University of Tartu
Faculty of Mathematics and Computer Science
Institute of Mathematical Statistics

DISTRIBUTIONS IN FINANCIAL MATHEMATICS (MTMS.02.023)

Lecture notes

Ants Kaasik, Meelis Käärik

Contents

Course material is based on the synthesis of several lecture notes about probability distributions and books that are available in the university library. This material includes most important aspects of the course but reading it is by no means equal to the experience that is gained in the lecture.

Some key textbooks:
Klugman, Panjer & Willmot (2004). *Loss Models: From Data to Decisions.*
Cooke & Nieboer (2011). *Heavy-Tailed Distributions: Data, Diagnostics, and New Developments.*
Balakrishnan, Nevzorov (2003). *A Primer on Statistical Distributions.*
Evans, Hastings & Peacock (2000). *Statistical Distributions.*
Embrechts, Mikosch & Klüppelberg (1997). *Modelling Extremal Events: for Insurance and Finance.*
Resnick (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*

1 Introduction

1.1 The data. Prediction problem

In many applications of the financial mathematics we want to look at the future (i.e. predict what are the possible scenarios and how likely they are). But this can only be based on the past (and present). So we train a model (estimate model parameters) using available data (training data). But it is typically not the model that we are interested in, rather we have a practical question that needs answering ("how probable is it that the investment will not return the required 5%"). However, to answer this question a model is still typically needed and when this is postulated (and the parameters estimated) then answering the initial question is straightforward.

So the first question is how the data is generated (what is the model responsible)? Usually this means that we pick a parametric model and estimate the model parameters using the training data. This seems reasonable but unfortunately reality is (typically) different – we are lucky if the model fits the data well but it is very unlikely that the model is true (if we do not consider simple physical experiments like die casting or measurement error distribution). Still, this approach might be the best we can (or are able to) do.

What is also important to remember: observations are typically not independent but if we must make this assumption then we must do our best so that the assumption would be violated by (our transformed) data "as little as possible". For example, it is very hard believe that stock prices (for a

particular stock) for successive days are independent, but changes in price are more likely to be.

1.2 The models

Let us consider some data (say, 100 observations) and think of the first model that crosses your mind? With financial data it is likely that simple models are not very realistic. At least not in the long run. This is probably so but we can also consider that the model parameters are time dependent (i.e. not constant in time) or a change to a different distribution occurs. A Hidden Markov Model (HMM) is already complex enough to be plausible. But even for this complex model any given observation is still a realization from some (univariate) probability distribution. So to be able to use this model we definitely need to know the possible candidates for univariate distributions. This is the reason why we consider (nothing but) *univariate probability distributions* in the course.

Of course if we are interested e.g. in a quantile only (and we have plenty of data) then it might be possible to not make an assumption of a specific underlying parametric distribution as we are not interested in it.

On rare occasions we might in fact find a theoretically justified distribution model (as is the case with die casting) Central Limit Theorem (CLT) is the most prominent example.

1.3 Summary of the introduction

We will use real data to study the particularities of the data that arises in financial mathematics and also consider how to use those particularities. To do this we need additional notations. This makes it easier to classify probability distributions. We consider different distributions that are possible models for financial data. We also consider the distribution free approach (which is based on either the empirical cumulative distribution function (cdf) or empirical probability density function (pdf)). Some attention is given to the practical ways of picking a suitable model (out of a list of candidate models).

2 Exploration and visualization of the data

2.1 Motivating example

Consider these (real) stock prices spanning 5 years (file *prices.csv*).

```
> setwd("C:\\Courses\\dfm")
> data=read.csv("prices.csv")
> #fix the date format
> lct = Sys.getlocale("LC_TIME"); Sys.setlocale("LC_TIME", "C")
> data$date=as.Date(data[,1],"%d-%B-%y")
> Sys.setlocale("LC_TIME", lct)
> data=data[,-1]
> data=data[order(data$date),]
> row.names(data)=NULL
```

There are several variables in the dataset.

```
> head(data)
```

```
   Open  High   Low Close   Volume       date
1 32.36 32.60 32.03 32.16 36810700 2004-01-02
2 32.52 32.97 32.40 32.91 41584300 2004-01-05
3 32.97 33.00 32.62 32.90 37038600 2004-01-06
4 33.31 34.00 33.22 33.99 62211000 2004-01-07
5 34.30 34.35 33.90 34.24 49040500 2004-01-08
6 33.80 34.60 33.80 33.97 56036400 2004-01-09
```

We are interested in the closing prices (variable *close*). Is it reasonable to believe that successive values are independent realizations?

```
> n=dim(data)[1]
> cor(data$Close[-1],data$Close[-n])
```

```
[1] 0.9924598
```

Perhaps things would be better if we would consider changes in closing prices (e.g. intuitively it should not be very easy to predict the sign of a price change).

Due to a lot of really large changes it is perhaps not that evident from the figure 2 but there is a problem...

What about relative changes (in percent)? Let us consider a density estimate based on relative changes in the closing prices (how this is achieved will be discussed in detail later in the course).

Is this good enough? Do we now have a (universal) symmetric law that should fit the data that is based on different stock prices? It seems that this graph is a little skewed to the right. In particular, the peak of the density (the mode) is positive real number, for sure. Is this how it is supposed to

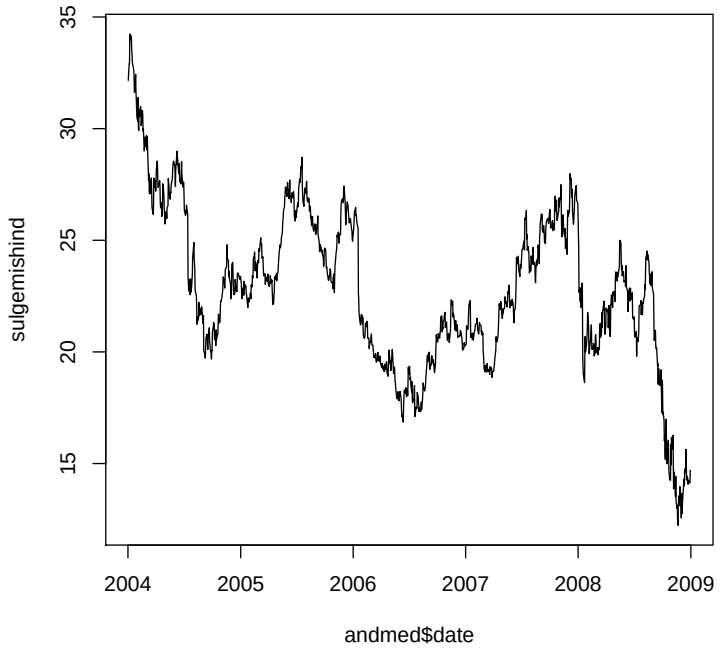

Figure 1: Closing prices

be? Perhaps it is just because the stock price has increased in the long run? Let us look at the figure 1 again.

Perhaps this asymmetric distribution can be explained with the fact that there are simply more large declines and to compensate many small increases are required. This is partly true in this case. But let us pause for a moment. Suppose the closing price on the first day is 100, on the second day it is 110 and on the third day it is 100 again. So the price has not actually changed over 2 days.But what about the *returns* or relative changes in price? For the first change it is 0.1 or ten percent but it is not -0.1 for the second change, because for this the drop must have been 11 units. Thus something closer to zero ($-1/11$ to be more precise). So the initial and final price are the same but we got two returns which both have different sign but the positive one is larger in absolute value. Is this a coincidence?

Home assignment 1. Show that for equal initial and final price ($p_t = p_{t+2}$) and some different price in between ($p_{t+1} \neq p_t$) we get two returns $(p_{t+2} - p_{t+1})/p_{t+1}$ and $(p_{t+1} - p_t)/p_t$ such that the positive return is always larger in absolute value. Show that it is not important whether the price increases ($p_{t+1} > p_t$) or decreases ($p_{t+1} < p_t$) at first. We consider that all

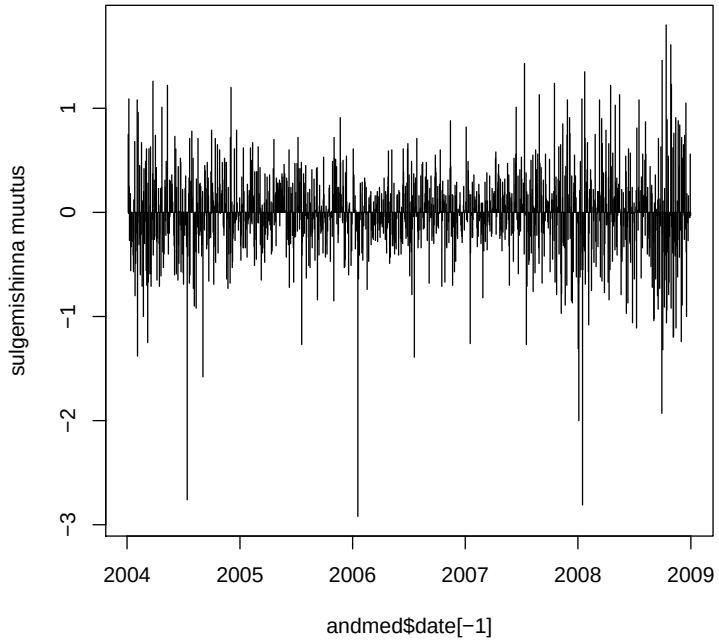

Figure 2: Changes in closing prices

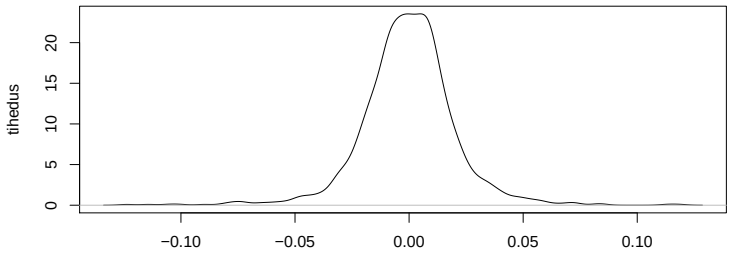

Figure 3: Density estimate of relative price changes

prices are positive.

This shows that it is hard to consider the returns as observations, because it seems reasonable that share price is constant on average (if share price is considered as a realization from a random process). Of course we could separate positive and negative returns. But in practice *log-returns*

$$\ln \frac{p_{t+1}}{p_t},$$

are considered instead. It is easy to see that if $p_t = p_{t+2}$ then $\ln(p_{t+1}/p_t)$ are $\ln(p_{t+2}/p_{t+1})$ equal in absolute value. Also, log-returns are very close to returns because

$$\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}, \text{ if } -1 < x \leqslant 1.$$

Ok, so now we can consider a symmetric random variable (rv). Also the set of initial realizations (stock prices) can be transformed (into log-returns) and back-transformed so no information is lost in the process. We use the sample of log-returns and estimate the parameters of a Gaussian distribution. That's it – we have the distribution that describes the changes in stock prices.

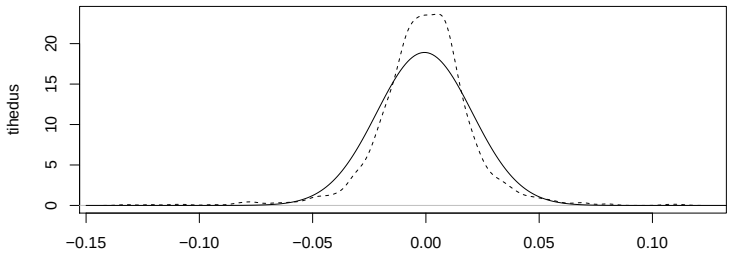

Figure 4: Approximating log-returns by a normal distribution

It seems that the model is not very good – the peak is not sharp enough and the tails are too light. This is exactly the case when the kurtosis of a distribution is too large. Thus the result: normal distribution has tails that are not heavy enough.

Home assignment 2. Make sure that this was not just a coincidence. Go to `http://www.google.com/finance` for data and pick a share. Choose "historical prices" and choose a time-series with a length of a few years. And finally choose "download to spreadsheet" which allows you to save a *.csv*-file. Complete all the steps that we did previously (including the graphs).

3 Heavy-tailed probability distributions

What could we choose as a distribution model? What distribution has a tail which is heavier than that of a normal (Gaussian) distribution. What about t-distribution?

Home assignment 3. Read R help for the function *dt* (i.e. the density function of a t-distribution) and pay attention how the variance is expressed. Use the data from the previous (log-returns) and estimate the variance. Explain why the method of moments cannot be used for fitting the distribution (we can make a simplification and say that the expectation is zero, then a single equation is needed).

Now, one of the models has a tail that is too light and another one has a tail that is too heavy. Thus, it would be nice to have something that would allow us to compare distributions based on their respective tails. Perhaps the easiest approach would be a comparison of two distributions.

Let us consider two probability density functions $f_1(x)$ and $f_2(x)$, respectively, and we compare their right tails (for simplicity assume that $\exists K \in \mathbb{R} : \forall x > K \ f_1(x) > 0, f_2(x) > 0$).

Definition 3.1. If it holds that

$$\lim_{x\to\infty} \frac{f_1(x)}{f_2(x)} = \infty,$$

then we say that *the distribution corresponding to probability density function $f_1(x)$ has a heavier tail than the distribution corresponding to probability density function $f_2(x)$.*

The definition for comparing left tails is similar.

Home assignment 4. Which distribution out of the three has the lightest tail and which one has the heaviest? Explain your answer!

1. $f_1(x) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}},\ \alpha > 0,\ \theta > 0,\ x > 0,$

2. $f_2(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}},\ \mu \in \mathbb{R},\ \sigma > 0,\ x > 0,$

3. $f_3(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k},\ \lambda > 0,\ k > 0,\ x > 0.$

To simplify, it is allowed to fix $\theta = 1$, $\mu = 0$, $\sigma = 1/\sqrt{2}$ and $\lambda = 1$.

Now we can fix some distribution F and say what distributions have a lighter and which ones a heavier tail than F. We could call the former light-tailed distributions and the latter heavy-tailed distributions. We consider three additional functions that help us describe a distribution.

Definition 3.2. *Tail distribution function* (also known as *survival function*) of a random variable X with cumulative distribution function $F(x)$ is the function

$$\bar{F}(x) = 1 - F(x) = \mathbb{P}(X > x).$$

By definition this is a non-increasing function that is left-continuous.

Intuitive way of comparing distributions would be to consider $\bar{F}_1(x)$ and $\bar{F}_2(x)$.

Definition 3.3. If

$$\lim_{x\to\infty} \frac{\bar{F}_1(x)}{\bar{F}_2(x)} = \infty,$$

then we say that the distribution corresponding to $\bar{F}_1(x)$ has a heavier tail than the distribution corresponding to $\bar{F}_2(x)$.

L'Hospital rule allows us to conclude that this definition is actually equal to the one given previously.

Definition 3.4. *Hazard function* (also known as *instantaneous failure rate*) of the random variable X is the function

$$h(x) = \frac{f(x)}{\bar{F}(x)},$$

where $f(x)$ is the probability density function of X and $\bar{F}(x)$ is the tail distribution function of X.

For a (continuous) random variable X the hazard function at x is thus the conditional density function of X at x under the condition that $X \geqslant x$. If the hazard function is a decreasing function then "for large values the probability of a particular value is decreasing and thus the probability of larger values increasing". This is characteristic to heavy-tailed distributions. If the hazard function is increasing then the tail of the distribution is light. This characterization places exponential distribution on the borderline – its hazard function is constant.

Definition 3.5. *Mean excess function* of a random variable X (also known as *mean residual life function*) is the conditional expectation

$$e(x) = \mathbb{E}(X - x | X > x).$$

Partial integration gives us

$$\begin{aligned} e(x) &= \frac{\int_x^\infty (y-x) f(y) dy}{\bar{F}(x)} = \frac{-(y-x)\bar{F}(y)|_x^\infty + \int_x^\infty \bar{F}(y) dy}{\bar{F}(x)} \\ &= \frac{\int_x^\infty \bar{F}(y) dy}{\bar{F}(x)}. \end{aligned}$$

If $e(x)$ is increasing then "the mean exceedance of a particular value is increasing for large values" and thus the random variable has a heavy tail. A decreasing mean excess function corresponds to a light tail. Similarly with the previous case the borderline distribution is exponential.

Home assignment 5. Prove that the mean excess function is constant for exponential distribution. Use the relationship between $e(x)$ and $\bar{F}(x)$ that was discovered previously.

4 Detecting the heaviness of the tail

4.1 Visual tests

Now we know that exponential distribution is on the borderline of light and heavy-tailed distributions. So perhaps it would be a good candidate for the log-returns data (it has a heavier tail than a normal distribution as can easily be seen when comparing the densities). Of course, exponential distribution has positive support so we can only use positive log-returns (at first).

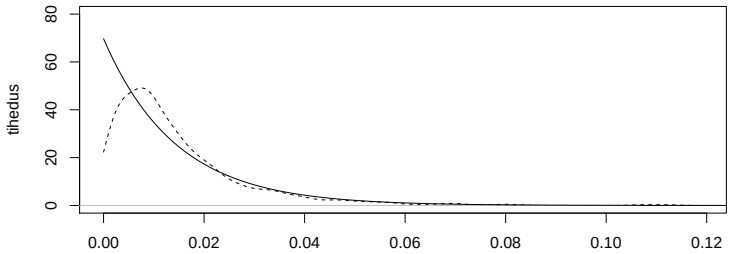

Figure 5: Approximating the positive log-returns with an exponential distribution

The fit seems quite decent, at least as far as the tail is concerned. Histogram is a useful graph to visualize the fit (in addition to the densities). Seems that the data is not from the exponential distribution.

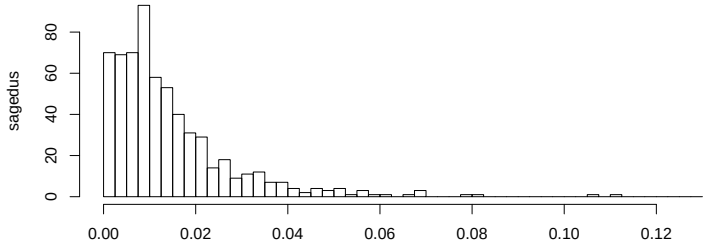

Figure 6: Histogram of the positive log-returns

It is important to note that even though it seems that extreme values from the tail are very unlikely to realize (usually the value of a share does not double in a day, but it actually can happen) it can be very important to predict correctly the frequency of 10% gains. But our financial data can be also of different nature (e.g. insurance claims) where the upper limit really is not clear (of course, the insurance can have a re-insurance contract but the tail behaviour is just as important – e.g. deductible can be a percentage of the total cost)

To compare the theoretical distribution with the data at hand we can also make use of a quantile-quantile graph (QQ-plot): one axis has the sample quantiles and the other the theoretical quantiles.

Consider an ordered sample $(x_{(1)}, \ldots, x_{(n)})$, i.e. $x_{(1)} \leq \ldots \leq x_{(n)}$ and let H be the theoretical candidate probability distribution function. Then we can plot

$$\left\{ x_{(i)}, H^{-1}\left(\frac{i}{n+1}\right) \right\}, \quad i = 1, \ldots, n.$$

Whether the theoretical quantiles are on the horizontal or vertical axis can vary but usually a line through the quartiles is plotted. It is important to note that we are not concerned with the location and scale parameter – these do not change the shape of a graph. This is true because such a linear transform does not change the order nor does it change how many times some value is larger than the other. So we can instead imagine that we plot

$$\left\{ a + bx_{(i)}, H^{-1}\left(\frac{i}{n+1}\right) \right\}, \quad i = 1, \ldots, n,$$

which only differs in scale.

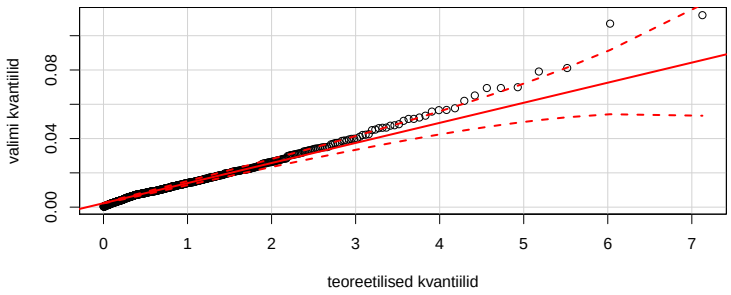

Figure 7: QQ-plot of positive log-returns and the exponential distribution

Home assignment 6. When drawing a QQ-plot why is $H^{-1}(i/(n+1))$ plotted instead of $H^{-1}(i/n)$?

As we see from 7 declaring a good fit was a bit too hasty – several points are outside of the confidence intervals and the slopes seem different. To be more precise: the tail of the exponential distribution is too light as it predicts smaller quantile values than they are in the actual sample. Figure 8 shows, that a t-distribution with 4 degrees of freedom fits the positive log-return data well. For the negative returns such a tail seems to light and thus a smaller degrees of freedom should be used (this means a heavier tail for t-distribution).

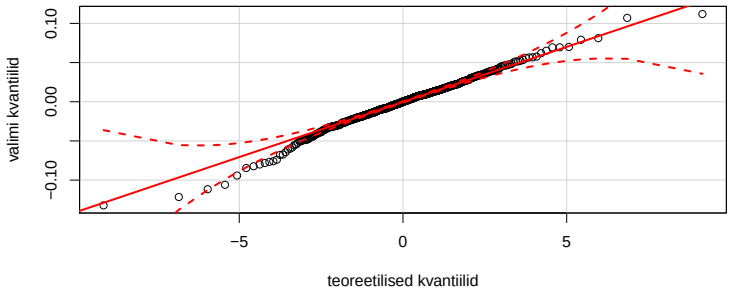

Figure 8: QQ-plot of log-returns and t-distribution (df=4)

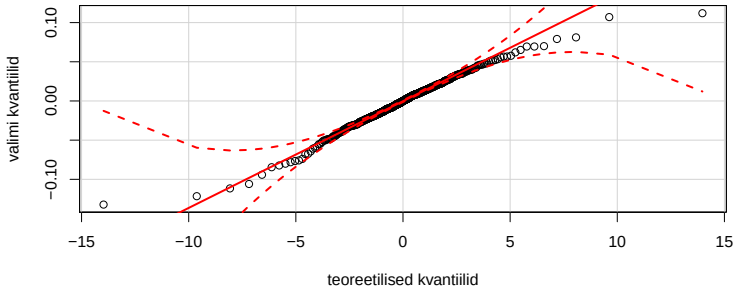

Figure 9: QQ-plot of log-returns and t-distribution (df=3)

As shown a QQ-plot is a useful tool for analyzing the weight of a tail. On the other hand – moments estimated from the sample should also give us similar information e.g. we know that some moments of a t-distribution might not exist (when the distribution has 1 degree of freedom (or less) then even the mean does not exist).

Home assignment 7. Show that a t-distribution with density

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-(\frac{\nu+1}{2})}$$

does not have a defined mean when $\nu = 1$. Hint: obviously we can divide the integral into two equal parts (except for the sign). Thus we must only show that one of them does not converge.

How should we use sample mean to judge the heaviness of a tail? Sample mean is always finite. That is why we cannot use it directly. The plot of a mean excess function, that are the points

$$\left\{x_{(i)}, \frac{1}{n-i}\left(x_{(i+1)}+\ldots+x_{(n)}\right)-x_{(i)}\right\}, \quad i=1,\ldots,n,$$

is an example. Figure 10 is for our positive log-returns, and it points at a heavy tail. However, very wide confidence intervals must be noted. This is typical because (however large the sample) as we move closer to the large values of the sample there will be little information left. Unfortunately this is exactly the most interesting (/important) part of the figure.

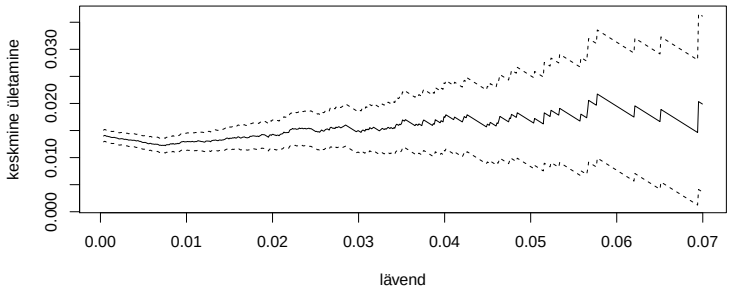

Figure 10: Empirical mean excess function for the positive log-returns

Figure 11 illustrates a heavy tail (t-distribution with 1 degree of freedom)

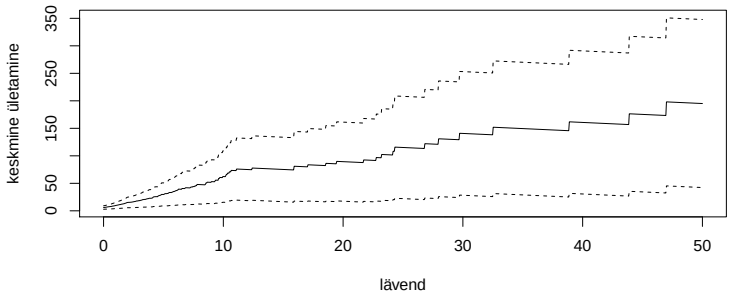

Figure 11: A mean excess function based on a sample with size 1000 from the t-distribution (df=1)

4.2 Maximum-sum ratio test

When a heavy tail is suspected then the existence of moments is not guaranteed. To test this for a sample (of positive values) the ratio between the maximum and the sum of a sample can be used.

Consider the ratio

$$M_n(p)/S_n(p), \qquad i = 1, \ldots, n,$$

where

$$p > 0 \text{ and } S_n(p) = X_1^p + \ldots + X_n^p, \quad M_n(p) = \max\{X_1^p, \ldots, X_n^p\}.$$

It holds that

$$\mathbb{P}\left(\frac{M_n(p)}{S_n(p)} \to 0\right) = 1 \Leftrightarrow \mathbb{E}|X|^p < \infty.$$

As can be seen from figures 12 and 13 it is pretty easy to test this for some moments (for our log-returns data) – mean is finite but the fifth moment probably does not exist.

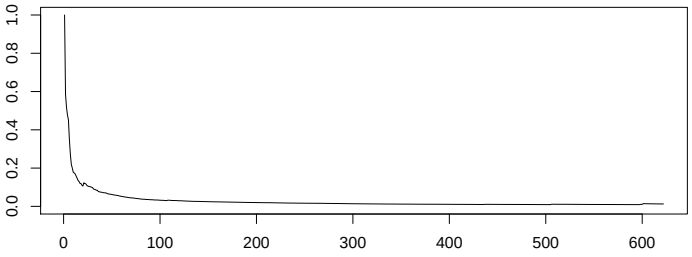

Figure 12: The ratio of maximum and sum for positive log-returns when $p = 1$

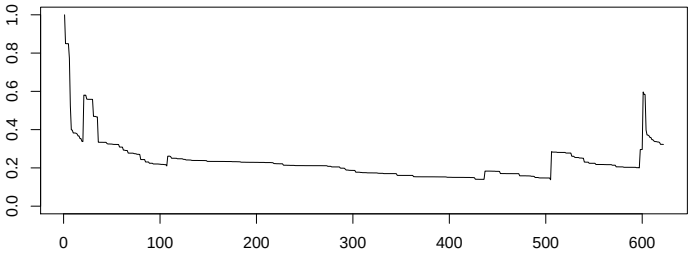

Figure 13: The ratio of maximum and sum for positive log-returns when $p = 5$

4.3 Records test

Before we analyzed the tail weight but what can be said about independence of the sample elements? If our realizations are not independent (and this is of course quite common for returns as clustering is common (bearish or bullish market)), it is useful to analyze the *records*.

In order to introduce the notion of records we need to bring in some additional notation. Let us consider a sequence of random variables, $X_1, X_2, \ldots,$

and denote the maximum of first n random variables by M_n:

$$M_1 = X_1, \quad M_n = \max_{i=1,\dots,n}(X_1, \dots, X_n), \quad n \geq 2.$$

Definition 4.1. Records r_i in sequence $X_1, X_2, \dots$ are defined by:

- $r_1 = X_1$,
- $r_i = X_n$ if $X_n > M_{n-1}$ (where i is the record index).

In other words, a record is a temporary maximum in the sequence of X_n.

The following result proves that records can be a useful tool for our needs: if a sample consists of independent realizations (from a common distribution) then it holds that

$$\mathbb{P}\left(\frac{R_n}{\log(n)} \to 1\right) = 1,$$

where R_n is a number of records for a sample of size n.

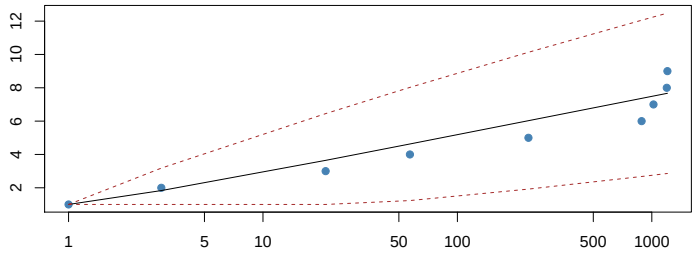

Figure 14: Number of records for the log-returns data

Home assignment 8. Find the distribution of number of records analytically (for an iid sample from a continuous distribution) when the sample size is 1, 2, 3 or 4.

Figure 14 seems to show that we cannot rule out independence. However, it seems a bit problematic that four records are clustered together "at the end of the sample".

4.4 Alternative definition of heavy tails

When a distribution has a heavy tail then (some of) the moments might or might not exist. As we have not previously defined the moment generating function (mgf) now is the time. We shall note the connection between the

existence of moments and the mgf and also show how mgf can be used as a divisor to distinguish light and heavy tailed distributions. In the following we deal with the right tail but the left tail can be handed in a similar fashion.

The moment generating function of a random variable X is the function $M(t) = \mathbb{E}(e^{tX})$.

Definition 4.2. We say that a random variable X has a heavy (right) tail when there is no $t^* > 0$, such that the mgf of X is finite in the range $(-t^*, t^*)$ that is

$$!\exists t^* > 0 : \quad \forall t \in (-t^*, t^*) \quad M(t) = \mathbb{E}(e^{tX}) < \infty.$$

If the random variable does not have a heavy (right) tail then we say that it has a light (right) tail.

Home assignment 9. Show that according to this latest definition exponential distribution with parameter $\lambda > 0$ is light tailed.

Moment generating function of a distribution is important because when it is finite (for some positive argument) then the distribution has finite moments of any order (and these can be easily found using the mgf). However, non-finite mgf does not mean that a distribution cannot have finite moments of any order – we shall see this later (e.g. for Weibull distribution which we encountered in an exercise where tails of some distributions were compared).

5 Creating new probability distributions. Mixture distributions

5.1 Combining distributions. Discrete mixtures

Let us return to figures 8 and 9. Left tail is heavier than the right and a (symmetrical) t-distribution would not fit for the full data. Yet there is an obvious way out – let us make a new distribution where the positive part comes from a t-distribution with 4 degrees of freedom and the negative part from a t-distribution with 3 degrees of freedom. Because of the symmetricity of the density of a t-distribution this can easily be done. Only inconvenience is that the density is no longer continuous at the origin (but it is not a major restriction in practice). As we see from figure 15 the combined distribution fits the data quite well.

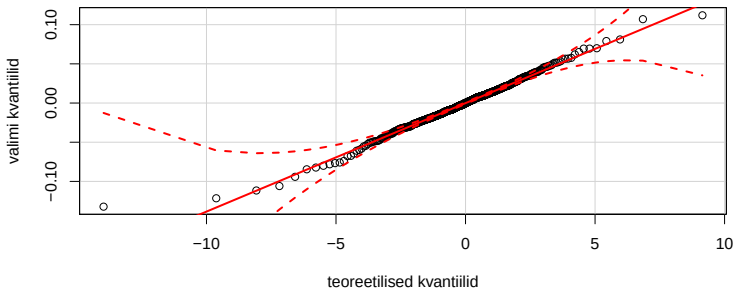

Figure 15: QQ-plot of log-returns and a combined t-distribution

Let us generalize the previous approach.

The easiest way of producing a random variable with (full) real support from a random variable X with support $(0, \infty)$ and density $f_X(x)$ is to use a random variable $-X$, for which we have $f_{-X}(x) = f_X(-x)$. If we define the new random variable Y as a discrete mixture of X and $-X$ then we have accomplished what we wanted to do and we have

$$f_Y(x) = af_X(x) + (1-a)f_X(-x), \quad a \in (0,1).$$

If $a = 1/2$ then we can call this process mirroring because

$$f_Y(x) = \frac{f_X(|x|)}{2}$$

or we simply use the vertical axis for mirroring the density (and later on we normalize). This is how a Laplace distribution is constructed from an

exponential distribution. It is obvious than when X has a heavy right tail then both right and left tail of Y are heavy (similarly light tail of X). Thus the Laplace distribution has light tails. Of course we can also paste together pieces that are not symmetrical (just like we did with the t-distribution).

Just as easy as this pasting operation is the division of a random variable that has support on the whole $\mathbb{R}$. If the density $f_Z(x)$ of a random variable Z is symmetrical then we can easily define a random variable W for which $f_W(x) = 2f_Z(x)I_{(0,\infty)}$.

Discrete mixture can be made from arbitrary distributions (we can even use both discrete distributions and continuous distributions). We can do this by making use of indicator functions that do not overlap. This is known as splicing. But we can use densities that have overlapping support and create, e.g., multimodal distributions.

Let $f_1(x),\ldots,f_i(x),\ldots,f_k(x)$ be density functions with non-overlapping supports $(c_0, c_1),\ldots,(c_{k-1}, c_k)$, respectively. If we take positive constants $a_1, \ldots, a_k$, such that $a_1 + \ldots + a_k = 1$, then a k-component spliced density is

$$f_X(x) = \begin{cases} a_1 f_1(x), & c_0 < x < c_1, \\ a_2 f_2(x), & c_1 < x < c_2, \\ \ldots & \\ a_k f_k(x), & c_{k-1} < x < c_k, \end{cases}$$

with support (c_0, c_k).

Typically the distributions used for splicing do not have such nice densities. Then we must normalize the probability density functions using the cumulative distribution functions. Thus we must replace $f_j(x)$ with the ratio $f_j(x)/[F_j(c_j) - F_j(c_{j-1})]$ for all $j = 1, \ldots, k$.

Most general way of defining a discrete mixture is the following.

Definition 5.1. Let $X_1, X_2, \ldots$ be random variables with respective cumulative distribution functions $F_{X_1}(x), F_{X_2}(x), \ldots$. We say that a random variable Y is a *discrete mixture* of $X_1, X_2, \ldots$ if it holds that

$$F_Y(x) = a_1 F_{X_1}(x) + a_2 F_{X_2}(x) + \ldots,$$

where $a_i \geqslant 0$ and $\sum a_i = 1$.

By the given definition, the derivation of random variable Y from random variables $X_1, X_2, \ldots$ is very straightforward. Thus the moments of Y are also easy to calculate.

5.2 Continuous mixtures

When defining a discrete mixture we do not require that the cumulative distribution functions $F_{X_1}(x), F_{X_2}(x), \ldots$ would be similar in any way. But we could consider a scenario where they would all be members of the same class. E.g. F_{X_i} could be a cumulative distribution function of a normal distribution with parameters μ_i ja $\sigma^2 > 0$ and the weights could be defined as $a_i = \frac{\lambda^i}{i!}e^{-\lambda}$, where $\lambda > 0$. Such construction is the basis for continuous mixtures where summation is replaced with integration and the weights are replaced with a density function.

Definition 5.2. Let us have a family of random variables X_λ such that their distribution depends on some parameter λ, which itself is a realization of some random variable Λ. In other words, for any fixed λ, the (conditional) density of X_λ is given by $f_{X_\lambda}(x)$. Let us assume that the density $f_\Lambda(\lambda)$ of random variable Λ is also known. Now, we can combine the random variables X_λ and Λ to obtain a new random variable X with the following density:

$$f_X(x) = \int f_{X_\lambda}(x) f_\Lambda(\lambda) d\lambda.$$

Random variable X is called a *continuous mixture* of distributions of X_λ and Λ (distribution of Λ is known as the *mixing distribution*).

Notice also that for any fixed λ we have $X_\lambda = (X|\Lambda = \lambda)$.

Also in this case the simulation principle is simple (as long as we know how to simulate both X_λ and Λ). The cdf can be derived as follows

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} \int f_{X_\lambda}(y) f_\Lambda(\lambda) d\lambda dy \\ &= \int \int_{-\infty}^{x} f_{X_\lambda}(y) f_\Lambda(\lambda) dy d\lambda \\ &= \int F_{X_\lambda}(x) f_\Lambda(\lambda) d\lambda. \end{aligned}$$

If the moments are of interest we can write:

$$\begin{aligned} \mathbb{E}(X^k) &= \int x^k \int f_{X_\lambda}(x) f_\Lambda(\lambda) d\lambda dx \\ &= \int \left(\int x^k f_{X_\lambda}(x) dx \right) f_\Lambda(\lambda) d\lambda \\ &= \int \mathbb{E}(X_\lambda^k) f_\Lambda(\lambda) d\lambda \\ &= \mathbb{E}[\mathbb{E}(X^k|\Lambda)]. \end{aligned}$$

Home assignment 10. Consider an exponential distribution with parameter $\lambda > 0$ and probability density function

$$f_1(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

Now let λ have a Γ-distribution with parameters $\alpha > 1$ and $\theta > 0$, i.e. its density function is

$$f_2(x) = \frac{\theta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\theta x}, \quad x > 0.$$

Find the cumulative distribution function, probability density function and mean of this continuous mixture distribution.

5.3 Completely new parts

Theoretically, there is no need to use any particular known distribution when constructing our approximation (perhaps it would be more precise to say "any particular known distribution that we are aware of" because probably somebody somewhere has already used such a distribution). We know that distributions with a continuous cumulative distribution function can be simulated by making use of an uniform distribution and applying the inverse of the cumulative distribution function to (the realization of) this distribution. This also means that we can make continuous distributions applying invertible functions to the uniform distribution. The cumulative distribution function of any such distribution is easy to write down and thus the probability density function is straightforward also.

E.g. we can consider the transformation

$$Y = \frac{(-\ln U)^{-\xi} - 1}{\xi},$$

where U is a random variable with standard uniform distribution and $\xi \in \mathbb{R}$. When $\xi = 0$ then we interpret this as

$$-\ln(-\ln U),$$

noting that $(1 + x/n)^n \to \exp(x)$ and for positive arguments $n(x^{1/n} - 1) \to \ln x$. The support of random variable Y is thus $(-\infty, -1/\xi)$, if $\xi < 0$ and $(-1/\xi, \infty)$ if $\xi > 0$. If $\xi = 0$, the support is the whole real line.

If we want to calculate the moments of Y then we should notice that $-\ln U$ is a random variable with standard exponential distribution. Using the New-

ton's binomial formula we obtain that

$$\begin{aligned}\mathbb{E}Y^n &= \mathbb{E}\left(\frac{(-\ln U)^{-\xi}-1}{\xi}\right)^n \\ &= \mathbb{E}\sum_{k=0}^{n} C_n^k \left[\frac{(-\ln U)^{-\xi}}{\xi}\right]^k \left(-\frac{1}{\xi}\right)^{n-k} \\ &= \sum_{k=0}^{n} C_n^k \left(-\frac{1}{\xi}\right)^n \mathbb{E}\left[(-\ln U)^{-\xi k}\right],\end{aligned}$$

i.e. we need to find the (not necessarily integer) moments of an exponential distribution and this allows us to calculate the moments of the new random variable.

5.4 Parameters of a distribution

In the previous we considered some possibilities of producing "new" distributions to get an approximation that would fit our data well. This idea is nice because using the tools from the preceding an array of very diverse distributions can be generated. There is a danger, however: if the model fits the (training) data well then we might discover that it does not fit the new (test) data. The reason for this is the stochasticity of a sample – every sample has its own particularities. If we include this into the model then the model will not fit any other sample. This is known as over-fitting. So how do we avoid this? By not using too complex models. But also by using common sense. What this means in particular will be discussed in what follows.

But a quick simple example can be given if we think about the previous examples and consider the amount of model parameters involved. For a combined t-distribution we have only two parameters – one for the right and the other for the left tail. But we could also consider that the weights would not be equal (e.g. more gains than losses). Then an additional weight parameter a is required:

$$f_Y(x) = af_{X_1}(x) + (1-a)f_{X_2}(x), \quad a \in (0,1).$$

and there are 3 parameters in total. Now perhaps we would like to allow that the "pasting point" (the point where the two pieces are joined) is not necessarily the origin so another parameter would then have to be added. This would mean 4 parameters and all of them should then (typically) be estimated from the data.

Home assignment 11. How many parameters are present in a (general) discrete mixture of five normal distributions? What is the minimal amount of parameters that can be retained in particular context?

6 Empirical distribution. Kernel density estimation

Let us recall that in our log-returns example we concluded that a normal distribution is not a suitable model. To illustrate this, densities were compared: we estimated the parameters of a normal distribution based on the data but we also used the data to estimate the density directly (using kernel density estimation). We will now pay more attention to the latter as it seems like a very appealing method – assumptions are minimal and we only "let the data speak". To explain the ideas of kernel density estimation we will first recall the concept of empirical distribution.

Empirical distribution is the distribution based on a sample and its cumulative distribution function can be found as follows.

Definition 6.1 (Empirical cumulative distribution function)**.** *Empirical cumulative distribution function* of a sample $x_1, \dots, x_n$ is defined as

$$F_n(x) = \frac{\{\#x_i : x_i \leqslant x\}}{n}.$$

The empirical cumulative distribution function is a step function like the cumulative distribution functions of discrete distributions. The reason for this is the fact that for any finite sample size n the empirical distribution is discrete.

Now, if we need to visualize the underlying theoretical density based on a sample, a natural starting point would be the probability mass function of the sample. Unfortunately, such graph would only consist of bars with equal height $\frac{1}{n}$ (if there are no duplicates in the sample). Obviously, more probable regions have more bars, but it is still hard to understand the specifics of the underlying distribution from such graph. Therefore, certain aggregation of some sample elements might be a reasonable choice.

We consider the following setup. First we find the sample minimum and maximum. Then we divide the interval between them into subintervals with equal length (the left-most is a closed interval and the others do not include the left endpoint). We count for each interval how many values fall into the interval and represent these frequencies by bars (more frequent classes are represented by higher bars). If the sample size and the number of intervals are both large then this graph (histogram) reminds us a density function which is divided into classes. This is how Figure 6 has been obtained based on the positive log-returns sample. Actually such a graph does not approximate a density function because the area under the graph is not equal to one. But if we were to increase the number of intervals then both would have the same shape (but not scale).

So, if we would like the histogram to represent or estimate the density, we need to transform the scale. But this is actually not the goal of a histogram, it is usually sufficient to get the idea about the shape.

Example 6.1. Let the interval between sample minimum and sample maximum be divided into k subintervals. Let the lengths of these intervals be $w_1, \ldots, w_k$ and respective probabilities $p_1, \ldots, p_k$. If we now draw the bars with respective heights $\frac{p_1}{w_1}, \ldots, \frac{p_k}{w_k}$, the area under the graph would be one, thus giving us an estimate for the density. An example of this is the figure 16.

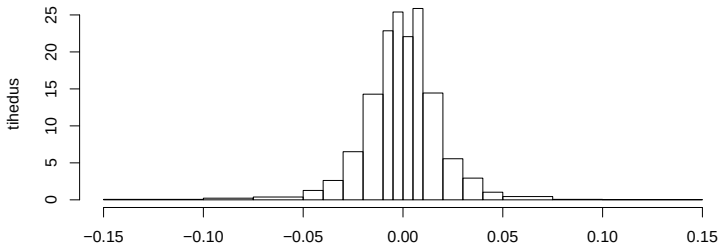

Figure 16: Density estimate of the log-returns data

Consider an empirical distribution with sample points $x_1, \ldots, x_n$ and respective probabilities $p(x_1) = \ldots = p(x_n) = \frac{1}{n}$. When we draw a histogram with unit area then these probabilities are "spread out". Namely, if a subinterval with length w contains k sample elements then the density estimate for that subinterval is $\frac{k}{nw}$ (regardless of how where these k points lie in this subinterval). This "spreading out" is also the basis of kernel density estimation.

Basically, we substitute the (fixed) probabilities of the empirical distribution by densities of continuous random variables, while (typically) keeping the means of these random variables where the actual sample elements occurred. (Notice that drawing a histogram can also be described as replacing the empirical probabilities by densities of uniform distributions, but we *do not keep the mean*.

Definition 6.2 (Kernel density estimate)**.** A *kernel density estimate* is constructed as

$$\hat{f}(x) = \sum_{j=1}^{n} p(x_j) K_h(x - x_j) = \frac{1}{h} \sum_{j=1}^{n} p(x_j) K(\frac{x - x_j}{h})$$

where $K(x)$ is called the *kernel* function and $K_h(x)$ is known as *scaled kernel*. The parameter h is fixed by the user and is known as *smoothing bandwidth*.

Thus, a scaled kernel is obtained from the "original" kernel using the following scaling:

$$K_h(u) = \frac{1}{h}K(\frac{u}{h}).$$

The kernel $K(u)$ itself is typically (but not always) a symmetrical density function with zero mean.

Definition 6.3 (Uniform (rectangular) kernel)**.** *Uniform kernel* or *rectangular kernel* is defined as

$$K(u) = \frac{1}{2}I_{|u|\leqslant 1},$$

and as such the uniform kernel distributes the probability equally around the original location.

Because a smoothing bandwidth 1 would be too large for our data, we will be using percent scale in the next figures (i.e., we plot the graph of *logreturns*100*).

Home assignment 12. Find analytically how many times the standard deviation of $K(x)$ increases when the smoothing bandwidth increases $h > 0$ times.

Home assignment 13. Study the command *?density* in R to learn how to set the smoothing bandwidth in R. How could we use the parameter *adjust* so that we would not need to transform the log-returns data into percent scale?

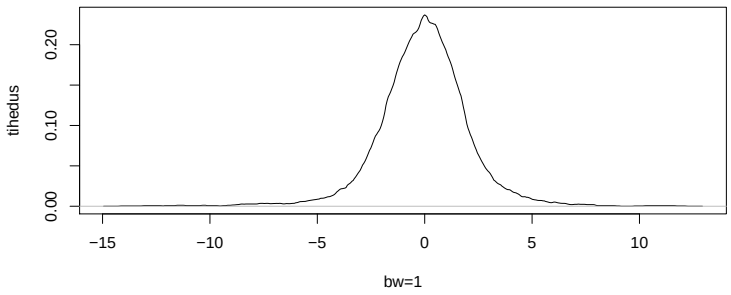

Figure 17: Kernel density estimate of log-returns using the uniform kernel (in percent scale)

Definition 6.4 (Triangular kernel)**.** *Triangular kernel* is defined as

$$K(u) = (1 - |u|)I_{|u|\leqslant 1},$$

thus it divides the probability so that there is linear decrease in both directions.

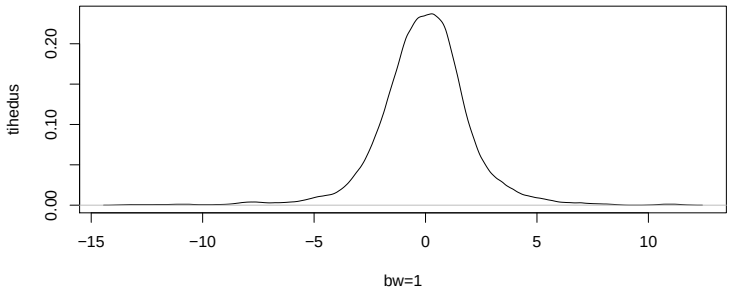

Figure 18: Kernel density estimate of log-returns using the triangular kernel (in percent scale)

Definition 6.5 (Epanechnikov kernel)**.** *Epanechnikov kernel* is defined as

$$K(u) = \frac{3}{4}(1 - u^2)I_{|u|\leqslant 1},$$

thus it divides the probability so that there is quadratic decrease in both directions.

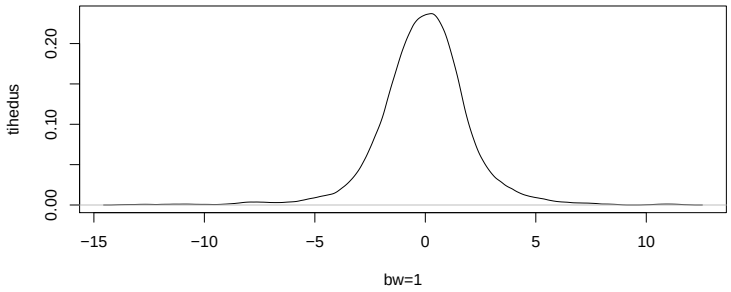

Figure 19: Kernel density estimate of log-returns using the Epanechnikov kernel (in percent scale)

Definition 6.6 (Gaussian kernel)**.** *Gaussian kernel* is defined as

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}u^2\right\},$$

thus it replaces the probability with a density of a normal distribution.

Thus we get a density estimate with bounded support if and only if the kernel itself has bounded support. But with unbounded kernel the tail weight (of the

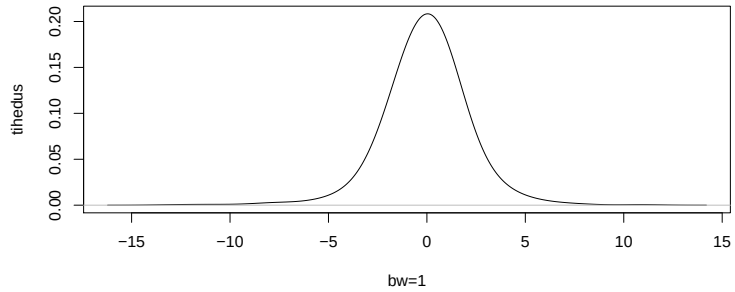

Figure 20: Kernel density estimate of log-returns using the Gaussian kernel (in percent scale)

density estimate) is defined by the tail weight of the kernel. So we basically make an arbitrary choice for the tail. For bounded kernels the bandwidth is much more important than the kernel itself. If h is very large then all the particularities of the sample will be smoothed out while a small h does not do enough smoothing and the result is graphically unconvincing.

Home assignment 14. Why do we often use zero-mean kernels in practice?

In R, instead of defining the actual bandwidth it is possible to calculate the optimal bandwidth using an algorithm. Choices *nrd0* and *nrd* are rules-of-thumb which assume that the distribution is similar to a normal distribution (unimodal, symmetric, light-tailed): Gaussian kernel will be used and the goodness of the result is defined by the integrated mean square error, that is

$$\int \left\{ \mathbb{D}(\hat{f}(x)) + [\mathbb{E}(\hat{f}(x)) - f(x)]^2 \right\} dx,$$

where $\hat{f}(x)$ is the kernel density estimate. The rule for finding the bandwidth is

$$c \cdot \min\{\hat{\sigma}, \frac{R}{1.34}\} n^{-1/5},$$

where $c = 0.9$ when we use *nrd0* and $c = 1.06$ when we use *nrd*, σ is the sample standard deviation and R the inter-quartile range.

Choices *ucv* and *bcv* use cross-validation and do not make assumptions about the shape of the original density (which produced our sample), and are thus a bit more general but they cannot be fully compared because the criterion for optimality is different. The choice *SJ* is also distribution free but the optimality criterion is minimal integrated mean square error.

In conclusion: it makes sense to use different criteria for defining the optimal smoothing bandwidth and compare the result obtained using the respective bandwidths.

7 Subexponential distributions

7.1 Preliminaries. Definition

Let us now consider distributions with support $(0, \infty)$. Positive random variables are very popular in applications (size of an insurance claim, number of claims etc) and it is often possible to represent the underlying random variable as a difference between two positive random variables (profit can be expressed as a difference between turnover and expenses).

In particular, we concentrate on a sub-class of heavy-tailed distributions: subexponential distributions. Each distribution in this class has many nice properties, yet the class is rich enough to include a large variety of models. In the following we study the properties of this class in more detail.

Let us first recall the concept of convolution of distributions.

Remark 7.1 (Convolution of distributions)**.** Let $X_i,\ i = 1, \ldots, k$ be independent random variables with distributions P_i, then the distribution of $X_1 + \ldots + X_n$ (say, P) is called the convolution of distributions P_i. Similar notion is used for distribution functions and probability density functions: the distribution function $F_{X_1+\ldots+X_n}$ is called the convolution of (independent) distribution functions F_{X_i} and denoted by

$$F_{X_1+\ldots+X_n} = F_{X_1} * F_{X_2} * \ldots * F_{X_n}.$$

If X_i-s have same distribution, the corresponding convolution is denoted by F^{*n}.

For two independent random variables X and Y:

- in general: $F_{X+Y}(s) = \int F_X(s-y)dF_Y(y)$;
- if X is continuous: $f_{X+Y}(s) = \int f_X(s-y)dF_Y(y)$;
- if both X and Y are continuous: $f_{X+Y}(s) = \int f_X(s-y)f_Y(y)dy$.

Definition 7.1 (Subexponential distributions)**.** Class

$$\mathcal{S} = \{X > 0 : \forall n \in \mathbb{N} \ \lim_{x\to\infty} \frac{\overline{F_X^{*n}}(x)}{\overline{F_X}(x)} = n\}$$

is called the *class of subexponential distributions*. Here $F^{*n}(x)$ denotes the n-fold convolution of a function $F(x)$.

The name of the class is due to the fact that if $X \in \mathcal{S}$ then $\forall \epsilon > 0$

$$\lim_{x\to\infty} e^{\epsilon x}\overline{F_X}(x) = \infty,$$

which means that the tail function goes to zero at a slower rate than any exponential function of the form $e^{-\epsilon x}$, where $\epsilon > 0$.

Definition 7.2 (Asymptotic equivalence (tail equivalence))**.** Let $F(x)$ and $G(x)$ be any two functions. We write $F(x) \sim G(x)$ and say that $F(x)$ and $G(x)$ are *asymptotically equivalent* if

$$\lim_{x\to\infty} \frac{F(x)}{G(x)} = 1.$$

7.2 Properties of subexponential class

To give some intuition about the class $\mathcal{S}$ let us consider a sequence $X_1, \ldots, X_n \in \mathcal{S}$ of *iid* random variables with cumulative distribution function $F(x)$. Let us also define $S_n = X_1 + \ldots + X_n$ and $M_n = \max(X_1, \ldots, X_n)$. Now, since $\mathbb{P}(M_n > x) = 1 - [F(x)]^n$ and because of the asymptotic equivalence $1 - [F(x)]^n \sim n\bar{F}(x)$, we obtain

$$\mathbb{P}(M_n > x) \sim \mathbb{P}(S_n > x).$$

Thus for subexponential distributions "a large sum is due to one large summand", which is very relevant in practice. Imagine the arrival of insurance claims. When the aggregate claim is very large then typically this is because of a single very large claim. This in turn means that there is no time to react because we do not know in advance when the large claim might occur.

Let $X_1 \in \mathcal{S}$ with cumulative distribution function $F_{X_1}(x)$ and $X_2 > 0$ with cumulative distribution function $F_{X_2}(x)$ be a random variable with a lighter tail i.e.

$$\lim_{x\to\infty} \frac{\overline{F_{X_1}}(x)}{\overline{F_{X_2}}(x)} = \infty.$$

Also let X_1 and X_2 be independent. Then

$$X_1 + X_2 \in \mathcal{S}$$

and

$$\overline{F_{X_1+X_2}}(x) \sim \overline{F_{X_1}}(x)$$

So a subexponential tail is still dominant in the sum.

Let $X_1 \in \mathcal{S}$ with cumulative distribution function $F_{X_1}(x)$ and $X_2 > 0$ with cumulative distribution function $F_{X_2}(x)$, such that for some $c > 0$

$$c\overline{F_{X_1}}(x) \sim \overline{F_{X_2}}(x).$$

Then also $X_2 \in \mathcal{S}$ and if X_1 and X_2 are independent then

$$X_1 + X_2 \in \mathcal{S}$$

and

$$\overline{F_{X_1+X_2}}(x) \sim (1 + c)\overline{F_{X_1}}(x)$$

So the class is closed with respect to addition.

Let $X_1, \ldots, X_n \in \mathcal{S}$ be *iid* random variables with cumulative distribution function $F(x)$. From previous we know that

$$\mathbb{P}(M_n > x) \sim \mathbb{P}(S_n > x) \sim n\bar{F}(x),$$

but due to the previous property we have that $S_n \in \mathcal{S}$ and $M_n \in \mathcal{S}$, so the class is also closed with respect to maximum.

Note that if the limit property in the definition of subexponentiality holds for some $n \geq 2$, i.e.

$$\lim_{x\to\infty} \frac{\overline{F^{*n}}(x)}{\overline{F}(x)} = n,$$

then the result holds for every $n \geq 2$. Thus, to check whether a random variable X belongs to the class of subexponential distributions, $X \in \mathcal{S}$, it is sufficient to find only one limit. But even this can be complicated in reality.

Often a Pitman condition is used instead to check if a distribution is in $\mathcal{S}$: Let $X > 0$ be a random variable with probability density function $f(x)$ and hazard function $h(x)$, which is eventually decreasing and for which we have $\lim_{x\to\infty} h(x) = 0$. If

$$\int_0^\infty e^{xh(x)} f(x)dx < \infty,$$

then $X \in \mathcal{S}$.

Home assignment 15. Let X_1 and X_2 be random variables with respective probability density functions $f_1(x)$ and $f_2(x)$. Show that $X_1, X_2 \in \mathcal{S}$.

1. $f_1(x) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}},\ \alpha > 0,\ \theta > 0,\ x > 0$

2. $f_2(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k},\ \lambda > 0,\ 0 < k < 1,\ x > 0$

It is allowed to take $\theta = 1$ and $\lambda = 1$. Using the knowledge about the tails, why does it make sense that a random variable X_3 with probability density function

$$f_3(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}},\ \mu \in \mathbb{R},\ \sigma > 0,\ x > 0,$$

is a subexponential distribution?

8 Well-known distributions in financial and insurance mathematics

In this section we recall some common distributions used in financial and insurance mathematics.

8.1 Exponential distribution

Exponential distribution is a widely used distribution in the theory of stochastic processes – it is the probability distribution that describes the time between events in a Poisson process (process in which events occur continuously and independently at a constant average rate).

For an exponentially distributed random variable X we write $X \sim \mathcal{E}(\lambda)$ (or $X \sim Exp(\lambda)$), where $\lambda > 0$ is a rate (or inverse scale) parameter.

The key property of exponential distribution is being *memoryless*. Only exponential and geometric distributions have the property of being memoryless:

$$\mathbf{P}\{X \geq t + s | X \geq s\} = \mathbf{P}\{X \geq t\}.$$

Main characteristics for the exponential distribution are:

- probability density function $f_X(x) = \lambda e^{-\lambda x},\ x \geq 0$,
- cumulative distribution function $F_X(x) = 1 - e^{-\lambda x},\ x \geq 0$,
- expectation $\mathbb{E}X = \frac{1}{\lambda}$,
- variance $VarX = \frac{1}{\lambda^2}$,
- mode argmax $f(x) = 0$,
- median med $X = \frac{\ln 2}{\lambda}$,
- the sum of independent exponential random variables with same parameter λ is a gamma-distributed random variable: $X_1, \ldots, X_n \sim \mathcal{E}(\lambda) \Rightarrow \sum_{k=1}^{n} \sim \Gamma(n, \lambda)$,
- the minimum of independent exponential random variables is also exponentially distributed: $X_1, \ldots, X_n \sim \mathcal{E}(\lambda) \Rightarrow \min\{X_1, \ldots, X_n\} \sim \mathcal{E}(\lambda_1 + \ldots + \lambda_n)$.

The following list contains few examples of applications where exponential distribution can be used.

- The exponential distribution occurs naturally when describing the lengths of the inter-arrival times in a homogeneous Poisson process. It describes the time for a continuous process to change state.
- In situations where certain events occur with a constant probability per unit length, for example, number of phone calls received in fixed time period.
- In queuing theory, the service times of agents in the system are often modeled as exponentially distributed variables.
- In physics, if you observe a gas at a fixed temperature and pressure in a uniform gravitational field, the heights of the various molecules also follow as approximate exponential distribution, known as Barometric formula.
- In hydrology, the exponential distribution is used to analyze extreme values of such variables as monthly and annual maximum values of daily rainfall and river discharge volumes.

8.2 Pareto distribution

The Pareto distribution is a power law probability distribution that is used in several models in economics and social sciences. The corresponding distribution family is a wide one, with several sub-families. We will first review the classical Pareto distribution and then focus on it's shifted version (so-called *American Pareto* distribution), which is most widely used in non-life insurance as a model for claim severity. The distribution is named after italian sociologist and economist Vilfredo Pareto (who is also the author of the 80-20 principle). Nowadays Pareto distribution is widely used in description of social, scientific, geophysical, actuarial, and many other types of observable phenomena

A. The classical Pareto, $Z \sim Pa^*(\alpha, \beta),\ \alpha,\ \beta > 0$

Main characteristics for the classical Pareto distribution are:

- probability density function $f_Z(z) = \frac{\alpha\beta^\alpha}{z^{\alpha+1}},\ z \geq \beta,$
- cumulative distribution function $F_Z(z) = 1 - \left(\frac{\beta}{z}\right)^\alpha,\ z \geq \beta,$
- expectation $\mathbb{E}Z = \frac{\alpha\beta}{\alpha-1},\ \alpha > 1,$
- variance $VarZ = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)},\ \alpha > 2,$
- moments $\mathbb{E}Z^n = \frac{\alpha\beta^n}{\alpha-n},\ \alpha > n,$ but $EZ^n = \infty,\ \alpha \leq n.$

B. American Pareto, $X \sim Pa(\alpha, \beta)$, $X = Z - \beta$, $Z \sim Pa^*(\alpha, \beta)$, $\alpha,\ \beta > 0$

The American Pareto distribution is obtained from classical Pareto distribution by shifting it to the origin.

Main characteristics for the American Pareto distribution are:

- probability density function $f_X(x) = \frac{\alpha\beta^\alpha}{(\beta+x)^{\alpha+1}},\ x \geq 0,$
- cumulative distribution function $F_X(x) = 1 - \left(\frac{\beta}{\beta+x}\right)^\alpha,\ x \geq 0,$
- expectation $\mathbb{E}X = \frac{\beta}{\alpha-1},\ \alpha > 1,$
- variance $VarX = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)},\ \alpha > 2,$
- moments $\mathbb{E}X^n = \frac{\beta^n n!}{\prod_{i=1}^n(\alpha-i)},\ \alpha > n$, but $\mathbb{E}X^n = \infty$, $\alpha \leq n,$
- $U \sim U(0,1) \Rightarrow \beta(U^{-1/\alpha} - 1) \sim Pa(\alpha, \beta),$
- mode argmax $f(x) = 0,$
- median med $X = \beta(2^{1/\alpha} - 1).$

Parameter estimation using the method of moments:

$$\hat{\alpha} = \frac{2(m_2 - m_1^2)}{m_2 - 2m_1^2},$$

$$\hat{\beta} = \frac{m_1 m_2}{m_2 - 2m_1^2},$$

where $m_1 = \sum_{j=1}^n \frac{x_j}{n}$ and $m_2 = \sum_{j=1}^n \frac{x_j^2}{n}$.

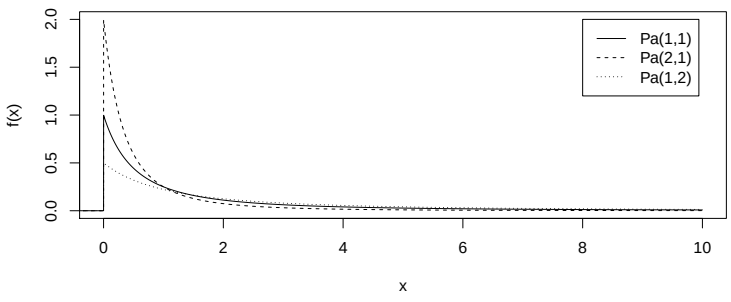

Figure 21: Pareto density under several parameter combinations

Home assignment 16. Suppose we have an *iid* sample $X_1, \ldots, X_n$ from $Pa(\alpha, \beta)$. What is the distribution of $\min\{X_1, \ldots, X_n\}$?

Pareto distribution is not limited to describing wealth or income, it is widely used in describing social, scientific, geophysical and many other types of observable phenomena, for example:

- the size of human settlement (many villages, few cities),
- file size of internet traffic that uses the TCP protocol,
- hard disk drive error rates,
- values of oil reserves in oil fields,
- standardized price returns on individual stocks,
- sizes of sand particles,
- sizes of meteorites,
- areas burnt in forest fires.

8.3 Weibull Distribution

The Weibull distribution is of interest in various fields, ranging from life data to weather data or observations made in economics and business administration, in hydrology, in biology or in the engineering science. The distribution is named after the Swedish physicist Waloddi Weibull who used it in reliability testing of materials.

For a random variable X that follows Weibull distribution we write $X \sim We(k, \lambda)$, where $k > 0$ is a shape parameter and $\lambda > 0$ is a scale parameter.

The main characteristics of the distribution are

- probability density function $f_X(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}$, $x \geq 0$,
- cumulative distribution function $F(x) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k}$,
- $\mathbb{E}(X^n) = \lambda^n \Gamma(1 + \frac{n}{k})$,
- $\mathbb{E}|X|^n < \infty$,
- $U \sim U(0,1) \Rightarrow \lambda(-\ln U)^{1/k} \sim We(k, \lambda)$,
- mode argmax $f(x) = \lambda\left(\frac{k-1}{k}\right)^{1/k} I_{\{k>1\}}$,
- median med $X = \lambda(\ln 2)^{1/k}$.

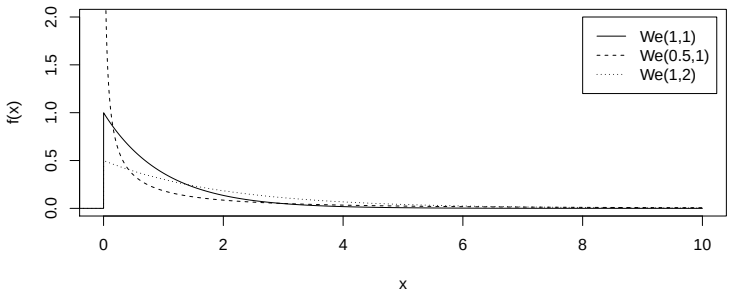

Figure 22: Weibull density under several parameter combinations

We also note that if

$k < 1$, then Weibull distribution is "between" exponential and Pareto;
$k > 1$, then Weibull distribution has lighter tail than exponential;
$k = 1$, then Weibull distribution reduces to exponential.

Remark. In order to stress the relation to exponential distribution, Weibull distribution is often parametrized by k and $\theta = \frac{1}{\lambda}$, implying the following:

- cumulative distribution function $F_X(x) = 1 - e^{-\theta x^k}$,
- probability density function $f_X(x) = k\theta^k x^{\alpha-1} e^{-\theta x^k}$.

8.4 Lognormal distribution

Lognormal distribution plays the role of normal distribution in the multiplicative central limit theorem (we multiply *iid* random variables). When the logarithm of a random variable has a normal distribution then the random variable itself has a lognormal distribution. Thus, the distribution function of lognormal distribution is found using the log transformation to reach normal distribution and standardization to reach standard normal distribution. Since the normal distribution is one of the most thoroughly studied distributions, the simple connection between lognormal and normal makes lognormal distribution also an appealing choice in different models.

For a lognormally distributed random variable X we write $X \sim LnN(\mu, \sigma)$ (or $X \sim Ln(\mu, \sigma)$), where $-\infty < \mu < \infty$ is a location parameter and $\sigma > 0$ is a scale parameter.

Main characteristics are:

- its connection to normal distribution, if $Y \sim N(\mu, \sigma)$ and $X = e^Y$, then $X = LnN(\mu, \sigma)$,

- cumulative distribution function $F_X(x) = F_Y(\ln x) = \Phi(\frac{\ln x - \mu}{\sigma})$,
- probability density function $f_X(x) = \frac{1}{x} f_Y(\ln x)$,
- expectation $\mathbb{E}X = e^{\mu + \frac{\sigma^2}{2}}$,
- variance $VarX = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$,
- moments $\mathbb{E}X^n = e^{n\mu + \frac{1}{2}n^2\sigma^2}$,
- $\mathbb{E}|X|^n < \infty$,
- $Y \sim N(0,1) \Rightarrow e^{\mu+\sigma Y} \sim LnN(\mu, \sigma^2)$,
- mode argmax $f(x) = e^{\mu-\sigma^2}$,
- median med $X = e^{\mu}$.

Parameter estimation:

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} \ln x_j,$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{j=1}^{n} [\ln(x_j - \hat{\mu})]^2.$$

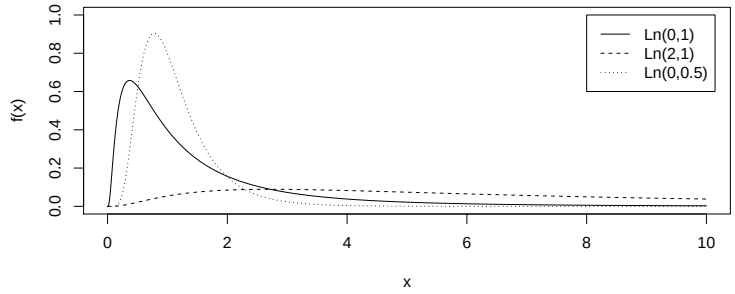

Figure 23: Log-normal density under several parameter combinations

8.5 Log-gamma distribution

A random variable is log-gamma distributed if its natural logarithm is gamma-distributed. Notice that as a gamma-distributed random variable can only take non-negative values, the support of log-gamma distribution is $[-1; \infty)$

For a log-gamma distributed random variable we write $X \sim Lg(\alpha, \beta)$, where $\alpha > 0$ is the rate parameter (of corresponding gamma distribution) and $\beta > 0$ is a shape parameter.

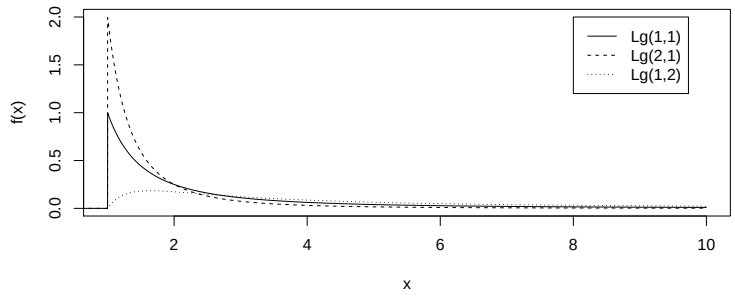

Figure 24: Log-gamma density under several parameter combinations

The main characteristics of the log-gamma distribution are:

- probability density function $f_X(x) = \frac{\alpha^\beta (\ln x)^{\beta-1}}{\Gamma(\beta) x^{\alpha+1}}$, $x \geq 1$,
- cumulative distribution function $F(x)$ does not have analytic expression,
- $\mathbb{E}(X^n) = \left(\frac{\alpha}{\alpha - n}\right)^\beta$,
- $\mathbb{E}|X|^n < \infty \Leftrightarrow n < \alpha$,
- $Y \sim \Gamma(\alpha, \beta) \Rightarrow e^Y \sim Lg(\alpha, \beta)$,
- mode argmax $f(x) = e^{\frac{\beta-1}{\alpha+1} I_{\{\beta>1\}}}$,
- median med X does not have analytic expression.

8.6 Burr distribution

In probability theory, statistics and econometrics, the Burr Type XII distribution or simply the Burr distribution is a continuous probability distribution for a non-negative random variable. It is also known as the Singh-Maddala distribution and is one of a number of different distributions, sometimes called the "generalized log-logistic distribution", as it contains the log-logistic distribution as a special case. It is most commonly used to model household income. The distribution is named after american statistician Irving W. Burr.

For a Burr distributed random variable X we write $X \sim Bu(c, \alpha)$, where $c > 0$ and $\alpha > 0$.

Main characteristiscs of the Burr distribution are:

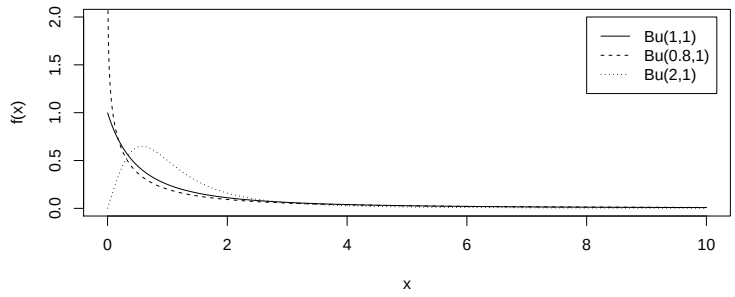

Figure 25: Burr density under several parameter combinations

- probability density function $f_X(x) = \frac{\alpha c x^{c-1}}{(1+x^c)^{\alpha+1}}$, $x \geq 0$,
- cumulative distribution function $F_X(x) = 1 - \left(\frac{1}{1+x^c}\right)^{\alpha}$,
- moments $\mathbb{E}(X^n) = \frac{\alpha\Gamma(\alpha - n/c)\Gamma(1+n/c)}{\Gamma(\alpha+1)}$,
- $\mathbb{E}|X|^n < \infty \Leftrightarrow n < c\alpha$,
- $U \sim U(0,1) \Rightarrow (U^{-1/\alpha} - 1)^{1/c} \sim Bu(c, \alpha)$,
- mode argmax $f(x) = \frac{c-1}{(c\alpha+1)^{1/c}} I_{\{c>1\}}$,
- median med $X = \left(2^{1/\alpha} - 1\right)^{1/c}$.

Home assignment 17. Let $X \sim Bu(c, 1)$, find the distribution of $\frac{1}{X}$.

9 Introduction to the extreme value theory (EVT)

9.1 Preliminaries. Max-stable distributions

Previously we have considered some heavy-tailed distributions which should fit financial data. Justification is based on the quality of the fit, typically there are no theoretical grounds (e.g. we probably do not believe that the t-distribution based model is actually the one generating the log-returns) and this is very common in modelling practice. We introduce the tools of EVT which actually have theoretical justification as models in certain (and not uncommon) situations.

Definition 9.1. We say that two random variables X and Y (and the corresponding distributions and distribution functions) are of the same type if there exist constants $a > 0$ and $b \in \mathbb{R}$ such that

$$Y \stackrel{d}{=} aX + b.$$

In other words, if $X \sim F_1(x)$ and $Y \sim F_2(x)$ then

$$\forall x \quad F_1(x) = F_2(ax + b).$$

Based on this definition the random variables X and $aX + b$ are *of the same type*, i.e. changing the scale and/or location parameter does not change the type of the distribution.

Home assignment 18. Are two normal distributions with different parameters of the same type? What about two Pareto distributions?

Definition 9.2. A non-degenerate random variable X (and the corresponding distributions and distribution functions) is called *max-stable* if

$$\max(X_1, \ldots, X_n) \stackrel{d}{=} c_n X + d_n$$

for *iid* $X, X_1, \ldots, X_n$, appropriate constants $c_n > 0$, $d_n \in \mathbb{R}$ and every $n \geq 2$.

For corresponding distribution function $F(x)$ this means

$$[F(x)]^n = F(c_n x + d_n).$$

So, a max-stable distribution is a distribution for which the maximum of an *iid* sample has distribution of the same type.

Theorem 9.1 (Convergence to types theorem)**.** *Let $X_1, X_2, \ldots$ be random variables and $a_n > 0$, $\alpha_n > 0$, b_n, $\beta_n \in \mathbb{R}$ be constants. Let us also assume that X and Y are non-degenerate and suppose the following holds*

$$\frac{X_n - b_n}{a_n} \stackrel{d}{\to} X.$$

Then

$$\frac{X_n - \beta_n}{\alpha_n} \xrightarrow{d} Y \tag{9.1}$$

holds if and only if

$$\frac{a_n}{\alpha_n} \to a \in (0, \infty), \qquad \frac{b_n - \beta_n}{\alpha_n} \to b \in \mathbb{R}.$$

If (9.1) holds then $Y \overset{d}{=} aX + b$*, i.e.* X *and* Y *are of the same type.*

This means that when we change the normalizing constants then we might obtain a different limiting distribution (i.e. different parameters) but this distribution is of the same type as the initial limiting distribution.

Let us now consider a common problem: given *iid* $X_1, \ldots, X_n$ we need to study the behaviour of $M_n = \max\{X_1, \ldots, X_n\}$. If $\mathbb{P}(X_i \leqslant x) = F(x)$, we have $\mathbb{P}(M_n \leqslant x) = [F(x)]^n$.

We can see that $M_n \to x_F = \sup\{x \mid F(x) < 1\}$ and the limiting distribution is degenerate.

A more viable option is to consider certain normalized random variable of the same type:

$$\frac{M_n - b_n}{a_n}$$

and study whether a sequence of such variables can have a non-degenerate limit. Let us call this limiting distribution $G(x)$ an *extreme value distribution.*

Remark 9.1. Recall that for sum $S_n = \sum_{i=1}^n X_i$ we also have $S_n \to \infty$, yet due to a normalization we still can obtain a non-degenerate limit. By subtracting the mean and dividing by the standard deviation we make sure that the expectation and variance of S_n do not grow out of control.

Theorem 9.2 (Max-stable distributions are weak limits of maxima)**.** *A non-degenerate distribution* $G(x)$ *is max-stable distribution if and only if it is an extreme value distribution (i.e. it is a limit of distributions for* $\frac{M_n - b_n}{a_n}$ *where* $M_n = \max(X_1, \ldots, X_n)$*,* $X_i \sim F$ *are* iid *random variables,* $a_n > 0$ *and* $b_n \in \mathbb{R}$*).*

Proof. A. Extreme value distributions are max-stable.

Suppose that $\mathbb{P}(\frac{M_n - b_n}{a_n} \leqslant x) \to G(x)$. Let us look at N independent samples from distribution F, each with n elements. So we have Nn elements from F and can derive

$$\mathbb{P}\left(\frac{M_{Nn} - b_n}{a_n} \leqslant x\right) = [\mathbb{P}(M_n \leqslant a_n x + b_n)]^N \to [G(x)]^N.$$

On the other hand, we can also write

$$\mathbb{P}\left(\frac{M_{Nn} - b_{Nn}}{a_{Nn}} \leqslant x\right) \to G(x).$$

Now, applying the convergence of types theorem, it follows that $[G(x)]^N = G(c_N x + d_N)$, where $a_n/a_{Nn} \to c_N > 0$ and $(b_n - b_{Nn})/a_{Nn} \to d_N$. So, an extreme value distribution is necessarily max-stable.

B. Max-stable distributions are extreme value distributions.

Now, let us consider a random variable Z with max-stable distribution $G(x)$. Let $X_1, \ldots, X_n$ be *iid* copies of Z. Then, by definition of max-stable distributions:

$$\max(X_1, \ldots, X_n) \stackrel{d}{=} c_n Z + d_n$$

for some $c_n > 0$, $d_n \in \mathbb{R}$ and every $n \geq 2$. Now, this implies that

$$\frac{M_n - d_n}{c_n} \stackrel{d}{=} Z,$$

and obviously Z is the limit of this sequence, which means that Z has extreme value distribution. □

9.2 Forms of max-stable distributions

Let us find the explicit forms of distribution functions of max-stable distributions. We depart from the equation $[G(x)]^N = G(c_N x + d_N)$, which holds for $\forall N \in \mathbb{N}$.

Let us first assume that $c_N = 1$ (for $\forall N$). In that case

$$[G(x)]^{NM} = [G(x + d_N)]^M = G(x + d_N + d_M)$$

and, on the other hand

$$[G(x)]^{NM} = G(x + d_{NM}),$$

which together imply that $d_N = K_1 \ln N$. Now taking a logarithm twice (note that $\ln G(x) \leqslant 0$) we get

$$\ln N + \ln(-\ln G(x)) = \ln(-\ln G(x + K_1 \ln N)),$$

so if the argument is increased by $K_1 \ln N$ then the function value is increased by $\ln N$. Thus

$$\ln(-\ln G(x)) = K_2 + \frac{x}{K_1},$$

where K_1 must be negative, because $\ln(-\ln G(x))$ is a decreasing function. Now the distribution function is given by $G(x) = \exp\Big(-\exp(\frac{x}{K_1} + K_2)\Big)$, which is of the same type as

$$G(x) = e^{-e^{-x}}.$$

Thus we have found the first type of max-stable distributions.

Now, let us start again with the equation $[G(x)]^N = G(c_N x + d_N)$, which holds for $\forall N \in \mathbb{N}$. If $c_N \neq 1$ then for $x = d_N(1 - c_N)^{-1}$ it holds that $x = c_N x + d_N$, i.e. $[G(x)]^N = G(x)$ holds. This implies that

$$G(d_N(1 - c_N)^{-1}) = \begin{cases} 0 \\ 1. \end{cases}$$

Let us first consider $G(d_N(1 - c_N)^{-1}) = 1$. Then the chosen point x must be the right endpoint of the distribution, let us denote it by x_F. We can take $x_F = 0$, because such a distribution $G(x)$ certainly exists in the family of distributions and we are only interested in the general form of the distribution. So $d_N = x_F(1 - c_N) = 0$. Now simple derivations

$$[G(x)]^{NM} = [G(c_N x)]^M = G(c_M c_N x)$$

and

$$[G(x)]^{NM} = G(c_{NM} x)$$

lead us to $c_N = N^{K_1}$. Taking the logarithm twice from the departing equation (note again that $\ln G(x) \leqslant 0$) we get

$$\ln N + \ln(-\ln G(x)) = \ln(-\ln G(N^{K_1} x))$$

or when the argument is increased by N^{K_1} times then the value of the function is increased by $\ln N$. Thus

$$\ln(-\ln G(x)) = \ln \sqrt[K_1]{xK_2},$$

where $K_2 < 0$ (because the term under the root must be positive) and $K_1 > 0$, because $\ln(-\ln G(x))$ is a decreasing function. We now get that $G(x) = \exp(-\sqrt[K_1]{xK_2})$ which is of the same type as

$$G(x) = e^{-(-x)^\alpha}, \qquad \alpha > 0, x < 0,$$

and we have found our second max-stable distribution.

Let us now consider the case when $G(d_N(1 - c_N)^{-1}) = 0$. This means that the corresponding point is the left endpoint of the distribution, let us denote

it by x_S. As previously, without the loss of generality we can take $x_S = 0$. This implies $d_N = x_S(1 - c_N) = 0$ and further

$$[G(x)]^{NM} = [G(c_N x)]^M = G(c_M c_N x)$$

and

$$[G(x)]^{NM} = G(c_{NM} x),$$

resulting in $c_N = N^{K_1}$. Now, similarly to previous, we get

$$\ln N + \ln(-\ln G(x)) = \ln(-\ln G(N^{K_1} x)),$$

from which

$$\ln(-\ln G(x)) = \ln \sqrt[K_1]{xK_2},$$

where $K_2 > 0$ and $K_1 < 0$, because $\ln(-\ln G(x))$ is a decreasing function. Now the formula for $G(x)$ is $G(x) = \exp(-\sqrt[K_1]{xK_2})$, which is of the same type as

$$G(x) = e^{-(x^{-\alpha})}, \qquad \alpha > 0, x > 0.$$

This is the third and final max-stable distribution.

In summary, we have covered all possible scenarios and thus there can be no more max-stable distributions (and thus also extreme value distributions).

9.3 Extreme value theorem. Examples

Based on the results obtained in previous subsection, we can formulate the following theorem, which is known as *the* basis of classical extreme value theory.

Theorem 9.3 (Extreme value theorem/ Extremal types theorem/ Fisher-Tippett theorem/ Fisher-Tippett-Gnedenko theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of* iid *random variables and let $M_n = \max(X_1, \ldots, X_n)$. If there exist norming constants $c_n > 0$, $d_n \in \mathbb{R}$ and some non-degenerate distribution function $G(x)$ such that*

$$\mathbb{P}\left(\frac{M_n - d_n}{c_n} \leqslant x\right) \to G(x) \tag{9.2}$$

then the limit distribution $G(x)$ belongs to the type of one of the following three distribution functions:

- *Fréchet:*

$$\Phi_\alpha(x) = \begin{cases} 0, & x \leq 0, \alpha > 0 \\ \exp(-x^{-\alpha}), & x > 0, \alpha > 0 \end{cases}$$

- *Weibull*
$$\Psi_\alpha(x) = \begin{cases} \exp(-(-x)^\alpha), & x \leq 0, \alpha > 0 \\ 1, & x > 0, \alpha > 0 \end{cases}$$
- *Gumbel*
$$\Lambda(x) = \exp(-e^{-x})$$

Generalized extreme value (GEV) distribution is a distribution that includes all three obtained classes and provides a unified parametrization for them.

Definition 9.3 ((Maximum) domain of attraction)**.** We say that a random variable X (and the corresponding distribution and distribution function) belong to the maximum domain of attraction of the extreme value distribution $G(x)$ if there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$ such that (9.2) holds.

Based on this definition and extreme value theorem, we can formulate the following result.

Corollary 9.1. *There exists a four-way division of distributions into classes that do not intersect:*

1. *distributions in the domain of attraction of the Fréchet extreme value distribution;*
2. *distributions in the domain of attraction of the Weibull extreme value distribution;*
3. *distributions in the domain of attraction of the Gumbel extreme value distribution;*
4. *distributions which do not have a non-degenerate limiting distribution for their sample maximum, regardless of the normalization.*

Example 9.1. Let $X, X_1, \ldots, X_n$ be *iid* uniformly distributed random variables, $X \sim U(a, b)$. Let us take $a_n = \frac{b-a}{n}$ and $b_n = b - \frac{b-a}{n}$. Then, for $1 - n < x < 1$ we can write

$$\begin{aligned} \mathbb{P}(M_n \leqslant a_n x + b_n) = (\mathbb{P}(X \leq a_n x + b_n))^n &= \left(\frac{a_n x + b_n - a}{b - a} \right)^n \\ &= \left(\frac{\frac{b-a}{n} x + b - \frac{b-a}{n} - a}{b - a} \right)^n \\ &= \left(1 - \frac{1 - x}{n} \right)^n \\ &\to \exp\left(- (1 - x) \right), \end{aligned}$$

which is a max-stable distribution known as the Weibull extreme value distribution.

Example 9.2. Let $X, X_1, \ldots, X_n$ be *iid* Pareto distributed random variables $X \sim Pa(\alpha, \theta)$, i.e. $F(x) = 1 - (\frac{\theta}{x+\theta})^\alpha$, $x > 0$. Let us take $a_n = \theta n^{1/\alpha}$ and $b_n = -\theta$. Then, for $x > 1/n^\alpha$ it holds that

$$\begin{aligned}\mathbb{P}(M_n \leqslant a_n x + b_n) &= \left[1 - \left(\frac{\theta}{\theta n^{1/\alpha} x - \theta + \theta}\right)^\alpha\right]^n \\ &= \left(1 + \frac{-(x^{-\alpha})}{n}\right)^n \\ &\to \exp\left(-\left(x^{-\alpha}\right)\right)\end{aligned}$$

We recognize a max-stable distribution called Fréchet' extreme value distribution.

Home assignment 19. Let $X_1, \ldots, X_n$ be *iid* exponentially distributed random variables, i.e. $F(x) = (1 - e^{-\lambda x})$, $x > 0$. As previously, let M_n denote the running maximum $M_n = \max\{X_1, \ldots, X_n\}$. Find the sequences of normalizing constants $a_n > 0$ and b_n, which give us the (distributional) convergence of

$$\frac{M_n - b_n}{a_n}$$

to a Gumbel extreme value distribution with distribution function

$$G(x) = e^{-e^{-x}}.$$

Also show analytically that with these normalizing constants indeed

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leqslant x\right) = G(x)$$

for any given value of x.

Remark 9.2. So far we only considered the right tail of a distribution, but actually the same models also apply for the left tail. This is due to simple derivation

$$\begin{aligned}\mathbb{P}(\min\{X_1, \ldots, X_n\} \leqslant x) &= \mathbb{P}(-\max\{-X_1, \ldots, -X_n\} \leqslant x) \\ &= \mathbb{P}(\max\{-X_1, \ldots, -X_n\} \geqslant -x)\end{aligned}$$

Thus, if we are indeed interested in the sample minimum then we just change the sign of our data and proceed as usual.

9.4 Generalized extreme value distribution

Distribution with a cumulative distribution function

$$F(x) = \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right),$$

where $\sigma > 0$ is called a generalized extreme value distribution (GEV(ξ, μ, σ)). Situation $\xi = 0$ is treated as a limit $\xi \to 0$. The support of the distribution is based on arguments for which $1+\xi\left(\frac{x-\mu}{\sigma}\right) \geqslant 0$.

GEV distribution with cumulative distribution function as specified above includes Frechet', Weibull and Gumbel distributions as special cases:

- $\xi = 0$ reduces to Gumbel distribution,
- $\xi > 0$ reduces to Frechet' distribution,
- $\xi < 0$ reduces to Weibull distribution.

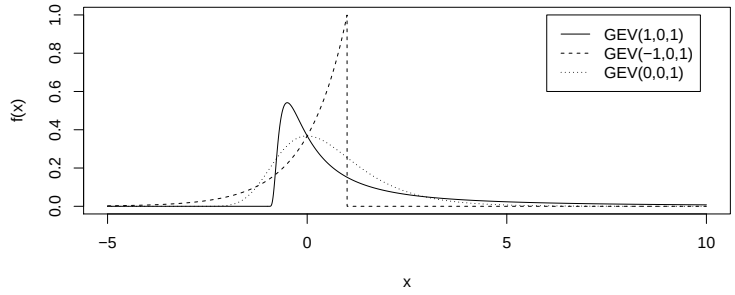

Figure 26: The probability density function of GEVD for different parameter value combinations

How to choose the constants a_n and b_n for a particular distribution (which has a cumulative distribution function $F(x)$ and right endpoint x_F)? It can be shown that if for the hazard function $h(x) = f(x)/\bar{F}(x)$ it holds that

$$\left(\frac{1}{h(x)}\right)' \to \xi, \text{ if } x \to x_F,$$

then selecting b_n as the $1-1/n$ quantile and a_n as $1/h(b_n)$ guarantees us a convergence to $GEV(\xi, 0, 1)$.

Of course we do not know the norming constants in practice. But we did see that if the distribution is in the domain of attraction of some extreme

value distribution then for some normalization we have that $a_n^{-1}(M_n - b_n) \xrightarrow{d} GEV(\xi, 0, 1)$. For large n thus

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leqslant x\right) \approx G(x).$$

which translates to

$$\mathbb{P}(M_n \leqslant y) \approx G\left(\frac{y - b_n}{a_n}\right),$$

that is the cumulative distribution function of $GEV(\xi, b_n, a_n)$ which means that the constants are not important to us as the scale and location parameter must be estimated anyhow.

So we saw that GEV distribution has theoretical justification as a model for *iid* sample maximum.

In practice, to apply GEV distribution model, the following steps can be taken:

1. We split the sample into blocks (block length is equal to the time horizon that we are interested in).
2. We find the maximal elements of each block and form a new sample from these.
3. We use this new sample to estimate the parameters of a GEV distribution.

If we can estimate the probability that $\{M_n > u\}$, then making use of independence allows us to estimate the complementary cumulative distribution function as well. GEV distribution has 3 parameters and estimation thus requires a large sample. Actually this is so also because the initial sample size is essentially divided by the block size. But the block size cannot be taken too small because then the asymptotics behind the theory of GEV distribution would not work.

Main properties of the GEV distribution are

- probability density function

  $$f_X(x) = \frac{1}{\sigma}\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-(\xi+1)/\xi} \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},$$

  where $\xi \in \mathbb{R},\ \mu \in \mathbb{R}, \sigma > 0$,

- cumulative distribution function $F_X(x) = \exp\left(-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right)$,

- moments $\mathbb{E}(X^n) = \sum_{k=0}^{n} C_n^k \left(\frac{\sigma}{\xi}\right)^k \left(\mu - \frac{\sigma}{\xi}\right)^{n-k} \Gamma(1 - k\xi)$,
- $\mathbb{E}|X|^n < \infty \Leftrightarrow \xi < \frac{1}{n}$,
- $U \sim U(0,1) \Rightarrow \mu + \frac{\sigma}{\xi}[(-\ln U)^{-\xi} - 1] \sim GEV(\xi, \mu, \sigma)$,
- mode argmax $f(x) = \mu + \frac{\sigma}{\xi}\left[(1+\xi)_+^{-\xi} - 1\right]$,
- median med $X = \mu + \sigma\frac{(\ln 2)^{-\xi} - 1}{\xi}$.

9.5 Generalized Pareto distribution (GPD)

In several situations, it is more important to study the tail behaviour of a distribution.

In that case, the approximation of main part is not that important and we can derive more efficient methods to describe the tail.

Suppose that for some μ', σ' and ξ we have

$$[F(a_n x + b_n)]^n \to \exp\left\{ -\left[1 + \xi \cdot \frac{x - \mu'}{\sigma'}\right]^{-1/\xi} \right\},$$

for all x. We can conclude that for some μ, σ and ξ it holds that

$$n \ln [F(x)] \approx -\left[1 + \xi \cdot \frac{x - \mu}{\sigma}\right]^{-1/\xi}$$

for all x. Now if x is close to the right endpoint of the distribution then $F(x) \approx 1$ and thus $\ln F(x) \approx F(x) - 1$. So, for large u we have

$$\bar{F}(u) \approx \frac{1}{n}\left[1 + \xi \cdot \frac{u - \mu}{\sigma}\right]^{-1/\xi}.$$

Thus, for large u and $y > 0$ we can write that

$$\begin{aligned}
\mathbb{P}(X > u + y | X > u) &\approx \left[\frac{1 + \xi(u + y - \mu)/\sigma}{1 + \xi(u - \mu)/\sigma}\right]^{-1/\xi} \\
&= \left[1 + \frac{\xi y}{\sigma + \xi(u - \mu)}\right]^{-1/\xi} \\
&= \left[1 + \frac{\xi y}{\sigma_u}\right]^{-1/\xi}.
\end{aligned}$$

So, after setting the threshold, **only two parameters remain**. The resulting distribution is known as the Generalized Pareto distribution (GPD).

Remark 9.3. Similarly to extreme value distributions, the GPD models can also applied to left tail. We can change the sign of our data (and also the threshold):

$$\mathbb{P}(X < u - x | X < u) = \mathbb{P}(-X > -u + x | -X > -u),$$

and obtain the familiar GPD model.

How large is large u or how do we choose the threshold? If u is too large then the approximation is good but our sample will be small and the parameter

estimates will thus have a large variance. On the other hand if u is not large enough then using the GPD as an approximation might not be justified. We saw above that if GPD model holds for threshold u, it also holds for any threshold $u' > u$ and the shape parameter does not change while the scale parameter increases linearly.

Similar test criterion can be obtained usingthe mean excess function. For large arguments the mean excess function is approximately equal to the mean of the GPD:

$$\mathbb{E}(X-u|\ X>u) \approx \frac{\sigma_u}{1-\xi} = \frac{\sigma+\xi(u-\mu)}{1-\xi} = \underbrace{\frac{\sigma-\xi\cdot\mu}{1-\xi}}_{const} + u\cdot\underbrace{\frac{\xi}{1-\xi}}_{const}.$$

Thus, one possible choice for the threshold is the point from where the (empirical) mean excess function can be considered linear.

We saw that GPD is a theoretically justified model for the tail of the distribution. In practice the following steps are taken:

1. We find a suitable threshold.
2. We select all sample elements that exceed the threshold and subtract the threshold value from them. This is our new sample.
3. The sample is used for estimating the parameters of the GPD.

The probability for exceeding the threshold can be estimated based on the original sample.

In conclusion, GPD is a limiting distribution for the values exceeding a threshold. If a random variable X follows generalized Pareto distribution, we write $X \sim GPD(\xi,\sigma)$. The support of the distribution is based on those positive arguments for which $1+\frac{\xi x}{\sigma} \geqslant 0$.

Main properties of the GPD distribution are:

- probability density function $f_X(x) = \frac{1}{\sigma}\left(1+\frac{\xi x}{\sigma}\right)^{-(\xi+1)/\xi}$, $\xi\in\mathbb{R}, \sigma>0$,
- cumulative distribution function $F(x) = 1-\left(1+\frac{\xi x}{\sigma}\right)^{-1/\xi}$,
- moments $\mathbb{E}(X^n) = \left(\frac{\sigma}{\xi}\right)^n \sum_{k=0}^{n} C_n^k \frac{(-1)^{n-k}}{1-k\xi}$,
- $\mathbb{E}|X|^n < \infty \Leftrightarrow \xi < \frac{1}{n}$,
- $U\sim U(0,1) \Rightarrow \frac{\sigma}{\xi}(U^{-\xi}-1) \sim GPD(\xi,\sigma)$,
- mode argmax $f(x) = -\frac{\sigma}{\xi} I_{\{\xi<-1\}}$,

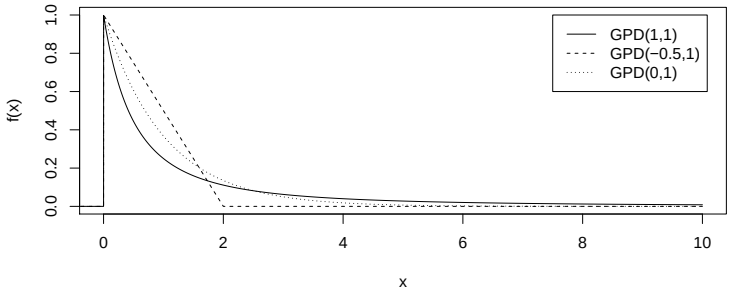

Figure 27: The probability density function of GPD for different parameter value combinations

- median med $X = \sigma \frac{2^{\xi}-1}{\xi}$.

Home assignment 20. In what way does the GPD generalize the Pareto distribution?

10 Stable distributions

Definition 10.1. We say that a random variable is stable (or has a stable distribution) if for the cumulative distribution function $F(x)$ and any $n \in \mathbb{N}$ there exist constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that for all x it holds that

$$F^{*n}(x) = F(a_n x + b_n).$$

This is property is known as sum stability – sum of the random variables is still in the same class.

We can conclude that if we have *iid* random variables $X, X_1, \ldots, X_n$ and if we denote $S_n = X_1 + \ldots + X_n$, then for every positive integer n we have

$$\frac{S_n - b_n}{a_n} \stackrel{d}{=} X,$$

meaning that stable distributions are a limit distribution for normed sums. As it was with max-stable distributions, it turns out that this is the only class of distributions appearing as a limit. Obviously, the most well known distribution belonging to this class is normal distribution.

The class has 4 parameters, the most important of them is the one that determines the heaviness (actually the lightness) of the tails $\alpha \in (0, 2]$. The skewness parameter $\beta \in [-1, 1]$ is also important. Scale and location parameters are respectively γ and δ. Even though all stable distributions are continuous, only three combinations of (α,β) allow us to write an analytic density:

- $\alpha = 2$, $\beta = 0$ gives us the normal distribution,
- $\alpha = 1$, $\beta = 0$ gives us the Cauchy distribution, and
- $\alpha = 1/2$, $\beta = 1$ is the Lévy distribution.

In other cases one has to deal with the characteristic function.

Standard stable distribution ($\gamma = 1$ and $\delta = 0$) has a characteristic function:

$$\varphi(t) = \begin{cases} \exp\{-|t|^\alpha [1 - i\beta \tan\left(\frac{\pi\alpha}{2}\right) \operatorname{sgn}(t)]\}, & \alpha \neq 1 \\ \exp\{-|t|[1 + i\beta \frac{2}{\pi} \operatorname{sgn}(t) \log|t|]\}, & \alpha = 1. \end{cases}$$

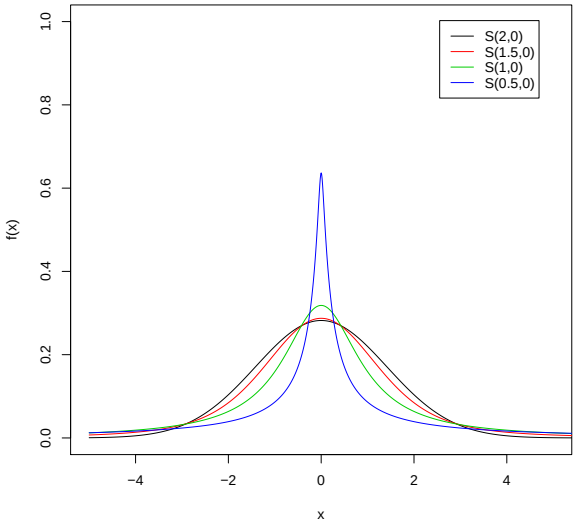

Figure 28: The effect of the shape parameter α to the tails of the stable distribution

The larger the β the more heavy the right tail is compared to the left tail (and vice versa). We can see from the characteristic function that if $X \sim S(\alpha, \beta)$ and $Y \sim S(\alpha, -\beta)$, then X and $-Y$ have the same distribution.

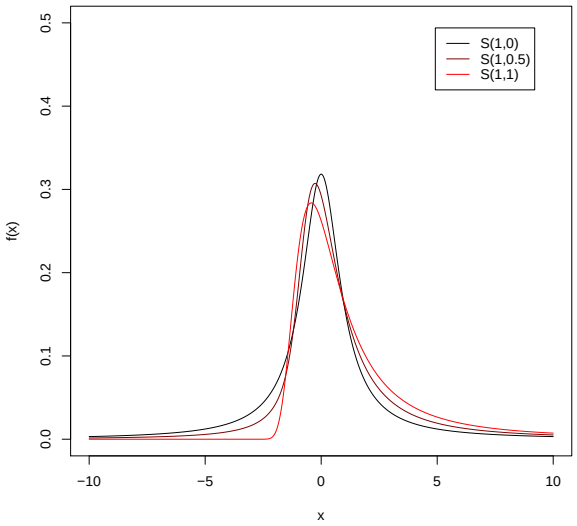

Figure 29: Effect of the skewness parameter β to the tails of the stable distribution

Main characteristics of a stable distribution are:

- probability density function and cumulative distribution function are not analytically expressible (except for some parameter values),
- mean $\mathbb{E}X = \delta$, if $\alpha > 1$, otherwise undefined,
- variance $VarX = 2\gamma^2$, if $\alpha = 2$, otherwise undefined,
- higher moments are undefined,
- median med $X = \delta$ if $\beta = 0$, otherwise not analytically expressible,
- mode argmax $f(x) = \delta$ if $\beta = 0$, otherwise not analytically expressible,
- moment generating function is undefined,
- skewness is 0 if $\alpha = 2$, otherwise undefined.

Let us now consider an important limit theorem, which is actually the generalization of the CLT. The latter requires finite variance, as we know. We now generalize the class of possible limit distributions and drop this requirement.

Theorem 10.1. *The class of stable (non-degenerate) distributions coincides with the class of all possible (non-degenerate) limit distributions for (properly normalized and centered) sums of* iid *random variables.*

In other words, if a distribution is stable, then it is a limit distribution for sums of *iid* random variables. Moreover, it is the only possible limit distribution for such sums.

We say that a distribution F is in the domain of attraction of an α-stable distribution if a stable distribution with shape parameter α is a possible limit of the (normed) sum of iid F distributed random variables. Thus all finite-variance distributions (random variables) are in the domain of attraction of a 2-stable distribution (normal distribution).

We know that when the variance is finite then the normalizing constants can be represented using the mean and standard deviation. This is out of the question when the second moment does not exist. Actually it is worth noticing that when normal distribution is the limit then a_n has the form $n^{1/2}C$. Unfortunately, it is not as simple as selecting a_n in the form of nC when there is finite mean.

More precisely, (the Marcinkiewicz-Zygmund) law of large numbers states that if and only if $\mathbb{E}(X)^p < \infty$ for some $p \in (0, 2)$, then for some b almost surely

$$\frac{S_n - b \cdot n}{n^{1/p}} \to 0.$$

But if $\mathbb{E}(X)^p = \infty$ for some $p \in (0, 2)$ then for every a almost surely

$$\limsup_{n\to\infty} \frac{S_n - b \cdot n}{n^{1/p}} = \infty.$$

So, roughly speaking, if we want a non-degenerate distribution then the scaling constant should be just of a lower order than the inverse of the rank of the last existing moment. So if $\mathbb{E}(X)^p < \infty$ if and only if $p < 1$, then the inverses of the existing moments form $(1, \infty)$ and we would expect the suitable scaling constant to be of order n (too low order would probably cause too large a variance and the limit would not exist).

Definition 10.2. We say that a function f is slowly varying if for every $t > 0$ we have that $f(tx)/f(x) \to 1$ when $x \to \infty$.

So, for example, a constant function and a logarithm function are slowly varying functions.

If a random variable X is in the domain of attraction of an α-stable distribution, we can choose the centering constant $b_n = \mathbb{E}(X)n$ when $\alpha > 1$ and $b_n = 0$ otherwise. The scaling constant has the form

$$a_n = n^{1/\alpha} L(n),$$

where L is a suitable slowly varying function.

11 Geometric stable distributions

In real life there are many random entities that are dependent on small impacts. We can imagine that a stock prices (or price change) is a result of many small factors and each factor has a small effect. The number of these small factors is probably not constant (when we compare days). Thus the number of random variables summed up should also be random. This leads us to geometric stable distributions.

Let a random variable N have a geometric distribution with parameter $p \in (0,1)$, i.e for each positive integer k we have

$$\mathbb{P}(N = k) = p(1-p)^{k-1}.$$

Definition 11.1 (Geometric stability)**.** We say that a random variable Y is geometric stable if for some sequence of *iid* random variables $X_1, X_2, \ldots$ that is also independent of N, there exist normalizing constants $a_p > 0$ and $b_p \in \mathbb{R}$ such that

$$\sum_{i=1}^{N} \frac{X_i - b_p}{a_p} \xrightarrow{d} Y$$

in the process $p \to 0$.

Let us look at standard exponential random variables in the role of the summands. Let $p \in (0,1)$ and

$$Y = p \sum_{i=1}^{N} X_i, \tag{11.1}$$

where N is geometrically distributed as before. So we are interested in finding a limit distribution (or the corresponding density function) of a geometric sum. Let us denote $S = \sum_{i=1}^{N} X_i$. Then, by (11.1), we have

$$\mathbb{P}(Y \leqslant x) = \mathbb{P}\left(S \leqslant \frac{x}{p}\right),$$

which for densities implies

$$f_Y(x) = \frac{1}{p} f_S\left(\frac{x}{p}\right).$$

Notice also that by the construction, the sum S is a discrete mixture, and for any fixed k, the conditional distribution of S given $N = k$ is gamma distribution (as it is sum of independent exponentials) with parameters k

and 1. Thus, we can retrieve the marginal density of Y through conditional distribution of $S|N$ and marginal distribution of N:

$$\begin{aligned} f_Y(x) &= \frac{1}{p}\sum_{k=1}^{\infty} f_{S|N=k}\left(\frac{x}{p}\right) f_N(k) \\ &= \frac{1}{p}\sum_{k=1}^{\infty} \frac{\left(\frac{x}{p}\right)^{k-1} e^{-x/p}}{\Gamma(k)} p(1-p)^{k-1} \\ &= e^{-x/p}\sum_{k=1}^{\infty} \frac{(1-p)^{k-1}\left(\frac{x}{p}\right)^{k-1}}{(k-1)!}, \end{aligned}$$

where we recognize the Taylor expansion of the exponential function and thus obtain

$$f_Y(x) = e^{-x/p}e^{(1-p)x/p} = e^{-x}$$

which means that the limit distribution is also standard exponential (and we even did not have to consider $p \to 0$) and exponential distribution is geometric stable.

In a more general case Rényi theorem states that if positive summands have finite mean then the normalized sum

$$p\sum_{i=1}^{N} X_i$$

always yields the exponential distribution as a limit in the process $p \to 0$. If the summands are symmetrical and have finite variance then the sum

$$\sqrt{p}\sum_{i=1}^{N} X_i$$

has a Laplace (distribution) limit.

So very often either exponential or Laplace distribution is the limit. But if we loosen the assumptions then the set of possible limit distributions expands. Let us still consider the case $b_p = 0$, i.e. the sum has the form

$$Y = \frac{1}{a_p}\sum_{i=1}^{N} X_i,$$

in the process $p \to 0$. If the summands are positive but we do not assume finite variance then the Mittag-Leffler class of distributions arises as the limit set. If the summands are symmetrical (but variance not necessarily finite) then the Linnik class of distributions arises as the limit set.

Just as with stable distributions, it is not possible to present the class of geometric stable distributions in a simpler way than using the characteristic function. The number of parameters is 4 and the function has the form

$$\varphi(t) = \begin{cases} \left[1 + \sigma^\alpha |t|^\alpha \left(1 - i\beta \operatorname{sgn}(t) \tan\left(\frac{\pi\alpha}{2}\right)\right) - i\mu t\right]^{-1}, & \alpha \neq 1 \\ \left[1 + \sigma^\alpha |t|^\alpha \left(1 + i\beta \operatorname{sgn}(t) \log|t| \frac{2}{\pi}\right) - i\mu t\right]^{-1}, & \alpha = 1, \end{cases}$$

where $\alpha \in (0, 2]$ is the shape parameter, $\beta \in [-1, 1]$ the skewness parameter, $\sigma \geqslant 0$ is scale parameter and μ is location parameter. Situation $\sigma = 0$ corresponds to an exponential distribution. The geometric stable distributions are continuous and have a heavy tail when $\alpha < 2$ (and $\sigma \neq 0$).

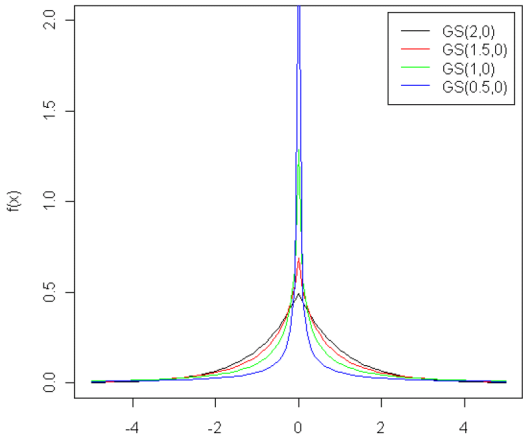

Figure 30: The effect of the shape parameter α to the tails of the geometric stable distribution

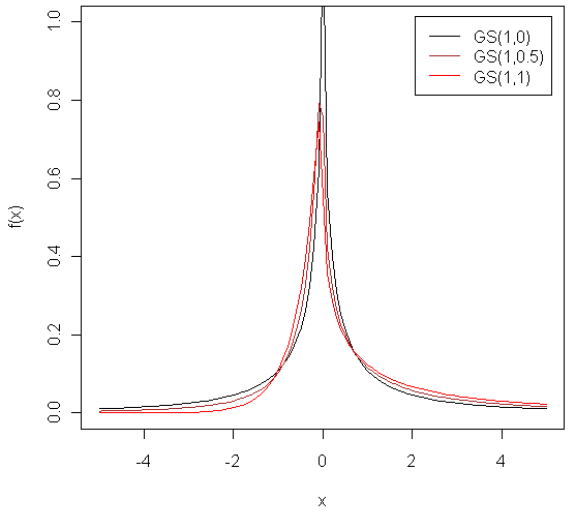

Figure 31: The effect of the skewness parameter β to the tails of the geometric stable distribution

If an analytic probability density function does not exist then the use of ML for parameter estimation is not straightforward (a grid method would be pretty imprecise and time consuming). An empirical characteristic function

$$\frac{1}{n}\sum_{i=1}^{n} e^{itx_i},$$

is constructed and then the real and imaginary parts are equated (for theoretical and empirical functions) to form the equations. Different (arbitrary) values of t_j are used for this. Obviously, the choice of the values t_j is still a problem to be addressed.

Alternative approach for finding parameters is to use the method of (non-integer) moments

This method is based on fact that for a geometric stable distribution, the moments $\mathbb{E}|X|^p$ are finite if $p < \alpha$

Then, the moments matching can be done through a 2-step process:

- detect the value of α (using the tools described previously)
- choose four constants less than that estimated value to form the equations.

12 Goodness of fit of a model

Any fitting problem with different candidate models sooner or later meets the question of how to measure which candidate is better, and, finally, which model should we choose. When one of the candidates is a special case of the other (nested models) then this is quite straightforward. If not then more thought is required. Likelihood function which is an intuitive starting point for comparing two models (especially when ML was used for parameter estimation) cannot be used directly.

12.1 Kolmogorov-Smirnov test

Let us assume that the distribution function F of a theoretical candidate distribution is continuous. To first answer whether the data fits the distribution a graphical method is used. QQ-plot is one of the options. When drawing the QQ-plot the role of the inverse cumulative distribution function is assumed by F^{-1}. In practice, the theoretical candidate is rarely fully specified. It is common to take a candidate from a specified class of distributions, but the parameters need to be estimated (from data). Let us denote the corresponding distribution by F^*.

Let F_n be the empirical cumulative distribution function. Clearly the function

$$D(x) := |F_n(x) - F(x)|$$

measures the goodness of fit. Thus the graph of this function is a possible tool for our purposes.

But we can also quantify the fit quality. For example, we might postulate the null hypothesis that the data comes from a known distribution F and the alternative hypothesis is thus the negation of this statement – data does not come from F. To test this hypothesis it is wise to make use of function $D(x)$.

Kolmogorov-Smirnov test is based on the statistic

$$\sup_x |F_n(x) - F(x)|,$$

the value of which is compared with a critical value. If it is exceeded then the null hypothesis is rejected. Although it might seem to be complicated, it is in fact easy to calculate the value of the statistic. This is so because the cumulative distribution function $F(x)$ is continuous and the empirical cumulative distribution function is a step function. This means that the supremum of $D(x)$ is always attained at some sample element or "just before it".

There is one thing that is very often overlooked – the definition of F must be independent from the sample. This means that we cannot use F^* in the role of F, because the parameters of it are estimated from the sample. If we do it then the test no longer works properly – null hypothesis is often not rejected (when it should be) and the power of the test is thus reduced.

The solution out of this is the splitting of the sample – first part is used for parameter estimation (i.e. finding F) and the other part is used for constructing F_n.

Let us illustrate the previous fact with a simulation. First let us generate 1000 samples, each with size $n = 100$, from standard normal distribution and each time find the Kolmogorov-Smirnov test statistic using the standard normal distribution in the role of F as well. For the second case we use the same generation rule but in the role of F we use a normal distribution, parameters of which are estimated using the same sample. For the third case again we use the same generation rule but the parameters are estimated using the first half of the sample (and this way F is chosen). The second half is used for constructing F_n.

From the first experiment we get that the 0.95 quantile is 0.133 (this is the critical value for this sample size because the test is distribution free and based only on the values of the cumulative distribution function). But in the second experiment this critical value is very rarely exceeded even though the candidate F is always wrong (even if just a little). In the third experiment the situation is more plausible – about half the simulations lead to rejecting the null hypothesis (as it should be). The fact that this does not happen more often is due to the power of the test.

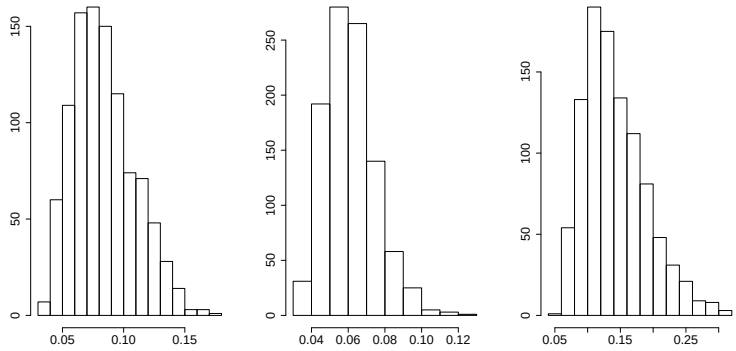

Figure 32: Distribution of the Kolmogorov-Smirnov statistic for different usage scenarios

12.2 Anderson-Darling test

Anderson-Darling test also makes use of $D(x)$, but extra weight is placed on the tails. Test statistic is formed as an integral

$$n \int \frac{(F_n(x) - F(x))^2}{F(x)\overline{F}(x)} f(x)dx,$$

where the integral is over the full support of the distribution and $f(x)$ is the probability density function corresponding to F. So essentially this is an average of $D^2(x)$, but the weights are not uniform – large and small arguments are up-weighted. On the other hand this is quite reasonable because as every cumulative distribution function begins with zero and ends with one and thus otherwise the difference near the tails would be small. For this test it also holds that F should be determined independently from the data.

12.3 Likelihood ratio test

Suppose now that we have 2 candidate models and A is a special case of B – B simplifies to A when we fix one of its parameters (we can just say that A is a simpler model than B). Then the value of the likelihood function for A (lets call it L_0) cannot be larger than the value for B (lets call it L_1), where in both cases ML is used. The ratio of the two values is connected to χ^2 distribution. The test statistic

$$2 \ln \left(\frac{L_1}{L_0} \right)$$

is approximately χ^2-distributed where the degrees of freedom is equal to the number of extra parameters that B has. Null hypothesis is thus that A is good enough and alternative hypothesis is that B is considerably better.

Remark 12.1. We can also use the test so that we fix, e.g., the mean of the distribution, and even if this depends on several parameters then the number of free parameters for A is one less than for B. Even though none of the parameters is strictly fixed, we can still use the likelihood ratio statistic.

In a situation where one model is not the extension of other, the likelihood ratio cannot be properly used.

Let us rewrite the likelihood ratio and look for possible generalizations. Let $c_{r,\alpha}$ be the 1-α-quantile of the distribution $\chi^2(r)$. Then the likelihood ratio statistic tells us to use a more complicated model (with r extra parameters) if

$$\ln L_1 > \ln L_0 + \frac{c_{r,\alpha}}{2},$$

where L_1 is the likelihood function value for the more complicated model and L_0 is the likelihood function value for the simpler model. Thus we see that there is also term that can be viewed as penalty – one extra parameter requires that the log-likelihood should increase by at least $c_{1,\alpha}/2$.

This argumentation leads to the general idea of information criteria – we compare the likelihood of different models and apply certain penalty function to a more complex model.

12.4 Information criteria. Akaike information criterion (AIC)

Information criteria (which are also based on log-likelihood) are also tools for measuring the goodness of a model. The most well-known of them is Akaike information criterion, which is a statistic defined by

$$AIC = -2\ln(\text{maximum likelihood}) + 2p,$$

where p is the number of parameters in a model under consideration.

So, AIC measures the goodness of fit as certain tradeoff between two components:

- goodness of fit (measured by the log-likelihood),
- model complexity (measured by the number of parameters p).

Information criteria can also take the sample size into account.

Bayesian IC (*aka* Schwarz IC) is calculated as

$$BIC = -2\ln L + p\ln n,$$

This means that when there are r extra parameters and the sample size is n then we should reduce the log-likelihood by $(r/2)\ln n$, or, in other words, each additional parameter is deemed worthy when the log-likelihood is increased by $0.5\ln n$. When this is not the case then a simpler model should be preferred.

The corrected AIC says that we should pick a model for which

$$AICc = AIC + \frac{2p(p+1)}{n-p-1} = -2\ln L + 2p + \frac{2p(p+1)}{n-p-1}$$

is smallest. One should also note that when the number of parameters is equal then the decision is essentially based on the likelihood function value.

13 Conclusion

Let us now try to summarize what we have learned.

1. What is our goal? Modelling the distribution is typically not the ultimate task but only a useful (and necessary) step to answer some question. This means that sometimes we do not need to model the whole distribution, perhaps only some part of it. So e.g. when we are interested in predicting the 0.99-quantile then it probably makes sense to just fit a GPD on the tail of the distribution and leave out the main part.

2. Are our data realizations of *iid* random variables? We have considered the test of records. We can also look at the signs of consecutive observations and other non-parametric tests. Also squaring the observations can be useful to study the correlation (and possibly rule out independence). We could also look at the behavior of the median (or some other quantile) in the sub-samples. Tests based on the mean are often not useful in the heavy-tailed data context.

   Some transformation might help us (log-returns seem less correlated than stock prices).

   Sometimes it can happen that we are not interested in independence – perhaps we only do wish to model the distribution. Then we have a situation similar to a Markov chain, where neighboring observations are clearly correlated but still are realizations from the distribution that we are interested in. (e.g. the Gibbs sampler). Then "obtaining independence" is also similar – we thin our observations by leaving out, say, nine out of ten and only retaining every tenth observation. This is of course very costly.

3. Modeling usually begins with the selection of the model class. We must ponder whether there is a clear basis to prefer one distribution to another (e.g. somehow it is known beforehand that a Pareto distribution can fit but Weibull definitely does not fit). Here the ratio of maximum and sum might help us. When we don't find a suitable candidate then it might be necessary to make one up, but this can lead to a multitude of complications. Sometimes the lack of data can also limit our choice for distributions. And if we are not certain that we have heavy-tailed data then this should also be tested for by visualizing the mean excess function or creating a suitable QQ-plot.

4. Depending on the class of distributions chosen different parameter estimation methods can be used. The best situation is the one where different estimation methods lead to similar parameter estimates. Even

though it is generally reasonable to ask for good asymptotic properties, we must realize that e.g. the consistency of an estimator does not guarantee us anything beneficial when we have a moderate-sized sample.

5. Sometimes it may turn out that a model which seemed reasonable does not give us a good fit. In this case the model must often be abandoned and the process must start over. Shortcomings of the model can sometimes be detected using the kernel density estimate for comparison (or perhaps the QQ-plot). Cross-validation or just separate training and test data are useful. This last scenario allows us to use the Kolmogorov-Smirnov test.

6. When we do not have one and only model in consideration and the goodness of a model is simply based on fit there still might be the desire to arrive at a single (final) model. Model selection is usually handled using information criteria. The question of how large a difference in those values is large enough to rule out worse models is still a critical question.

7. When we have a parametric model then usually answering any particular questions using this model is quite straightforward. Sometimes a sensitivity analysis might be required (to show how much the changes in parameter estimates affect the answer to our question). This is also the case when our modelling process leads to several competing models.