = x/\sigma$.

Consider the case of uncorrelated (in a normal model equivalent to independent) variables. Here $\Sigma = I_N$, the identity matrix, implying $MD = ||x||$. Thus, MD is to be regarded as a generalized, *weighted* Euclidean distance.

In the IID case $MD^2 \in \chi^2(n)$. According to our assumption, and to the fact that financial log-return series (not the absolute values or the squares!) are usually close to white noise, we set

$$c = \chi^2_{1-\alpha}(n)$$

Hence, we have defined a multivariate outlier.

For $m = 4$ in the case of $n = 15$ for a group of fundamentally related (and correlated) instruments, a QQ-plot of the Mahalanobis distance against the $\chi^2(4)$-distribution tells us that the approximated model seems good enough (see figure 12).

Of course, it is important to have good and robust estimates of μ and Σ. μ can be robustly estimated by the median, but we assume $\mu = 0$. Rousseeuw (1985) proposes the so-called Minimum Volume Ellipsoid as robust estimate of σ. If non-robust estimates are used, the masking and swamping effect may be severe. This is especially true in the multivariate case.

9.2 Phase Space outliers

Gather (1999) has a very interesting idea. The univariate series is transformed into a multidimensional sample by a phase-space reconstruction. Then, the described multivariate technique can be used.

The so-called phase space reconstruction is a simple but fundamental tool introduced at the beginning of the 80's to analyze nonlinear deterministic, especially chaotic, systems in theoretical physics.

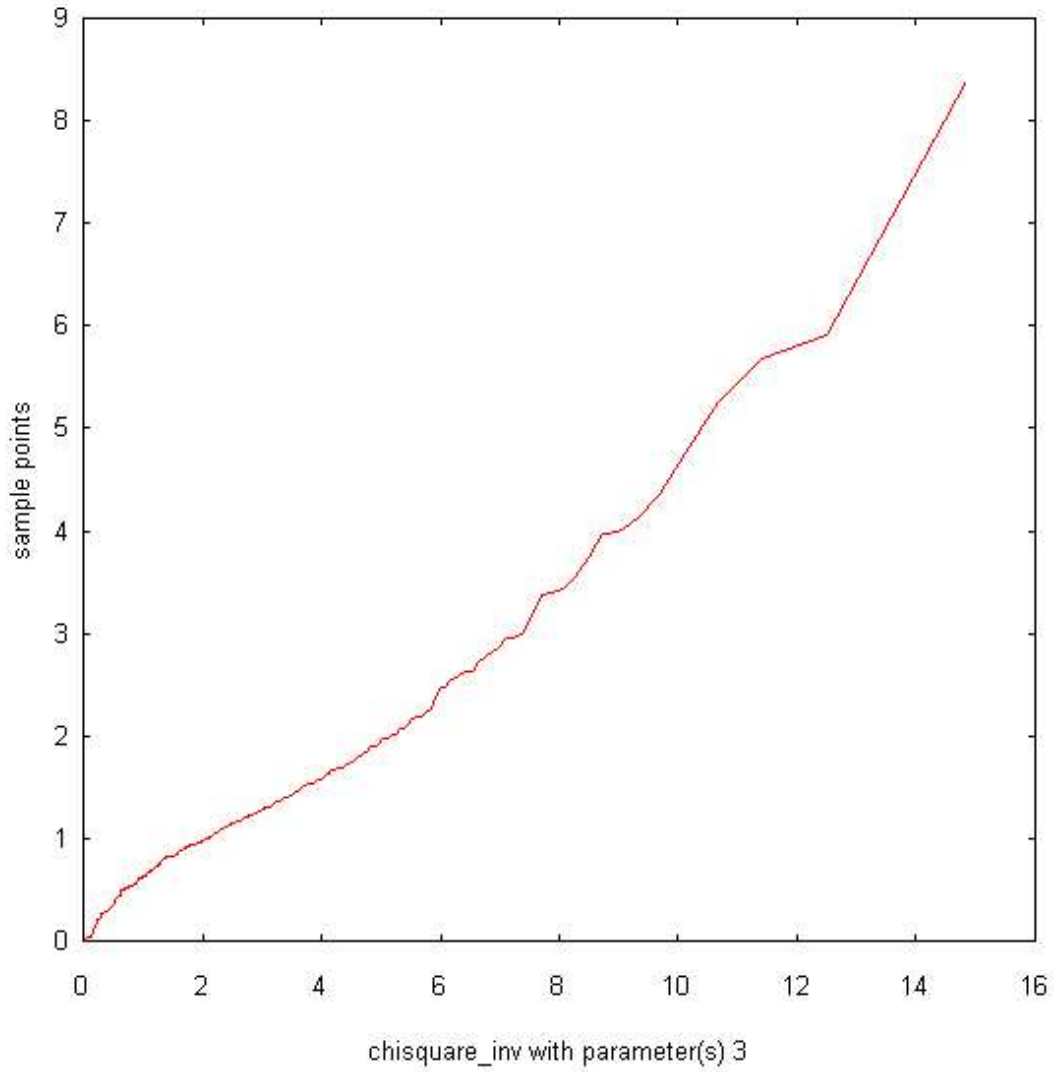

Figure 12: QQ-plot of the Mahalanobis distance against the $\chi^2(4)$-distribution (for $N = 4, NN = 15$).

9.2.1 Phase space reconstruction

Let $\{X_t\}_{t=1,...,N}$ be a univariate time series. Consider the set of m-dimensional vectors, where the components are the time delayed elements of the time series:

$$\bar{X}_t := (X_t, X_{t+T}, X_{t+2T}, ..., X_{t+(m-1)T})^T$$

$$\bar{X}_t \in \Re^m, T, m \in \{t = 1, 2, ..., N - (m-1)T\}$$

with $m, T \ll N$. m is called the embedding dimension and T the time delay. The dynamical information of the time series is thus transformed into a m-dimensional space, called the phase-space. The set

$$\{\bar{X}_t | t = 1, 2, ..., N - (m-1)T\}$$

is called phase space reconstruction or embedding. This phase space vectors have certain geometrical properties connected to the original system. Notice that different choices of T and m lead to different phase space transformations.

In the 2-dimensional (and possibly 3-dimensional) case, the phase space can be visualized. Inspired of Gather (1999), we construct 3 examples, all with $m = 2$ and $T = 1$:

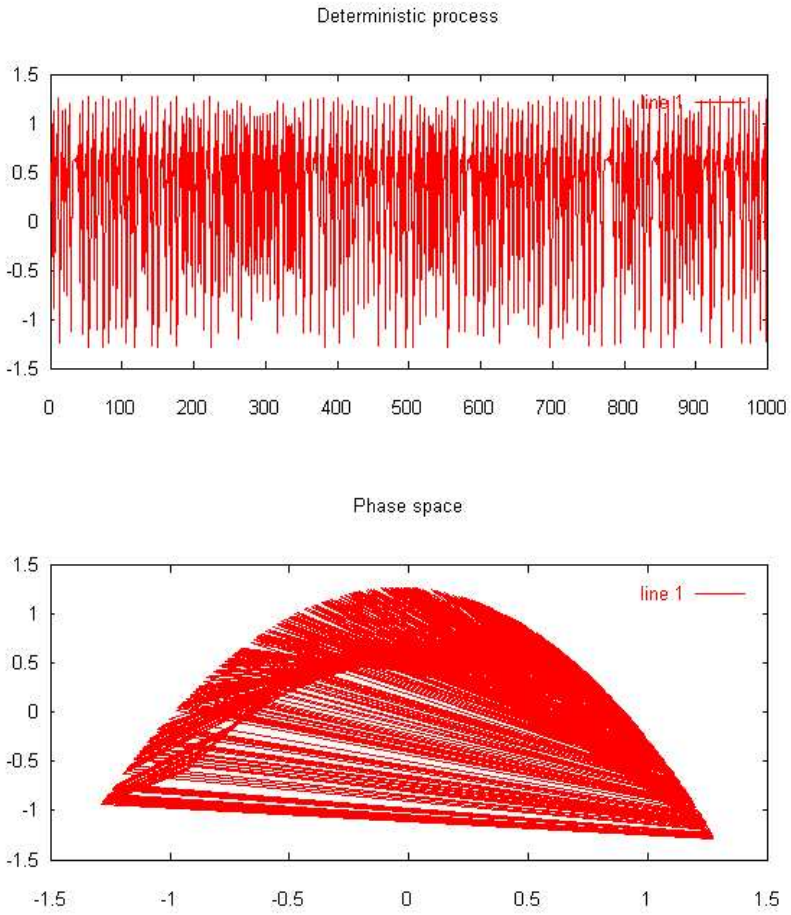

Figure 13: A nonlinear deterministic process. The deterministic structure of the process is easy to recognize

9.2.2 Example 1 - a nonlinear deterministic process

Consider the difference equation

$$X_t = 1 - 1.4X_t^2 + 0.3Y_{t-2}$$

A simulated trajectory and its 2-dimensional phase space with $T = 1$ is shown in figure 13. The deterministic structure of the process is easy to recognize.

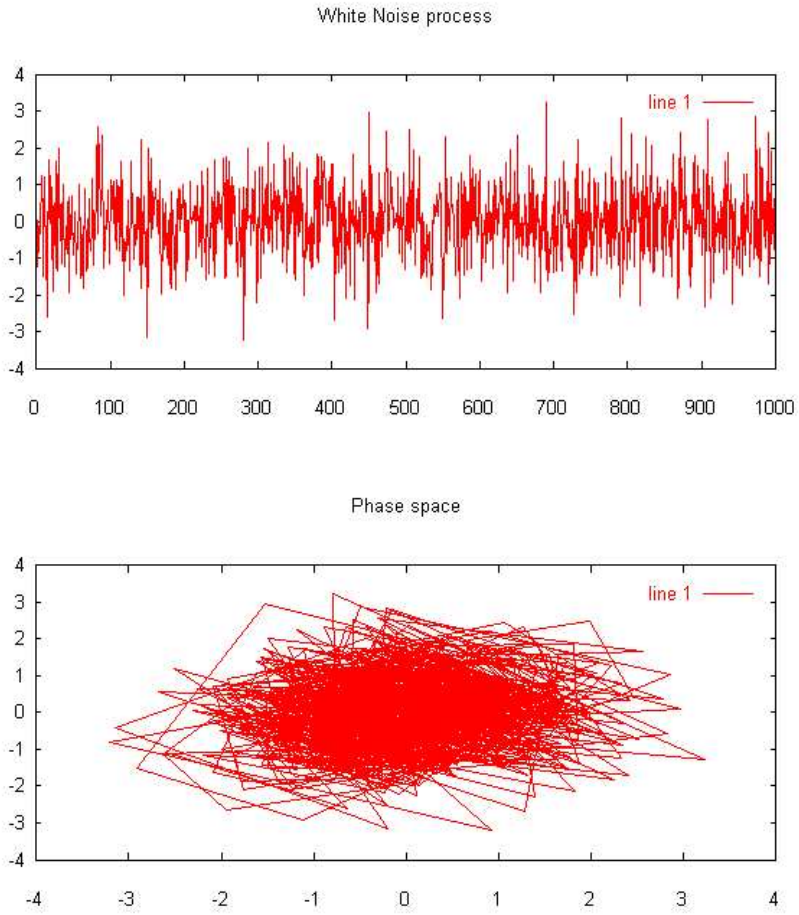

Figure 14: Simulated White Noise process with Gaussian noise. The phase space vectors are contained in a spherical cloud and $N(\mu, I)$-distributed.

9.2.3 Example 2 - a pure white noise process

We now consider a pure WN process with mean zero and ACF $\delta(h)$. (also for $T = 1, m = 2$). Now, no structure in the phase space is visible - see figure 14. The phase space vectors are contained in a spherical cloud.

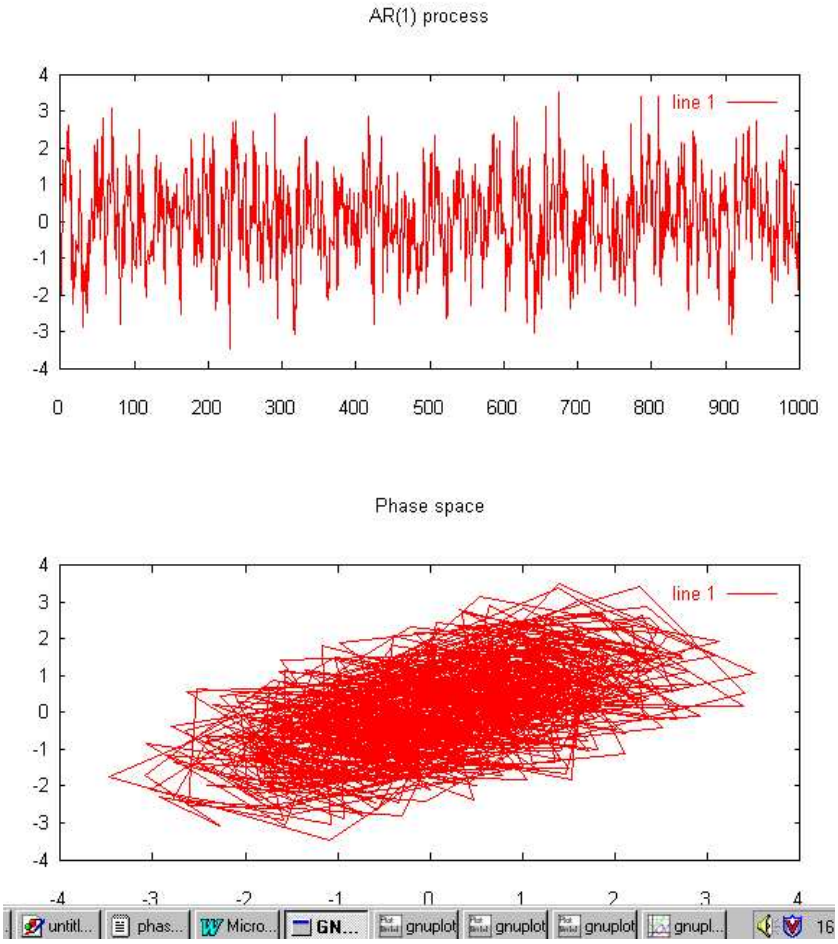

Figure 15: Simulated AR(1) process with Gaussian noise. The phase space vectors are contained in an elliptical cloud and $N(\mu, \Sigma)$-distributed for some $\Sigma \neq I$.

9.2.4 Example 3 - a stationary AR(1) process

For a stationary AR(1) process

$$X_t = 0.5X_{t-1} + Z_t, Z_t \in \text{IID } N(0,1)$$

the phase space vectors are contained in an elliptical cloud, which can be seen in figure 15. Similar as for elliptical multivariate distributions, the ellipse reduces to a circle if there is no dependence.

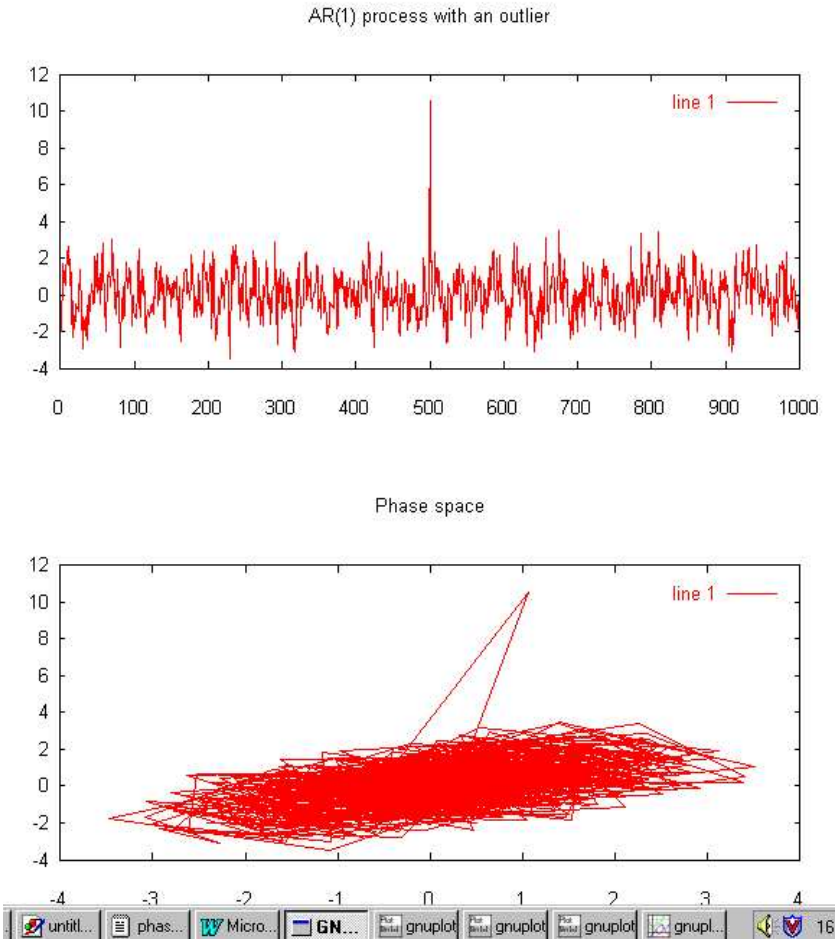

Figure 16: An AR(1) process with an outlier

9.2.5 Example 4 - a stationary AR(1) process with an outlier

For the same AR(1) process, we simulate an outlier by setting a value in the middle to $2max(|X_t|)$.

$$X_t = 0.5X_{t-1} + Z_t, Z_t \in \text{IID } N(0,1)$$

We see that the outlier causes a deviating movement in the phase space.

9.2.6 An AR(p) model

The AR(1) example can be extended to AR(p) models and higher m. The phase space is now an m-dimensional ellipsoid. The smaller the dependence in the series, the more the ellipsoid will resemble a spherical cloud. This is also the case when we do have a process with memory, but choose T such that $ACF(T) = Corr(X_t, X_{t-T}) = 0$. Hence, T should be set to 1.

Consider a stationary Gaussian process

$$\{X_t\}_{t\in\mathbb{Z}}, X_t \in N(\mu, \sigma^2)$$

with absolutely summable ACF $\gamma(h)$

$$\sum_{h=1}^{\infty} \gamma(h) < \infty$$

Gather (1999) recommends to choose embedding dimension m as

$$m = 1 + \max_{\tau}\{\tau; |\rho(\tau)| > 0\}$$

where $\rho(\cdot)$ is the partial autocorrelation function PACF (can be estimated from the series).

By the normality assumption, it now follows that the sample phase space vectors follow a multivariate normal distribution, and the Mahalanobis distance technique for outlier detection can be applied to this series!

Gather (1999) shows that the Mahalanobis distance of the new series (with the corresponding sample covariance matrix for the phase space "series" (vector)), is still asymptotically χ^2-distributed! Further, an improved identification procedure for the dependent data is developed, see Gather (1999) for details.

9.3 Outliers in multivariate time series

Extending the work of Fox (1972), four types of outliers have been proposed for univariate time series analysis as described previously. These are additive outliers (AO), innovative outliers (IO), level shifts (LS) and temporary changes (TC). A generalization to the multivariate case for ARIMA models has been done recently by Tsay et al (1999). They also discuss the interaction of outliers in different components within this framework and finally develop an iterative detection procedure for estimating multivariate outliers. This procedure combines two test-statistics, one Mahalanobis distance based test statistic for the cross-sectional direction, and one simple absolute distance (a quantile) for a marginal component. This method was not implemented in this work, since our simple MD-criterion works well, and since the authors recommend the readers not to use the procedure in the case of conditional heteroskedasticity, since this case is not considered in their procedure and would lead to a large number of "identified" outliers. Here, only the definitions are given, since very little work has been done on the subject. For details of the detection procedure, see Tsay et al (1999).

Following their notations, let $x_t = (x_{1t}, ..., x_{kt})^T$ be a k-dimensional time series that follows a vector autoregressive integrated moving-average (ARIMA) model

$$\Phi(B)x_t = \theta(B)z_t$$

where

$$\Phi(B) = I - \Phi_1 B - ... - \Phi_p B^p$$

and

$$\theta(B) = I - \theta_1 B - ... - \theta_q B^q$$

are $k \times k$ matrix polynomials if finite degrees p and q, B is the time backshift operator, $z_t = (z_{1t}, ..., z_{kt})^T$ is IID $N(0,1)$ We assume that all zeros of the determinants $|\Phi(B)|$ and $|\theta(B)|$ are outside or on the unit circle. In the latter case, we also assume that the series start at a fixed time point t_0 with fixed initial values and initial innovations. Otherwise, the series is asymptotically stationary.

The autoregressive representation is

$$\Pi(B)x_t = z_t$$

where $\Pi(B) = I - \sum_{i=1}^{\infty} \Pi_I B^i = \{\theta(1)\}^{-1}\Phi(B)$

The moving-average representation is

$$x_t = \psi(B)z_t$$

This gives that

$$\Pi(B)\psi(B) = \psi(B)\Pi(B) = I$$

(I is the identity matrix).

Now, denote the observed time series with $x_t^* \in Re^k$ and let $\omega \in \Re^k$ be the size of an outlier on the underlying series $\{x_t\}$. The definitions of AO,IO,LS and TC in univariate time series (section 6.2) can now be directly generalized:

Definition 9.2 (Multivariate AO's, IO's, TC's and LS's)

For the described k-dimensional time series x_t, an outlier $\omega \in \Re^k$ at time τ affects the observed series x_t^* as

$$x_t^* = x_t + \alpha(B)\omega 1_{t=\tau}$$

where

$$\alpha(B) = \begin{cases} \psi(B) & \text{for a multivariate IO} \\ I_k & \text{for a multivariate AO} \\ (1-B)^{-1}I_k & \text{for a multivariate LS} \\ \{D(\delta)\}^{-1} & \text{for a multivariate TC} \end{cases}$$

where $D(\delta)$ is a $k \times k$ matrix with diagonal elements $(1 - \delta B), \delta \in (0,1)$

10 Extreme Value Theory

Extreme value theory (EVT) is a tool for estimating the *tails* of a distribution. One could say that this theory is in contrast to the *mean based* theory, with Brownian motions and stochastic volatility models. One main reference, or perhaps the main reference, on this subject is the book from Embrechts, Klüppelberg and Mikosch (1997). This section first gives a brief overview of the basic theory, then describes an application for finance, by combining EVT with a GARCH volatility model.

10.1 Value-at-Risk and Expected Shortfall

A *very* common risk measure in the financial world is the Value-at-Risk (VaR). This is in fact nothing else but a quantile of the distribution of losses, typically at 95 %. For a distribution function F, VaR is the qth quantile of F:

$$VaR_q := F^{-1}(q)$$

for some q, typically in the domain $q \in (0.95, 1)$.

VaR provides an upper bound for a loss that is only exceeded on a small proportion of occasions, sometimes referred to as a confidence level.

VaR has been criticized as a risk measure, since it is not necessarily subadditive: There are cases where a portfolio can be split into sub-portfolios such that the sum of the VaR for the sub-portfolios is smaller than the VaR for the total portfolio. Further, VaR gives no information of the potential *size* of the loss exceeding VaR.

Therefore, it has been proposed to use the *expected* shortfall (ES) or "tail conditional expectation" instead of VaR. ES is the expected size of a loss exceeding VaR:

$$ES_q := E[X|X > VaR_q]$$

10.2 POT-methods

The most common group of models are the Peaks-Over-Threshold (POT) models. These are models for all large observations that exceed a high threshold. The POT models are generally considered to be the most useful for practical applications. With the POT class of models, one may further distinguish two styles of analysis. These are semi-parametric models, built around the so called Hill estimator (and its relatives), and the fully parametric models, based on the generalized Pareto distribution (GPD). Both classes are theoretically justified and empirically useful when used correctly.

10.2.1 Step 1: The distribution of exceedances

Let $X_1, X_2..., X_n$ be identically distributed (not independent) random variables with unknown distribution function $F(x) = P\{X \leq x\}$. Given a high threshold u_n, we index each observation exceeding u_n and obtain an other sample $\{Y_1, Y_2, ..., Y_{N_u}\}, N_u \leq n$.

Consider the IID case. Each point has the same chance to exceed the threshold with success probability $P\{X_i > u_n\}, i = 1, ..., n$. Hence, the number of exceeding observations is

$$N_{u_n} := \#\{i : X_i > u_n, i = 1, ..., n\} = \sum_{i=1}^{n} 1_{X_i > u_n}$$

N_{u_n} follows a binomial distribution with parameters n and $P(X_i > u_n)$. Now a limit process can be derived by letting the sample size n tend to infinity and, simultaneously, increasing u_n in the correct proportion: If for some $\tau > 0$

$$nP\{X_i > u_n\} \to \tau, n \to \infty$$

Then, by a classical theorem

$$N_{u_n} \to^d Po(\tau)$$

If $X_i, i = 1, ..., n$ come from an absolutely continuous distribution, a suitable series u_n can be found for every $\tau > 0$. (See Embrechts,Klüppelberg and Mikosch (1997), chapter 3).

Indexing all points $\{i : X_i > u_n, i = 1, ..., n\}$ in the interval $[0, n]$, this interval will grow larger and larger whereas the indexed points will become sparser and sparser as u_n increases with n.

10.2.2 Step 2: Generalized Pareto Distribution

We are not only interested in when and how often the exceedences occur, but also in how large the excess $X - u|X > u$ is. Consider the conditional CDF of the excess observations $X - u$

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

or in terms of the underlying F as

$$F_u(x) = \frac{F(x + u) - F(u)}{1 - F(u)}$$

An important result in EVT is that for a very large class of distributions, it can be shown that

$$\lim_{u \to \infty} \sup_{x \geq 0} |F_u(x) - G_{\xi,\beta(u)}| = 0$$

where G is the CDF of the *Generalized Pareto Distribution* (GPD):

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \xi x/\beta)^{-1/\xi}, & \xi \neq 0 \\ 1 - e^{-x/\beta}, & \xi = 0 \end{cases}$$

If $\xi \geq 0$, the support of this distribution is $[0, \infty)$, for $\xi < 0$, the support is a compact interval. This distribution is generalized in the sense that it subsumes other distributions under a common parametric form. ξ is the important shape parameter. The case $\xi = 0$ corresponds to the exponential distribution $exp(\beta)$. The case $\xi < 0$ known as a Pareto II distribution.

If $\xi > 0$, then $G_{\xi,\beta}$ is a reparameterized version of the ordinary Pareto distribution, that has a long history in actuarial mathematics. This is because the GPD is *heavy-tailed* when $\xi > 0$. Whereas a normal distribution has finite moments of all orders, a heavy-tailed distribution does not, as already mentioned, possess a complete set of moments. In the case of a GPD with $\xi > 0$, it is found that $E(X^k) = \infty$ for $k \geq 1/\xi$. When $\xi = 1/2$, the GPD has an infinite second moment, variance. When $\xi = 1/4$, the GPD has an infinite fourth moment. Empirically, as mentioned in the section Empirical properties of financial time series, it is often found that our series have an infinite fourth moment. The normal distribution can not model these phenomena, but the GPD can be used to capture exactly this behaviour.

To summarise, the excess distribution converges to a GPD. We have not defined the "very large class" of distributions for which this is true, but for our purposes it is enough to say that this holds for *all* the common parametric continuous distributions (such as normal, lognormal, χ^2, t,F,gamma etc). Hence, the GPD is *the natural model* for the unknown excess distribution!

10.2.3 Parameter estimation

Finally, the parameters ξ and β must be estimated. A standard method is Maximum Likelihood (ML), where the joint PDF is maximized. However, in practice, this might be numerically troublesome if the data set is small, and one cannot rely on the asymptotic optimality properties of the ML-estimators. Recall, that only the excess fraction of the set is used, and the used data set depends of course on the choice of threshold u. In our case, we typically have a complete data set of about 250 data, which is not enough.

For the choice of threshold u, the mean excess function

$$e(u) = E(X - u | X > u)$$

is a useful tool. It can be estimated by the empirical function

$$e_n(u) = \frac{1}{\#\{i : X_i > u, i = 1, ..., n\}} \sum_{i=1}^{n} (X_i - u)^+$$

For fat tails, $e_n(u)$ tends to infinity. For the GPD with $\xi > 0$ it can be shown that $e_n(u)$ is a linearly increasing function. Hence, a possible choice of u is given by the value for which $e_n(u)$ is approximately linear. In practice, this often gives values such that the GPD is a good model for, very roughly, half the sample.

10.2.4 Step 3: Tail estimation

Now, these results can be used to estimate tails and quantiles. Denote the tail of F by

$$\bar{F} := 1 - F$$

This yields

$$\bar{F}_u(y) = P\{X - u > y | X > u\} = \frac{\bar{F}(u+y)}{\bar{F}(u)}, y \geq 0$$

or

$$\bar{F}(u+y) = \bar{F}_u(y)\bar{F}(u), y \geq 0$$

Hence, an estimator of the tail $\bar{F}_u(y)$ (for values greater than u) can be obtained by estimating the tails $\bar{F}_u(y)$ and $\bar{F}(u)$.

$\bar{F}(u)$ can be estimated by its empirical counterpart

$$\bar{F}^*(u) = \frac{1}{n}\sum_{i=1}^{n} I(X_i > u) = N_u/n$$

and $\bar{F}_u(y)$ by the GPD, where the scaling function $a(u)$ has to be taken into account. This gives

$$\bar{F}_u^*(y) \approx (1 + \xi^* y/\beta^*)^{-1/\xi^*}$$

The two parameters ξ and β have to be estimated, which is (theoretically) best done using ML.

Now, for a given u this gives the tail estimator

$$\bar{F}^*(u+y) = \frac{1}{n}\sum_{i=1}^{n} (1 + \xi^* y/\beta^*)^{-1/\xi^*}$$

For a given $q \in (0,1)$, this function can be inverted to give an estimator of the q-quantile:

$$x_q^* = VaR_q^* = u + \frac{\beta^*}{\xi^*}\Big((\frac{n}{N_u}(1-q))^{-\xi^*} - 1\Big)$$

Hence, for a given probability $q > F(u)$, this is VaR estimate VaR_q^*.

10.2.5 Estimation of ES

Expected Shortfall is related to VaR by

$$ES_q = VaR_q + E(X - VaR_q | X > VaR_q)$$

The second term is simply the mean of the excess distribution $F_{VaR_q}(x)$ over the threshold VaR_q. The model for the excess distribution has a good stable property: If we take any higher threshold, such as VaR_q for $q > F(x)$, then the excess distribution above the higher threshold is also GPD with the same shape parameter, but with another scale parameter!

For our model

$$F_u(x) = G_{\xi,\beta}(x)$$

a consequence is that

$$F_{VaR_q}(x) = G_{\xi,\beta+\xi(VaR_q-u)}(x)$$

Hence, we have a simple explicit model for the excess losses above the VaR! Provided that $0 < \xi < 1$, the mean of $F_{VaR_q}(x)$

$$E(F_{VaR_q}(x)) = \frac{\beta + \xi(VaR_q - u)}{1 - \xi}$$

Next, we find that

$$\frac{ES_q}{VaR_q} = \frac{1}{1-\xi} + \frac{\beta - \xi u}{(1-\xi)VaR_q}$$

or

$$ES_q = \frac{VaR_q}{1-\xi} + \frac{\beta - \xi u}{1-\xi}$$

Substituting VaR_q for the estimate VaR_q^* and (ξ, β) for the ML-estimates (ξ^*, β^*) gives the final estimate

$$ES_q^* = \frac{VaR_q^*}{1-\xi^*} + \frac{\beta^* - \xi^* u}{1-\xi^*}$$

Typical values for real-world financial data could be $\frac{1}{1-\xi^*} \approx 2$ and $\frac{\beta^* - \xi^* u}{1-\xi^*} \approx 4$. Roughly, this means that the ES estimate is obtained from the VaR estimate by doubling it.

To summarise, the POT-method consists of 4 steps:

1. For a sample $\{X_1, ..., X_n\}$, select a high threshold u and obtain the exceedances $\{Y_1, Y_2, ..., Y_{N_u}\}$
2. Fit a GPD to the excesses: Estimate β and ξ (e g with MLE)
3. Estimate the CDF $F(x)$ with for $x > u$ as $F^*(x)(1 - F_n(u))G_{\xi^*,\beta^*}(x) + F_n(u)$, where $F_n(u) = \frac{N_u}{n}$
4. Invert $F^*(x)$ to obtain the quantile estimates.

10.3 Combining a GARCH model with EVT

A fact is that none of the described EVT methods give VaR estimates which reflect the current volatility background, which is a major drawback. To cope with this problem, Frey and McNeil (2000) proposes to use the POT-method on the residuals of a fitted GARCH model, which ideally should be an IID series. This theoretically interesting approach actually combines the usage of historical simulation for the central parts of the distribution, with EVT for the tails, working on GARCH-residuals.

Being at the end of day t, we are interested in the *conditional* return distribution

$$F_{X_{t+1}|\mathcal{F}_t}(x)$$

Especially, we want to derive VaR and ES for $F_{X_{t+1}|\mathcal{F}_t}(x)$ ad denote these risk measures ES_q^t and VaR_q^t. This is in contrast to the "ordinary" *unconditional* distribution $F_{X_{t+1}}(x)$.

For the standard model with zero mean $X_t = \sigma_t Z_t$, we have

$$VaR_q^t = \sigma_{t+1} VaR(Z)_q$$

$$ES_q^t = \sigma_{t+1} ES(Z)_q$$

where $VaR(Z)_q$ and $ES(Z)_q$ denotes the unconditional VaR and ES for the noise Z.

Hence, we are back in the familiar unconditional framework! A model must be assumed for the conditional volatility σ_t, and as described in chapter 1 the GARCH(1,1) or IGARCH ("RiskMetrics EWMA") models are common choices. Which model we choose is unimportant for the EVT analysis, what *is* important is that the residuals Z_t is an approximately IID series.

The authors propose to use the GARCH(1,1) model

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

where $\alpha_0, \alpha_1, \beta_1 > 0$ and $\alpha_1 + \beta_1 < 1$. However, in this work it was usually found that the residuals do *not* constitute an IID series; They certainly reduce the autocorrelation, but it was in practice usually not possible to estimate a model with such a perfect fit. Frey and McNeil, however, claim to "have found no evidence against the IID hypothesis for the residuals".

The authors also compare the GARCH residual approach with an "ad-hoc" model of the residuals (i.e. the "standard" EVT approach) as t-distributed. As mentioned, the value of the shape parameter ξ in the limiting GPD approximation of the excess distribution for a $t(\nu)$ distribution is actually the reciprocal $1/\nu$. Hence, the results for the assumed t-distribution (with ν ML-estimated) are similar to those of the GPD. However, the t-distribution is symmetric, and when the distribution of X is asymmetric, the GPD gives better results. Moreover, the GPD can be theoretically motivated, without any ad-hoc assumptions, and is therefore to be recommended.

10.4 Quantile based methods

Fitting a GPD can be somewhat risky in practice, since it requires a numerical optimization of the likelihood function. For small data sets, this may be troublesome and the asymptotic properties may not longer be true. An alternative is to use a so-called quantile based method, whereas the most common is tool is the Hill estimator.

We assume the tails (only the tails are now considered!) are of so-called Pareto-type and decay as a power function:

$$\bar{F}(x) = 1 - F(x) = x^{-\alpha} L(x), \alpha > 0$$

where $L(x)$ is a slowly varying function:

$$\lim_{t \to \infty} \frac{L(tx)}{L(x)} = 1 \; \forall x$$

α is called the *tail index*. The reciprocal $\gamma = 1/\alpha$ is called the *extreme value index* or the *extremal index*.

Examples of heavy-tailed distributions are the *Pareto* distribution

$$\bar{F}_\alpha(x) = x^{-\alpha}$$

and the now familiar t-distribution

$$\bar{F}_\alpha(x) = x^{-\alpha}\alpha^{(\alpha-1)/2}$$

Notice that the tail index is equivalent to the degree of freedom for the latter.

Equivalently to $\bar{F}(x)$ we have, for some other slowly varying function $l(x)$ the quantile function

$$Q(1-p) = \bar{F}^{-1}(1-p)$$

For our types, we have

$$Q(1-p) = p^{-\gamma}l(1/p)$$

With

$$P(X > u) = u^{-\alpha}L(u)$$

we have that (since $log(\cdot)$ is monotone and increasing)

$$P(log(X) > u) = e^{-\alpha u}L(e^u)$$

or

$$log[Q(1-p)] = -\gamma log(p) + log[l(1/p)]$$

Denote now $X_i$as the order statistics of $\{X\}$:

$$X_1 \leq X_2 \leq ... \leq X_n$$

A natural estimator for the $1 - \frac{j}{n+1}$-quantile is simply to take the ordered observation nr $(n-j)+1$, i.e. $Q(1-\frac{j}{n+1})$ can be estimated with X_{n-j+1}. It can be shown that this is a consistent estimator.

We test the method on a time series of 7 years daily for the foreign exchange (FX) rate SEK/USD (figure 10.4). To investigate the estimator, we plot the empirical quantiles X_{n-j+1} against $j/(n+1)$ on a log-log scale for fixed n (figure 18)

As seen, this gives an increasing function. For $j > k$, for some k, it becomes essentially linear. Now, assume k is known (at least it can be found empirically be inspecting the plot). Now the slope of the line (the linear part) can be estimated with

$$\gamma^* = \frac{\frac{1}{k}\sum_{j=1}^{k}\Big(log(X_{n-j+1}) - log(X_{n-k})\Big)}{\frac{-1}{k}\sum_{j=1}^{k}\Big(log(\frac{j}{n+1}) - log(\frac{k}{n+1})\Big)}$$

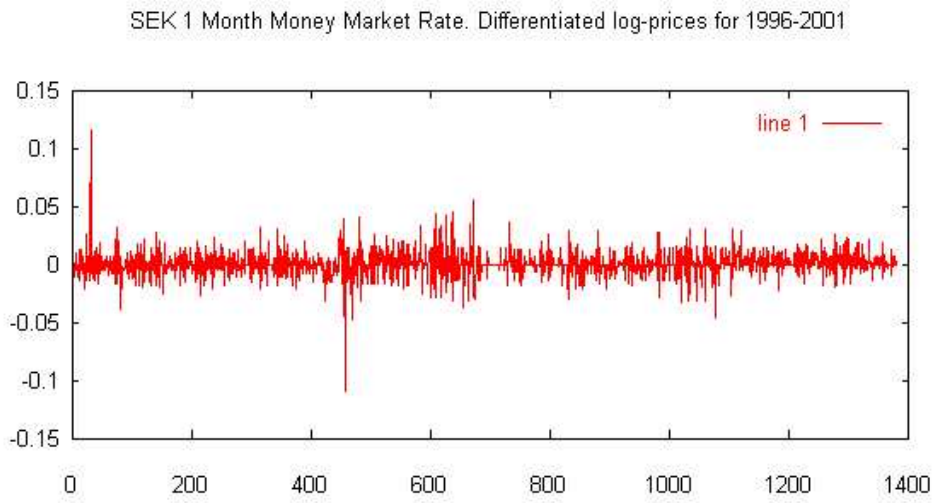

Figure 17: The log-differentiated SEK/USD FX time series.

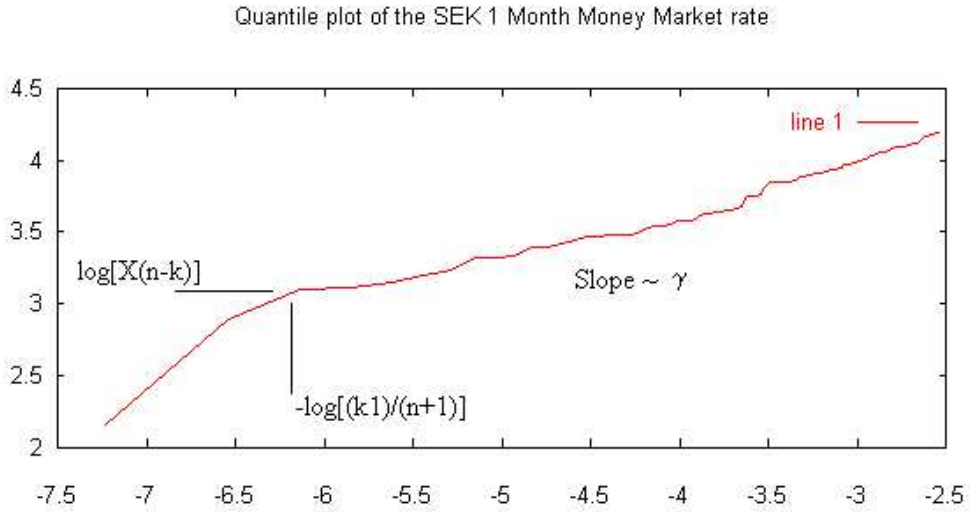

Figure 18: Quantile plot of the SEK/USD log-differentiated FX time series.

For k large, the denominator is approximately 1. This leads to the famous *Hill estimator:*

$$\gamma^*(k,n) = \frac{1}{k}\sum_{j=1}^{k}\Big(log(X_{n-j+1}) - log(X_{n-k})\Big)$$

Finally, to obtain a quantile estimate $Q^*(1-p)$, we simply extrapolate along the fitted line. Using the Hill estimate γ^* for the slope, the equation for a straight line in the log-log scale is

$$log(Q^*(1-p)) = log(X_{n-k}) + \gamma^*\Big(-log(p) + log(\frac{k+1}{n+1})\Big)$$

or

$$Q^*(1-p) = X_{n-k}\Big(\frac{k+1}{p(n+1)}\Big)^{\gamma^*}$$

10.4.1 Choosing k - a tradeoff problem

In the same spirit as the threshold choice for the POT method, the choice of k, the number of observations to include, is crucial to obtain good (unbiased) estimates. The fundamental tradeoff problem is that the *efficiency* of the estimator increases (the variance of the estimator decreases) with k, but so does also the bias!

Various methods have been proposed to estimate k. See for example Huisman (1997) for a discussion. They tell us that some simulation studies have shown that k should not exceed $0.1n$. Other studies obtain k from a Monte Carlo simulation: They draw k from an a priori known distribution with known tail index (such as the t-distribution, where the degree of freedom is equal to the tail index). Then, k is selected such that the mean square error of the Hill estimates is minimized. In this case, the results depend on the underlying distribution.

Following the advice $k < 0.1n$, we plot $\gamma(k, 1400)$ for the same SEK/USD FX series (with $n = 1400$) for $k = 1, 2, ..., 140$.

As can be seen (figure 19), $\gamma(k)$ is stable for, approximately, $k > 40$. Hence, in this case, a choice of k s.t., approximately, $0.03n < k < 0.1n$ would work, giving $\gamma(k) \approx 0.35$ and tail index $1/\gamma \approx 2.9$.

For the following class of distribution functions

$$F(x) = 1 - ax^{-\alpha}(1 - bx^{-\beta}), \alpha, \beta > 0$$

an asymptotic approximation of the bias in the Hill estimator is presented by Dacorogna et al. (1995). They note that for almost any fat-tailed distribution, $F(x)$ provides the second order asymptotic expansion of the CDF. Dacorogna et al. (1995) show that the asymptotic expected value of the Hill estimator for a given k is approximated by

$$E[\gamma^*(k,n)] = \frac{1}{\alpha} - \frac{b\beta}{\alpha(\alpha+\beta)}a^{-\frac{\beta}{\alpha}}(\frac{k}{n})^{\beta/\alpha}$$

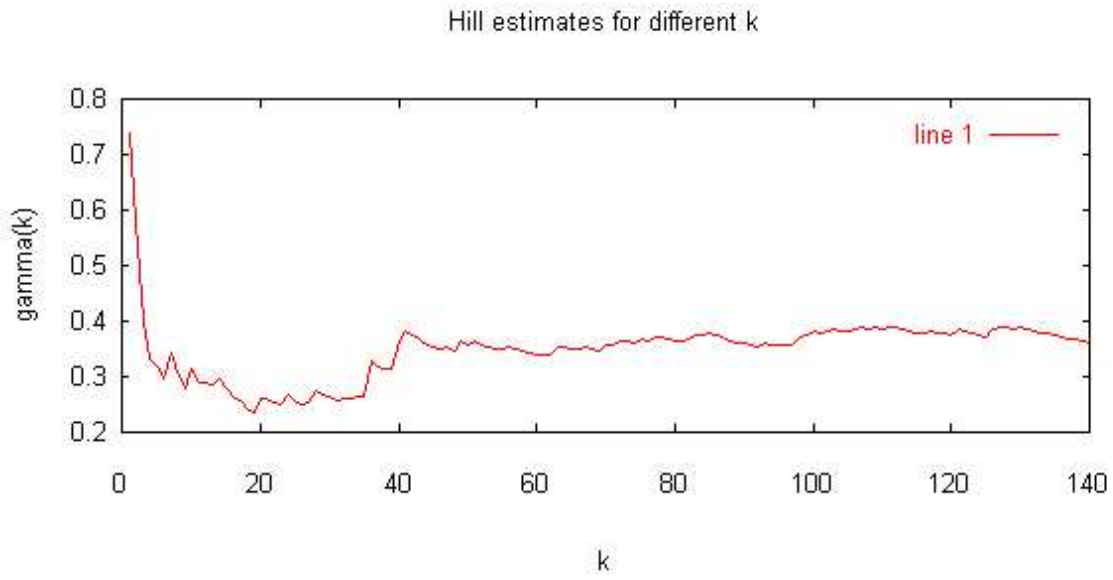

Figure 19: The Hill estimator $\gamma^*(k)$ for $k = 1, ..., n/10$. It is stable for $0.03n < k < 0.1n$

and the asymptotic variance of the Hill estimator by

$$var[\gamma^*(k,n)] \approx \frac{1}{k\alpha^2}$$

From these two equations, we see the problem: From the first, we see that the bias increases in k. From the latter, we see that the variance decreases in k.

10.4.2 An improvement

For the efficiency, a large amount of data is a benefit. However, for our purposes we are restricted to the small sample size of approximately 250 due to the BIS agreement.

Huisman et al (1997) propose an alternative way of improving the Hill estimate in small fat-tailed samples. Their method is computationally heavy, but seem to give better estimates in small samples. They provide some backtesting results on financial data.

Their method is a modification of the conventional Hill estimator. They notice that the bias is an almost linear function of k. Several conventional Hill estimates are computed, and the final estimate is a weighted average of these estimates, with weights obtained by using least square techniques. See their paper for details.

11 Robust estimates

In this paper the focus is on modelling time series model parameter changes, denoted "outliers". Then, if the model is good, these modelled changes can be used to detect real-world suspicious events for some reason of interest. In the multivariate section, this is getting complicated, so we also consider the time independent Mahalanobis distance metric. Assuming the returns to be IID, we can construct an approximate confidence interval. In one dimension, this practically boils down to the naive Z-score "method". Even though simple, this would work fine for a stationary series with constant variance, if the moments were known. The problem is that the standard estimators are heavily biased by (large) outliers. This is also a problem for estimating the models for the described time series outlier detection procedures. This makes the need for robust estimators unavoidable.

11.1 Univariate data

The core of computation of most statistical properties is the summation operator. In contrast, robust operators may consider the *ordered* statistics. A good example is the median: For a symmetrical distribution, the median is a very robust location estimator, and no summation at all is done. More important is to estimate σ robustly, since the sample standard deviation is a very non-robust estimate. An often used alternative is the median absolute deviation (MAD) defined as

$$MAD(\{X_t\}) := median|\{X_t - median(\{X_t\})|$$

This is the sample standard deviation, but with both mean operators replaced by medians, and the norm lowered from 2 to 1. For $x \in N(\mu, \sigma^2)$, MAD must be multiplied with 1.4826 to be a consistent estimator. The MAD is a scale invariant estimator, i.e. if x_t is multiplied by a positive constant c, also the MAD has to be multiplied with c.

For linear regression there are several ways proposed for identifying influential observations, for example the-so-called Cook's distance, built on the leave-one-out concept. Usually, the regression parameters are obtained with the ordinary least squares (OLS) estimator. In the Gaussian case, the OLS estimator can be shown to equal the ML-estimator. As for σ, summation of squares is highly non-robust. Also here, a more robust estimate is obtained by replacing the summation of squared residuals with a summation of their absolute values (L_1-regression). Another alternative is to keep the squares, but to replace the sum with the median. This gives the often used Least Median of Squares (LMS). This estimator has a so-called breakdown point at 0.5, which is the highest possible, meaning that more than half of the points must be outliers w.r.t to true regression line for the estimator to fail and "break down".

The square-minimizing estimators OLS and LMS both belong to a class of estimators called S-estimators. The mean of squared residuals produces the traditional estimate of the variance for a sample with known mean zero. Similarly, the median of squared residuals produces the square of the MAD for a sample with known median zero. Hence, LMS and OLS both minimize an estimator for the scale of $\{X_t\}$, any other estimator for the scale

may be used as objective function. This produces the class of S estimators. They all have a high breakdown point (i.e. they are very robust against outliers), but are therefore not very effective.

Another important class of robust estimator is the class of generalized maximum likelihood (GM) estimators. For a regression, the GM estimator assigns weights from a chosen weight function (usually a distance measure) to the regressors. If the weights are dropped, the class reduces to the class of M estimators. In finance or econometrics, this class is sometimes called Quasi Maximum Likelihood (QML) or Pseudo Maximum Likelihood (PML) estimators. Several weight functions are proposed in the literature, such as the so-called Huber function, the bisquare function and many more. In general, the GM estimators have a lower breakdown point than the S-estimators, but are also more efficient. The MAD is an example of a GM-estimator.

11.2 Multivariate data

The M estimators have been generalized to the multivariate case, but have a quite low breakdown point. The perhaps most used high-breakdown estimator for multivariate scatter is the Minimum Volume Ellipsoid, introduced 1985 by Rousseeuw. This estimator resembles the LMS estimator. It looks for the ellipsoid with the smallest volume covering at least half of the observations. The center is taken as location estimator, and the matrix defining the ellipsoid is used to construct an estimate of the covariance matrix. Another high breakdown estimator for multivariate scatter is the minimum covariance matrix determinant (MCD). A drawback is MVE and MCD are both *very* computationally heavy and quite ineffective. See Rousseeuw (1985) for more details about MVE and Rousseeuw (1999) for details and a recently suggested faster algorithm for computing the MCD.

11.3 Empirical results for the unconditional volatility

A more robust estimate of the unconditional variance or standard deviation (the volatility) σ was needed. To construct such an estimator, several ideas were used.

- **Trimming**
  First, instead of the series $\{X_t\}, t = 1..n$, we considered the "trimmed" series $\{Y_t\}, t = 1..n - \tau$, which is the same series but with the 1 % (absolute) largest values removed. This would give the estimate (assuming mean zero)

  $$\sigma^* = \frac{1}{n - \tau - 1} \Sigma_t Y_t^2$$

  Trimming is a standard trick in robust statistics, but not always a very popular method among statisticians, since we actually *change* the series. However, it could be argued, that if there is one or a few big outliers, those will certainly bias the normal moment estimate heavily. But in case they are not outliers, it should not make an all too big difference to remove them.

- $L_2 \rightarrow \text{Ł}_1$
  Notice that the sample standard deviation volatility σ is the Euclidean norm (2-norm) of the vector $\{Y_t\}$, i.e.
  $$\sigma = ||Y||_2$$
  A natural idea is to lower the norm
  $$\sigma = (y_1^p, y_2^p, ..., y_n^p)^{1/p}$$
  to some $p \in (0, 2)$. It was found that for our purposes, a more robust but still reasonable estimate of $\sigma$was obtained for values of p around 1, and $p = 1$ was used. This further diminishes the sensitivity to big outliers, and a big outlier now only has an effect of order $1/n$. We obtain
  $$\sigma^* = \frac{1}{n - \tau - 1}\Sigma_t |Y_t|$$
  Remark: At this point, an even more robust estimate would have been to take the median instead of the mean of $\{|X_t|\}$. This is the popular MAD-estimator. This was tried, but it turned out that the MAD-estimator was too ineffective to be useful.

- **Trim bias compensation**
  Next, we compensate for the trimming, which of course in any case, outlier or not, will bias the estimate downwards. Under the null hypothesis of no outliers, empirical results on real-world time series (manually inspected to be clean) and some Monte-Carlo simulated $t(\nu)$ series, $\nu \in (3, 5)$, showed that this estimate decreases fairly linearly in τ with an approximate slope of 0.0002. Certainly, this will not always be very accurate, but the gain of punishing out the effects of the 1 % biggest values might be very high. Hence, we compensate the estimate for trimming as
  $$\sigma^{**} = \sigma^* + 0.0002\tau$$

- **Scaling**
  Now, with the trim effect taken care of, the estimate has to be re-scaled to be unbiased and consistent with the standard L_2-estimate. With or without trimming (has a roughly equal effect on both estimates), for a standard normal $N(0, 1)$-variable, σ becomes approximately 20 % underestimated in L_1. However, our more heavy-tailed series becomes about 30 % underestimated (under the null hypothesis of no big outliers) and thus
  $$\sigma^{***} = \frac{1}{0.7}\sigma^{**}$$

- **Missing values/zero returns** Finally, the downbiasing effect of low liquidity, i.e. missing values or constant values (implying a zero logreturn), is compensated. (Usually, missing values are replaced by the preceding). Under the null hypothesis of no

outliers, it was found out empirically that the bias can be roughly approximated by a linear function. Linear regressions of a variety of instruments gave the bias model

$$B(L_Y) = 1 - 0.6(1 - L_Y)$$

where L_Y (for liquidity) was defined as the ratio

$$L_Y = \frac{\#\{t : Y_t = 0\} + \#\{t : Y_t = NA\}}{n}$$

(where n is the total sample size and NA means missing).

Modifying the estimate by multiplying this factor gives the final estimate

$$\hat{\sigma} = \frac{1 - 0.6(1 - L_Y)}{0.7}\Big(0.0002\tau + \frac{1}{n - \tau - 1}\Sigma_t|Y_t|\Big)$$

11.3.1 Example:

A comparison of the described robust estimate of σ and the standard sample standard deviation was made:

ADO	Robust	Non-robust
R_AED1MD=	0.0227	0.0272
R_AED1YD=	0.0236	0.0261
R_AED2MD=	0.0229	0.0247
R_AED3MD=	0.0204	0.0287
R_AED6MD=	0.0222	0.0240
R_AED9MD=	0.0246	0.0260
R_AUD1MD=	0.0128	0.0128
R_AUD1YD=	0.0143	0.0138
R_AUD2MD=	0.0117	0.0114

Consider the fourth instrument, 3 months AED. The robust estimate is 40 % smaller - why? We inspect the series and find that a big outlier in the price series has given rise to 2 outliers in the differentiated return series. If this obvious outlier is removed, the non-robust estimate decreases about 40 %. This illustrates the benefit of robust estimates, and explains why it might be worth using this somewhat fuzzy and empirical measure.

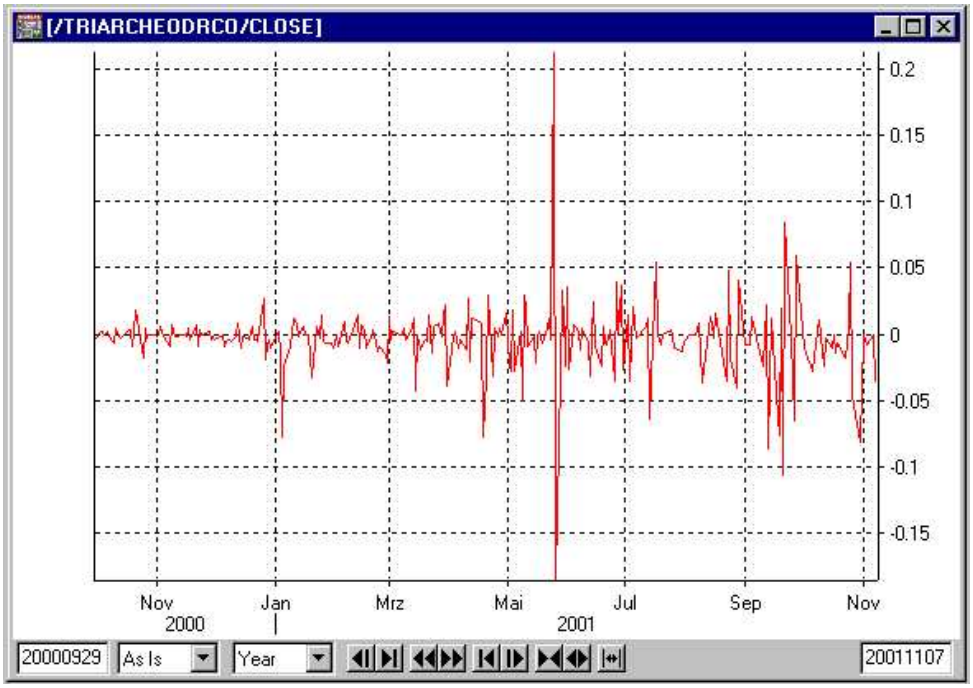

Figure 20: An outlier: The robust and the standard estimate of σ differ by more than 40 %

12 Interpolation

For a given time series model, there is always a predictor that is optimal, in a certain sense, with respect to the model.

The interpolation problem is, more generally, the same problem as the task of finding the best predictor, i.e. the predictor that minimizes the expected least square error. This means (compare to least squares estimates!) that the predictor is orthogonal to the space of observed variables.

Now, in the stochastic world, the covariance operator $Cov(x, y)$ can be though of as an inner product. Then, just as for the usual (deterministic) least squares estimates, we have the orthogonality condition

$$Cov(Y^* - Y, W_i) = 0 \forall i$$

or (assuming mean 0)

$$E[(\text{error}) \times (\text{predictor variable})] = E[(Y^* - Y)W_i] = 0, \forall i$$

where Y^* is the predictor of Y and W_i are the observed variables. (the projection theorem). Or, more directly, Y^* is directly defined as the orthogonal projection of the observed variables $\{W_i, i = 1, ..., n\}$ on the spanned Hilbert space:

$$Y_0^* := Proj_{\bar{sp}}\{W_i, i = 1, ..., n\}$$

where $Proj_a b$ denotes the orthogonal projection of b on a.

Another fundamental result in time series analysis is that, given the existence of $f(\lambda) > 0$, X_t has a spectral representation (an inverse Fourier transform)

$$X_t = \int_{-\pi}^{\pi} e^{itu} f(u) du$$

Now, consider a time series $\{X_t, t \in \mathcal{Z}\}$ with mean 0 and spectral density $f(\lambda)$, where $f(\lambda) > 0 \forall \lambda \in [-\pi, \pi]$, which has been observed except at the time point $t = 0$. The *best linear interpolator* X_0^* of X_0 is then defined by

$$X_0^* = Proj_{\bar{sp}}\{X_i, i \neq 0\}$$

This has a unique solution, and it turns out that it is easiest to find in the spectral domain, and we obtain X^* as the spectral representation

$$X_0^* = \int_{[-\pi,\pi]} \Big(1 - \frac{2\pi}{k f(\lambda)}\Big) f(\lambda) d\lambda$$

where

$$k = \int_{[-\pi,\pi]} \frac{du}{f(u)}$$

giving the expected mean square interpolation error

$$E[(X_0^* - X_0)^2] = 2\pi k$$

See Grandell, p 57-58 for the detailed derivation. In practice, $f(\lambda)$ (or the ACVF) can of course be estimated by its empirical counterpart, if no model is assumed.

13 Summary

The aim of this paper was to give a review of methods for modelling tails of financial returns, together with some simulation results and with emphasis on practical real-world problems.

A system for detecting values out of their expected position was implemented and calibrated from a large variety of different financial time series, such as equities, indices and FX rates. The final decision of the credibility of a found suspect must be based on external, situation-dependent information and be decided from case to case, except from extremely aberrant values.

We classify three main categories of outliers:

- **Multivariate outliers** A multivariate outlier is a time point in a multivariate time series being *jointly* much out of its expected position, as predicted by the rest of the series.
- **Univariate outliers** A multivariate outlier in the special case of one dimension boils down to a univariate outlier.
- **Marginal outliers** A marginal outlier is time point in a scalar component series. Compared to all other components at this time point, it is much out of its expected position.

Hence, multivariate outliers (and the special case univariate outliers) look in the time direction only, whereas marginal outliers look in the cross-sample direction, orthogonal to the time direction.

These types interact and give rise to so-called masking effects and swamping effects, which occur in different ways in the univariate and the multivariate case.

The outlier detection problem is closely connected to the prediction problem, since an outlier is always defined as being much different from a value predicted by some model.

- **Univariate case**

  We define four types of time series outliers, AO's, IO's, TC's and LS's, and how they can be modelled. Then, we describe a method for joint estimation of outlier effects and model parameters for these outlier types occurring in ARMA time series models. A modification to GARCH models is described as well. Extra attention is paid to the practical problem of estimating the GARCH model parameters, and a solution is proposed.

  We also describe how to use a moving median smoothing process, together with some empirically found results of how the unconditional standard deviation can be estimated robustly.

  Then, in contrast to the mean-based theory, we also use EVT for tail estimation. The POT-method and the Hill estimator are described. We also show an application

that combines EVT with a stochastic volatility model, in order to give conditional VaR estimates.

We also give an explicit formula for optimal interpolation, given a time series model.

- **Multivariate case**
  We use the Mahalanobis distance as joint distance measure, and also show how this multivariate technique can be useful for a univariate time series as well by considering the phase space. The definitions of AO's, IO's, TC's and LS's are generalized to the multivariate case.

- **Marginal components considerations**
  In order to predict the marginal components, the conditional marginal distribution of a multivariate elliptical model was used. We combine the lepokurtosis property of an ad-hoc assumed $t(4)$-distribution with the useful analytical properties of the multivariate normal distribution. The popular EM-algorithm is also described.

  We also give a brief overview of robust estimators and construct an empirically well working (robust but not too inefficient) volatility estimator.

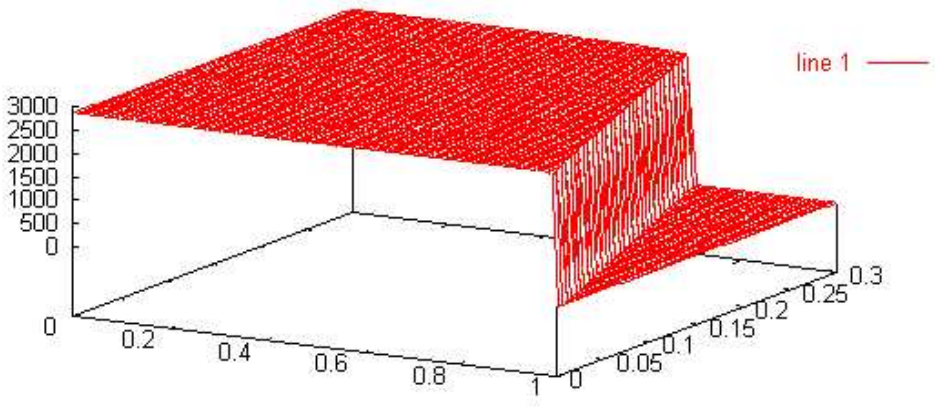

Figure 21: The likelihood surface $L(\alpha, \beta)$

A Estimating the GARCH(1,1)

The unconditional variance σ^2 of σ_t is

$$\sigma^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta}$$

A reduced, more robust problem is therefore obtained by estimating σ^2 with the usual sample variance, leaving us with 2 variables, α_1 (from now on just called α) and β. α and β should be estimated by maximum likelihood. The likelihood function to maximize is

$$L(\alpha, \beta) = -\sum_{t=1}^{N}(log(\sigma_t^2) + Z_t)$$

under the stationarity restriction

$$\alpha + \beta < 1$$

This maximization may in practice be *very* troublesome due to two reasons:

1. A globally very flat surface.
2. A locally *not* very flat surface.

This unfortunately combination makes it hard for any gradient method to find the maximum.

A.0.2 Problem 1

Zumbach (2000) proposes the coordinate transformation $(\alpha, \beta) \rightarrow (z_{corr}, z_{ema})$:

$$z_{corr} = log(-log(\mu_{corr}))$$

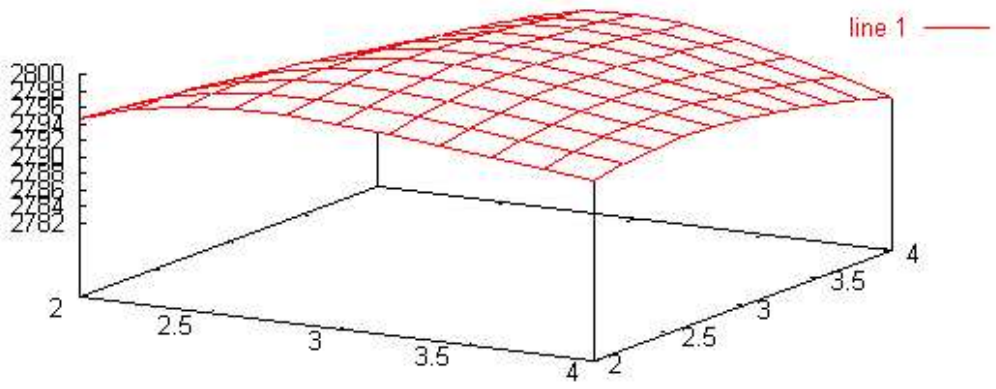

Figure 22: The likelihood surface $L(z_{corr}, z_{ema})$

$$z_{ema} = log(-log(\mu_{ema}))$$

where

$$\mu_{ema} = \alpha + \beta$$
$$\mu_{corr} = \frac{\alpha}{1 - \mu_{ema}}$$

This simple transformation drastically improves the situation. First, because the surface in the new coordinate system is not at all as flat as in the original coordinates. Second, because in the new coordinates, no constraint is needed. This solves problem 1. It seems as without this transformation, no reliable estimates are generally obtained, making the proposed estimation procedure in the standard literature of Hull useless.

A.0.3 Problem 2

Even though an unambiguous global maximum now exists, the surface is locally still not smooth, and any gradient search method may still fail. The *only* really safe method is to discretize the area and evaluate the function at every point. One other possible solution to this would be to apply a smoothing filter, but then one would still have to evaluate the function in every point, and then we have already found the maximum. The estimation function in the GARCH filter first makes such a global grid search, with grid spacing h. Then, a local Gauss-Newton gradient search with numerical derivatives is done locally around this point. Even if the gradient search does not converge and gets stuck in a local maximum, the remaining distance to the global maximum can be at most h.

The fundamental reason for problem 2 is a too small sample size.

The likelihood L depends on α and β in a complicated way, but gets smooth if fed with a long series. For $N = 250$, the gradient may often not converge. This, however, is

not a big problem. First, because the global grid search already has found a point close to the maximum. Second, because the Gauss-Newton maximization converges quadratically, and just one iteration in the right direction will take us very close to the maximum. If the second iteration diverges is of minor importance.

If the global optimization would fail, two extra precautions are used, making the procedure completely safe:

1. The average GARCH volatility $\sqrt{mean(\{\sigma_t(\alpha,\beta)\})}$is compared with the standard equally weighted volatility. If they differ too much, the estimation is aborted.

2. If the Goodness-of-Fit (GOF) value is too low, the estimation is aborted.

If the estimation fails by any criterion, the default values of α and β are used. This makes the estimation completely safe, in the meaning that it will never give totally unrealistic values.

For a more complete investigation of the numerical problems connected with ML-estimation of GARCH-processes to real-world financial data, see Zumbach (2000).

B References

- **Beirlant, J. and Matthys, G. (2000)**
  Quantile estimation for heavy-tailed data
  www.inrialpes.fr/is2/pub/work/extreme/jb.ppt

- **Blasnik (1995)**
  The Need for Statistical Analysis and Reporting Requirements: Some Suggestions for Regression Models
  Project report presented at 1995 Energy Program Evaluation Conference, Chicago
  www.proctoreng.com/reports/need.html

- **Bollerslev, T (1986)**
  Generalized autoregressive conditional heteroskedasticity
  Journal of Econometrics, 52, p. 5-59

- **Brockwell,P.J. and Davis,R.A (1996)**
  Introduction to time series and forecasting
  Springer-Verlag, New York

- **Chen, C. and Liu, L (1993)**
  Joint estimation of model parameters and outlier effects in time series
  Journal of the American Statistical Association, Vol. 88 1993, No. 421

- **Chu, C.-S. J. (1995)**
  Detecting parameter shifts in GARCH models
  Econometric reviews 14: 241-266

- **Dacorogna, M., Müller, O., Pictet and C. de Vries (1995)**
  The distribution of extremal foreign exchange rate returns in extremely large data sets
  Tinbergen institute discussion paper no. 95-70

- **Embrechts P, McNeil AJ and Straumann D (2000)**
  Correlation and dependency in risk management: properties and pitfalls
  In Risk management: value at risk and beyond, edited by Dempster M and Moffatt HK, published by Cambridge University Press (yet to appear). Available at
  http://www.math.ethz.ch/m̃cneil/pub_list.html

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

- **Engle, R F (1982)**
  Autoregressive conditional heteroskedasticity with estimates of the variance of UK

inflation
Econometrica, 150 p. 987-1007

- **Engle, R.F. and Ng, V.K (1993)**
Measuring and testing the impact of news on volatility
Journal of finance 48: 1749-1777

- **Fallon, A. and Spada, C.**
Detection and accommodation of outliers in normally distributed data sets
Virginia Polytechnic Institute and State University
http://www.ce.vt.edu/program_areas/environmental/teach/smprimer/outlier/outlier.html

- **Fox, A. J (1972)**
Outliers in time series
Journal of the Royal Statistical Society, series B 34, p 350-363

- **Franses, P. and van Dijk, D. (2000)**
Outlier detection in the GARCH(1,1) model.
Econometric Institute Research Report 9926/A
www.few.eur.nl/few/people/djvandijk/papers/publications.htm

- **Gather, U., Bauser, M. and Imhoff, M. (1999)**
The Identification of Multiple Outliers in Online Monitoring Data
Technical Report, University of Dortmund.
Available at http://www.statistik.uni-dortmund.de/sfb475/

- **Grandell, J (1999)**
Time series Analysis
Lecture notes, KTH Stockholm

- **Huisman et al (1997)**
Fat tails in small samples
Journal of Business Economics and Statistics.

- **Hull, J. (1997)**
Options, futures and other derivatives
Prentice Hall, third edition. ISBN 0-13-264367-7

- **Kaiser, T. (1996)**
One-Factor-GARCH Models for German Stocks: Estimation and Forecasting
Working paper, Eberhard-Karls-Universität Tübingen
Available at http://econpapers.hhs.se/paper/wpawuwpem/9612007.htm

- **Klüppelberg, Emmer and Trüstedt (1998)**
VaR - a measure for the extreme risk
Center for mathematical sciences, Munich University of technology

Available at
http://www-m4.mathematik.tu-muenchen.de/m4/lect-conf/ims_paper/cklu_1.html

- **Li, W.K. and Mak, T.K. (1994)**
On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity
Journal of Time Series Analysis 15: 627-636

- **Lin, S.-J. and Yang, J. (1999)**
Testing shift in financial models with conditional heteroskedasticity; An empirical distribution function approach
University of Technology, Sidney, Quantitative Finance Research Group, Research paper no. 30.

- **Lucas, A (1996)**
Outlier robust unit root analysis
Ph.D thesis, Vrije Universiteit, Dept. Finance and Financial Sector Management
Available at http://staf.feweb.vu.nl:81/alucas/thesis/default.htm

- **Lundbergh, S. and Teräsvirta, T. (2000)**
Evaluating GARCH models
Stockholm School of Economics, Working Paper no. 292
www.hhs.se/stat/research/nonlinear.htm

- **McNeil A.J. (1999)**
Extreme value theory for risk managers
Internal Modelling and CAD II published by RISK Books, 93-113
Available at http://www.math.ethz.ch/~mcneil/pub_list.html

- **McNeil, A. J. and Frey, R (2000)**
Estimation of tail-related risk-measures for heteroskedastic financial time series: an extreme value approach
Research paper, ETHZ. Published in Journal of Empirical Finance, 7: 271-300 Available at
http://www.math.ethz.ch/~mcneil/pub_list.html

- **Mikosch, T. (2000)**
Modeling financial time series
Lecture notes, University of Copenhagen. Available at
www.math.ku.dk/~mikosch/Preprint/GARCH/SVM/svm.ps.gz

- **Mikosch, T. and Starica, C. (2000)**
Change of Structure in financial time series, long range dependence and the GARCH model
Working Paper 58, University of Aarhus, Aarhus Scool of Business Available at
http://citeseer.nj.nec.com/mikosch99change.html

- **Morgan Guaranty, (1996)**
  RiskMetrics Technical Document
  Morgan guaranty Trust Company of New York, 4th edition

- **Müller, U. (2000)**
  The Olsen filter for data in finance
  Internal document, Olsen and Associates.

- **Røyslien,J. (1998)**
  Random t-distributed Fields
  Trondheim University of Science and Technology
  Available at http://www.math.ntnu.no/preprint/statistics/1998/

- **Rousseeuw, P.J. (1985)**
  Multivariate estimation with high breakdown point
  Mathematical statistics and applications, volume B, W. Grossmann, G. Phlug, I. Vincze, and W. Wertz (eds.), Dordrecht: Reidel Publishing, 283-297

- **Rousseeuw, P.J. and Leroy, A.M. (1987)**
  Robust regression and outlier detection
  New York: Wiley

- **Rousseeuw, P J and Van Driessen, K (1999)** *A fast algorithm for the minimum covariance determinant estimator*
  Technometrics, 41,212-223. Available at http://win-www.uia.ac.be/u/statis/publicat/fastmcd_rea

- **SAS Software tutorial**
  Multivariate analysis concepts
  Available at http://www.sas.com/service/doc/pubcat/chaps/56903.pdf

- **Strandberg, F (2001)**
  Univariate outlier detection in Asset Control
  Internal document, RCO 31, HypoVereinsbank Munich

- **Sundberg, R. (1997)**
  Tillämpad matematisk statistik
  Lecture notes, Mathematical statistics, Stockholm University

- **Tsay, Pena and Pankratz (1999)**
  Outliers in multivariate time series
  Available at http://halweb.uc3m.es/esp/Personal/personas/dpena/curri.html

- **Tolvi, J. (1999)**
  Outliers in time series: A review
  University of Turku, Department of Economics, Research reports No. 76.
  Available at http://aws.tt.utu.fi/tolvi1.html

- **Tolvi, J. (2001)**
  Nonlinear model selection in the presence of outliers
  University of Turku, Department of Economics, Research reports No. 90.
  Available at http://aws.tt.utu.fi/tolvi4.html

- **Woodward, Wayne, Stephan, Sain, Gray (1999)**
  Outlier Testing When Some Data Are Missing
  Available at http://www.ctbt.rnd.doe.gov/Symposium1999/proceedings.html

- **Wrinkler, R**
  Introduction to robust statistics and data filtering
  U.S Naval observatory. Available at
  http://people.ne.mediaone.net/rile/ROBSTAT.htm

- **Zumbach, G. (2000)**
  The Pitfalls in Fitting GARCH(1,1) Processes
  Published in "Advances in Quantitative Asset Management" Studies in Computational Finance edited by Christian L. Dunis, Kluwer academic publishers, pp. 179-200, ISBN 0-7923-7778-8, SICF1 0-7923-7778-8.

- **Quinn, A. and Tesar, L. (2000)**
  A survey of techniques for preprocessing in high dimensional data clustering
  Institute of Information Theory and Automation, Academy of Science of the Czech Republic, Department of Adaptive Systems
  Extended abstract for the conference "Cybernetics and Informatics Eurodays Young Generation Viewpoint" September 26-30, 2000, Marianska, Czech Republic