Quantitative Risk Management (QRM) 2023/2024

Lecture notes

Rasmus Frigaard Lemvig (rfl@math.ku.dk)

First edition

Preface to the first edition

These lecture notes were written for the course *Quantitative Risk Management* (abbreviated *QRM*) at the University of Copenhagen in the winter of 2023/2024. They are based on the lectures by Jeffrey F. Collamore with some of my own additions, mostly in the form of supplementary examples, exercises and an appendix. I also added a few more results that I felt were helpful when solving the mandatory exercises. The notes should be sufficient to cover the entire syllabus, and if the reader wants to dive into certain topics in more detail, the notes and comments at the end of each week provide further references.

The notes are built up in the following way: Each week in these notes corresponds roughly to a week of teaching. Some subsections have been moved to make the presentation more clear, but the reader can expect the notes to follow the teaching almost one to one. The appendix at the end deserves an explanation. The appendix contains three sections. The first is very short and concerns generalised inverses. This section is essential for the course, and it should be clear in the text when the reader should consult this part of the appendix. As for the other two sections, probability theory and calculus, they are purely supplementary. I wrote them with the intend of making it easer to look up some basic notions if necessary. The course relies on a general understanding of probability theory, and sometimes tools like conditioning arguments and integration by parts (with respect to functions) are used without explanation. For students with less training in such computations, it may be a good idea to take a brief look at the relevant subsections.

Feedback in general is very appreciated. The exercises are my own addition (well, most of them), and any suggestions on how these can be improved to be more interesting etc. are very welcome. Last but not least, there is likely a bunch of typos remaining and maybe even a few mathematical errors. These are all due to me. Please don't hesitate to let me know, if you find any.

Rasmus Frigaard Lemvig
January 2024

Table of contents

Table of contents

Week 1 - Risk measures

1 Introduction

What is quantitative risk management about?

Quantitative risk management aims at describing and understanding risk in a financial context. To motivate our discussion, let us start with a basic example. Suppose we have a stock with value S_n at time n (discrete time units). Assume the Bernoulli model

$$S_0 = 1, \quad P(S_{n+1} = 2S_n) = \frac{2}{3} \quad \text{and} \quad P(S_{n+1} = 0.5S_n) = \frac{1}{3} \quad \text{for} \quad n > 0.$$

In this model there are some basic questions we may ask. For example, what is the risk in this investment policy? What are the expected returns? We can compute

$$E[S_n \mid S_{n-1}] = \frac{2}{3} \cdot 2S_{n-1} + \frac{1}{3} \cdot \frac{1}{2}S_{n-1} = \frac{3}{2}S_{n-1}$$

and it follows that

$$M_n = \left(\frac{2}{3}\right)^n S_n$$

is a martingale, and

$$1 = M_0 = E[M_n] = \left(\frac{2}{3}\right)^n E[S_n] \quad \text{implying} \quad E[S_n] = \left(\frac{3}{2}\right)^n \to \infty.$$

However, there is still a risk that we lose money in this investment policy. Note that we can write

$$S_n = R_1 \cdots R_n$$

for Bernoulli variables R_i with $P(R_i = 2) = 2/3$ and $P(R_i = 1/2) = 1/3$. Taking log yields

$$\log S_n = \sum_{i=1}^{n} \log R_i$$

which is a risk process that can be studied using ruin theory. If $E[e^{-\alpha \log S_1}] = 1$ for some $\alpha > 0$, we obtain the Cramér-Lundberg estimate (for some threshold $u > 0$)

$$P(\log S_n < -u \text{ for some } n) \sim Ce^{-\alpha u}$$

for a constant C. In Cramér-Lundberg theory, there is a tradeoff between profit and risk and the same rule applies to risk management in finance. This tradeoff can be studied from several viewpoints. Sometimes this tradeoff is handled using a utility function U and then maximizing the expected utility $E[U(S_n)]$. In this course we follow another approach via so-called risk measures. This approach allows us to compute loss probabilities directly. All these terms will get a more rigorous definition later on.

Some historical remarks

An essential part of risk management is to determine the capital needed to withstand shocks for financial institutions. History has many examples of banks and other institutions failing due to insufficient financial coverage. We go through some of these examples to add more context to the following (more mathematical) discussion.

Example 1.1. Barings Bank was founded in 1762 and was one of the UK's oldest and most respected banks. In 1995 the bank collapsed despite having more than $900 million in capital. This was due to unauthorized trading by a single employee. He bought straddles (selling both one call option and put option) which in a typical market will usually expire and provide a gain. However, market instability was caused by a Japanese earthquake, and the bank suffered a loss of more than $1 billion, resulting in a bankruptcy. ○

Example 1.2. In 1994, the hedge fund LTCM (Long-Term Capital Management) was founded. The employees were experienced traders and academics. $1.3 billion were invested with returns after two years close to 40 %. Early in 1998 net assets were $4 billion, but by the end of the year the fund was close to default. The U.S. Federal Reserve managed a $3.5 billion rescue package to avoid a systematic crisis in the world financial system. The triggering event for this disaster was the devaluation of the ruble by Russia. ○

Example 1.3. In 2023, Silicon Valley Bank collapsed. The bank owned low interest bonds and paid even lower interest to the depositors. When the depositors withdrew their capital, this required selling the low-interest treasury bonds, whose market prices had decreased sharply. ○

Example 1.4. The final example concerns a particular individual, Jesse Livermore. Livermore is considered the greatest short-seller in history. He succesfully shorted:

- The 1906 earthquake (through a railway investment).
- The 1907 market crash (”panic of 1907”).
- The 1929 market crash (through ca. 100 shorts, netting about $100 million).

He also had numerous successful ”long” investments. He went bankrupt a number of times however. He went bankrupt in 1901, 1908 and again in 1934. His book on trading remains a classic to this day. ○

2 Loss random variables

In describing risk, we focus on the loss instead of the profit. The following definition sets the stage for our discussion in the first weeks of the course.

Definition 2.1. Let $t_0, t_1, ...$ denote discrete timepoints (days, weeks, months or years for example) and let $\Delta t_n = t_{n+1} - t_n$ denote the time passed between timepoint $n+1$ and n. We let V_n denote the *capital* at time t_n, and we let L_{n+1} define the loss between t_n and t_{n+1} e.g.

$$L_{n+1} = -(V_{n+1} - V_n).$$

We let a general loss random variable be denoted by L. Many models consist of assumptions on L. Let us consider some motivating examples. Note that in these examples, it makes sense to think of V_n as a portfolio.

Example 2.2 (**Stock investment**)**.** Let $V_n = S_n$ with S_n the price of a certain stock at time t_n. We let $X_{n+1} = \log S_{n+1} - \log S_n$ denote the log returns. Hence

$$e^{X_{n+1}} = \frac{S_{n+1}}{S_n}.$$

Historically, X_{n+1} has often been given a normal distribution. This is motivated by the Black-Scholes model, see the end of the examples for a concise explanation of this model. In this model, the change of the stock price (in continuous time t) is described by the dynamics

$$\frac{dS(t)}{S(t)} = rdt + \sigma dW(t)$$

with $W(t)$ denoting a Brownian motion. Solving the equation explicitly, assuming that we are currently at time t so that $S(t)$ is known, yields the expression for the stock price at time $T > t$

$$S(T) = S(t)e^{\left(r-\frac{\sigma^2}{2}\right)(T-t)+\sigma(W(T)-W(t))}.$$

The discrete time analogue is

$$S_{n+1} = S_n e^{\left(r-\frac{\sigma^2}{2}\right)(t_{n+1}-t_n)+\sigma\sqrt{t_{n+1}-t_n}Z}$$

with $Z \sim \mathcal{N}(0,1)$. The log return becomes

$$X_{n+1} = \log S_{n+1} - \log S_n = \left(r - \frac{\sigma^2}{2}\right)(t_{n+1} - t_n) + \sigma\sqrt{t_{n+1} - t_n}Z$$

which is normal distributed. Note that the loss L_{n+1} can be written as

$$L_{n+1} = -(S_{n+1} - S_n) = -S_n(e^{X_{n+1}} - 1).$$

If we are at time n, the value S_n is known. A key goal of this course is to model the unknown part X_{n+1} and thereby make inference about the behaviour of the process V_n at a future time. Note also how this approach differs from the one in classical ruin theory where the entire positive timeline is considered. Here we model the change in capital one step at a time. ○

Example 2.3 (**Stock investment with more assets**)**.** The preceding example can be generalised. Assume that we have d stocks. We can then form a portfolio at time n by

$$V_n = \sum_{i=1}^{d} \alpha_i S_n^{(i)}$$

with α_i the number of units bought of stock i and $S_n^{(i)}$ the value of stock i at time n. Using the previous example, the loss is given by

$$L_{n+1} = -\sum_{i=1}^{d} \alpha_i S_n^{(i)} (e^{X_{n+1}^{(i)}} - 1), \quad X_{n+1}^{(i)} = \log S_{n+1}^{(i)} - \log S_n^{(i)}.$$

Again we need to come up with a model for the log returns $X^{(1)}, ..., X^{(d)}$, where we can write

$$\mathbf{X}_{n+1} = (X_{n+1}^{(1)}, ..., X_{n+1}^{(d)}).$$

Very often the variables in $\mathbf{X}_{n+1}$ are dependent. Think for example of a portfolio of stocks in the same type of companies. This dependence structure is crucial to understand and capture when estimating risk in a portfolio.

○

Example 2.4 (**Bond investment**). Consider a *zero-coupon bond.* Such an asset pays one unit at a fixed time T. Let us briefly consider such a bond in continuous time. We have an interest rate r_t at time t, and we let B_t denote the price of the bond at time t (this price will depend on T). The price of the bond can be described by the dynamics

$$dB_t = r_t B_t dt,$$

and using the boundary condition $B_T = 1$, we can solve the above equation and get

$$1 = B_T = B_t e^{\int_t^T r_s ds},$$

and we can rewrite this expresion in terms of B_t as

$$B_t = e^{-\int_t^T r_s ds} = e^{-(T-t)y(t,T)}, \quad \text{where} \quad y(t,T) = \frac{1}{T-t}\int_t^T r_s ds.$$

$y(t,T)$ is called the *yield* of the bond. Let us now consider discrete time, and say that the current time is t_n. The loss is

$$L_{n+1} = -(B_{t_{n+1}} - B_{t_n}) = -B_{t_n}\left(\frac{B_{t_{n+1}}}{B_{t_n}} - 1\right).$$

Let us fix some notation. Denote the yield at time n by $Z_n = y(t_n, T)$, and let $X_{n+1} = Z_{n+1} - Z_n$. Note that Z_n is known at time t_n while X_{n+1} again needs to be modelled. We can rewrite the expression $\frac{B_{t_{n+1}}}{B_{t_n}}$ as

$$\begin{aligned} \frac{B_{t_{n+1}}}{B_{t_n}} &= e^{-(T-t_{n+1})y(t_{n+1},T)+(T-t_n)y(t_n,T)} \\ &= e^{-(T-t_n-\Delta t_n)(Z_n+X_{n+1})+(T-t_n)Z_n} \\ &= e^{\Delta t_n Z_n-(T-t_{n+1})X_{n+1}}. \end{aligned}$$

This expression makes it clear how the unknown variable X_{n+1} enters the loss.

○

In the above examples, we ended up with an expression containing a term of the form $e^{(\cdots)} - 1$. This makes it tempting to use a Taylor approximation since $e^x \approx 1 + x$ for small x, and historically, such an approximation was often considered to ease computations. Taking the example with the stock portfolio, we let

$$L_{n+1}^{\Delta} = -\sum_{i=1}^{d} \alpha_i S_n^{(i)} X_{n+1}^{(i)}$$

denote the *linerarized log returns*. This approximation can often be problematic however since it is of interest to consider large losses (which we will do next week).

A brief rundown of the Black-Scholes model

Consider a risk free asset with price process denoted by B (a bank account) given by the continuous dynamics

$$dB_t = r_t B_t dt$$

where it is often assumed that $B_0 = 1$. We can explicitly solve for the price and obtain

$$B_t = B_0 e^{\int_0^t r_s ds}.$$

r_t is called the interest rate and is assumed to be an adapted process. We model a risky asset (such as a stock) with price process S_t by a *stochastic differential equation* (SDE) of the form

$$dS_t = \mu(t, S_t)dt + \sigma(t, S_t)dW_t$$

with deterministic functions μ and σ and a Brownian motion W. μ is called the *local mean rate of return* for S_t while σ is called the *volatility* of S_t. For all the necessary results on SDEs, consult chapter 4 and 5 of [6]. For our purposes, it suffices to know the model on an intuitive basis. The Black-Scholes model is a special case of the above model.

Definition 2.5. The (Standard) Black-Scholes model consists of two assets with dynamics

$$\begin{aligned} dB_t &= rB_t dt, \\ dS_t &= \mu S_t dt + \sigma S_t dW_t \end{aligned}$$

with r, μ and σ constants.

In the language of SDEs, a process with the dynamics of S_t is called a *Geometric Brownian motion* (GBM). Such an SDE can be solved explicitly. In our case, we may write

$$S_t = S_0 e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t}.$$

The Black-Scholes model can of course be extended to include more risky assets with the same type of dynamics as S_t. Such a model is naturally referred to as a multidimensional Black-Scholes model.

The goal of arbitrage theory is to price financial derivatives i.e. products based on the price of underlying assets. We think intuitively of an arbitrage as a money machine/free lunch i.e. as a portfolio of assets that costs nothing and produces a positive amount of money with probability one. It turns out that the absence of arbitrage is the same as the existence of a so called *equivalent martingale measure* (EMM) i.e. a measure Q equivalent to the underlying measure P (Q and P have the same null sets) and such that the discounted price processes

$$\frac{S_t}{B_t}$$

are martingales under Q. Q is also referred to as a *risk neutral measure.* Under the measure Q, the dynamics of S_t change to

$$dS_t = rS_t dt + \sigma S_t dW_t^Q$$

where W^Q is a Brownian motion under the Q measure. Note that the volatility is unchanged while the local mean rate of return becomes the interest rate times S_t. Say we have a derivative which expires at time T, the current time is $t < T$ and that the derivative pays $X = \Phi(S_T)$ at time T. Note that the payout is a function of the price of the risky asset at time T (such a derivative is called *simple*). If $\Pi_t[X]$ denotes the (arbitrage free) price of X at time t, arbitrage theory yields the following formula.

Theorem 2.6 (**Risk Neutral Valuation**). *The arbitrage free price of $\Phi(S_T)$ at time $t < T$ is given by*

$$\Pi_t[X] = e^{-r(T-t)} E^Q[\Phi(S_T) \mid \mathcal{F}_t].$$

By an arbitrage free price we mean a price process that doesn't introduce an arbitrage into the market. Note that the above formula says that the price is given by the discounted expected value under the risk neutral measure (given the information we currently have available).

Example 2.7 (**European call option**). A European call option gives the holder the right (but not the obligation) to buy one stock at time T at price K (the *strike price*). The payout is $(S_T - K)^+ = \max\{S_T - K, 0\}$ since if $S_T > K$, we get the payout $S_T - K$ while if $S_T \leq K$, the option is worthless. In the Black-Scholes model, we can solve for S_T as

$$S_T = S_t e^{\left(r - \frac{\sigma^2}{2}\right)(T-t) + \sigma(W_T^Q - W_t^Q)}$$

since we are free to choose the current starting value (so here we choose to consider the start value at time t). If $C(t,T)$ denotes the price at time t, we have by the Risk Neutral Valuation formula that

$$C(t,T) = e^{-r(T-t)} E^Q[(S_T - K)^+ \mid \mathcal{F}_t]$$

and this can be computed explicitly[1]. The result is known as the *Black-Scholes formula.* It says that

$$C(t,T) = S_t \Phi(u) - K e^{-r(T-t)} \Phi(v)$$

where

$$u = \frac{\log(S_t/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}, \quad v = u - \sigma\sqrt{T-t}.$$

[1] The brave reader can carry out this computation. It involves a lot of integration by substitution.

Say we have a portfolio consisting of one European call option, so that the value is $V_n = C(t_n, T)$. We note that the risk in this portfolio can be explained by the three quantities

$$\mathbf{Z}_n = (Z_n^{(1)}, Z_n^{(2)}, Z_n^{(3)}) := (\log S_{t_n}, r_n, \sigma_n)$$

since the volatility and interest rate are not constant in real life. One typically needs to model the change in these factors (called risk factors, see the discussion below), namely $\mathbf{X}_{n+1} = \mathbf{Z}_{n+1} - \mathbf{Z}_n$. We can for example consider the linearized loss

$$L_{n+1}^{\Delta} = -\left(\frac{\partial C}{\partial t}\Delta t + \frac{\partial C}{\partial S} X_{n+1}^{(1)} + \frac{\partial C}{\partial r} X_{n+1}^{(2)} + \frac{\partial C}{\partial \sigma} X_{n+1}^{(3)} \right).$$

In mathematical finance, these first order derivatives have names. $\partial C/\partial t$ is called "theta", $\partial C/\partial S$ "delta", $\partial C/\partial r$ "rho" and $\partial C/\partial \sigma$ "vega". Together these quantities are referred to as the *greeks*, see chapter 10 in [6].

○

A general risk model

All examples considered above had the same form for V_n. We could write $V_n = f(t_n, \mathbf{Z}_n)$ for some (measurable) function f and $\mathbf{Z}_n$ suitable random variables.

Definition 2.8. For a portfolio of the form $V_n = f(t_n, \mathbf{Z}_n)$ with $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ a measurable function and $\mathbf{Z}_n = (Z_{n,1}, ..., Z_{n,d})$ a random vector, we call the variables in $\mathbf{Z}$ *risk factors.* We call $\mathbf{X}_{n+1} = \mathbf{Z}_{n+1} - \mathbf{Z}_n$ the *risk-factor changes* at time $n+1$.

In the previous examples we stressed that we only need to model the change in risk factors $\mathbf{X}_{n+1}$ since the value $\mathbf{Z}_n$ is already known at time t_n. Hence we can write the loss entirely in terms of the change in risk factors. Explicitly,

$$L_{n+1} = -(V_{n+1} - V_n) = -(f(t_{n+1}, \mathbf{Z}_n + \mathbf{X}_{n+1}) - f(t_n, \mathbf{Z}_n)) =: l_{[n]}(\mathbf{X}_{n+1}).$$

We refer to $l_{[n]}$ as the *loss operator.* By considering the linearized loss

$$L_{n+1}^{\Delta} = -\frac{\partial f}{\partial t}(t_n, \mathbf{Z}_n)\Delta t - \sum_{i=1}^{d} \frac{\partial f}{\partial z_i}(t_n, \mathbf{Z}_n)\mathbf{X}_{n+1}^{(i)}$$

obtained by applying a first order Taylor expansion, we can similarly define the *linearized loss operator*

$$l_{[n]}^{\Delta}(x) := -\frac{\partial f}{\partial t}(t_n, \mathbf{Z}_n)\Delta t - \sum_{i=1}^{d} \frac{\partial f}{\partial z_i}(t_n, \mathbf{Z}_n)x^{(i)}.$$

3 Risk measures

We need some notion of the "size" of a risk in order to quantify the risk of a loss.

Definition 3.1. Let L denote a loss. A *risk measure* ρ associates a real number to L denoted by $\rho(L)$.

A way to make the definition more formal is to let $\mathcal{G}$ denote the set of all measurable real-valued functions on the background probability space. A risk measure is then a mapping $\rho : \mathcal{G} \to \mathbb{R}$. We will not worry about these details in this course. We now go through some essential examples of risk measures. In the following, let L denote some loss random variable.

Example 3.2. For $\alpha \in (0,1)$, let $\rho = \inf\{x \in \mathbb{R} : P(L > x) \leq 1 - \alpha\}$. This risk measure is called the *Value at Risk* at level α. One can intuitively think of $\mathrm{VaR}_\alpha(L)$ as the smallest value of x such that $P(L \leq x) \geq \alpha$. We can rewrite

$$\begin{aligned}\mathrm{VaR}_\alpha &= \inf\{x \in \mathbb{R} : 1 - P(L \leq x) \leq 1 - \alpha\} \\ &= \inf\{x \in \mathbb{R} : F_L(x) \geq \alpha\} \\ &= F_L^{\leftarrow}(\alpha) =: q_\alpha(F_L)\end{aligned}$$

with F_L denoting the distribution function of L and $F_L^{\leftarrow}$ the generalised inverse of F_L. Since F_L is a distribution function, the generalised inverse coincides with the quantile function $q_{(\cdot)}(L)$. So a statistician would simply call VaR_α the α-quantile of L. See the appendix for more information on generalised inverses and their properties. ○

Example 3.3. For $\alpha > 0$ and F_L continuous and strictly increasing, we define

$$\overline{\mathrm{ES}}_\alpha(L) = E[L \mid L \geq \mathrm{VaR}_\alpha(L)]$$

called the *Expected Shortfall* at level α. This is the expected loss given that the loss has surpassed the Value at Risk. We immediately see that $\overline{\mathrm{ES}}_\alpha(L) \geq \mathrm{VaR}_\alpha(L)$ and that $\overline{\mathrm{ES}}_\alpha$ takes into account the severity of the loss in comparison to VaR_α. ○

We want to generalize the Expected Shortfall to also be valid for non-continuous distribution functions. We will need the following lemma.

Lemma 3.4. *If U is uniformly distributed on $(0,1)$ and L is a random variable with continuous distribution function F_L, then $L \overset{\mathrm{d}}{=} F_L^{\leftarrow}(U)$.*

Proof. Let $Y = F_L^{\leftarrow}(U)$. A property of generalised inverses (see the appendix) is that $u \leq F_L(x)$ if and only if $F_L^{\leftarrow}(u) \leq x$. Hence

$$P(Y \leq x) = P(F_L^{\leftarrow}(U) \leq x) = P(U \leq F_L(x)) = F_L(x)$$

since U is uniformly distributed on $(0,1)$. This proves $Y \overset{\mathrm{d}}{=} L$ as desired. ■

The following proposition tells us how to generalize the notion of Expected Shortfall.

Proposition 3.5. *Let L be a loss variable with F_L continuous and strictly increasing. Then*

$$\overline{\mathrm{ES}}_\alpha(L) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(L)du.$$

Proof. Because F_L is continuous and strictly increasing, the generalised inverse is a proper inverse. Hence

$$P(L \geq \mathrm{VaR}_\alpha(L)) = 1 - \alpha$$

so by the previous lemma,

$$\begin{aligned}\overline{\mathrm{ES}}_\alpha &= \frac{1}{P(L \geq \mathrm{VaR}_\alpha(L))} E[L 1_{\{L \geq \mathrm{VaR}_\alpha(L)\}}] = \frac{1}{1-\alpha} E[L; L \geq \mathrm{VaR}_\alpha(L)] \\ &= \frac{1}{1-\alpha} E[F_L^{\leftarrow}(U); F_L^{\leftarrow}(U) \geq F_L^{\leftarrow}(\alpha)] = \frac{1}{1-\alpha} E[F_L^{\leftarrow}(U); U \geq \alpha] \\ &= \frac{1}{1-\alpha} E[\mathrm{VaR}_U(L); U \geq \alpha] = \frac{1}{1-\alpha} \int_\alpha^1 \mathrm{VaR}_u(L) du.\end{aligned}$$

■

Most problems include distributions that have continuous and strictly increasing distribution functions, but it is also useful to have a definition of Expected Shortfall for general distributions. The following definition summarises the above considerations.

Definition 3.6. Let $\alpha \in (0,1)$ and L a loss variable. We let

$$\mathrm{VaR}_\alpha(L) = \inf\{x \in \mathbb{R} : P(L > x) \leq 1 - \alpha\}$$

denote the Value at Risk at level α. If $E[|L|] < \infty$, we define

$$\mathrm{ES}_\alpha(L) = \frac{1}{1-\alpha} \int_\alpha^1 \mathrm{VaR}_u(L) du$$

called the Expected Shortfall at level α.

Properties/axioms of risk measures

It is natural to ask what characterises a good risk measure. We think intuitively of a risk measure as an amount of capital needed by a financial institution to withstand large shocks. In the above examples with VaR and ES, we had a level α, and we often think of α as large, for example $0.95, 0.99$ or 0.995. In an article by Artzner, Delbaen, Eber and Heath [1], certain desirable properties of risk measures are suggested. They are as follows.

Definition 3.7. For a risk measure ρ and loss variables L, L_1, L_2, we consider the following axioms/properties:

1. *Translation invariance*: $\rho(L + a) = \rho(L) + a$ for every constant $a \in \mathbb{R}$.
2. *Subadditivity*: $\rho(L_1 + L_2) \leq \rho(L_1) + \rho(L_2)$.
3. *Positive homogeneity*: $\rho(\lambda L) = \lambda \rho(L)$ for all $\lambda > 0$.
4. *Monotonicity*: $L_1 \leq L_2$ implies $\rho(L_1) \leq \rho(L_2)$.

A risk measure satisfying all these axioms is called *coherent*.

The rationale for translation invariance is that adding a deterministic quantity to the loss should increase the capital we need to set aside by exactly that amount. The rationale for subadditivity is that diversification should reduce risk. Positive homogeneity makes sense since if we invest more money into the samme asset, the amount of capital we need to set aside should be multiplied by the same factor. Monotonocity also clearly makes sense. Note also the similarity to the pricing principles from non-life insurance.

Value at Risk and Expected Shortfall are the two most popular choices of risk measures. One reason many prefer Expected Shortfall over Value at Risk is that Expected Shortfall in general satisfies all the above axioms, while Value at Risk only satisfies three.

Proposition 3.8. *Let $\alpha \in (0,1)$ and let L be a loss variable.*

(i) For constants $a > 0, b \in \mathbb{R}$, we have

$$\mathrm{VaR}_\alpha(aL+b) = a\,\mathrm{VaR}_\alpha(L) + b$$

so in particular, VaR_α satisfies translation invariance and positive homogeneity.

(ii) VaR_α satisfies monotonicity.

Proof. For (i), we compute

$$\begin{aligned}\mathrm{VaR}_\alpha(aL+b) &= \inf\{x \in \mathbb{R} : P(aL+b > x) \leq 1-\alpha\} \\ &= \inf\{x \in \mathbb{R} : P(L > (x-b)/a) \leq 1-\alpha\} \\ &= \inf\{ax+b \in \mathbb{R} : P(L > x) \leq 1-\alpha\} \\ &= a\inf\{x \in \mathbb{R} : P(L > x) \leq 1-\alpha\} + b = a\,\mathrm{VaR}_\alpha(L) + b.\end{aligned}$$

(ii) If $L_1 \leq L_2$ then $\{x \in \mathbb{R} : P(L_2 > x) \leq 1-\alpha\} \subseteq \{x \in \mathbb{R} : P(L_1 > x) \leq 1-\alpha\}$ and the claim follows. ■

Value at Risk is not a coherent risk measure in general. A counterexample can be found in [17], example 2.25. The reader can construct a continuous counterexample as an exercise.

Theorem 3.9. *Expected Shortfall is a coherent risk measure.*

Proof. Translation invariance, positive homogeneity and monotonicity follow immediately by the previous proposition and the definition of Expected Shortfall in terms of the Value at Risk. Subadditivity is harder to prove. We only prove it in the case where L_1 and L_2 have continuous distribution functions. The proof of the general case can be found in [17], see Theorem 8.14. The following proof is from Example 2.26 of the same book. Recall from earlier computations that for L with a continuous distribution function,

$$\mathrm{ES}_\alpha(L) = \frac{1}{1-\alpha}E[L1_{\{L \geq \mathrm{VaR}_\alpha(L)\}}].$$

Define $I_i := 1_{\{L_i \geq \mathrm{VaR}_\alpha(L_i)\}}$ for $i = 1,2$ and $I_{12} := 1_{\{L_1+L_2 \geq \mathrm{VaR}_\alpha(L_1+L_2)\}}$. We compute

$$\begin{aligned}(1-\alpha)(\mathrm{ES}_\alpha(L_1) + \mathrm{ES}_\alpha(L_2) - \mathrm{ES}_\alpha(L_1+L_2)) &= E[L_1I_1] + E[L_2I_2] - E[(L_1+L_2)I_{12}] \\ &= E[(L_1(I_1 - I_{12}))] + E[(L_2(I_2 - I_{12}))].\end{aligned}$$

We now consider two cases for L_1. If $L_1 \geq \mathrm{VaR}_\alpha(L_1)$, then $I_1 - I_{12} \geq 0$ and hence $L_1(I_1 - I_{12}) \geq \mathrm{VaR}_\alpha(L_1)(I_1 - I_{12})$. If $L_1 < \mathrm{VaR}_\alpha(L_1)$, then $I_1 - I_{12} \leq 0$ so again we have

$L_1(I_1 - I_{12}) \geq \mathrm{VaR}_\alpha(L_1)(I_1 - I_{12})$. Applying the same reasoning to L_2, we get

$$\begin{aligned}(1-\alpha)(\mathrm{ES}_\alpha(L_1) + \mathrm{ES}_\alpha(L_2) - \mathrm{ES}_\alpha(L_1 + L_2)) &\geq E[\mathrm{VaR}_\alpha(L_1)(I_1 - I_{12}) + \mathrm{VaR}_\alpha(L_2)(I_2 - I_{12})] \\ &= \mathrm{VaR}_\alpha(L_1)E[I_1 - I_{12}] + \mathrm{VaR}_\alpha(L_2)E[I_2 - I_{12}] \\ &= \mathrm{VaR}_\alpha(L_1)((1-\alpha) - (1-\alpha)) \\ &+ \mathrm{VaR}_\alpha(L_2)((1-\alpha) - (1-\alpha)) \\ &= 0\end{aligned}$$

implying that $\mathrm{ES}_\alpha(L_1) + \mathrm{ES}_\alpha(L_2) \geq \mathrm{ES}_\alpha(L_1 + L_2)$ which is the desired statement.

■

Computing VaR and ES

We start by considering VaR and ES in some concrete examples. Afterwards, we consider methods for computing these risk measures directly from data and how we can form confidence intervals.

Example 3.10 (Stock investment). Recall the example on stock investments from earlier. If S_n denotes the price of the stock at time n, we had the loss variable $L_{n+1} = -S_n(e^{X_{n+1}}-1)$ where $X_{n+1} = \log S_{n+1} - \log S_n$ denotes the log return. We assume $X_{n+1} \sim \mathcal{N}(\mu, \sigma^2)$. In this case, L_{n+1} has a non-zero density, so we can compute $\mathrm{VaR}_\alpha(L)$ by solving the equation $P(L_{n+1} > x) = 1 - \alpha$:

$$\begin{aligned}1 - \alpha = P(-S_n(e^{X_{n+1}} - 1) > x) &= P\left(X_{n+1} < \log\left(1 - \frac{x}{S_n}\right)\right) \\ &= P\left(\frac{X_{n+1} - \mu}{\sigma} < \frac{\log\left(1 - \frac{x}{S_n}\right) - \mu}{\sigma}\right) = \Phi\left(\frac{\log\left(1 - \frac{x}{S_n}\right) - \mu}{\sigma}\right)\end{aligned}$$

with Φ denoting the distribution function of a standard normal variable. Hence

$$\frac{\log\left(1 - \frac{x}{S_n}\right) - \mu}{\sigma} = \Phi^{-1}(1-\alpha)$$

and we can solve for x explicitly as follows:

$$\begin{aligned}\sigma\Phi^{-1}(1-\alpha) + \mu = \log\left(1 - \frac{x}{S_n}\right) \quad &\Leftrightarrow \quad e^{\sigma\Phi^{-1}(1-\alpha)+\mu} = 1 - \frac{x}{S_n} \\ &\Leftrightarrow \quad x = S_n - S_n e^{\sigma\Phi^{-1}(1-\alpha)+\mu}.\end{aligned}$$

If we consider more stocks, the expression becomes a lot more complicated. It gets even worse with a more diverse portfolio (with stocks, bonds and call options for example).

○

Example 3.11. This is from Example 2.11 and 2.14 from [17]. Let $\alpha \in (0,1)$ and assume that the loss distribution F_L is normal distributed with mean μ and variance σ^2. Since F_L is continuous and strictly increasing, we have

$$\mathrm{VaR}_\alpha(L) = \mu + \sigma\Phi^{-1}(\alpha)$$

using the properties of Value at Risk from before. We can now compute the Expected Shortfall, where we again use that F_L is continuous and strictly increasing,

$$\begin{aligned}\mathrm{ES}_\alpha(L) = E[L \mid L \geq q_\alpha(L)] &= \mu + \sigma E\left[\frac{L-\mu}{\sigma} \mid \frac{L-\mu}{\sigma} \geq q_\alpha\left(\frac{L-\mu}{\sigma}\right)\right] \\ &= \mu + \sigma E\left[\frac{L-\mu}{\sigma} \mid \frac{L-\mu}{\sigma} \geq \Phi^{-1}(\alpha)\right] \\ &= \mu + \frac{\sigma}{1-\alpha}\int_{\Phi^{-1}(\alpha)}^{\infty} x\varphi(x)dx\end{aligned}$$

with φ denoting the density of a standard normal variable. Note that

$$\varphi'(x) = \frac{d}{dx}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right) = -x\varphi(x)$$

so that $x\varphi(x)$ has $-\varphi(x)$ as antiderivative. Hence

$$\mathrm{ES}_\alpha(L) = \mu + \frac{\sigma}{1-\alpha}[-\varphi(x)]_{\Phi^{-1}(\alpha)}^{\infty} = \mu + \sigma\frac{\varphi(\Phi^{-1}(\alpha))}{1-\alpha}.$$

○

We now consider methods for computing VaR and ES directly from data. Up to time t_n we have the empirical observations $L_1, ..., L_n$. For this discussion, we make the (not so realistic) assumption that $L_1, ..., L_n$ is an iid sample. Order the sample and let

$$L_{1,n} \geq L_{2,n} \geq ... \geq L_{n,n}$$

denote the corresponding order statistics. We define the empirical distribution function

$$F_L^{(n)}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{[L_i,\infty)}(x)$$

that assigns equal weight to each observation. To estimate the risk measures of interest, an idea is to replace F_L (which is unknown) with the empirical distribution function $F_L^{(n)}$. This idea is justified by e.g. the Strong Law of Large Numbers which implies that

$$F_L^{(n)}(x) \to P(L_1 \leq x) \quad \text{a.s. as} \quad n \to \infty$$

for each x or the even stronger result known as the *Glivenko-Cantelli Theorem* which states that

$$\sup_{x\in\mathbb{R}} |F_L^{(n)}(x) - F_L(x)| \to 0 \quad \text{a.s. as} \quad n \to \infty.$$

We can form the natural estimator of the Value at Risk given by

$$\widehat{\mathrm{VaR}}_\alpha(L) = \inf\{x \in \mathbb{R} : 1 - F_L^{(n)} \leq 1 - \alpha\}.$$

A problem with this approach is that it is often very difficult to infer the tail behaviour of F_L from $F_L^{(n)}$ since we often do not have enough data in the tail of F_L. However, one nice thing about this approach is that we can explicitly solve for the estimator. We have

$$\widehat{\mathrm{VaR}}_\alpha(L) = L_{[n(1-\alpha)]+1,n} = \widehat{q}_\alpha(F_L)$$

i.e. the empirical quantile. Here $[n(1-\alpha)]$ denotes the largest integer less than or equal to $n(1-\alpha)$. Similarly, if F_L is continuous and strictly increasing, we may compute an estimate for $\mathrm{ES}_\alpha(L)$, namely

$$\widehat{\mathrm{ES}}_\alpha(L) = \frac{\sum_{i=1}^{[n(1-\alpha)]+1} L_{i,n}}{[n(1-\alpha)]+1}$$

i.e. the average of the observations greater than or equal to $\widehat{\mathrm{VaR}}_\alpha(L)$. Again we stress that this approximation is often insufficient since we usually do not have many observations in the tails of F_L. We now turn to the problem of computing confidence intervals where we focus on the Value at Risk. Let $\beta \in (0,1)$ denote some "small" value. Our approach is to find $\widehat{A}$ and $\widehat{B}$ such that

$$P(\widehat{A} < \mathrm{VaR}_\alpha(L) < \widehat{B}) \geq 1-\beta$$

by determining $\widehat{A}$ and $\widehat{B}$ in such a way that

$$P(\mathrm{VaR}_\alpha(L) \leq \widehat{A}) \leq \frac{\beta}{2} \quad \text{and} \quad P(\mathrm{VaR}_\alpha(L) \geq \widehat{B}) \leq \frac{\beta}{2}.$$

Assume that L has a density so that for the true $\mathrm{VaR}_\alpha(L)$ we have $P(L > \mathrm{VaR}_\alpha(L)) = 1-\alpha$. For the observed data, each data point can land on either side of $\mathrm{VaR}_\alpha(L)$. Hence we have a sequence of Bernoulli trials with probability $1-\alpha$ of landing to the right of $\mathrm{VaR}_\alpha(L)$ (a success) and probability α of landing to the left of $\mathrm{VaR}_\alpha(L)$ (a failure). Formally, let $Z = 1_{\{L > \mathrm{VaR}_\alpha(L)\}}$, then

$$q := P(Z=0) = \alpha, \quad p := P(Z=1) = 1-\alpha.$$

Let Y denote the number of successes in n trials i.e. $Y = \sum_{i=1}^n Z_i$ for $Z_i = 1_{\{L_i > \mathrm{VaR}_\alpha(L)\}}$ where L_i is the ith loss. Then $Y \sim \mathrm{Bin}(n,p)$ and

$$P(Y \geq j) = \sum_{k=j}^{n} \binom{n}{k} p^k q^{n-k}.$$

Note that $L_{j,n} \geq \mathrm{VaR}_\alpha(L)$ if and only if $Y \geq j$ so using the above, we can find the smallest j such that $P(Y \geq j) \leq \beta/2$. Then $P(L_{j,n} \geq \mathrm{VaR}_\alpha(L)) = P(Y \geq j) \leq \beta/2$ and so we set $\widehat{A} = L_{j,n}$. Conversely, we can find the largest i such that $P(Y \leq i) \leq \beta/2$. Then $P(L_{i,n} \leq \mathrm{VaR}_\alpha(L)) \leq \beta/2$ so set $\widehat{B} = L_{i,n}$.

Notes and comments

Chapter 1 of [17] contains a longer informal introduction to the field of quantitative risk management as well as more historical perspectives. Section 2.3 contains more perspectives on different types of risk measurements. These include the notational-amount approach and the scenario-based risk measures. [6] is a good reference on the Black-Scholes model and a very readable introduction to pricing of financial derivatives.

Exercises

Exercise 1.1:
Consider a loss variable L which is exponential distributed i.e. $L \sim \text{Exp}(\lambda)$ for $\lambda > 0$.

1)Compute $\text{VaR}_\alpha(L)$.

2)Compute $\text{ES}_\alpha(L)$.

Exercise 1.2:
Consider a loss variable L with distribution function F given by

$$F(x) = \begin{cases} 0, & \text{if } x < 1 \\ 1 - \frac{1}{1+x}, & \text{if } 1 \leq x < 3 \\ 1 - \frac{1}{x^2}, & \text{if } x \geq 3 \end{cases}$$

1)Compute $\text{VaR}_{0.85}(L)$.

2)Compute $\text{VaR}_{0.95}(L)$ and $\text{VaR}_{0.99}(L)$.

3)Compute $\text{ES}_{0.85}(L)$.
Hint: Draw the graph of F.

Exercise 1.3:
Let L be a loss variable with distribution function

$$F(x) = \frac{1}{1 + e^{-\frac{x-\mu}{s}}}$$

where $\mu \in \mathbb{R}$ and $s > 0$ parameters. This distribution is called the logistic distribution with location μ and scale s. Let $\alpha \in (0, 1)$.

1)Compute $\text{VaR}_\alpha(L)$.

2)Compute $\text{ES}_\alpha(L)$.

Exercise 1.4:
Consider the risk measure $\rho(L) = E[L]$. Show/convince yourself that ρ is a coherent risk measure. Explain why ρ may still be a bad risk measure.

Exercise 1.5:
In this exercise, we will show that the stochastic process given by

$$S_t = se^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t}$$

is a solution to the stochastic differential equation

$$dS_t = \mu S_t dt + \sigma S_t dW_t, \quad S_0 = s.$$

We will apply the *Itô formula*. The Itô formula says that if we have a continuous time stochastic process X with differential

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t,$$

and a C^2 function f, then the process $Z_t = f(t, X_t)$ has stochastic differential given by

$$
\begin{aligned}
dZ_t = & \left(\frac{\partial f}{\partial t}(t, X_t) + \mu(t, X_t) \frac{\partial f}{\partial x}(t, X_t) + \frac{1}{2} \sigma(t, X_t)^2 \frac{\partial^2 f}{\partial x^2}(t, X_t) \right) dt \\
& + \sigma(t, X_t) \frac{\partial f}{\partial x}(t, X_t) dW_t.
\end{aligned}
$$

1)Identify the function f such that $S_t = f(t, W_t)$. Compute $\frac{\partial f}{\partial t}$, $\frac{\partial f}{\partial x}$ and $\frac{\partial^2 f}{\partial x^2}$.
2)Apply the Itô formula to show that S_t satisfies the stochastic differential equation.

Exercise 1.6:
Consider the Standard Black-Scholes model and the derivative that pays $X = \log S_T$ at time T. Determine the arbitrage free price of this derivative at time $t < T$. Assume the natural filtration generated by the Brownian motion. Hint: It may be helpful to consult the subsection in the appendix on stochastic processes.

Week 2 - Methods for computing VaR and extreme value theory

4 Computing Value at Risk

Last week we introduced methods to obtain estimates for the Value at Risk and the Expected Shortfall from data. We continue this discussion where we focus on the Value at Risk. We will introduce four methods, namely

(i) The Variance-Covariance (Var-Cov) method,

(ii) Monte Carlo simulation,

(iii) Importance Sampling and

(iv) Bootstrapping.

The motivating example to keep in mind is the one with d investments from last week. Recall that the loss was given by

$$L_{n+1} = -\sum_{i=1}^{d} \alpha_i S_n^{(i)} \left(e^{X_{n+1}^{(i)}} - 1 \right)$$

with $S_n^{(i)}$ the value of the ith asset at time n, $X_{n+1} = \log S_{n+1} - \log S_n$ the log return and α_i the number of assets bought of asset i. We now go through the different methods.

The Var-Cov method

Consider the linearized loss

$$L_{n+1}^{\Delta} = -\sum_{i=1}^{d} \alpha_i S_n^{(i)} X_{n+1}^{(i)}$$

obtained by the Taylor approximation $e^x \approx 1 + x$. In the Variance-Covariance method, we assume that $\mathbf{X}_{n+1} \sim \mathcal{N}(\mathbf{m}, \Sigma)$ i.e. a multivariate normal distribution with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Letting

$$\mathbf{X}_{n+1} = \begin{pmatrix} X_{n+1}^{(1)} \\ \vdots \\ X_{n+1}^{(d)} \end{pmatrix}, \quad \mathbf{w}_n = \begin{pmatrix} \alpha_1 S_n^{(1)} \\ \vdots \\ \alpha_d S_n^{(d)} \end{pmatrix},$$

we may rewrite $L^{\Delta}_{n+1} = -\langle \mathbf{w}_n, \mathbf{X}_{n+1}\rangle = -\mathbf{w}_n^T\mathbf{X}_{n+1}$. By the properties of the multivariate normal distribution (see the appendix), we have

$$L^{\Delta}_{n+1} \sim \mathcal{N}(-\mathbf{w}_n^T\mathbf{m}, \mathbf{w}_n^T\Sigma\mathbf{w}_n).$$

Let $\mu_n = -\mathbf{w}_n^T\mathbf{m}$ and $\sigma_n^2 = \mathbf{w}_n^T\Sigma\mathbf{w}_n$, so that we may write $L^{\Delta}_{n+1} \overset{\text{d}}{=} \mu_n + \sigma_n Z$ for $Z \sim \mathcal{N}(0,1)$. From last week, we can compute $\mathrm{VaR}_\alpha(L^{\Delta}_{n+1})$ as

$$\mathrm{VaR}_\alpha(L^{\Delta}_{n+1}) = \mu_n + \sigma_n\Phi^{-1}(\alpha).$$

Finally, we can use data to obtain estimates of μ and σ^2 i.e. of $\mathbf{m}$ and Σ. A virtue of this method is how simple it is to use. We get an exact analytical expression for $\mathrm{VaR}_\alpha(L^{\Delta}_{n+1})$. The problem is the approximation $e^x \approx 1 + x$ behind the method. This is not very precise for large losses, and often we are interested in the tail of the distribution where x is large. Also, the normality assumption is often problematic with real data.

Monte Carlo simulation

Monte Carlo simulation is a purely computational method. Suppose we simulate N samples of $\mathbf{X}_{n+1}$ and call these $\mathbf{x}_1, ..., \mathbf{x}_N$. We work with these simulated values as if they were empirical samples. From these we can form "empirical" samples of L_{n+1}. Call these $l_1, ..., l_N$. We use these to compute the empirical Value at Risk. Formally, start by ordering the samples to obtain the order statistics

$$l_{1,N} \geq ... \geq l_{N,N}.$$

Then we can compute the estimate

$$\mathrm{VaR}_\alpha(L_{n+1}) \approx \widehat{\mathrm{VaR}}_\alpha(L_{n+1}) = l_{[N(1-\alpha)]+1,N}$$

per the discussion last week. This method is a lot more precise compared to the Variance-Covariance method since it does not rely on the approximation $e^x \approx 1+x$. Another obvious advantage is flexibility. The method works for any distribution (at least if we can efficiently simulate large samples from that distribution). One problem is that the rate of convergence of the estimate will be slow since we are working with the tail of the loss distribution. Hence we often have to generate a very large number of samples to obtain a robust estimate. This is especially problematic if we have a large number of assets.

Rare event simulation

Before moving on to importance sampling, we take a brief detour and consider a problem closely related to the one above for the Monte Carlo method, namely estimating $p_x = P(L > x)$ for large values of x. To estimate p_x, we generate a computational iid sample $l_1, ..., l_N$ of L. We can then form the indicators $1_{\{l_1 > x\}}, ..., 1_{\{l_N > x\}}$. By the SLLN,

$$\frac{1}{N}\sum_{i=1}^{N} 1_{\{l_i > x\}} \to E[1_{\{L > x\}}] = P(L > x) = p_x \quad \text{a.s.}$$

Define the natural estimator

$$\widehat{p}_x^{(N)} = \frac{1}{N}\sum_{i=1}^{N} 1_{\{l_i > x\}}.$$

As shown above, this estimator is consistent (it converges in probability to the true underlying parameter, in this case p_x). A natural question to ask is how fast $\widehat{p}_x^{(N)}$ converges. If

$$S_N = \frac{1}{N}\sum_{i=1}^{N} 1_{\{l_i > x\}},$$

then the CLT applies to show that

$$\frac{S_N - Np_x}{\sigma_x\sqrt{N}} \xrightarrow{\text{d}} Z \sim \mathcal{N}(0,1)$$

where σ_x denotes the variance of $1_{\{l_1 > x\}}$. Note that $N\widehat{p}_x^{(N)} = S_N$ so for large N, we get (in distribution) that

$$Z \approx \frac{S_N - Np_x}{\sigma_x\sqrt{N}} = \frac{\widehat{p}_x^{(N)} - p_x}{\sigma_x/\sqrt{N}}.$$

We can rewrite this relation and obtain the approximation (in distribution)

$$\widehat{p}_x^{(N)} \approx p_x + \frac{\sigma_x}{\sqrt{N}}Z.$$

We can form an asymptotic confidence interval for p_x as follows. Let $\beta \in (0,1)$ be some "small" value and let $z_{\beta/2}$ denote the $\beta/2$-quantile of $\mathcal{N}(0,1)$ so that $P(Z > z_{\beta/2}) = \beta/2$. Thus, with the "high" probability $1-\beta$, we have

$$p_x \in \left(\widehat{p}_x^{(N)} - \frac{\sigma_x}{\sqrt{N}}z_{\beta/2}, \widehat{p}_x^{(N)} + \frac{\sigma_x}{\sqrt{N}}z_{\beta/2}\right).$$

While the error $\sigma_x/\sqrt{N}$ goes to zero for $N \to \infty$, the probability p_x is often also small, so it can happen that the error still dominates the estimate $\widehat{p}_x^{(N)}$, even when N is very large. To make this precise, define the relative error

$$\text{RE} = \frac{\sigma_x z_{\beta/2}}{\sqrt{N}p_x} = \frac{\sigma_x}{p_x}C(N)$$

with $C(N)$ some constant depending on N. We compute

$$\sigma_x^2 = \text{Var}[1_{\{L>x\}}] = E[1_{\{L>x\}}^2] - E[1_{\{L>x\}}]^2 = p_x - p_x^2.$$

We now have for the relative error that

$$\frac{\sigma_x}{p_x} = \frac{\sqrt{p_x - p_x^2}}{p_x} = \sqrt{\frac{1}{p_x} - 1} \to \infty \quad \text{as} \quad x \to \infty.$$

So the relative error explodes as x gets large. Hence we need more sophisticated techniques so that the relative error is bounded. We present one such method very briefly now.

Importance sampling

To remedy the issue of diverging relative errors, we briefly introduce the main ideas of importance sampling. Consider the setup $L = f(\mathbf{X})$ where $\mathbf{X}$ is an $\mathbb{R}^d$-valued random variable with distribution function F (see the appendix for a brief discussion on multidimensional distribution functions) and f is a deterministic function. We refer to F as the *true* distribution of $\mathbf{X}$. Define the *moment-generating function* (mgf) of $\mathbf{X}$ as

$$\kappa(\xi) = E\left[e^{\langle \xi, \mathbf{X} \rangle}\right] \quad \text{for} \quad \xi \in \mathbb{R}^d.$$

Consider the *shifted distribution* given by

$$dF_\xi(x_1, ..., x_d) = \frac{e^{\langle \xi, \mathbf{X} \rangle}}{\kappa(\xi)} dF(x_1, ..., x_d)$$

for all $\xi \in \mathbb{R}^d$ such that the moment-generating function is finite. We note the following properties of F_ξ:

(i) F_ξ is a probability distribution.

(ii) $E_\xi[\mathbf{X}] = \nabla \Lambda(\xi)$ where $\Lambda(\xi) = \log \kappa(\xi)$ is the *cumulant-generating function* of X and E_ξ indicates the expectation taken with respect to the shifted measure.

Property (i) follows easily by the calculation

$$\int dF_\xi(\mathbf{x}) = \frac{1}{\kappa(\xi)} \int e^{\langle \xi, \mathbf{X} \rangle} dF(\mathbf{x}) = \frac{\kappa(\xi)}{\kappa(\xi)} = 1.$$

The idea is now to choose ξ in a good way such that $L > x$ occurs frequently i.e. so that $P_\xi(L > x)$ is large, where P_ξ denotes the probability under the shifted distribution. To relate simulations under the shifted distribution with parameter ξ to the original probability, we apply a representation formula,

$$\begin{aligned} p_x = P(L > x) &= \int_{\{\mathbf{y}: f(\mathbf{y}) > x\}} dF(\mathbf{y}) = \int_{\{\mathbf{y}: f(\mathbf{y}) > x\}} \frac{dF}{dF_\xi}(\mathbf{y}) dF_\xi(\mathbf{y}) \\ &= E_\xi\left[1_{\{f(\mathbf{X}) > x\}} \frac{dF}{dF_\xi}(\mathbf{X})\right] \end{aligned}$$

with $\frac{dF}{dF_\xi}$ the Radon-Nikodym derivative.

Bootstrapping

The idea of bootstrapping is to use the existing data to generate new data by resampling. We sample with replacement and if we have N data points $x_1, ..., x_N$, we choose a member with probability $1/N$ N times to get a new data set of the same size. We can do this procedure b times to obtain the bootstrap samples

$$\begin{gathered} x_1^{(1)}, ..., x_N^{(1)} \\ x_1^{(2)}, ..., x_N^{(2)} \\ \vdots \\ x_1^{(b)}, ..., x_N^{(b)}. \end{gathered}$$

If we have some parameter θ that we want to estimate, we can compute an empirical estimate $\widehat{\theta}$ from the original data and estimates $\widehat{\theta}_j$ from each of the b bootstrap samples. We can then use the $\widehat{\theta}_j$ to say something about the distribution of $\widehat{\theta}$. We may for example generate confidence intervals by computing empirical quantiles using the $\widehat{\theta}_j$. Explicitly, define the residuals $R_j = \widehat{\theta} - \widehat{\theta}_j$ and order them from smallest to largest, $R_{1,b} \geq R_{2,b} \geq ... \geq R_{b,b}$. The confidence bounds are then given by

$$\left[\widehat{\theta} + R_{[b(1-\beta/2)],b}, \widehat{\theta} + R_{[b(\beta/2)]+1,b}\right].$$

These bounds improve classical estimates based on the CLT which converge more slowly.

For small probabilities, one considers a modification of this idea, namely *smoothed bootstrap*. In smoothed bootstrapping, one smoothes the data around the tail values. Namely, given the original sample $x_1, ..., x_N$, sample from the density

$$g(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

where K is a smooth function, e.g.

$$K(t) = \frac{1}{2\pi} e^{-t^2/2}.$$

One can then employ importance sampling to shift the smoothed distribution so that more samples will be in the tail.

5 Extreme value theory

We now move on to the other topic for this week, namely extreme value theory. We start by briefly discussing a method for assessing a distributional assumption on the underlying data, namely the QQ plot. Afterwards, we briefly introduce distributions with regularly varying tails and we discuss methods of doing statistics with such distributions.

Data analysis

Let $x_1, ..., x_n$ be observed data. If we want to use this data to make predictions about future values, it is natural to propose a distribution F and assume that $x_1, ..., x_n$ are realizations of iid variables $X_1, ..., X_n$ with distribution F. The QQ plot (quantile-quantile plot) can be used to determine whether F is a reasonable choice of distribution. The QQ plot consists of the points

$$\left\{\left(F^{\leftarrow}\left(\frac{n-k+1}{n+1}\right), x_{k,n}\right) : k = 1, ..., n\right\}$$

where as usual, $x_{1,n} \geq ... \geq x_{k,n}$ denotes the order statistics of the sample. Recall that

$$F_n^{\leftarrow}\left(\frac{n-k+1}{n+1}\right) = x_{k,n}$$

where F_n is the empirical distribution function. If F is the true underlying distribution function for the data, $F_n \to F$ a.s., so that

$$x_{k,n} = F_n^{\leftarrow}\left(\frac{n-k+1}{n+1}\right) \approx F^{\leftarrow}\left(\frac{n-k+1}{n+1}\right).$$

Hence we expect the QQ plot to be a straight line through 0 and with slope 1 if the data truly comes from the reference distribution. The plot tells a bit more however. If the true underlying distribution is an affine transformation of F, the plot will still be a straight line. If that is the case, we can estimate proper location and scale parameters. The plot also indicates whether the reference distribution has lighter or heavier tails than the empirical sample. See the plots below.

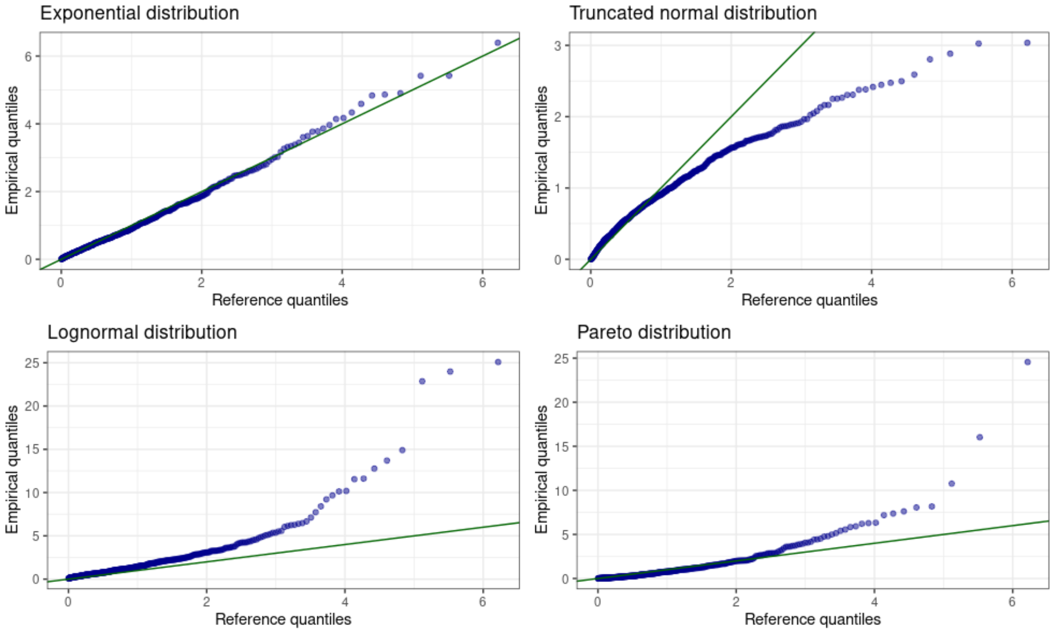

Figure 1: Four examples of QQ-plots. We have 500 simulated values from the following distributions: Standard exponential (upper left), folded/truncated normal (i.e. $|X|$ for $X \sim \mathcal{N}(0,1)$) (upper right), standard lognormal (i.e. $\log X \sim \mathcal{N}(0,1)$, lower left) and the Pareto distribution with $\kappa = 1$ and $\alpha = 2$ (lower right). The reference distribution is standard exponential. The green line has slope one and intercept zero.

Consider the plots for a moment. The plot in the upper left corner shows that the data follows a straight line with slope one and intercept zero which is expected, since the reference distribution and empirical distribution are the same. For the truncated normal, we see that the data curves downwards, indicating lighter tails than the exponential distribution. For the lognormal and Pareto distributions, the data curves upwards, indicating heavier tails than the exponential distribution.

Standard distributions in statistics include the normal, exponential and gamma distributions. These all have in common that they are light-tailed in the sense that the mgf of these variables exists in some neighbourhood around zero. This is not typical behaviour for log returns in a financial situation. These distributions are usually a lot more heavy-tailed. Heavy-tailed distributions include the regularly varying distributions such as the Pareto along with moderately heavy-tailel distributions such as the lognormal. In these cases, the mgf does not exist. We will now scratch the surface of the theory of regularly varying distributions.

The regularly varying class

Definition 5.1. A measurable function $h : (0, \infty) \to (0, \infty)$ is called *regularly varying* if there is a $\rho \in \mathbb{R}$ such that for all $t > 0$,

$$\lim_{x\to\infty} \frac{h(tx)}{h(x)} = t^\rho.$$

In this case, we write $h \in \mathrm{RV}_\rho$. If $h \in \mathrm{RV}_0$, we call h *slowly varying.*

Example 5.2. Constant functions and log are examples of slowly varying functions. ○

We will often use the following characterisation of regular variation.

Proposition 5.3. *$h \in RV_\rho$ if and only if $h(x) = L(x)x^\rho$ for L a slowly varying function.*

Proof. Assume $h \in \mathrm{RV}_\rho$. Define the function $L(x) = h(x)x^{-\rho}$. Then for $t > 0$, we have

$$\lim_{x\to\infty} \frac{L(tx)}{L(x)} = \lim_{x\to\infty} \frac{h(tx)(xt)^{-\rho}}{h(x)x^{-\rho}} = \lim_{x\to\infty} \frac{h(tx)}{h(x)} t^{-\rho} = t^\rho t^{-\rho} = 1$$

so L is slowly varying and satisfies $h(x) = L(x)x^\rho$. The converse implication is also easy and is left to the reader. ■

Definition 5.4. A distribution function F is *regularly varying* if $\overline{F} = 1 - F \in \mathrm{RV}_{-\alpha}$ for some $\alpha > 0$. α is called the *index* of F.

The definition says that F is regularly varying of index $\alpha > 0$ if for all $t > 0$, it holds that

$$\frac{P(X > tx)}{P(X > x)} \to t^{-\alpha} \quad \text{as} \quad x \to \infty$$

where X has distribution function F. Equivalently by the above proposition, $P(X > x) = L(x)x^{-\alpha}$ for a slowly varying function L. Let us consider a very important example of a regularly varying distribution.

Definition 5.5. The *Generalised Pareto Distribution* (GPD) has distribution function

$$G_{\gamma,\beta}(x) = 1 - \left(1 + \frac{\gamma x}{\beta}\right)^{-1/\gamma}, \quad x > 0$$

where $\gamma > 0$ and $\beta > 0$ are parameters.

Lemma 5.6. *The Generalised Pareto Distribution with parameters $\beta, \gamma > 0$ is regularly varying with index $1/\gamma$.*

Proof. Simply note that

$$\overline{G}_{\gamma,\beta}(x) = \left(1 + \frac{\gamma x}{\beta}\right)^{-1/\gamma}$$

so for $t > 0$, we have

$$\lim_{x\to\infty} \frac{\overline{G}_{\gamma,\beta}(tx)}{\overline{G}_{\gamma,\beta}(x)} = \lim_{x\to\infty} \left(\frac{1 + \frac{\gamma t x}{\beta}}{1 + \frac{\gamma x}{\beta}}\right)^{-1/\gamma} = \lim_{x\to\infty} \left(\frac{\frac{\beta}{\gamma x} + t}{\frac{\beta}{\gamma x} + 1}\right)^{-1/\gamma} = t^{-1/\gamma}.$$

■

From a statistical perspective, it is of interest to determine the tail parameter α if the underlying distribution of the data is assumed have a regularly varying distribution. We present two such methods. In the following, we assume $\overline{F} \in \mathrm{RV}_{-\alpha}$ so that $\overline{F}(x) = L(x)x^{-\alpha}$, and the goal is to estimate α.

The Hill estimator

Deriving the Hill estimator is based on the following result by Karamata.

Theorem 5.7 (**Karamata's Theorem**). *If L is slowly varying and $\beta < -1$, then*

$$\int_u^\infty x^\beta L(x)dx \sim -\frac{1}{\beta+1}u^{\beta+1}L(u), \quad u \to \infty.$$

In words, $\int_u^\infty x^\beta L(x)dx$ behaves asymptotically like the integral of a power function. The slowly varying function plays little role asymptotically. We can now derive the Hill estimator. Let $\beta = -\alpha - 1$ where $\alpha > 0$. Karamata's Theorem implies

$$\int_u^\infty x^{-\alpha-1} L(x)dx \sim \frac{1}{\alpha}u^{-\alpha}L(u), \quad u \to \infty$$

so for $\overline{F}(x) = L(x)x^{-\alpha}$, we have

$$\int_u^\infty x^{-1}\overline{F}(x)dx \sim \frac{1}{\alpha}\overline{F}(u), \quad u \to \infty.$$

We rewrite the left hand side. First note that $(\log x - \log u)' = 1/x$. We can now apply integration by parts (see the appendix for a review) and obtain

$$\int_u^\infty x^{-1}\overline{F}(x)dx = \left[(\log x - \log u)\overline{F}(x)\right]_u^\infty - \int_u^\infty (\log x - \log u)d\overline{F}(x).$$

$\overline{F}(x)$ decays like $x^{-\alpha}$ and hence decays to zero faster than $\log x$ grows to ∞ and thus the first term is zero. By using that $d\overline{F}(x) = d(1 - F(x)) = -dF(x)$, we get

$$\int_u^\infty x^{-1}\overline{F}(x)dx = \int_u^\infty (\log x - \log u)dF(x)$$

and so

$$\frac{1}{\overline{F}(u)}\int_u^\infty (\log x - \log u)dF(x) \to \frac{1}{\alpha}, \quad u \to \infty.$$

This is a theoretical result. To turn this into an estimator, we have to use the empirical distribution function. Suppose we have iid data $x_1, ..., x_n$ distributed according to F. Order the samples, $x_{1,n} \geq ... \geq x_{n,n}$. Let F_n denote the empirical distribution function. We can then approximate F by F_n. Replacing F by F_n in the above expression yields

$$\frac{1}{\alpha} \approx \frac{1}{\overline{F}_n(u)}\int_u^\infty (\log x - \log u)dF_n(x).$$

for sufficiently large u. We want to simplify this expression. Let N_u denote the number of observations greater than u i.e.

$$N_u = \#\{i : x_i > u\},$$

then $\overline{F}_n(u) = N_u/n$. Also recall that $\overline{F}_n(x_{k,n}) = (k-1)/n$. Now choose a "small" k and set $u = x_{k,n}$. Then

$$\begin{aligned}\frac{1}{\alpha} &\approx \frac{n}{k-1}\int_u^\infty (\log y - \log x_{k,n})dF_n(y) = \frac{n}{k-1}\sum_{j=1}^{k}\frac{1}{n}(\log x_{j,n} - \log x_{k,n}) \\ &= \frac{1}{k-1}\sum_{j=1}^{k}(\log x_{j,n} - \log x_{k,n})\end{aligned}$$

since each jump of F_n is of size $1/n$ and the jth jump of F_n occurs at $x_{j,n}$ (see the appendix). Replacing $k-1$ by k gives us the Hill estimator.

Definition 5.8. Let $x_1, ..., x_n$ be a sample from a regularly varying distribution with index α. Let $x_{1,n} \geq ... \geq x_{n,n}$ denote the order statistics. We call

$$\widehat{\alpha}_k = \left(\frac{1}{k}\sum_{j=1}^{k}(\log x_{j,n} - \log x_{k,n})\right)^{-1}$$

the *Hill estimator* of α.

Remark 5.9. Note that $\widehat{\alpha}_k$ depends on k i.e. the threshold.

It is natural to ask how we choose a good value of k. Choosing k small, we get few data points which increases the variance of the estimator, so the estimate is not sufficiently robust. On the other hand, choosing k large makes the approximation based on Karamata's Theorem imprecise. Furthermore, it often happens with real data that the center and the tail have very different distributions, so taking too many data points close to the center makes the estimator biased. One often chooses k based on a *Hill plot.* A Hill plot consists of the value pairs

$$\{(k, \widehat{\alpha}_k) : k = 2, ..., n\}$$

and the value of k is chosen in a region where the estimator looks stable. We stress that this does not always occur for real data.

Example 5.10. A very classical data set in extreme value theory is the Danish fire insurance data. The data consists of large Danish fire insurance claims from 1980 to 1990. The data is available in the R package `evir` which has functions to compute estimates and make plots related to extreme value theory. We present some plots of the data below.

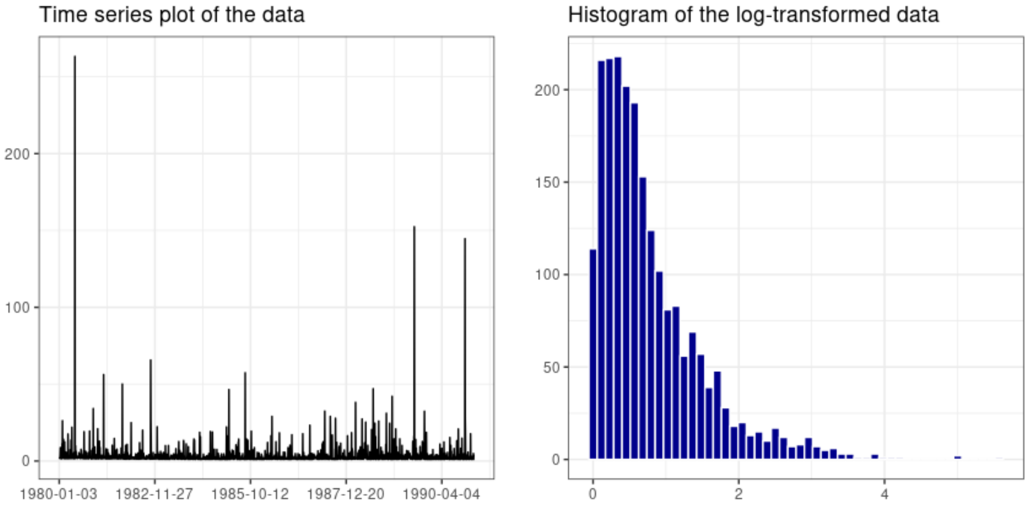

Figure 2: Exploratory plots of the Danish fire insurance data.

Below is a Hill plot of the data:

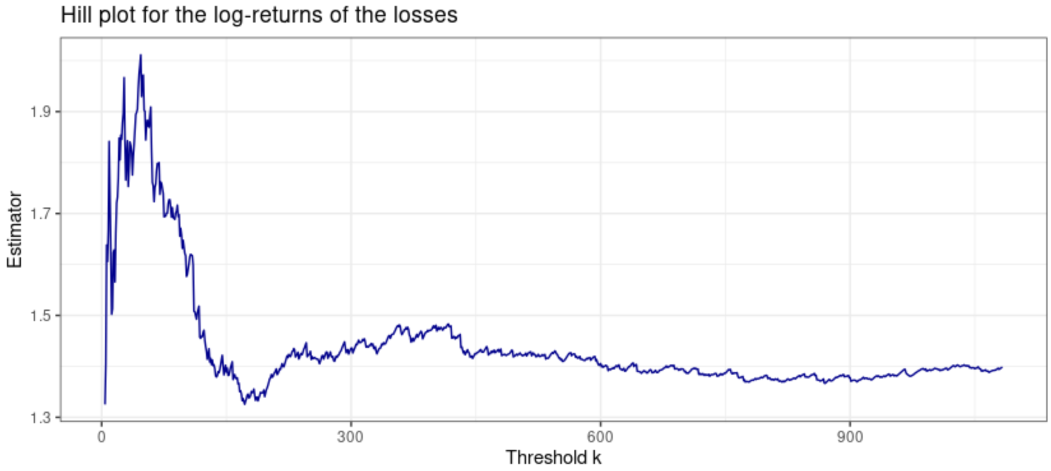

Figure 3: A Hill plot of the Danish fire insurance data.

The plot looks stable from around $k = 300$. We have $\widehat{\alpha}_{300} = 1.4357$, and this is one of many estimates we can choose to report. There is no single correct answer for the choice of k and very often, real life data is much more ugly than this particular data set. This illustrates the importance of applying different tools in extreme value theory before drawing a conclusion. ○

Using the Hill estimator, we can compute the Value at Risk at level β, VaR_β (the letter α is already used). We want to solve for x in the equation $P(X > x) = 1 - \beta$. Choose $k, x_{k,n}$ by looking at the Hill plot. Since X is regularly varying,

$$\overline{F}(x) = \overline{F}\left(\frac{x}{x_{k,n}} x_{k,n}\right) \sim \left(\frac{x}{x_{k,n}}\right)^{-\alpha} \overline{F}(x_{k,n}) \quad \text{as} \quad x \to \infty.$$

Now replace F by the empirical distribution function F_n. Since $\overline{F}_n(x_{k,n}) = (k-1)/n$, we obtain

$$\overline{F}(x) \approx \frac{k-1}{n}\left(\frac{x}{x_{k,n}}\right)^{-\widehat{\alpha}_k}.$$

Now solve for x and set $\mathrm{VaR}_\beta(X) = x$.

The Peaks over threshold (POT) method

Real data often has two ”components”, namely a center component and a tail component. Often the tail of the data is described by a different distribution than the values close to the center. The following definition captures the idea of considering the tail component.

Definition 5.11. Given a distribution function F and a positive threshold u, we define the *excess distribution function* F_u via the tail

$$\overline{F}_u(x) = P(X > u + x \mid X > u) = \frac{\overline{F}(u+x)}{\overline{F}(u)}, \quad x \geq 0.$$

Rearranging, the above definition can be expressed as $\overline{F}(u+x) = \overline{F}(u)\overline{F}_u(x)$ for $x \geq 0$. If we have a sample $x_1, ..., x_n$ and $N_u = \#\{i : x_i > u\}$, then $\overline{F}(u) \approx N_u/n$. To estimate $\overline{F}_u(x)$, we apply the generalized Pareto distribution. By the assumption of regular variation, we have

$$\frac{\overline{F}(tu)}{\overline{F}(u)} \to t^{-\alpha} \quad \text{for} \quad u \to \infty.$$

Hence for large u,

$$\overline{F}_u(x) = \frac{\overline{F}\left(\left(1+\frac{x}{u}\right)u\right)}{\overline{F}(u)} \approx \left(1 + \frac{x}{u}\right)^{-\alpha}.$$

There is some cheating involved here. $t = 1 + x/u$ depends on u, but since $t \to 1$ as $u \to \infty$, it is not really an issue. Recall that the GPD with parameters $\beta, \gamma > 0$ has tail

$$\overline{G}_{\gamma,\beta}(x) = \left(1 + \frac{\gamma x}{\beta}\right)^{-1/\gamma} = \left(1 + \frac{x}{u}\right)^{-\alpha} \quad \text{for} \quad x \geq 0$$

where $\gamma = 1/\alpha$ and $\beta = \beta(u) = u/\alpha$ (while β depends on u, one chooses a fixed u so that β is also fixed). We can now describe the POT method in two steps:

(i) Estimate $\overline{F}(u) \approx N_u/n$.

(ii) Approximate $\overline{F}_u(x)$ by a GPD with parameters $\gamma, \beta > 0$.

There are two things we need to elaborate on concerning step (ii). First of all, we need to choose a threshold u. Second, we need methods for estimating β and γ. Let us first address the second issue. If we have a sample $x_1, ..., x_n$, we start by discarding all $x_i \leq u$. We are then left with a subsample $z_1, ..., z_{N_u}$ with $z_i > u$ for all $i = 1, ..., N_u$. We then estimate β and γ via maximal likelihood based on this subsample. The likelihood function is

$$L(\gamma, \beta; z_1, ..., z_{N_u}) = \prod_{i=1}^{N_u} g_{\gamma,\beta}(z_i) \quad \text{for} \quad g_{\gamma,\beta}(x) = \frac{d}{dx} G_{\gamma,\beta}(x).$$

We can be specific and compute

$$g_{\gamma,\beta}(x) = \frac{1}{\beta}\left(1 + \frac{\gamma x}{\beta}\right)^{-1/\gamma - 1}, \quad x > 0.$$

We can then consider the log-likelihood

$$\begin{aligned} l(\gamma, \beta; z_1, ..., z_{N_u}) &= \log L(\gamma, \beta; z_1, ..., z_{N_u}) = \sum_{i=1}^{N_u} \log g_{\gamma,\beta}(z_i) \\ &= \sum_{i=1}^{N_u} \left(-\log \beta - \frac{\gamma + 1}{\gamma} \log \left(1 + \frac{\gamma z_i}{\beta}\right)\right) \\ &= -N_u \log \beta - \frac{\gamma + 1}{\gamma} \sum_{i=1}^{N_u} \log \left(1 + \frac{\gamma z_i}{\beta}\right) \end{aligned}$$

and maximising this equation numerically in terms of β and γ yields the maximal likelihood estimators $\hat{\beta}_n$ and $\hat{\gamma}_n$. It is possible to construct asymptotic confidence intervals by using the result (valid for $\gamma > -1/2$)

$$\sqrt{N_u}\left(\widehat{\gamma}_n - \gamma, \frac{\widehat{\beta}_n}{\beta} - 1\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, M^{-1}) \quad \text{for} \quad N_u \to \infty$$

where

$$M^{-1} = (1 + \gamma)\begin{pmatrix} 1 + \gamma & -1 \\ -1 & 2 \end{pmatrix}.$$

We now return to the first problem of determining a good value of u. As usual there is a tradeoff between getting sufficiently many datapoints and choosing a value large enough so that the asymptotics "kick in". The main tool for this job is the mean-excess function.

Definition 5.12. Let X be an integrable random variable with distribution F. The *mean-excess function* of X is defined as

$$e(u) = E[X - u \mid X > u].$$

In the following we assume that the tail parameter α satisfies $\alpha > 1$. We compute

$$e(u) = \int_0^\infty (x - u) dP(X \leq x \mid X > u) = \frac{1}{\overline{F}(u)} \int_u^\infty (x - u) dF(x)$$

and using integration by parts, we get

$$\int_u^\infty (x-u)dF(x) = -\int_u^\infty (x-u)d\overline{F}(x) = \left[-(x-u)\overline{F}(x)\right]_u^\infty + \int_u^\infty \overline{F}(x)dx = \int_u^\infty \overline{F}(x)dx$$

since $\overline{F}(x)$ decays to zero slower than x by the assumption $\alpha > 1$. Using Karamata's Theorem, Theorem 5.7, we get

$$\int_u^\infty \overline{F}(x)dx = \int_u^\infty L(x)x^{-\alpha}dx \sim -\frac{L(u)u^{-\alpha+1}}{-\alpha+1} \quad \text{for} \quad u \to \infty$$

so

$$e(u) \sim \frac{1}{L(u)u^{-\alpha}} \frac{L(u)u^{-\alpha+1}}{\alpha-1} = \frac{u}{\alpha-1} \quad \text{as} \quad u \to \infty$$

which shows that $e(u)$ becomes linear asymptotically. This is a crucial observation and hence we state the above result as a proposition.

Proposition 5.13. *If F is regularly varying with index $\alpha > 1$, the mean-excess function $e(u)$ satisfies*

$$e(u) \sim \frac{u}{\alpha-1} \quad \textit{as} \quad u \to \infty.$$

To see how the mean-excess function helps in determining a suitable u, we consider the empirical mean-excess function. The idea is to replace F and $\overline{F}$ by their empirical counterparts. The *empirical mean-excess function* is given by

$$\begin{aligned} e_n(u) &= \frac{1}{N_u/n}\int_u^\infty (x-u)dF_n(x) = \frac{n}{N_u}\sum_{j=1}^n (x_{j,n}-u)1_{\{x_{j,n}>u\}}\frac{1}{n} \\ &= \frac{1}{N_u}\sum_{j=1}^n (x_{j,n}-u)1_{\{x_{j,n}>u\}}. \end{aligned}$$

If we set $u = x_{k,n}$ for some $k = 2,3,...,n$ ($k = 1$ is excluded since $e_n(u) = 0$ in this case), we can simplify the above expression to

$$e_n(x_{k,n}) = \frac{1}{k-1}\sum_{j=k+1}^n (x_{j,n} - x_{k,n}).$$

Using the empirical mean-excess function we can construct a *mean-excess plot* by plotting the values

$$\left\{(x_{k,n}, e_n(x_{k,n})) : k = 2,3,...,n\right\}.$$

If the values $x_1,...,x_n$ come from a regularly varying distribution, the plot will roughly look like a straight line for large thresholds. For distributions with lighter tails, the mean-excess function will either decrease or remain roughly constant. It is left to the reader to compute some examples of mean-excess functions in the exercises. Some examples of mean-excess plots are below.

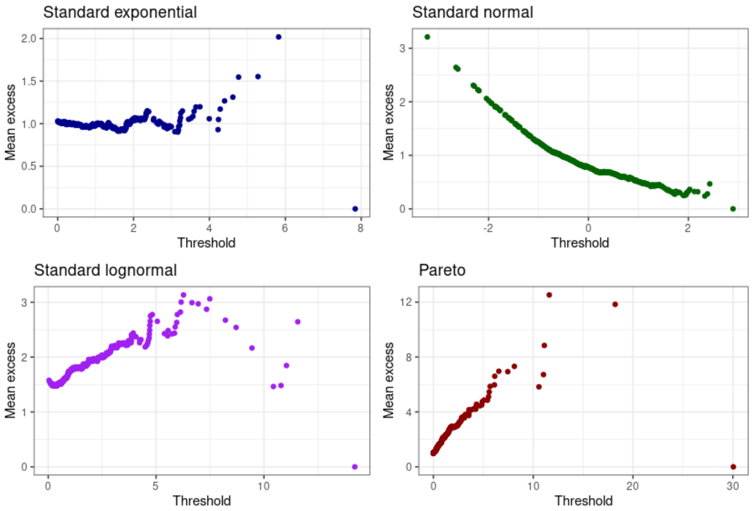

Figure 4: Examples of mean-excess plots based on 500 simulated values from the following distributions: Standard exponential (upper left), standard normal (upper right), standard lognormal (lower left) and the Pareto distribution with $\kappa = 1$ and $\alpha = 2$ (lower right).

The plots should serve not just as an example but also as a warning. The behaviour of the large values in the plots are very chaotic, and one should be cautious in the interpretation of such plots. To illustrate this further, the following plots are made with the exact same distributions but with a different simulated sample.

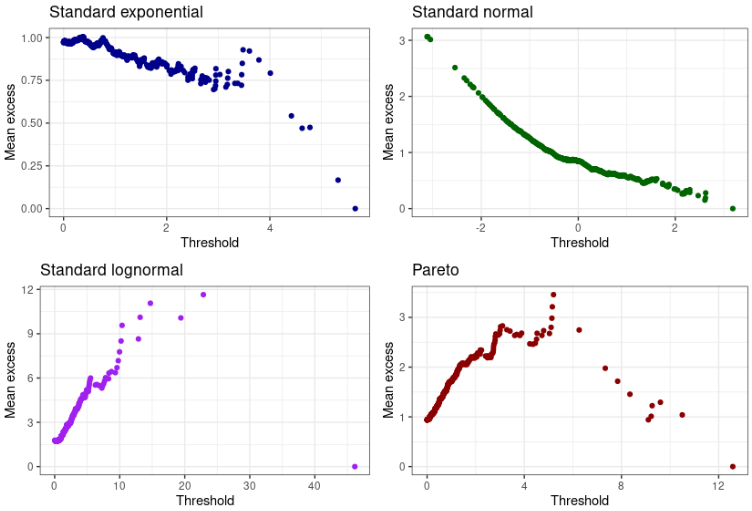

Figure 5: More mean-excess plots with the same distributions as before, but with a new sample.

It often occurs in practice that the tail and center of the data have different distributions.

For example, the tail could have a regularly varying distribution while the center has a light-tailed distribution such as a normal or gamma distribution. In that case, the mean-excess plot will probably only become linear for large values of u. In any case, one should use the plot to find a value of u where the points begin to form a straight line. We illustrate this with an example.

Example 5.14. Let us again consider the Danish fire insurance data. We want to model the excesses using the POT method. We wish to determine a proper threshold and therefore make a mean-excess plot:

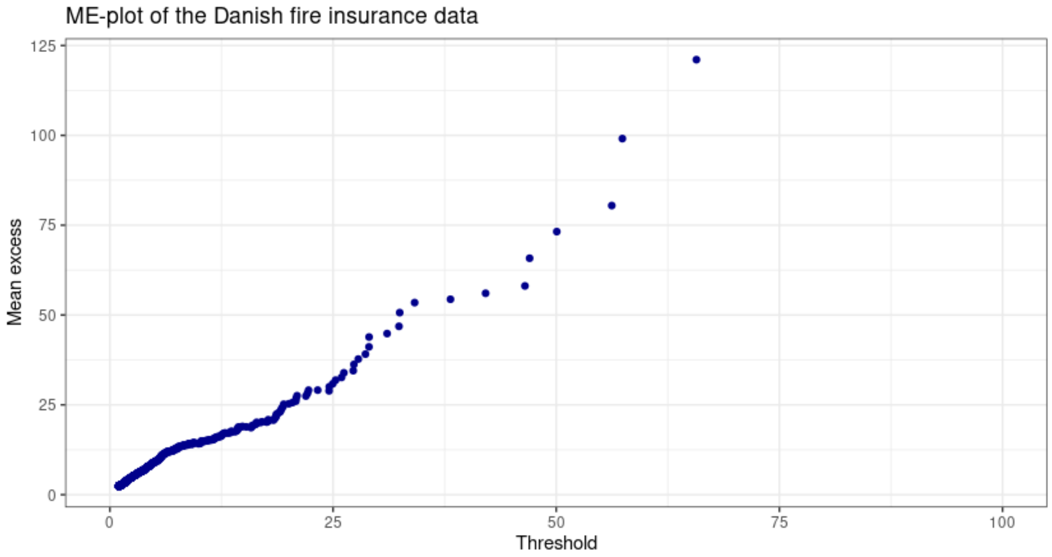

Figure 6: Mean-excess plot of the Danish fire insurance data.

The plot looks very linear for all thresholds, indicating Pareto tails. A word of warning: This is typically not the case for real data! The Danish fire insurance data is in a sense too nice to illustrate this point. Based on the plot, we choose the threshold $u = 4$. While the linear trend may begin for slightly larger values, this choice also gives us sufficient data to work with. To fit the generalized Pareto distribution, we apply maximum likelihood estimation using the `gpd` function in the `evir` package as follows:

```r
data(danish)
u <- 4
gpdfit <- gpd(danish, threshold = u, method = "ml")
gpdfit$par.ests
```

This gives the estimates $\widehat{\gamma} = 0.7209$ and $\widehat{\beta} = 2.6291$. We now plot the empirical tail and the tail from the POT approximation:

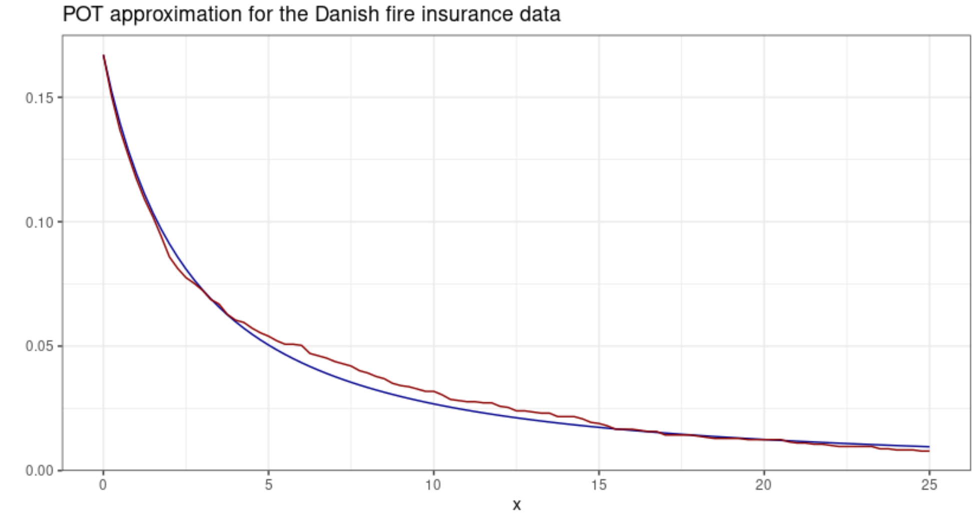

Figure 7: Blue: The tail from the fitted generalized Pareto distribution. Red: The empirical tail from the data.

The approximation looks very good. Usually, one is not so lucky. If one is interested in a statistical goodness of fit, a possible way (which is also implemented in `evir` if one uses `plot` around a fitted gpd object) is to consider the excesses $z_i = x_i - u$ which should be approximately $G_{\widehat{\gamma},\widehat{\beta}}$ distributed. Hence $-\log G_{\widehat{\gamma},\widehat{\beta}}(z_i)$ (which we call the *generalised residuals*) should be approximately standard exponential distributed. We make a residual plot and a QQ-plot of the $-\log G_{\widehat{\gamma},\widehat{\beta}}(z_i)$ against a standard exponential distribution and get the following:

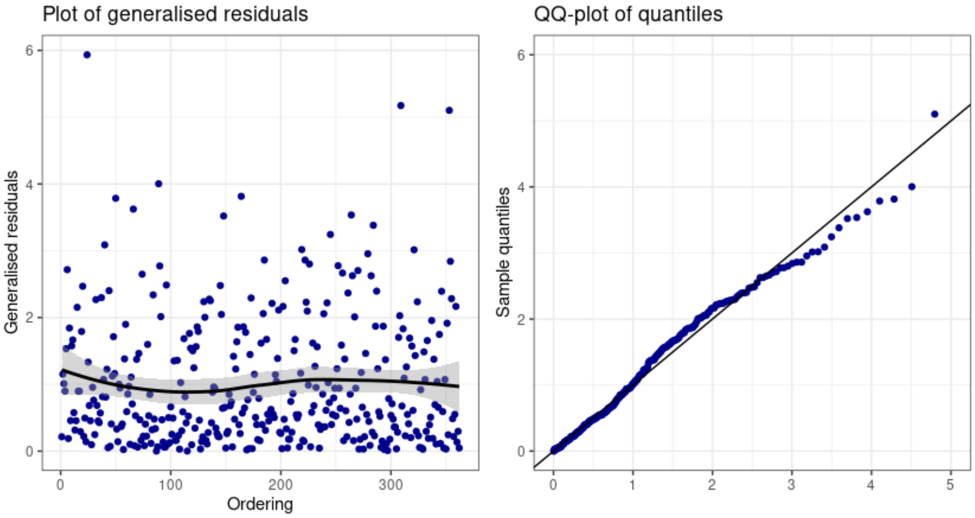

Figure 8: Diagnostic plots for the fitted GPD.

We see no clear tendencies of the generalised residuals. Some of them are quite large, but otherwise the left plot looks good. The right plot also looks good. While some of the sample quantiles are below the line with slope one and intercept zero, the residuals overall seem to follow a standard exponential distribution. We conclude that the model is an adequate fit. ◦

Notes and comments

The idea of bootstrapping was proposed in the famous paper by B. Efron, [8]. For more information on importance sampling and Monte Carlo methods, consult [12] and [2]. [5] contains all the information one could wish for concerning regular variation, including a proof of Karamata's Theorem. [10] is an excellent source for extreme value theory. Chapter 3.4 covers the GPD and chapter 6 covers statistical methods, including the Hill estimator and the POT method.

Exercises

Exercise 2.1:
For the shifted distribution

$$dF_\xi(x_1, ..., x_d) = \frac{e^{\langle \xi, \mathbf{X} \rangle}}{\kappa(\xi)} dF(x_1, ..., x_d)$$

as presented in the subsection on importance sampling above, prove that

$$E_\xi[\mathbf{X}] = \nabla\Lambda(\xi).$$

Exercise 2.2:
In this exercise, we will get more comfortable with the concept of regular and slow variation.

1)Prove the remaining implication in Proposition 5.3.

2)Verify that the function $h(x) = \log\log x$ for x sufficiently large is slowly varying.

3)Show that the Pareto distribution with parameters $\kappa > 0$ and $\alpha > 0$ is regularly varying. What is the index? Recall that the distribution function is given by

$$1 - \left(\frac{\kappa}{\kappa + x}\right)^\alpha, \quad x > 0.$$

Exercise 2.3:
Consider the Student t distribution with $\nu > 0$ degrees of freedom, i.e. the distribution with density

$$g_\nu(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

Show that this distribution is regularly varying and determine the corresponding index.

Exercise 2.4:
Consider a regularly varying distribution function F supported on $[0, \infty)$ of index $\alpha > 0$. Let $X \sim F$. Prove that $E[X^\beta] < \infty$ if $\beta < \alpha$. Hint: Apply Karamata's Theorem. Recall also the formula $E[X] = \int_0^\infty P(X > t)dt$ for a positive random variable X.

One can use a different form of the Karamata theorem to show that $E[X^\beta] = \infty$ when $\beta > \alpha$.

Exercise 2.5:
Karamata's Representation Theorem says that if L is slowly varying, then we can write

$$L(x) = c_0(x) e^{\int_{x_0}^x \frac{\varepsilon(t)}{t} dt}$$

where $c_0(x) \to c_0 > 0$ for $x \to \infty$ and $\varepsilon(x) \to 0$ for $x \to \infty$ for some $x_0 \geq 0$. If L is written in this form, we call the above a *Karamata representation* for L.

1)Find a Karamata representation for log.

2)Prove that a function of the form above is slowly varying.

Exercise 2.6:
Consider the distribution F with tail $\overline{F}(x) = x^{-2}$ for $x > 1$. Then $\overline{F} \in \mathrm{RV}_{-2}$.

1)Implement the Hill estimator (in R for example) and a function that can simulate values from F.

2)Simulate 50 values of F and make a Hill plot using your function from before. Also plot the true value of the index as a line in the plot.

3)Repeat for 100, 250 and 1000 values. Comment on the results.

Exercise 2.7:
Recall that we proved the following formula for the mean-excess function of an integrable random variable X:

$$e(u) = \frac{1}{\overline{F}(u)} \int_u^\infty \overline{F}(x)dx.$$

1)Let $X \sim \mathrm{Exp}(\lambda)$. Prove that $e(u) = 1/\lambda$.

2)Let X be Pareto distributed with parameters $\kappa > 0$ and $\alpha > 1$. Prove that

$$e(u) = \frac{\kappa + u}{\alpha - 1}.$$

3)Relate the results to the mean-excess plots in the discussion above.

Exercise 2.8:
The goal of this exercise is to prove that if $X \geq 0$ is a random variable with distribution function F and

$$\lim_{x\to\infty} \frac{\overline{F}(x-y)}{\overline{F}(x)} = e^{\gamma y}, \quad y > 0$$

for some $\gamma \in (0, \infty)$, then

$$e(u) \to \gamma^{-1} \quad \text{for} \quad u \to \infty.$$

1)Prove that $\overline{F} \circ \log \in \mathrm{RV}_\gamma$.

2)Prove the above result. Hint: Karamata's Theorem.

3)The above result also holds for $\gamma \in \{0, \infty\}$, and you can use this without proof. Prove that if X is standard normal, then $e(u) \to 0$. Relate this to the plots in the discussion above.

Week 3 - Spherical and elliptical distributions

Multivariate random vectors: dependence

Consider again the "canonical example" of stock returns where the risk factors are the log returns $\mathbf{X}_{n+1}$ with $X_{n+1}^{(i)} = \log S_{n+1}^{(i)} - \log S_n^{(i)}$. What is the distribution of $\mathbf{X}_{n+1}$? Inspired by the Black-Scholes model, we could assume $\mathbf{X}_{n+1} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. There are several problems with the normal assumption, however. A typical problem is that assets are very often correlated in such a way that high (low) returns for one asset correlates with high (low) returns for another. The normal distribution is very light-tailed, so such correlations are often not captured. The plots below illustrate this issue.

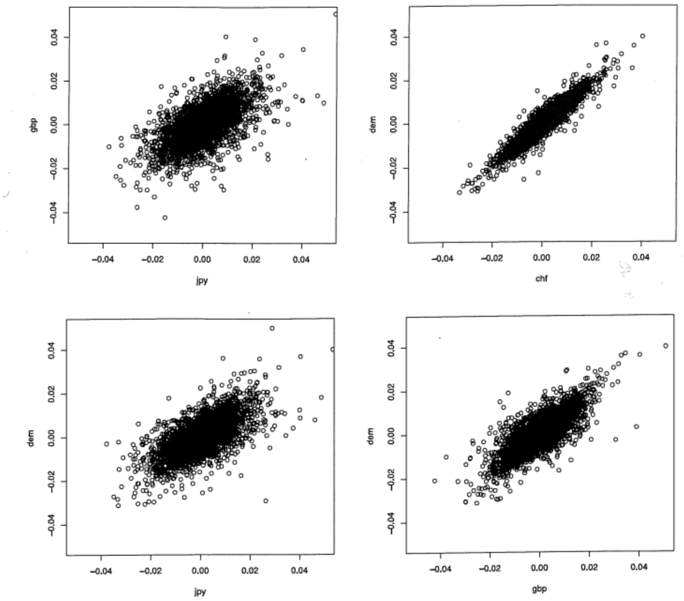

Figure 9: Log returns of foreign exchange rates quotes against the US dollar.

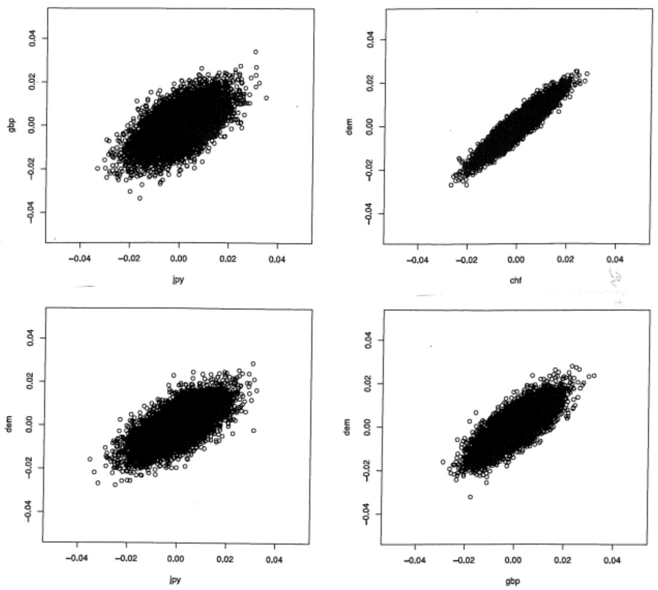

Figure 10: Simulated foreign exchange rates using a bivariate normal distribution with estimated means and covariance matrix.

From the plots, it is evident that the normal distribution fails to capture the dependency in the tails. Furthermore, the probability mass is too concentrated around the mean. The dependency in tails is a very typical phenomenon in financial data as illustrated in the plot below.

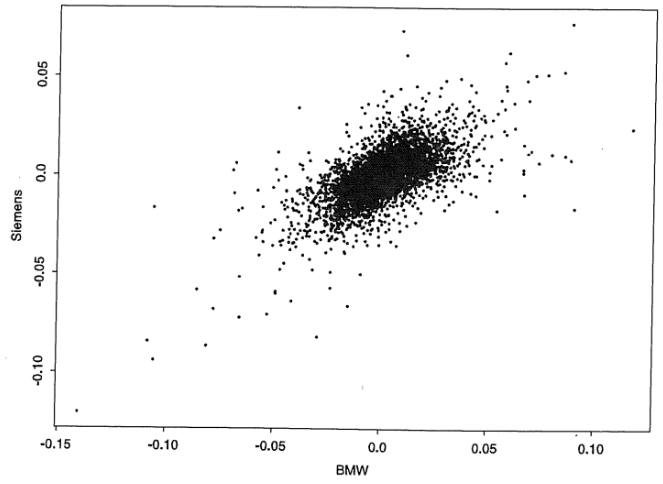

Figure 11: Log returns from BMW and Siemens stocks.

To remedy the issues with the normal distribution, we introduce a class of distributions that in some way resembles the normal distribution and shares a lot of its properties while also being more flexible in terms of modelling tail behaviour. This is the class of spherical and elliptical distributions.

6 Spherical and elliptical distributions

To motivate the spherical and elliptical distributions, we first briefly consider the multivariate normal distribution. If $\mathbf{X} \sim \mathcal{N}(0, I_d)$ (I_d denotes the d-dimensional identity matrix), then $\mathbf{X}$ has density (see the appendix)

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\sum_{i=1}^d x_i^2} = \frac{1}{(2\pi)^{d/2}} e^{-r^2/2}$$

with $r^2 = x_1^2 + \cdots + x_d^2$. Hence the density only depends on $\|\mathbf{x}\| = r$. Graphically, the level sets of f are spheres (or circles in two dimensions). One can say that $\mathcal{N}(0, I_d)$ is spherically symmetric/rotationally invariant. Define the random variable R by $R^2 = X_1^2 + \cdots + X_d^2$, then $R^2 \sim \chi^2(d)$ i.e. R^2 is Chi-square distributed with d degrees of freedom. We call R the *radial component* of $\mathbf{X}$. Intuitively, we can decompose $\mathbf{X}$ as $\mathbf{X} \overset{\text{d}}{=} R\mathbf{S}$ with $\mathbf{S}$ uniformly distributed on the d-dimensional unit sphere $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \sum_{i=1}^d x_i^2 = 1\}$. While this is an informal approach, it gives us the idea on how to proceed formally.

Definition 6.1. For a d-dimensional random vector $\mathbf{X}$, we define the *characteristic function* of $\mathbf{X}$ as

$$\Phi_{\mathbf{X}}(\mathbf{t}) = E[e^{i\mathbf{t}^T\mathbf{X}}], \quad \mathbf{t} \in \mathbb{R}^d.$$

Remark 6.2. Note the similarity to the moment-generating function $\kappa_{\mathbf{X}}(\mathbf{t}) = E[e^{\mathbf{t}^T\mathbf{X}}]$. These two transforms satisfy similar properties. For example, two random variables have the same distribution if and only if their characteristic functions are equal. See the appendix for more background on these functions. The characteristic function has the advantage that it always exists (since the integrand is bounded by one in norm) but it provides less information about the tail behaviour than the moment-generating function.

Example 6.3. If $\mathbf{X} \sim \mathcal{N}(0, I_d)$, simple calculations yield

$$\Phi_{\mathbf{X}}(\mathbf{t}) = e^{-\frac{1}{2}\mathbf{t}^T\mathbf{t}}.$$

Note that we can write $\Phi_{\mathbf{X}}(\mathbf{t}) = \psi(\|\mathbf{t}\|^2)$ for the function $\psi : \mathbb{R} \to \mathbb{R}$ given by $\psi(t) = e^{-t/2}$. This is a formal way of stating that $\Phi_{\mathbf{X}}$ doesn't depend on the direction of $\mathbf{t}$. ○

We can now introduce spherical distributions.

Definition 6.4. A random vector $\mathbf{X}$ in d dimensions has a *spherical distribution* if

$$\Phi_{\mathbf{X}}(\mathbf{t}) = \psi(\|\mathbf{t}\|^2) = \psi(t_1^2 + \cdots + t_d^2), \quad \mathbf{t} \in \mathbb{R}^d$$

for some univariate function ψ. ψ is called the *characteristic generator* of $\mathbf{X}$, and we write $\mathbf{X} \sim S_d(\psi)$.

If $\mathbf{X} \sim S_d(\psi), \mathbf{t} \in \mathbb{R}^d$ and $\mathbf{X}^\theta$ denotes $\mathbf{X}$ rotated by θ (and similarly for $\mathbf{t}^\theta$), we have

$$\Phi_{\mathbf{X}^\theta}(\mathbf{t}^\theta) = E\left[e^{i\langle \mathbf{t}^\theta, \mathbf{X}^\theta\rangle}\right] = E\left[e^{i\langle \mathbf{t}, \mathbf{X}\rangle}\right] = \Phi_{\mathbf{X}}(\mathbf{t}) = \psi(\|\mathbf{t}^\theta\|^2) = \Phi_{\mathbf{X}}(\mathbf{t}^\theta)$$

which is true for all $\mathbf{t}^\theta$. By the uniqueness of the characteristic function, $\mathbf{X}^\theta \overset{\mathrm{d}}{=} \mathbf{X}$. This gives a formal argument for the intuition of $\mathbf{X}$ being rotationally invariant. The following result gives an equivalent formulation of spherical distributions.

Proposition 6.5. *The following are equivalent:*

(i) $\mathbf{X}$ *has a spherical distribution.*

(ii) $\mathbf{X} \overset{\mathrm{d}}{=} R\mathbf{S}$ *with* $\mathbf{S}$ *uniformly distributed on the unit sphere* $\mathbb{S}^{d-1} := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$ *and* R *is a one-dimensional random variable independent of* $\mathbf{S}$.

Proof. We first prove that (ii) implies (i). We have

$$\begin{aligned}\Phi_{\mathbf{X}}(\mathbf{t}) = \Phi_{R\mathbf{S}}(\mathbf{t}) &= E\left[e^{i\langle \mathbf{t}, R\mathbf{S}\rangle}\right] = E\left[E\left[e^{i\langle \mathbf{t}, R\mathbf{S}\rangle} \mid R\right]\right] \\ &= E\left[E\left[e^{i\langle R\mathbf{t}, \mathbf{S}\rangle} \mid R\right]\right] = E[\Phi_{\mathbf{S}}(R\mathbf{t})]\end{aligned}$$

and since $\mathbf{S}$ is uniformly distributed on the unit sphere, $\Phi_{\mathbf{S}}$ only depends on the length and not the direction. Hence $\Phi_{\mathbf{X}}(\mathbf{t})$ also only depends on the length of $\mathbf{t}$ and $\mathbf{X}$ has a spherical distribution. We now show that (i) implies (ii). We have $\Phi_{\mathbf{X}}(\mathbf{t}) = \psi(\|\mathbf{t}\|^2)$. Set $\mathbf{s} = \mathbf{t}/\|\mathbf{t}\|$, then

$$\Phi_{\mathbf{X}}(\mathbf{t}) = E\left[e^{i\|\mathbf{t}\|\langle \mathbf{s}, \mathbf{X}\rangle}\right],$$

and by assumption, this does not depend on $\mathbf{s}$, only $\|\mathbf{t}\|$. Let $\mathbf{S}$ be uniformly distributed on the unit sphere with distribution function $F_{\mathbf{S}}$. Since $\Phi_{\mathbf{X}}(\mathbf{t})$ is constant in $\mathbf{s}$, we have

$$\begin{aligned}\Phi_{\mathbf{X}}(\mathbf{t}) &= \int_{\mathbb{S}^{d-1}} E\left[e^{i\|\mathbf{t}\|\langle \mathbf{s}, \mathbf{X}\rangle}\right] dF_{\mathbf{S}}(\mathbf{s}) = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}^d} e^{i\|\mathbf{t}\|\langle \mathbf{s}, \mathbf{x}\rangle} dF_{\mathbf{X}}(\mathbf{x}) dF_{\mathbf{S}}(\mathbf{s}) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{S}^{d-1}} e^{i\|\mathbf{t}\|\langle \mathbf{s}, \mathbf{x}\rangle} dF_{\mathbf{S}}(\mathbf{s}) dF_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbb{R}^d} \int_{\mathbb{S}^{d-1}} e^{i\langle \mathbf{s}, \|\mathbf{t}\|\mathbf{x}\rangle} dF_{\mathbf{S}}(\mathbf{s}) dF_{\mathbf{X}}(\mathbf{x}) \\ &= \int_{\mathbb{R}^d} E\left[e^{i\langle \|\mathbf{t}\|\mathbf{x}, \mathbf{S}\rangle}\right] dF_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbb{R}^d} E\left[e^{i\langle \|\mathbf{x}\|\mathbf{t}, \mathbf{S}\rangle}\right] dF_{\mathbf{X}}(\mathbf{x}) \\ &= E\left[E\left[e^{i\langle \|\mathbf{X}\|\mathbf{t}, \mathbf{S}\rangle} \mid \mathbf{X}\right]\right] = E\left[e^{i\langle \mathbf{t}, \|\mathbf{X}\|\mathbf{S}\rangle}\right] = E\left[e^{i\langle \mathbf{t}, R\mathbf{S}\rangle}\right] = \Phi_{R\mathbf{S}}(\mathbf{t})\end{aligned}$$

where we have defined $R := \|\mathbf{X}\|$. Hence $\mathbf{X} \overset{\mathrm{d}}{=} R\mathbf{S}$ where R and $\mathbf{S}$ have the desired properties. ■

While characterisation (ii) is more intuitive, it is easier to work with definition (i) when one wants to prove properties of spherical distributions. The following corollary tells how to compute R and $\mathbf{S}$ when we know that $\mathbf{X}$ is spherical.

Corollary 6.6. *Let* $\mathbf{X} \overset{\mathrm{d}}{=} R\mathbf{S}$ *be spherical. Then*

$$\left(\|\mathbf{X}\|, \frac{\mathbf{X}}{\|\mathbf{X}\|}\right) \overset{\mathrm{d}}{=} (R, \mathbf{S}).$$

Proof. The proof is from [17], see Corollary 6.22. Let $f_1(\mathbf{x}) = \|\mathbf{x}\|$ and $f_2(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$. Since $\mathbf{X} \overset{\mathrm{d}}{=} R\mathbf{S}$, we have

$$\left(\|\mathbf{X}\|, \frac{\mathbf{X}}{\|\mathbf{X}\|}\right) = (f_1(\mathbf{X}), f_2(\mathbf{X})) \overset{\mathrm{d}}{=} (f_1(R\mathbf{S}), f_2(R\mathbf{S})) = (R, \mathbf{S})$$

as desired. ■

We now turn to a generalisation of spherical distributions, namely elliptical distributions.

Definition 6.7. A d-dimensional random vector $\mathbf{X}$ has an *elliptical distribution* if $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Y}$ with $\mathbf{Y} \sim S_k(\psi)$ and A is a $d \times k$ matrix. We write $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ for $\Sigma = AA^T$. We call $\boldsymbol{\mu}$ the *location parameter* and Σ the *dispersion matrix.*

As a motivation for this definition, suppose $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where Σ is positive definite. From linear algebra (see for example chapter 7 in [3]), we know that there exists some matrix A such that $AA^T = \Sigma$. If $Y \sim \mathcal{N}(0, I_d)$, then

$$\mathbf{X} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + A\mathbf{Y}.$$

There exist methods to find A such that $AA^T = \Sigma$. One such method is the *Cholesky factorisation.* This factorisation determines a lower triangular matrix A such that $AA^T = \Sigma$. In detail,

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & & \vdots \\ \vdots & \vdots & & \vdots \\ a_{d1} & a_{d2} & \cdots & a_{dd} \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{d1} \\ 0 & a_{22} & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{dd} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1d} \\ \Sigma_{21} & \Sigma_{22} & & \vdots \\ \vdots & \vdots & & \vdots \\ \Sigma_{d1} & \Sigma_{d2} & \cdots & \Sigma_{dd} \end{pmatrix}.$$

The algorithm can (somewhat informally) be described as follows: Since $\Sigma_{11} = a_{11}^2$, Σ_{11} determines a_{11}. Since $\Sigma_{21} = a_{11}a_{21}$, Σ_{21} determines a_{21} and so on. Since we can go back and forth between the matrix A and the matrix Σ, the notation $E_d(\boldsymbol{\mu}, \Sigma, \psi)$ makes sense. Using the characterisation of spherical distributions in terms of a radial component, we can also write

$$\mathbf{X} = \boldsymbol{\mu} + RA\mathbf{S}$$

for $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$. We now turn to some properties of elliptical distributions. Afterwards we will look at some examples.

7 Properties of elliptical distributions

We start by computing the characteristic function for an elliptical distribution.

Lemma 7.1. *If $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$, then*

$$\Phi_{\mathbf{X}}(\mathbf{t}) = e^{i\langle \mathbf{t}, \boldsymbol{\mu}\rangle} \psi(\mathbf{t}^T \Sigma \mathbf{t}).$$

Proof. The proof is a straightforward computation. Write $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Y}$ with $\mathbf{Y} \sim S_k(\psi)$. We have

$$\Phi_{\mathbf{X}}(\mathbf{t}) = E\left[e^{i\langle \mathbf{t}, \mathbf{X}\rangle}\right] = e^{i\langle \mathbf{t}, \boldsymbol{\mu}\rangle} E\left[e^{i\langle \mathbf{t}, A\mathbf{Y}\rangle}\right]$$

and since $\langle \mathbf{t}, A\mathbf{Y}\rangle = \mathbf{t}^T A\mathbf{Y} = (A^T\mathbf{t})^T\mathbf{Y} = \langle A^T\mathbf{t}, \mathbf{Y}\rangle$, we have

$$\begin{aligned} E\left[e^{i\langle \mathbf{t}, A\mathbf{Y}\rangle}\right] &= E\left[e^{i\langle A^T\mathbf{t}, \mathbf{Y}\rangle}\right] = \psi(\|A^T\mathbf{t}\|^2) = \psi((A^T\mathbf{t})^T A^T\mathbf{t}) \\ &= \psi(\mathbf{t}^T AA^T\mathbf{t}) = \psi(\mathbf{t}^T\Sigma\mathbf{t}) \end{aligned}$$

as desired. ■

We want to be able to relate covariances to the dispersion. The following proposition shows how these are related for elliptical distributions.

Proposition 7.2. *If $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$, the covariance between the components is given by*

$$\operatorname{Cov}(X_j, X_l) = -2\psi'(0)\Sigma_{jl}.$$

Proof. Since covariance does not depend on the mean, we can without loss of generality assume that $E[X_j] = 0$ for all $j = 1, ..., d$. We have

$$\begin{aligned}
\frac{\partial}{\partial t_j}\frac{\partial}{\partial t_l}\Phi_{\mathbf{X}}(\mathbf{t})\bigg|_{\mathbf{t}=0} &= \frac{\partial}{\partial t_j}\frac{\partial}{\partial t_l}E\left[e^{i\langle\mathbf{t},\mathbf{X}\rangle}\right]\bigg|_{\mathbf{t}=0} = \frac{\partial}{\partial t_j}\frac{\partial}{\partial t_l}E\left[e^{i(t_1X_1+\cdots t_dX_d)}\right]\bigg|_{\mathbf{t}=0} \\
&= E\left[(iX_j)(iX_l)e^{i(t_1X_1+\cdots t_dX_d)}\right]\bigg|_{\mathbf{t}=0} = -E\left[X_jX_le^{i(t_1X_1+\cdots t_dX_d)}\right]\bigg|_{\mathbf{t}=0} \\
&= -E\left[X_jX_l\right] = -\operatorname{Cov}(X_j, X_l)
\end{aligned}$$

and thus by the previous lemma,

$$\operatorname{Cov}(X_j, X_l) = -\frac{\partial}{\partial t_j}\frac{\partial}{\partial t_l}\Phi_{\mathbf{X}}(\mathbf{t})\bigg|_{\mathbf{t}=0} = -\frac{\partial}{\partial t_j}\frac{\partial}{\partial t_l}\psi(\mathbf{t}^T\Sigma\mathbf{t})\bigg|_{\mathbf{t}=0}.$$

Let us for simplicity assume $d = 2$. Then

$$\mathbf{t}^T\Sigma\mathbf{t} = \begin{pmatrix} t_1 & t_2 \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = t_1^2\Sigma_{11} + 2t_1t_2\Sigma_{12} + t_2^2\Sigma_{22} =: w(\mathbf{t})$$

and thus

$$\begin{aligned}
\frac{\partial}{\partial t_1}\frac{\partial}{\partial t_2}\psi(\mathbf{t}^T\Sigma\mathbf{t})\bigg|_{\mathbf{t}=0} &= \frac{\partial}{\partial t_1}\left(\psi'(w(\mathbf{t}))(2t_1\Sigma_{12} + 2t_2\Sigma_{22})\right)\bigg|_{\mathbf{t}=0} \\
&= \psi''(w(\mathbf{t}))(2t_1\Sigma_{11} + 2t_2\Sigma_{12})(2t_1\Sigma_{12} + 2t_2\Sigma_{22}) + \psi'(w(\mathbf{t}))2\Sigma_{12}\bigg|_{\mathbf{t}=0} \\
&= 2\psi'(0)\Sigma_{12}.
\end{aligned}$$

This calculation can be generalised so that $\frac{\partial}{\partial t_j}\frac{\partial}{\partial t_l}\psi(\mathbf{t}^T\Sigma\mathbf{t})\Big|_{\mathbf{t}=0} = 2\psi'(0)\Sigma_{jl}$. We conclude that

$$\operatorname{Cov}(X_j, X_l) = -2\psi'(0)\Sigma_{jl}.$$

■

Example 7.3. If $\mathbf{Y} \sim \mathcal{N}(0, I_d)$, then $\psi(r) = e^{-r/2}$ as seen earlier. We see that $\psi'(r) = -\frac{1}{2}e^{-r/2}$, so $\psi'(0) = -\frac{1}{2}$. If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, the above proposition tells us that $\operatorname{Cov}(X_j, X_l) = -2\psi'(0)\Sigma_{jl} = \Sigma_{jl}$ as expected. ○

We list some further properties of elliptical distributions.

Theorem 7.4. *Let $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Y} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$.*

(i) (Linear combinations). If B is a $k \times d$ matrix and $\mathbf{b} \in \mathbb{R}^k$, then

$$B\mathbf{X} + \mathbf{b} \sim E_k(B\boldsymbol{\mu} + \mathbf{b}, B\Sigma B^T, \psi).$$

(ii) (Marginal distributions). The marginals $X_1, ..., X_d$ also have elliptical distributions with the same characteristic generator. Explicitly, $X_i \sim E_1(\mu_i, \Sigma_{ii}, \psi)$.

(iii) (Convolutions). If $\tilde{\mathbf{X}} \sim E_d(\tilde{\boldsymbol{\mu}}, \Sigma, \tilde{\psi})$ is another elliptical distribution independent of $\mathbf{X}$ with the same dimension and dispersion matrix, then $\mathbf{X} + \tilde{\mathbf{X}} \sim E_d(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \Sigma, \bar{\psi})$ with $\bar{\psi}(u) = \psi(u)\tilde{\psi}(u)$.

(iv) (Quadratic forms). We have

$$R^2 = \|\mathbf{Y}\|^2 = (\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}).$$

Here, R is called the Mahalanobis distance*.*

Proof. See the exercises. ■

While elliptical distributions are quite flexible as seen in the examples earlier, the above theorem also illustrates a drawback. The flexibility of elliptical distributions are limited by the fact that if $\mathbf{X}$ is elliptical, so is any coordinate. In real data, it is often the case that the marginals have very different types of distributions. This motivates the topic for next week, namely copulas. We now consider some examples.

Example 7.5 (Normal Variance Mixture Models). Let $\mathbf{Z} \sim \mathcal{N}(0, I_k)$, $W \geq 0$ a random variable and A a fixed $d \times k$ matrix. A normal variance mixture model is a model of the form

$$\mathbf{X} = \boldsymbol{\mu} + \sqrt{W} A \mathbf{Z}.$$

One can show, conditional on $W = w$ that

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, w\Sigma), \quad \Sigma = AA^T.$$

Thus, $\mathbf{X}$ is obtained by drawing from a collection of normal random variables with random covariance $W\Sigma$. ○

Example 7.6 (t-distribution). Here we consider a special case of a normal variance mixture model, namely if we let

$$W \sim \mathrm{Ig}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$$

where $\mathrm{Ig}(\alpha, \beta)$ denotes the *inverse gamma distribution* with density

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{x}\right)^{\alpha+1} e^{-\beta/x}.$$

In particular, we have $\nu/W \sim \chi^2_\nu$. For $AA^T = \Sigma$, we write

$$\mathbf{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$$

and we call this distribution the multivariate t distribution.

○

We end this week with an application.

An application: Portfolio investment theory

Say we want to minimise the risk in a portfolio with d assets with returns $\mathbf{X} = (X_1, ..., X_d)$, $\mu_i = E[X_i]$. The following ideas go back to Marcowicz. Let R_p denote the total returns i.e.

$$R_p = \sum_{i=1}^{d} w_i X_i = \mathbf{w}^T \mathbf{X}$$

where w_i are the weights of the portfolio i.e. $\sum_{i=1}^{d} w_i = 1$. If we fix the expected total returns $E[R_p] = \sum_{i=1}^{d} w_i \mu_i = \mathbf{w}^T \boldsymbol{\mu}$, we want to minimize the risk in the sense of minimising the variance (note that this approach is different from the strategy in this course, where we focus on risk measures). If $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Y}$ where $\mathbf{Y} \sim \mathcal{N}(0, I_d)$, then

$$\mathrm{Var}(R_p) = \mathrm{Var}(\mathbf{w}^T(\boldsymbol{\mu} + A\mathbf{Y})) = \mathrm{Var}(\mathbf{w}^T A\mathbf{Y})$$

and since $\mathbf{Y}$ is spherical, the variance is minimised whenever $\|A^T\mathbf{w}\|$ is minimized with respects to the weights w_i.

To transfer these ideas to the setting in this course, let ρ be a risk measure which satisfies monotonicity and translation invariance. Assume more generally that $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Y}$ is elliptical. The goal now is to minimise $\rho(L)$ where $L = -\mathbf{w}^T\boldsymbol{\mu} - \mathbf{w}^T A\mathbf{Y}$ is the loss. We have

$$\rho(L) = -\mathbf{w}^T\boldsymbol{\mu} + \rho(-\mathbf{w}^T A\mathbf{Y}).$$

As $\mathbf{Y}$ is spherical, we have $\mathbf{Y} \overset{\mathrm{d}}{=} -\mathbf{Y}$, so we can remove the minus in the risk measure and obtain

$$\rho(L) = -\mathbf{w}^T\boldsymbol{\mu} + \rho(\mathbf{w}^T A\mathbf{Y}) = -E[R_p] + \rho(\mathbf{w}^T A\mathbf{Y}).$$

As $E[R_p]$ is fixed, $\rho(L)$ is minimised whenever $\rho(\mathbf{w}^T A\mathbf{Y})$ is minimised. As $\mathbf{Y}$ is spherical, $\rho(\mathbf{w}^T A\mathbf{Y})$ is minimised when $\|A^T\mathbf{w}\|$ is minimised. Thus the answer remains the same as in the classical case above.

Notes and comments

Chapter 6 of [17] discusses spherical and elliptical distributions. The chapter contains a few more details and examples. Section 6.3.4 is dedicated to estimation of dispersion and correlation in elliptical distributions.

Exercises

Exercise 3.1:
Prove Theorem 7.4. Hint: Characteristic functions!

Exercise 3.2:
Let $X \sim \Gamma(\alpha, \beta)$ i.e. let X have a gamma distribution with parameters $\alpha, \beta > 0$. Verify that $1/X \sim \mathrm{Ig}(\alpha, \beta)$. This provides an explanation for the name "inverse gamma".

Exercise 3.3:

1)Without using an R package, simulate 500 values of the multivariate t distribution with

$$\nu = 3, \quad \boldsymbol{\mu} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}.$$

Make a plot of the result. Hint: Use the result of the previous exercise.

2)Now simulate 500 values of the multivariate normal distribution with the same $\boldsymbol{\mu}$ and Σ and plot the result. Compare the plot to the one for the t distribution.

Exercise 3.4:
Without using an R package, write an R function to simulate from the multivariate normal distribution with mean $\boldsymbol{\mu} = (1, 0, 2)$ and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0 & -2 \\ 0 & 3 & 5 \\ -2 & 5 & 1 \end{pmatrix}.$$

Exercise 3.5:
Let $\mathbf{S}$ be uniformly distributed on the unit circle $\mathbb{S}^1$. Simulate 500 values of $\mathbf{S}$ and plot the result. Hint: Corollary 6.6.

Week 4 - Copulas I

8 Copulas: Basic properties

We need to be able to work with distributions that are less rigid than the elliptical distributions. Let us set the stage. We have data of the form $\mathbf{X} = (X_1, ..., X_d)$ (for example log returns) and the goal is to find a suitable joint distribution function F for $\mathbf{X}$. We don't want to exclude the possibility that the X_i have different types of distributions. Before proceeding, we encourage the reader to recall the basic properties of generalised inverses as described in the appendix.

Recall that if X_i has distribution function F_i, then $F_i^{\leftarrow}(U_i) \overset{\mathrm{d}}{=} X_i$ with U_i a Unif$(0, 1)$ variable. This is the "inverse transform method" used in simulation. Recall also that $F_i(U_i) \overset{\mathrm{d}}{=} X_i$ whenever F_i is continuous.

To study the problem of determining the joint distribution F, we assume that the marginal distribution functions F_i are known. The transformation $U_i := F_i(X_i)$ in the continuous case is illustrated below.

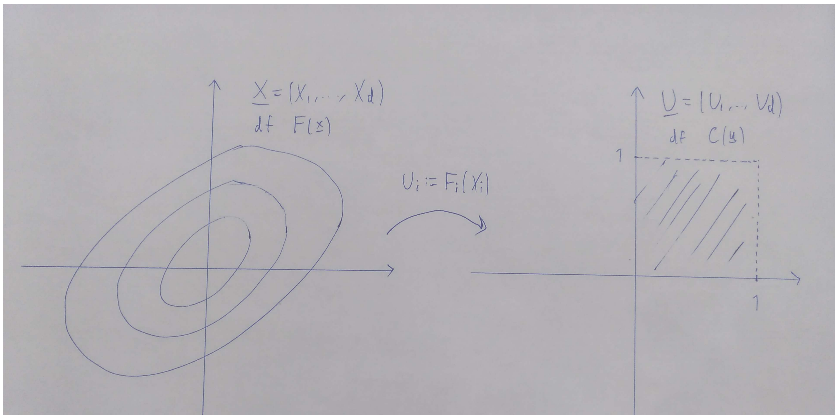
$\underline{X} = (X_1, \ldots, X_d)$
df $F(\underline{x})$
$U_i := F_i(X_i)$
$\underline{U} = (U_1, \ldots, U_d)$
df $C(\underline{u})$
1
1

Figure 12: An illustration (in two dimensions) of the idea of a "copula space". The "original space" on the left is where the variables live, and we wish to transform them into a collection of uniform variables with support on $[0, 1]^d$.

Definition 8.1. A *copula* C is a distribution function on $[0,1]^d$ such that all marginals are $\mathrm{Unif}(0,1)$ distributed.

To be able to go back and forth between the "original space" and the "copula space", Sklar's Theorem is an essential tool.

Theorem 8.2 (**Sklar's Theorem**). *Let F be the joint distribution function of the random vector $\mathbf{X} = (X_1, ..., X_d)$ with marginal distribution functions $F_1, ..., F_d$. There exists a copula C such that*

$$F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)). \tag{3.1}$$

If $F_1, ..., F_d$ are continuous, C is unique. Conversely, given a copula C and marginal distribution functions $F_1, ..., F_d$, then F as defined in (3.1) is a joint distribution function with marginals $F_1, ..., F_d$.

Proof. For the sake of simplicity, assume that the F_i are continuous. Then $U_i := F_i(X_i) \sim \mathrm{Unif}(0,1)$. Suppose $\mathbf{X} \sim F$ with marginals $X_i \sim F_i$. Let $\mathbf{U} = (F_1(X_1), ..., F_d(X_d))$ and let C be the distribution function of $\mathbf{U}$. By construction and the continuity assumption on the F_i, C is a copula. We compute

$$\begin{aligned} C(F_1(x_1), ..., F_d(x_d)) &= P(U_1 \leq F_1(x_1), ..., U_d \leq F_d(x_d)) \\ &= P(F_1(X_1) \leq F_1(x_1), ..., F_d(X_d) \leq F_d(x_d)) \\ &= P(X_1 \leq x_1, ..., X_d \leq x_d) = F(x_1, ..., x_d) \end{aligned}$$

which shows that the copula C has the desired properties. As for uniqueness, by continuity of the F_i, we have $F_i(F_i^{\leftarrow}(u_i)) = u_i$ for all $u_i \in [0,1]$. Letting $x_i = F_i^{\leftarrow}(u_i)$ in the expression above, we get

$$C(u_1, ..., u_d) = F(F_1^{\leftarrow}(u_1), ..., F_d^{\leftarrow}(u_d))$$

and any copula $\tilde{C}$ satisfying $\tilde{C}(F_1(x_1), ..., F_d(x_d)) = F(x_1, ..., x_d)$ must satisfy the same relation. Uniqueness now follows. To prove the converse statement, let C be a copula and $F_1, ..., F_d$ univariate distribution functions. Let $\mathbf{U} = (U_1, ..., U_d)$ have distribution function C and define $X_i := F_i^{\leftarrow}(U_i)$, $\mathbf{X} := (X_1, ..., X_d)$. We know that $X_i \sim F_i$, so the marginal distributions are correct. Also,

$$\begin{aligned} C(F_1(x_1), ..., F_d(x_d)) &= P(U_1 \leq F_1(x_1), ..., U_d \leq F_d(x_d)) \\ &= P(F_1^{\leftarrow}(U_1) \leq x_1, ..., F_d^{\leftarrow}(U_d) \leq x_d) \\ &= P(X_1 \leq x_1, ..., X_d \leq x_d) \end{aligned}$$

which shows that $F(x_1, ..., x_d) := C(F_1(x_1), ..., F_d(x_d))$ is the distribution function for $\mathbf{X}$. ■

Sklar's Theorem provides a recipe for constructing copulas using a known joint distribution function. We call such copulas *implicit copulas.* Different examples of copulas will be given in the next section. We first consider some more theoretical properties. We start with the following useful characterisation of copulas.

Proposition 8.3. *A function $C : [0,1]^d \to [0,1]$ is a copula if and only if*

(i) $C(u_1, ..., u_d) = 0$ if $u_i = 0$ for any $i = 1, ..., d$.

(ii) $C(1, ..., 1, u_i, 1, ..., 1) = u_i$ for any $i = 1, ..., d$ and $u_i \in [0,1]$.

(iii) For all $(a_1, ..., a_d), (b_1, ..., b_d) \in [0,1]^d$ with $a_i \leq b_i$, we have

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots i_d} C(u_{1i_1}, ..., u_{di_d}) \geq 0$$

where $u_{j1} = a_j$ and $u_{j2} = b_j$ for all $j = 1, ..., d$.

The first two properties are self-explanatory. The third property is called the *rectangle inequality* and can be interpreted as follows: For uniform variables $(U_1, ..., U_d)$, then $P(a_1 \leq U_1 \leq b_1, ..., a_d \leq U_d \leq b_d) \geq 0$.

Proposition 8.4 (Fréchet bounds). *For every copula C, we have $W(\mathbf{u}) \leq C(\mathbf{u}) \leq M(\mathbf{u})$ where*

$$W(\mathbf{u}) = \max\left\{\sum_{i=1}^{d} u_i + 1 - d, 0\right\} \quad \text{and} \quad M(\mathbf{u}) = \min_{i=1,...,d} u_i.$$

Proof. Let $\mathbf{U}$ have distribution function C. For every $u_i \in [0,1]$, we have

$$C(\mathbf{u}) = P(U_1 \leq u_1, ..., U_d \leq u_d) \leq P(U_i \leq u_i) = u_i.$$

The bound $C(\mathbf{u}) \leq M(\mathbf{u})$ now follows by minimising over all i. Conversely, for $\mathbf{u} \in [0,1]^d$,

$$\begin{aligned} 1 - C(\mathbf{u}) = 1 - P(U_1 \leq u_1, ..., U_d \leq u_d) &= P\left(\bigcup_{i=1}^{d} \{U_i > u_i\}\right) \\ &\leq \sum_{i=1}^{d} P(U_i > u_i) = \sum_{i=1}^{d} (1 - u_i) = d - \sum_{i=1}^{d} u_i. \end{aligned}$$

Thus $-C(\mathbf{u}) \leq d - 1 - \sum_{i=1}^{d} u_i$ implying $C(\mathbf{u}) \geq \sum_{i=1}^{d} u_i + 1 - d$. Since $C(\mathbf{u}) \geq 0$ always holds, the lower bound follows. ■

What kinds of random vectors produce these upper and lower bounds? It turns out that the function W is not a copula for $d > 2$, see Example 7.24 in [17]. We can however produce M as a copula for any d and W for $d = 2$. To do so, we introduce the concepts of comonotonicity and countermonotonicity.

Definition 8.5. We say that $X_1, ..., X_d$ are *comonotone* if $(X_1, ..., X_d) = (\alpha_1(Z), ..., \alpha_d(Z))$ for some univariate variable Z and nondecreasing functions $\alpha_1, ..., \alpha_d$. We say that (X_1, X_2) are *countermonotonic* if $(X_1, X_2) = (\alpha(Z), \beta(Z))$ for some univariate variable Z, some nondecreasing function α and some nonincreasing function β.

The following proposition shows that the lower and upper Fréchet bounds arise from countermonotonic and comonotonic random vectors, respectively.

Proposition 8.6. *A comonotone bundle $(X_1, ..., X_d)$ has M as a copula, and a countermonotonic bundle (X_1, X_2) in $d = 2$ dimensions has W as a copula.*

Proof. Consider a comonotone bundle $(X_1, ..., X_d)$ and assume for simplicity that the functions α_i are strictly increasing and continuous. If F is the joint distribution function, we have

$$\begin{aligned} F(x_1, ..., x_d) &= P(X_1 \le x_1, ..., X_d \le x_d) = P(\alpha_1(Z) \le x_1, ..., \alpha_d(Z) \le x_d) \\ &= P(Z \le \alpha_1^{\leftarrow}(x_1), ..., Z \le \alpha_d^{\leftarrow}(x_d)) = P\left(Z \le \min_{i=1,...,d} \alpha_i^{\leftarrow}(x_i)\right) \\ &= \min_{i=1,...,d} P(Z_i \le \alpha_i^{\leftarrow}(x_i)) = \min_{i=1,...,d} P(\alpha_i(Z) \le x_i) \\ &= \min_{i=1,...,d} P(X_i \le x_i) = \min_{i=1,...,d} F_i(x_i) = M(F_1(x_1), ..., F_d(x_d)) \end{aligned}$$

so M is a copula for $(X_1, ..., X_d)$. Now consider a countermonotonic bundle $(X_1, X_2) = (\alpha(Z), \beta(Z))$. Assume for simplicity that α is strictly increasing and continuous, β is strictly decreasing and continuous and that Z is continuous. If F is the distribution function of (X_1, X_2), we have

$$\begin{aligned} F(x_1, ..., x_d) &= P(\alpha(Z) \le x_1, \beta(Z) \le x_2) = P(Z \le \alpha^{\leftarrow}(x_1), Z \ge \beta^{\leftarrow}(x_2)) \\ &= P(Z \le \alpha^{\leftarrow}(x_1)) - P(Z \le \alpha^{\leftarrow}(x_1), Z < \beta^{\leftarrow}(x_2)) \\ &= F_1(x_1) - \min\{F_1(x_1), 1 - F_2(x_2)\} \\ &= \max\{F_1(x_1) - F_1(x_1), F_1(x_1) - 1 + F_2(x)\} \\ &= \max\{0, F_1(x_1) + F_2(x_2) + 1 - 2\} = W(F_1(x_1), F_2(x_2)) \end{aligned}$$

so W is a copula for (X_1, X_2). ■

Remark 8.7. The implications in the above proposition are biimplications, see the notes and comments at the end of the chapter.

What happens when we take monotone transformations of a copula? While the distribution itself may change, the copula does not change under strictly increasing transformations as the following result shows.

Proposition 8.8. *Consider a random vector $(X_1, ..., X_d)$ with continuous marginal distributions F_i and copula C. Let $T_1, ..., T_d$ be strictly increasing continuous functions. Then $(T_1(X_1), ..., T_d(X_d))$ also has copula C.*

Proof. Let $\tilde{X}_i := T_i(X_i), \tilde{X}_i \sim \tilde{F}_i$ and $(\tilde{X}_i, ..., \tilde{X}_d) \sim \tilde{F}$. Let $\tilde{C}$ be the copula of $(\tilde{X}_i, ..., \tilde{X}_d)$. By Sklar's Theorem,

$$\begin{aligned} \tilde{C}(\tilde{F}_1(x_1), ..., \tilde{F}_d(x_d)) &= \tilde{F}(x_1, ..., x_d) = P(T_1(X_1) \le x_1, ..., T_d(X_d) \le x_d) \\ &= P(X_1 \le T_1^{\leftarrow}(x_1), ..., X_d \le T_d^{\leftarrow}(x_d)) \\ &= F(T_1^{\leftarrow}(x_1), ..., T_d^{\leftarrow}(X_d)) = C(F_1(T_1^{\leftarrow}(x_1)), ..., F_d(T_d^{\leftarrow}(x_d))). \end{aligned}$$

We now claim that $\tilde{F}_i = F_i \circ T_i^{\leftarrow}$. By definition,

$$\tilde{F}_i(x) = P(\tilde{X}_i \le x) = P(T_i(X_i) \le x) = P(X_i \le T_i^{\leftarrow}(x)) = F_i(T_i^{\leftarrow}(x))$$

as claimed. F_i is continuous by assumption and $T_i^{\leftarrow}$ is continuous since T_i is strictly increasing. Hence $\tilde{F}_i$ is continuous. We have now proved that

$$\tilde{C}(\tilde{F}_1(x_1), ..., \tilde{F}_d(x_d)) = C(\tilde{F}_1(x_1), ..., \tilde{F}_d(x_d))$$

and so $\tilde{C} = C$ by the uniqueness part of Sklar's Theorem. ■

9 Examples of copulas

We will study three types of copulas, namely fundamental copulas, implicit copulas and explicit copulas.

Fundamental copulas

Fundamental copulas arise from theoretical considerations. We have already seen two examples from the Fréchet bounds.

Definition 9.1. For any $d > 1$, we call

$$M(\mathbf{u}) = \min_{i=1,\dots,d} u_i$$

the *comonotonicity copula.* For $d = 2$, we call

$$W(u_1, u_2) = \max\{u_1 + u_2 - 1, 0\}$$

the *countermonotonicity copula.*

The Fréchet bound copulas are not the only "theoretical" copulas.

Example 9.2 (The independence copula). For any $d > 1$, the *independence copula* is given by

$$C(u_1, ..., u_d) = \prod_{i=1}^{d} u_i.$$

Unsurprisingly, C arises from independent variables. Suppose $X_1, ..., X_d$ are independent with continuous distribution functions F_i. Then

$$\begin{aligned} C(F_1(x_1), ..., F_d(x_d)) &= \prod_{i=1}^{d} F_i(x_i) = \prod_{i=1}^{d} P(X_i \le x_i) \\ &= P(X_1 \le x_1, ..., X_d \le x_d) = F(x_1, ..., x_d) \end{aligned}$$

so by the uniqueness from Sklar's Theorem, C is the copula for $X_1, ..., X_d$.

○

Implicit copulas

Implicit copulas arise from known joint distribution functions. Let F be a given joint distribution function. If the marginal distribution functions F_i are continuous, we can use Sklar's Theorem to construct a copula C via

$$C(u_1, ..., u_d) = F(F_1^{\leftarrow}(u_1), ..., F_d^{\leftarrow}(u_d)).$$

We give a concrete example.

Example 9.3 (**The Gaussian copula**)**.** Let $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$ be a d-dimensional normal vector with distribution function Φ_Σ where

$$\Sigma = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1d} \\ \rho_{12} & 1 & \cdots & \\ \vdots & & & \vdots \\ \rho_{1d} & \cdots & \cdots & 1 \end{pmatrix}.$$

One can think of Σ as a correlation matrix. Note that the marginals X_i are standard normal, so that they have common distribution function Φ. We can then construct the *Gaussian copula*

$$C_\Sigma^{\text{Ga}}(u_1, ..., u_d) = \Phi_\Sigma(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d)).$$

What if we have a general mean and covariance matrix? Let $\mathbf{Y} = \boldsymbol{\mu} + B\mathbf{X}$ be an affine transformation of $\mathbf{X}$ where

$$B = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_d \end{pmatrix}$$

is a diagonal matrix with $\sigma_i > 0$ for all i. Then $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, B\Sigma B^T)$, and any desired covariance matrix can be written in the form $B\Sigma B^T$. Note that $Y_i = \mu_i + \sigma_i X_i$ is a strictly increasing and continuous transformation of X_i, so Proposition 8.8 implies that $\mathbf{Y}$ and $\mathbf{X}$ have the same copula, namely C_Σ^{Ga}. Simulating from this copula is easy. The trick is to follow the construction in Sklar's Theorem. To simulate a sample from C_Σ^{Ga}, follow the steps:

(i) Simulate $\mathbf{X}$ from the multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$.

(ii) Set $U_i = \Phi(X_i)$ for $i = 1, ..., d$.

(iii) Now $\mathbf{U} = (U_1, ..., U_d)$ has the distribution function C_Σ^{Ga}.

○

Explicit copulas

Explicit copulas are given by a concrete formula. Some well-known examples are the following.

Example 9.4 (**Gumbel copula**)**.** The Gumbel copula is given by

$$C_\theta^{\text{Gu}}(u_1, u_2) = \exp\left(-\left((-\log u_1)^\theta + (-\log u_2)^\theta\right)^{1/\theta}\right)$$

where $1 \leq \theta < \infty$ is a parameter. ○

Example 9.5 (**Clayton copula**)**.** The Clayton copula is given by

$$C_\theta^{\text{Cl}}(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$$

where $0 < \theta < \infty$ is a parameter. ○

Both of the above examples are so-called *Archimedean copulas*.

Definition 9.6. Let $\varphi : [0,1] \to [0,\infty]$ be continuous, strictly decreasing and convex with $\varphi(0) = \infty$ and $\varphi(1) = 0$. Then

$$C(u_1, u_2) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2))$$

is called the *Archimedean copula* with generator φ.

We see that by choosing $\varphi(t) = (-\log t)^\theta$ for $\theta \geq 1$, we obtain the Gumbel copula. Choosing $\varphi(t) = \frac{1}{\theta}(t^{-\theta} - 1)$ gives the Clayton copula.

Example 9.7 (**Generalised Clayton copula**). Choosing $\varphi(t) = \theta^{-\delta}(t^{-\theta} - 1)^\delta$ for $\theta > 0$ and $\delta \geq 1$ gives the *Generalised Clayton copula*. The special case $\delta = 1$ corresponds to the Clayton copula from above. ○

Example 9.8 (**Frank copula**). The Frank copula is the Archimedean copula with generator

$$\varphi(t) = -\log\left(\frac{e^{-\theta t} - 1}{e^\theta - 1}\right)$$

where $\theta \in \mathbb{R} \setminus \{0\}$ is a parameter. ○

It should of course be verified that an Archimedean copula is a copula. We state the relevant result without proof.

Theorem 9.9. *Let $\varphi : [0,1] \to [0,\infty]$ be a continuous, strictly decreasing function such that $\varphi(1) = 0$, $\varphi(0) = \infty$. The function $C : [0,1]^2 \to [0,1]$ given by*

$$C(u_1, u_2) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2))$$

is a copula if and only if φ is convex.

Proof. See [19], Theorem 4.1.4. ■

While the method of constructing Archimedean copulas is certainly practical, there are some limitations worth mentioning. An Archimedean copula is seen to be symmetric in the two arguments which is clearly a constraint in modelling. Furthermore, the structure of an Archimedean copula is maybe too simple since it is two dimensional but has a one-dimensional generator.

Notes and comments

Chapter 7 in [17] covers copulas in more generality than here. In particular, see section 7.2.1 for a complete characterisation of the copulas of countermonotonic and comonotonic random vectors. If the reader reads the section on Archimedean copulas (section 7.4), they should be aware that φ and φ^{-1} are switched. For more details on copulas as a subject, we refer to the book by Nelsen, [19]. Chapter 4 about Archimedean copulas contains all the theoretical details on Archimedean copulas as well as a table of copulas with the corresponding generators, see page 116-119. Some of the exercises below are also borrowed from [19].

Exercises

Exercise 4.1:
Write out the rectangle inequality (see Proposition 8.3) in the case where $d = 2$. Use it to verify that the *Morgenstern copula* given by

$$C(u_1, u_2) = u_1 u_2 (1 + \delta(1 - u_1)(1 - u_2)),$$

for $\delta \in [-1, 1]$ a parameter, is indeed a copula.

Exercise 4.2:
Let C and $\tilde{C}$ be copulas and $\theta \in [0, 1]$. Verify that $\theta C + (1 - \theta)\tilde{C}$ is also a copula.

Exercise 4.3:
Let (X, Y) have joint distribution function

$$H(x, y) = (1 + e^{-x} + e^{-y})^{-1}$$

for all $(x, y) \in \mathbb{R}^2$.

1)Verify that X and Y both have the logistic distribution i.e.

$$F_X(x) = (1 + e^{-x})^{-1}, \quad F_Y(y) = (1 + e^{-y})^{-1}.$$

2)Show that (X, Y) has the copula

$$C(u_1, u_2) = \frac{u_1 u_2}{u_1 + u_2 - u_1 u_2}.$$

Exercise 4.4:
Let X and Y be random variables with continuous distribution functions F_X and F_Y. Let α, β be functions, C the copula of (X, Y) and $\tilde{C}$ the copula of $(\alpha(X), \beta(Y))$.

1)If α is strictly increasing and β strictly decreasing, prove that

$$\tilde{C}(u_1, u_2) = u_1 - C(u_1, 1 - u_2).$$

2)If α is strictly decreasing and β is strictly increasing, prove that

$$\tilde{C}(u_1, u_2) = u_2 - C(1 - u_1, u_2).$$

3)If α and β are strictly increasing, prove that

$$\tilde{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2).$$

Exercise 4.5:
Recall that the Frank copula has generator

$$\varphi(t) = -\log\left(\frac{e^{-\theta t} - 1}{e^{\theta} - 1}\right)$$

with $\theta \in \mathbb{R} \setminus \{0\}$ a parameter.

1)Verify that the Frank copula is indeed a copula using Theorem 9.9.

2)Write the Frank copula in explicit form.

Exercise 4.6:
Consider the function

$$\varphi(t) = \log(1 - \theta \log t)$$

for $\theta \in (0, 1]$ a parameter.

1)Verify that φ is a valid generator for an Archimedean copula.

2)Write the corresponding copula in explicit form.

Exercise 4.7:
The d-dimensional t copula is given by

$$C^t_{\nu,\Sigma}(u_1, ..., u_d) = t_{\nu,\Sigma}(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_d))$$

with $t_{\nu,\Sigma}$ the distribution function of $t(\nu, 0, \Sigma)$ and t_ν the univariate t distribution function with ν degrees of freedom.

1)Describe a procedure to simulate from the t copula.

2)Simulate 1000 samples from the t copula with $\nu = 3$ and

$$\Sigma = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}.$$

3)Simulate 1000 samples from the Gaussian copula with Σ above. Plot the two simulated samples and compare. Try choosing different marginal distributions for these two copulas and see what happens.

Week 5 - Copulas II

10 Archimedean copulas in higher dimensions

Last week we introduced Archimedean copulas as a recipe for constructing copulas in two dimensions. How can we generalise this construction to higher dimensions? The logical next step would be to propose

$$C(u_1, ..., u_d) = \varphi^{-1}(\varphi(u_1) + \cdots \varphi(u_d)) \tag{4.2}$$

where $\varphi : [0, 1] \to [0, \infty]$ has the same properties as in the two-dimensional case i.e. φ is strictly decreasing, convex and satisfies $\varphi(0) = \infty$ and $\varphi(1) = 0$. Is C as constructed above a copula? The answer is no in general. One issue is that C is not even a distribution function in general for dimensions higher than two. In order to answer the question of when C is a copula, we need the following definition.

Definition 10.1. A decreasing function f is *completely monotonic* on $[a, b]$ if

$$(-1)^k \frac{d^k}{dt^k} f(t) \geq 0 \quad \text{for} \quad k = 1, 2, ... \quad \text{and} \quad t \in (a, b).$$

It turns out that the property of being completely monotonic determines whether C defined by equation (4.2) is a copula.

Theorem 10.2. *Let $\varphi : [0, 1] \to [0, \infty]$ be a continuous strictly decreasing function such that $\varphi(0) = \infty$ and $\varphi(1) = 0$. C defined by (4.2) is a copula for all $d \geq 2$ if and only if φ^{-1} is completely monotonic on $[0, \infty)$.*

Proof. See Theorem 4.6.2 in [19] and the references in the paragraph above the theorem. ■

In principle one should be able to check whether φ^{-1} is completely monotonic, but it is a tedious procedure. As an alternative to verifying complete monotonicity, one can apply Laplace transforms of distribution functions. We recall the definition.

Definition 10.3. Let G be a distribution function on $[0, \infty)$ with $G(0) = 0$. The Laplace transform of G is

$$\psi(t) = \int_0^\infty e^{-tx} dG(x).$$

Remark 10.4. Let G be a distribution function on $[0, \infty)$ with $G(0) = 0$ and Y a random variable distributed according to G. We then have the following relationship between the Laplace transform ψ of G and the moment-generating function κ_Y given by

$$\psi(t) = \kappa_Y(-t).$$

Lemma 10.5. *A function ψ on $[0, \infty)$ is the Laplace transform of a distribution function G if and only if ψ is completely monotonic and $\psi(0) = 1$.*

The above results provide a strategy to verify that φ^{-1} is completely monotonic. It suffices to show that φ^{-1} is a Laplace transform of some distribution function G.

Example 10.6. Consider the Clayton copula with generator $\varphi(t) = (t^{-\theta} - 1)/\theta$. We can solve for φ^{-1} and get

$$\varphi^{-1}(t) = (\theta t + 1)^{-1/\theta} = \frac{(1/\theta)^{1/\theta}}{(t + 1/\theta)^{1/\theta}}.$$

Now recall that the gamma distribution with parameters $\alpha, \beta > 0$ has the moment-generating function

$$\kappa(t) = \left(\frac{\beta}{\beta - t}\right)^{\alpha}$$

so if we let $\alpha = \beta = 1/\theta$, we have that the Laplace transform of this distribution is given by $\varphi^{-1}(t)$. We conclude that the Clayton copula extends to any dimension. Explicitly,

$$C_\theta^{\mathrm{Cl}}(u_1, ..., u_d) = (u_1^{-\theta} + u_2^{-\theta} + \cdots + u_d^{-\theta} - d + 1)^{-1/\theta}$$

is a copula on $[0, 1]^d$ for any $d \geq 2$.

○

Archimedean copulas constructed using Laplace transforms of distributions deserve their own name.

Definition 10.7. An *LT-Archimedean copula* is an Archimedean copula if the generator φ satisfies $\varphi^{-1} = \psi$, where ψ is the Laplace transform of some distribution function G on $[0, \infty)$.

How do we simulate random vectors with a given copula? If we have an LT-Archimedean copula, the following proposition provides a recipe.

Proposition 10.8. *Let G be a distribution function on $[0, \infty)$ and $V \sim G$. Let $\psi = \varphi^{-1}$ denote the Laplace transform of V. Suppose we have variables $W_1, ..., W_d$ which are conditionally independent given V with conditional distribution function*

$$F_{W_i|V=v}(u) = e^{-v\varphi(u)}.$$

Then the distribution function of $\mathbf{W}$ satisfies $F_{\mathbf{W}}(\mathbf{u}) = C(\mathbf{u})$.

Proof. The proof is a straightforward computation using a conditioning argument and conditional independence:

$$\begin{aligned} F_{\mathbf{W}}(\mathbf{u}) &= P(W_1 \leq u_1, ..., W_d \leq u_d) = \int_0^\infty P(W_1 \leq u_1, ..., W_d \leq u_d \mid V = v) dG(v) \\ &= \int_0^\infty \prod_{i=1}^d P(W_i \leq u_i \mid V = v) dG(v) = \int_0^\infty \prod_{i=1}^d e^{-v\varphi(u_i)} dG(v) \\ &= \int_0^\infty e^{-v(\varphi(u_1) + \cdots + \varphi(u_d))} dG(v) = \psi(\varphi(u_1) + \cdots + \varphi(u_d)) \\ &= \varphi^{-1}(\varphi(u_1) + \cdots + \varphi(u_d)) = C(\mathbf{u}). \end{aligned}$$

■

The proposition tells us that if we want to simulate from the copula $C(\mathbf{u}) = \varphi^{-1}(\varphi(u_1) + \cdots + \varphi(u_d))$, we should apply the following steps:

1. Identify the distribution G having the Laplace transform $\psi = \varphi^{-1}$.
2. Simulate $V \sim G$.
3. Generate iid $U_1, ..., U_d \sim \text{Unif}(0,1)$ and apply the inverse transform method i.e. $W_i = F^{\leftarrow}_{W_i|V=v}(U_i)$ with v equal to the simulated value of V from step 2.

We can actually be more specific in step 3. $F_{W_i|V=v}$ has a proper inverse which we solve for as follows:

$$e^{-v\varphi(F^{\leftarrow}_{W_i|V=v}(u))} = u \quad \Leftrightarrow \quad -\log u = v\varphi(F^{\leftarrow}_{W_i|V=v}(u)) \quad \Leftrightarrow \quad F^{\leftarrow}_{W_i|V=v}(u) = \varphi^{-1}\left(-\frac{\log u}{v}\right).$$

Hence W_i in step 3 should be set to

$$W_i = \psi\left(-\frac{\log U_i}{V}\right).$$

11 Fitting copulas to data

If a copula C has been chosen for the data, one can apply maximal likelihood methods in the case where C depends on a parameter θ. If c_θ denotes the density of C, then the maximal likelihood estimator $\hat{\theta}$ would maximise the log-likelihood

$$\log L(\theta; \mathbf{u}_1, ..., \mathbf{u}_n) = \sum_{i=1}^{n} \log c_\theta(\mathbf{u}_i)$$

for iid data $\mathbf{u}_1, ..., \mathbf{u}_n$. This of course requires that we choose a specific copula.

To determine a proper copula for a data set, we should think of properties like correlation, tail dependence and symmetry. The tail behaviour is especially relevant in a risk management context since we want to model large losses. Large losses also tend to ”move together”. When one loss is large, the other losses tend to be large as well. This can for example happen in a portfolio with stocks in similar companies. When we make the transformation from the original space of our data to the copula space, the size of the data is not preserved, only the ordering. Hence we need notions of correlation that only depends on ordering. This leads to different notions of rank correlation. We will study two types, namely Kendall's τ and Spearman's ρ. Afterwards, we will consider tail dependence via the coefficient of upper (respectively lower) tail dependence.

Kendall's τ

Definition 11.1. *Kendall's* τ for (X_1, X_2) is defined as

$$\rho_\tau(X_1, X_2) = P((X_1 - Y_1)(X_2 - Y_2) > 0) - P((X_1 - Y_1)(X_2 - Y_2) < 0)$$

where (Y_1, Y_2) is independent of (X_1, X_2) with $(Y_1, Y_2) \overset{\text{d}}{=} (X_1, X_2)$.

Kendall's τ gives an indication of whether X_1 and X_2 get large together or if one tends to get larger when the other gets smaller. If X_1 and X_2 tend to move together, the event $\{(X_1-Y_1)(X_2-Y_2)>0\}$ will have high probability, resulting in a value of $\rho_\tau(X_1,X_2)$ close to 1, and if X_1 and X_2 move in opposite directions, $\{(X_1-Y_1)(X_2-Y_2)<0\}$ will have high probability so that $\rho_\tau(X_1,X_2)$ is close to -1. To make this precise, we can make the following definition.

Definition 11.2. Consider two points $(x_1,x_2),(y_1,y_2)\in\mathbb{R}^2$. We say that (x_1,x_2) and (y_1,y_2) are *concordant* if $(x_1-x_2)(y_1-y_2)>0$ and *discordant* if $(x_1-x_2)(y_1-y_2)<0$.

Intuitively, $\rho_\tau(X_1,X_2)=1$ should correspond to comonotonicity while $\rho_\tau(X_1,X_2)=-1$ should correspond to countermonotonicity. This is indeed the case. We want to be able to estimate Kendall's τ from data $\{(X_{t,1},X_{t,2}):t=1,...,n\}$. To do so, we need to compare all pairs $(X_{t,1},X_{t,2})$ and $(X_{s,1},X_{s,2})$ with one another. In the case of concordance, the pair should yield the value 1 (indicating a positive sign) and -1 in the case of discordance. Since there are $\binom{n}{2}$ pairs in total, we obtain the estimator

$$\hat{\rho}_\tau=\binom{n}{2}^{-1}\sum_{1\le t\le s\le n}\operatorname{sign}((X_{t,1}-X_{s,1})(X_{t,2}-X_{s,2}))$$

where

$$\operatorname{sign}(x)=\begin{cases}1, & x>0\\ 0, & x=0\\ -1, & x<0\end{cases}.$$

We have the following results on Kendall's τ for copulas. For their proofs, we refer to chapter 5 of [19].

Theorem 11.3. *The following hold:*

(i) Let (X_1,X_2) be continuous variables with copula C. Then

$$\rho_\tau(X_1,X_2)=4\int_{[0,1]^2}C(u_1,u_2)dC(u_1,u_2)-1=4E[C(U_1,U_2)]-1$$

for uniform variables U_1,U_2 on $[0,1]$ with joint distribution function C.

(ii) Let (X_1,X_2) be random variables with Archimedean copula with generator φ. Then

$$\rho_\tau(X_1,X_2)=1+4\int_0^1\frac{\varphi(t)}{\varphi'(t)}dt.$$

We leave it as an exercise for the reader to verify that if (X_1,X_2) have the Clayton copula, $\rho_\tau(X_1,X_2)=\theta/(\theta+2)$ while for the Gumbel copula, $\rho_\tau(X_1,X_2)=1-1/\theta$.

Spearman's ρ

Another measure of rank correlation is Spearman's ρ defined as follows:

Definition 11.4. Consider a pair of variables (X_1,X_2) and introduce independent copies $(Y_1,Y_2)\overset{\mathrm{d}}{=}(Z_1,Z_2)\overset{\mathrm{d}}{=}(X_1,X_2)$. *Spearman's ρ* of (X_1,X_2) is given by

$$\rho_S(X_1,X_2)=3(P((X_1-Y_1)(X_2-Z_2)>0)-P((X_1-Y_1)(X_2-Z_2)<0)).$$

The intuition for Spearman's ρ is the same as for Kendall's τ.

Theorem 11.5. *Let X_1 and X_2 be continuous variables with copula C. Then*

$$\rho_S(X_1, X_2) = 12 \int_{[0,1]^2} C(u_1, u_2) dC(u_1, u_2) - 3 = 12E[C(U_1 U_2)] - 3$$

where (U_1, U_2) have distribution function C.

Remark 11.6. It is not hard to verify that (see the exercises) that

$$12E[U_1U_2] - 3 = \frac{E[(U_1 - EU_1)(U_2 - EU_2)]}{\sqrt{\mathrm{Var}(U_1)}\sqrt{\mathrm{Var}(U_2)}} = \rho_L(U_1, U_2)$$

which is the ordinary (linear) correlation for U_1 and U_2 with joint distribution function C.

The following result is useful for relating the rank correlations for the Gaussian copula.

Theorem 11.7. *Let $\mathbf{X} = (X_1, X_2)$ have a bivariate Gaussian distribution with copula C^{Ga}_Σ where*

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Then the ordinary linear correlation ρ of X_1 and X_2 is related to ρ_τ and ρ_S via

$$\rho_\tau(X_1, X_2) = \frac{2}{\pi} \arcsin \rho, \quad \rho_S(X_1, X_2) = \frac{6}{\pi} \arcsin \frac{\rho}{2}.$$

Proof. See Theorem 7.42 in [17]. ■

Tail dependence

Definition 11.8. The *coefficient of upper tail dependence* of X_1 and X_2 is given by

$$\lambda_U(X_1, X_2) = \lim_{u \uparrow 1} P(X_2 > F_2^{\leftarrow}(u) \mid X_1 > F_1^{\leftarrow}(u))$$

where F_i denotes the distribution function of X_i, $i = 1, 2$.

λ_U essentially measures how X_2 behaves when X_1 gets large. We leave it as an exercise for the reader to verify that in the extreme case where X_1 and X_2 are comonotone, then $\lambda_U(X_1, X_2) = 1$. To compute the coefficient of upper tail dependence, the following result is often useful.

Proposition 11.9. *Assume X_1 and X_2 have continuous distribution functions F_1 and F_2 and unique copula C. Then*

$$\lambda_U(X_1, X_2) = \lim_{u \uparrow 1} \frac{1 - 2u + C(u, u)}{1 - u}.$$

Proof. Since F_1 and F_2 are continuous, we have

$$\lambda_U(X_1, X_2) = \lim_{u \uparrow 1} P(F_2(X_2) > u \mid F_1(X_1) > u) = \lim_{u \uparrow 1} P(U_2 > u \mid U_1 > u).$$

Note that we have the following:

$$P(U_1 > u) + P(U_2 > u) + P(U_1 \le u, U_2 \le u) = 1 + P(U_1 > u, U_2 > u).$$

To see this, it may help to draw a figure. The equation can be rewritten as

$$(1 - u) + (1 - u) + C(u, u) = 1 + P(U_1 > u, U_2 > u)$$

so that

$$P(U_2 > u \mid U_1 > u) = \frac{P(U_1 > u, U_2 > u)}{P(U_1 > u)} = \frac{1 - 2u + C(u, u)}{1 - u}$$

from which the claim follows. ■

In the exercises below and in the mandatory assignments, we will see examples of computing the coefficient of upper tail dependence. We now introduce the corresponding notion for the lower tail.

Definition 11.10. The *coefficient of lower tail dependence* of X_1 and X_2 is given by

$$\lambda_L(X_1, X_2) = \lim_{u \downarrow 0} P(X_2 \le F_2^{\leftarrow}(u) \mid X_1 \le F_1^{\leftarrow}(u)).$$

We have an analogous result to the one above for continuous marginal distribution functions.

Proposition 11.11. *Assume X_1 and X_2 have continuous distribution functions F_1 and F_2 and unique copula C. Then*

$$\lambda_L(X_1, X_2) = \lim_{u \downarrow 0} \frac{C(u, u)}{u}.$$

Proof. The proof is essentially a simpler version of the one given for the coefficient of upper tail dependence. We compute

$$\begin{aligned}\lambda_L(X_1, X_2) &= \lim_{u \downarrow 0} P(X_2 \le F_2^{\leftarrow}(u) \mid X_1 \le F_1^{\leftarrow}(u)) = \lim_{u \downarrow 0} P(F_2(X_2) \le u \mid F_1(X_1) \le u) \\ &= \lim_{u \downarrow 0} P(U_2 \le u \mid U_1 \le u) = \lim_{u \downarrow 0} \frac{P(U_2 \le u, U_1 \le u)}{P(U_1 \le u)} = \lim_{u \downarrow 0} \frac{C(u, u)}{u}\end{aligned}$$

as desired. ■

Example 11.12. Let $(X_1, X_2) \sim \mathcal{N}(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Assume $\rho \in (-1, 1)$ (the cases $\rho = \pm 1$ are left as exercises). We want to show that the coefficient of upper tail dependence $\lambda_U(X_1, X_2)$ is zero. Consider a $t > 1$ such that $t\rho < 1$ and make the bound

$$\begin{aligned}P(X_2 > F_2^{\leftarrow}(u) \mid X_1 > F_1^{\leftarrow}(u)) \le\ & P(X_2 > F_2^{\leftarrow}(u), X_1 \le tF_1^{\leftarrow}(u) \mid X_1 > F_1^{\leftarrow}(u)) \\ &+ P(X_1 > tF_1^{\leftarrow}(u) \mid X_1 > F_1^{\leftarrow}(u)).\end{aligned}$$

Consider the second term first and let $v := F_1^{\leftarrow}(u)$. We have by L'Hospital's rule

$$\begin{aligned}\lim_{u\uparrow 1} P(X_1 > tF_1^{\leftarrow}(u) \mid X_1 > F_1^{\leftarrow}(u)) = \lim_{v\to\infty} \frac{P(X_1 > tv)}{P(X_1 > v)} &= \lim_{v\to\infty} \frac{\int_{tv}^{\infty} e^{-x^2/2}dx}{\int_{v}^{\infty} e^{-x^2/2}dx} \\ &= \lim_{v\to\infty} \frac{e^{-(tv)^2/2}}{e^{-v^2/2}} = 0.\end{aligned}$$

We now consider the limit of the first term i.e.

$$\lim_{u\uparrow 1} P(X_2 > F_2^{\leftarrow}(u), X_1 \leq tF_1^{\leftarrow}(u) \mid X_1 > F_1^{\leftarrow}(u)) = \lim_{v\to\infty} \frac{\int_{v}^{tv} P(X_2 > v \mid X_1 = x)d\Phi(x)}{\int_{v}^{\infty} d\Phi(x)}.$$

We now use that $X_2 \mid X_1 = x \sim \mathcal{N}(\rho x, 1-\rho^2)$. Letting $Y \sim \mathcal{N}(\rho x, 1-\rho^2)$ and setting $Z = (Y - \rho x)/\sqrt{1-\rho^2} \sim \mathcal{N}(0,1)$, we can rewrite

$$P(X_2 > v \mid X_1 = x) = P\left(Z > \frac{v - \rho x}{\sqrt{1-\rho^2}}\right).$$

Consider $x \in [v, tv]$. If $x = v$, then $v - \rho x = v - \rho v > 0$. Similarly, if $x = tv$, we have (due to the assumption $t\rho < 1$) that $v - \rho x = v - t\rho v > 0$. We conclude that

$$\lim_{v\to\infty} P\left(Z > \frac{v - \rho x}{\sqrt{1-\rho^2}}\right) = 0$$

uniformly for $x \in [v, tv]$. It follows that the limit of the first term goes to zero as well which establishes $\lambda_U(X_1, X_2) = 0$. We refer to this property as *asymptotic independence* in the tails. No matter how large the correlation ρ is in the range $(-1, 1)$, if we go far enough into the tails, extreme events occur independently in X_1 and X_2. This illustrates a potential problem with the Gaussian copula for modelling financial data. In such data, large losses in different variables are often correlated, and the normal copula may fail to capture this. ○

Notes and comments

The book by Nelsen, [19], contains all the information about copulas that we use in this course. The interested reader can find many supplementary examples of computations of the different rank correlations. Most of the proofs of the results this week can be found there as well, and if not, references are given.

Exercises

Exercise 5.1:

1)Recall that the Clayton copula has generator

$$\varphi(t) = \frac{1}{\theta}(t^{-\theta} - 1)$$

Verify that $\rho_\tau = \theta(\theta + 2)$.

2)Recall that the Gumbel copula has generator

$$\varphi(t) = (-\log t)^\theta.$$

Verify that $\rho_\tau = 1 - 1/\theta$.

Exercise 5.2:
Recall the Clayton copula

$$C_\theta^{\mathrm{Cl}}(u_1, u_2) = (u_1^{-1/\theta} + u_2^{-1/\theta} - 1)^{-1/\theta}, \quad 0 < \theta < \infty$$

and Gumbel copula,

$$C_\theta^{\mathrm{Gu}}(u_1, u_2) = \exp\left(-\left((-\log u_1)^\theta + (-\log u_2)^\theta\right)^{1/\theta}\right), \quad 1 \leq \theta < \infty.$$

1)Compute λ_U for the Clayton copula.

2)Compute λ_U for the Gumbel copula.

Exercise 5.3:
Prove the claim in Remark 11.6.

Exercise 5.4:
Let (X_1, X_2) be comonotone. Show that $\lambda_U(X_1, X_2) = 1$.

Exercise 5.5:
Consider the Archimedean copula with generator

$$\varphi(t) = \log\left(\frac{1 - \theta(1 - t)}{t}\right)$$

for $\theta \in [-1, 1)$ a parameter.

1)Write the copula explicitly.

2)Compute λ_U for this copula.

Week 6 - Credit risk I

12 Portfolio credit risk

Introduction and setup

Credit risk is the risk associated with loans and other obligations, namely the risk that a financial party is not able to pay what it owes to another party (for a loan, this is called *default*). Simply put, we have a situation where a financial institution (such as a bank) called the *lender* lends money to another party called the *obligor*,

$$\text{Lendor} \longrightarrow \text{Obligor}$$

In this course, we consider a very simple case. We assume the following:

- We have a one-period model. Namely, we will consider the loss that occurs over a single timestep such as a year, a month etc.
- We have n total loans.
- We have a probability of default p_i for loan i. p_i will depend on external factors, and these should be incorporated into the model.

Definition 12.1. The total one-period loss (for the bank, say) is given by

$$L = \sum_{i=1}^{n} X_i L_i (1 - \lambda_i)$$

where L_i is the size of the ith loan, λ_i is the *recovery rate* for the ith loan and

$$X_i = \begin{cases} 1, & \text{if default} \\ 0, & \text{if no default} \end{cases}.$$

The recovery rate λ_i is a number between $[0, 1]$ that describes how much the bank can recover in case of default. If, for example, a third of the loan is already repaid by the time of default, the bank will only lose two thirds of the loan. Before considering concrete models for credit risk, we make two important remarks.

(i) Defaults will be dependent since defaults are often triggered by external factors, for example increases in interest rates.

(ii) Large losses are usually not caused by one default but by a large number of defaults, even though the individual losses are often not large.

We will study the following models:

- The Merton model/KMV model.
- The Probit normal mixture model.
- The Bernoulli mixture model.
- The Poisson mixture model.

The Merton model is an example of a *structural model*. The three others are examples of *reduced form models*.

The Merton model

Merton's model considers a firm as an obligor. The model consists of a process V_A in continuous time describing the total value of the assets of the firm by time t and a fixed number K called the debt to be paid at time T. According to the model, V_A behaves like the assets in the Black-Scholes model:

$$dV_A(t) = \mu_A V_A(t)dt + \sigma_A V_A(t)dW(t)$$

where W is a standard Brownian motion. The reader can consult the rundown of the Black-Scholes model in week 1. Since V_A is a Geometric Brownian motion, we have the explicit solution for the assets at time T in terms of the value at the current time t, namely

$$V_A(T) = V_A(t)e^{\left(\mu_A - \frac{\sigma_A^2}{2}\right)(T-t)+\sigma_A(W(T)-W(t))}.$$

Note that $Z := W(T) - W(t) \sim \mathcal{N}(0, T-t)$. Default in this model means by definition that $V_A(T) < K$. We can rewrite this as follows:

$$\begin{aligned} V_A(T) < K \quad &\Leftrightarrow \quad \log V_A(t) + \left(\mu_A - \frac{\sigma_A^2}{2}\right)(T-t) + \sigma_A Z < \log K \\ &\Leftrightarrow \quad \frac{\log K - \log V_A(t) + \left(\frac{\sigma_A^2}{2} - \mu_A\right)(T-t)}{\sigma_A} > Z \\ &\Leftrightarrow \quad \frac{\log K - \log V_A(t) + \left(\frac{\sigma_A^2}{2} - \mu_A\right)(T-t)}{\sigma_A\sqrt{T-t}} > Y \end{aligned}$$

where $Y := Z/\sqrt{T-t} \sim \mathcal{N}(0,1)$. Note that in the above expression, Y is the only source of randomness since the value $V_A(t)$ is known at time t. We summarise our findings in the following definition.

Definition 12.2. In the above setup of Merton's model, the quantity

$$DD := -\frac{\log K - \log V_A(t) + \left(\frac{\sigma_A^2}{2} - \mu_A\right)(T-t)}{\sigma_A\sqrt{T-t}}$$

is called the *distance to default*. The probability of default can thus be written

$$P(\text{Default}) = P(Y < -DD)$$

with $Y \sim \mathcal{N}(0,1)$.

To estimate $P(\text{Default})$, we need to estimate DD. One problem is that V_A is not observable. It is hard to put a number on the value of every asset of a company (buildings, furniture, machines, patents etc.). However, the *equity* V_E is observable. We define the equity at time T to be

$$V_E(T) = (V_A(T) - K)^+$$

since if $K > V_A(T)$ the company defaults and the equity is zero. We recognise the above as a call option with strike K. If we also assume a constant interest rate r, we can apply the Black-Scholes formula,

$$V_E(t) = V_A(t)\Phi(z) - Ke^{-r(T-t)}\Phi(y)$$

where

$$z = \frac{\log V_A(t) - \log K + \left(r + \frac{\sigma_A^2}{2}\right)(T-t)}{\sigma_A\sqrt{T-t}}, \quad y = z - \sigma_A\sqrt{T-t}.$$

This formula relates V_A to V_E. A specific form of the Merton model used in the industry is the *KMV model.* In this model, one estimates the volatility of V_E, $\sigma_E = g(V_A(t), \sigma_A, r)$. Given μ_A and σ_A, the firm estimates the DD and hence the probability of default. All these estimates may be very unreliable and so in the actual KMV procedure, further steps are taken, namely:

- Consider n other firms.
- Calculate DD_i, $i = 1, ..., n$, where DD_i is the distance to default for firm i.
- Compare with past empirical observations with the same DD_i. Use the empirical frequency of default from these past loans and compare with the predictions from the estimates.

The description of the KMV model above is vague on purpose. Since it is an industry model, the exact procedures are not public and may change from firm to firm.

The Merton model we have considered so far only involves one firm, but the model is easily extended to an arbitrary number of firms.

The multivariate Merton model

In the multivariate Merton model, we consider n firms with asset processes

$$dV_{A,i}(t) = \mu_{A,i}V_{A,i}(t)dt + V_{A,i}(t)\sum_{j=1}^{m}\sigma_{A,i,j}dW_j(t), \quad i = 1, ..., n$$

with m independent Brownian motions $W_1, ..., W_m$. Note that all the Brownian motions appear in all the asset processes. This makes the $V_{A,i}$ dependent. One can think of the Brownian motions as underlying risk factors (for example fluctuations in interest rates). Letting $\sigma_{A,i}^2 = \sum_{j=1}^m \sigma_{A,i,j}^2$, we can explicitly solve for each $V_{A,i}$ and obtain

$$V_{A,i}(T) = V_{A,i}(t)e^{\left(\mu_{A,i} - \frac{\sigma_{A,i}^2}{2}\right)(T-t) + \sum_{j=1}^m \sigma_{A,i,j}(W_j(T) - W_j(t))}$$

which is very reminiscent of the univariate case. Letting

$$Z_i = \sum_{j=1}^{m} \sigma_{A,i,j}(W_j(T) - W_j(t)),$$

we have $Z_i \sim \mathcal{N}(0, \sigma_{A,i}^2(T-t))$. Like in the one-dimensional case, we can solve for Z_i and get

$$Z_i = \log V_{A,i}(T) - \log V_{A,i}(t) + \left(\frac{\sigma_{A,i}^2}{2} - \mu_{A,i}\right)(T-t).$$

The ith firm defaults if $V_{A,i}(T) < K$ where K is some threshold which we think of as the debt of the company. We can rewrite $V_{A,i}(T) < K$ as

$$Y_i := \frac{Z_i}{\sigma_{A,i}\sqrt{T-t}} < \frac{1}{\sigma_{A,i}\sqrt{T-t}}\left(\log K - \log V_{A,i}(t) + \left(\frac{\sigma_{A,i}^2}{2} - \mu_{A,i}\right)(T-t)\right) =: -DD_i$$

DD_i is the distance to default for the ith firm, and hence default for company i means $Y_i < -DD_i$. Since $Y_i \sim \mathcal{N}(0,1)$, we have

$$P(\text{Default for company } i) = P(Y_i < -DD_i) = \Phi(-DD_i).$$

Note that the Y_i are **dependent**. We want to describe this dependence. Note that we can write

$$Y_i = \frac{1}{\sigma_{A,i}\sqrt{T-t}} \sum_{j=1}^{m} \sigma_{A,i,j}(W_j(T) - W_j(t)) = \sum_{j=1}^{m} \frac{\sigma_{A,i,j}}{\sigma_{A,i}} \frac{W_j(T) - W_j(t)}{\sqrt{T-t}}$$

so the Y_i are weighted sums of the same iid $\mathcal{N}(0,1)$-variables. More generally, we can observe that the Y_i have the form

$$Y_i = \sum_{j=1}^{m} c_{i,j} R_j$$

for constants $c_{i,j}$ specific to the ith loan and $R_1, ..., R_m$ iid $\mathcal{N}(0,1)$. This leads us to consider models for the Y_i called *factor models*. In a factor model, we assume that the Y_i have the form

$$Y_i = \sum_{j=1}^{k} a_{ij} U_j + b_i \mathcal{W}_i.$$

The U_j are common stochastic factors that affect all loans, and we assume that $U_1, ..., U_k$ are iid $\mathcal{N}(0,1)$. The variable $\mathcal{W}_i$ is a firm specific stochastic factor, and the a_{ij} and b_i are constants. In order to estimate the default probability, we will need estimates of a_{ij}, b_i and DD_i. This is not an easy task. These quantities are not observable from data, and hence we cannot apply ordinary statistical methods to estimate them. Another approach is usually to divide the loans into different "classes" where for a specific class, a_{ij} and b_i do not depend on i but only on the class. This means that in a specific class, we have

$$Y_i = \sum_{j=1}^{k} a_j U_j + b \mathcal{W}_i.$$

A standard approach at this point is to normalize the constants so that $\sum_{j=1}^{k} a_j^2 + b^2 = 1$. Then

$$\sum_{j=1}^{k} a_j U_j \sim \mathcal{N}(0, \|\mathbf{a}\|^2), \quad \mathbf{a} = (a_1, ..., a_k)$$

from which it follows that $Y_i \overset{\mathrm{d}}{=} \|\mathbf{a}\| Z + b\mathcal{W}_i$ for $Z \sim \mathcal{N}(0,1)$. Since $\|\mathbf{a}\|^2 + b^2 = 1$, we can write $\|\mathbf{a}\| = \sqrt{\rho}$ and $b = \sqrt{1-\rho}$ for some $\rho \in (0,1)$. Y_i can thus be written as

$$Y_i \overset{\mathrm{d}}{=} \sqrt{\rho} Z + \sqrt{1-\rho}\mathcal{W}_i.$$

A model with Y_i of this form is called a *probit normal mixture model.*

Estimating VaR in the probit normal mixture model

Let $Z, \mathcal{W}_1, ..., \mathcal{W}_n$ be iid $\mathcal{N}(0,1)$ and $Y_i = \sqrt{\rho} Z + \sqrt{1-\rho}\mathcal{W}_i$ for $\rho \in (0,1)$. Default for the ith loan means $Y_i < d_i$ for some threshold d_i (earlier denoted by $-DD_i$). Now let

$$X_i = \begin{cases} 1, & \text{if } Y_i < d_i \\ 0, & \text{if } Y_i \geq d_i \end{cases}$$

then $N_n = \sum_{i=1}^{n} X_i$ is the total number of defaults in the portfolio of loans. We focus on the number of defaults and not the total loss, and the goal is to compute/estimate the VaR. We carry out this computation in several steps. Define $p_i(Z) = P(X_i = 1 \mid Z)$ i.e. the probability that the ith loan defaults given Z. Then

$$\begin{aligned} p_i(Z) &= P(Y_i < d_i \mid Z) = P(\sqrt{\rho} Z + \sqrt{1-\rho}\mathcal{W}_i < d_i \mid Z) \\ &= P\left(\mathcal{W}_i < \frac{d_i - \sqrt{\rho} Z}{\sqrt{1-\rho}} \mid Z\right) = \Phi\left(\frac{d_i - \sqrt{\rho} Z}{\sqrt{1-\rho}}\right) \end{aligned}$$

since $\mathcal{W}_i \sim \mathcal{N}(0,1)$. As $Z \overset{\mathrm{d}}{=} -Z$, we can rewrite the above to

$$\Phi\left(\frac{d_i - \sqrt{\rho} Z}{\sqrt{1-\rho}}\right) = \Phi\left(\frac{d_i}{\sqrt{1-\rho}} + \frac{\sqrt{\rho}}{\sqrt{1-\rho}} Z\right) = \Phi(a_i + bZ)$$

where

$$a_i = \frac{d_i}{\sqrt{1-\rho}}, \quad b = \frac{\sqrt{\rho}}{\sqrt{1-\rho}}.$$

Our goal is to compute VaR_α for N_n. To do so, we first compute the VaR for $p_i(Z)$. This amounts to solving the equation $1 - \alpha = P(p_i(Z) \geq x_i)$ for x_i:

$$\begin{aligned} 1 - \alpha &= P(p_i(Z) \geq x_i) = P(\Phi(a_i + bZ) \geq x_i) = P(a_i + bZ \geq \Phi^{-1}(x_i)) \\ &= P\left(Z \geq \frac{\Phi^{-1}(x_i) - a_i}{b}\right) = 1 - \Phi\left(\frac{\Phi^{-1}(x_i) - a_i}{b}\right) \end{aligned}$$

hence we need to solve

$$\alpha = \Phi\left(\frac{\Phi^{-1}(x_i) - a_i}{b}\right) \quad \Leftrightarrow \quad \Phi^{-1}(\alpha) = \frac{\Phi^{-1}(x_i) - a_i}{b} \quad \Leftrightarrow \quad a_i + b\Phi^{-1}(\alpha) = \Phi^{-1}(x_i)$$

which yields the final answer $x_i = \Phi(a_i + b\Phi^{-1}(\alpha))$. Hence

$$\mathrm{VaR}_\alpha(p_i(Z)) = \Phi(a_i + b\Phi^{-1}(\alpha)) = \Phi\left(\frac{d_i}{\sqrt{1-\rho}} + \sqrt{\frac{\rho}{1-\rho}}\Phi^{-1}(\alpha)\right).$$

Let us make the simplifying assumption that $d_i = d$ for all i, i.e. that the distances to default are identical for all loans. This also implies that $p_i(Z)$ is the same for all i since the Y_i are iid conditional on Z. This allows us to write $p(Z)$ without the subscript i. Furthermore, we will write $N_n(Z)$ to stress that N_n depends on Z. We claim that, conditional on Z,

$$\frac{N_n(Z)}{n} \xrightarrow{\mathrm{P}} p(Z)$$

where $\xrightarrow{\mathrm{P}}$ denotes convergence in probability. See the appendix for a review if necessary. To be precise, we claim that for every $\varepsilon > 0$,

$$P\left(\left|\frac{N_n(Z)}{n} - p(Z)\right| > \varepsilon \mid Z\right) \to 0$$

uniformly in Z. Let $\varepsilon > 0$ be given. We start by noting that

$$p(Z) = P(X_i = 1 \mid Z) = E[X_i \mid Z] = E\left[\frac{N_n(Z)}{n} \mid Z\right].$$

This comes from the fact that now,

$$X_i = \begin{cases} 1, & \text{if } Y_i < d \\ 0, & \text{if } Y_i \geq d \end{cases}$$

which implies that the X_i are identically distributed so that

$$E[X_1 \mid Z] = E[X_2 \mid Z] = \cdots = E[X_n \mid Z]$$

and thus

$$E[N_n(Z) \mid Z] = \sum_{j=1}^{n} E[X_j \mid Z] = nE[X_i \mid Z]$$

for any i. The rest of the proof is a straightforward application of Chebyshev's inequality. We get

$$\begin{aligned} P\left(\left|\frac{N_n(Z)}{n} - p(Z)\right| > \varepsilon \mid Z\right) &= P(|N_n(Z) - E[N_n(Z) \mid Z]| > n\varepsilon) \\ &\leq \frac{1}{n^2\varepsilon^2}\operatorname{Var}(N_n(Z) \mid Z) = \frac{p(Z)(1-p(Z))}{n\varepsilon^2}. \end{aligned}$$

The final inequality is a consequence of the X_i being iid given Z (since this is true for the Y_i) and from observing that conditional on Z, $N_n(Z)$ has a binomial distribution with parameters n and $p(Z)$. The above converges to zero uniformly in Z which proves the claim. We are now close to the goal of providing a formula for estimating the VaR. Assume $f(x) := P(p(Z) > x)$ is continuous. Then we have for $\alpha \in (0,1)$ that

$$1 - \alpha = P(p(Z) > \mathrm{VaR}_\alpha(p(Z)))$$

and using the result $N_n(Z)/n \xrightarrow{\mathrm{P}} p(Z)$, we have the approximation

$$1 - \alpha \approx P\left(\frac{N_n(Z)}{n} > \mathrm{VaR}_\alpha(p(Z))\right)$$

which yields the *Basel formula*

$$\mathrm{VaR}_\alpha(N_n(Z)) \approx n\,\mathrm{VaR}_\alpha(p(Z)) = n\Phi\left(\frac{d}{\sqrt{1-\rho}} + \sqrt{\frac{\rho}{1-\rho}}\Phi^{-1}(\alpha)\right).$$

Notes and comments

See section 10.3 in [17] for more on the Merton model. Section 10.1 gives an informal introduction to credit risk.

Exercises

Exercise 6.1:
Upgrade the statement $N_n(Z)/n \xrightarrow{\mathrm{P}} p(Z)$ to $N_n(Z)/n \to p(Z)$ almost surely (conditional on Z). Hint: Use the Markov inequality and A.2.27. The fourth central moment of the binomial distribution with parameters n and p is given by

$$np(1-p)(1+(3n-6)p(1-p)).$$

Week 7 - Credit risk II and further topics

13 Portfolio credit risk continued

The Bernoulli mixture model

In this model, we assume common factors $\mathbf{Z} = (Z_1, ..., Z_m)$. These could for example be economic factors such as interest rates. We again let $p_i(\mathbf{Z}) = P(X_i = 1 \mid \mathbf{Z})$ with $X_i = 1$ when company i defaults and $X_i = 0$ otherwise just like before. We assume that the defaults (i.e. the X_i) are independent given $\mathbf{Z}$. We again let $N_n = \sum_{i=1}^{n} X_i$ and note that the X_i are Bernoulli variables conditional on $\mathbf{Z}$ with success probability $p_i(\mathbf{Z})$. This also implies that N_n is binomial with parameters n and $p_i(\mathbf{Z})$ conditional on $\mathbf{Z}$.

We will consider a simplified version of this model, namely a so-called *one-factor model.* In such a model, we assume that $\mathbf{Z}$ is one-dimensional (so we write Z instead of $\mathbf{Z}$) and that $p_i(Z) = p(Z)$ is the same for all i. Note that this model is a generalisation of the probit normal mixture model. Indeed, the probit normal mixture model has this exact setup but with a specific X_i, namely

$$X_i = \begin{cases} 1, & \text{if } \sqrt{\rho} Z + \sqrt{1-\rho}\mathcal{W}_i < d \\ 0, & \text{if } \sqrt{\rho} Z + \sqrt{1-\rho}\mathcal{W}_i \geq d \end{cases}$$

for $Z, \mathcal{W}_1, ..., \mathcal{W}_n$ iid and standard normal as we saw last week. Returning to the one-factor Bernoulli mixture model, we have

$$P(N_n = k \mid Z) = \binom{n}{k} p(Z)^k (1 - p(Z))^{n-k}$$

as was also noted before. If Z has distribution function G, we can write

$$P(N_n = k) = \binom{n}{k} \int_{\mathbb{R}} p(z)^k (1 - p(z)^{n-k} dG(z).$$

What are choices of G and $p(Z)$ that make the above expression mathematically tractible? Consider the special case $Z \sim \text{Beta}(a, b)$ and $p(Z) = Z$. This model is called the *Beta mixture model.* The density of Z is

$$g(z) = \frac{1}{\beta(a,b)} z^{a-1} (1-z)^{b-1}, \quad 0 \leq z \leq 1 \quad \text{where}$$

$$\beta(a,b) = \int_0^1 z^{a-1}(1-z)^{b-1}dz = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

With these assumptions, we can compute $P(N_n = k)$ explicitly as follows:

$$\begin{aligned} P(N_n = k) &= \binom{n}{k}\int_0^1 z^k(1-z)^{n-k}g(z)dz = \binom{n}{k}\int_0^1 \frac{1}{\beta(a,b)}z^{k+a-1}(1-z)^{n-k+b-1}dz \\ &= \binom{n}{k}\frac{\beta(a+k,b+n-k)}{\beta(a,b)}. \end{aligned}$$

Since we now have an explicit expression for the density of N_k, we can compute all sorts of risk measures such as VaR, ES etc. explicitly as well. While this is a nice property of the model, we stress that the motivation for choosing $Z \sim \text{Beta}(a,b)$ and $p(Z) = Z$ is purely mathematical. Nevertheless, we continue our study of the model. How do we estimate a and b? While a and b are not directly observed from data, we can relate it to quantities that are observed. We know the number of defaults that have occured at a given time which allows us to estimate the probability of default p, at least in a portfolio of homogeneous loans (think for example of a portfolio consisting only of AAA rated loans, only B rated loans etc.). We can also estimate the linear correlation ρ_L of the X_i. We now determine the relations between these quantities and a and b in the Beta mixture model. We have

$$\begin{aligned} p &= P(\text{Default}) = E[P(\text{Default} \mid Z)] = E[p(Z)] = E[Z] = \frac{a}{a+b}, \\ \rho_L &= \frac{\text{Cov}(X_i,X_j)}{\text{Var}(X_i)} = \frac{E[(X_i-p)(X_j-p)]}{p(1-p)} = \frac{E[X_iX_j] - 2pE[X_i] + p^2}{p(1-p)} \end{aligned}$$

and

$$\begin{aligned} E[X_iX_j] &= P(X_iX_j = 1) = P(X_i = 1, X_j = 1) = E[P(X_i = 1, X_j = 1 \mid Z)] \\ &= E[P(X_i = 1 \mid Z)P(X_j = 1 \mid Z)] = E[Z^2] = \frac{a(a+1)}{(a+b)(a+b+1)}. \end{aligned}$$

We hence have two equations in two unknowns. Solving these yields the relations

$$a = (1-p)\frac{1-\rho_L}{\rho_L}, \quad b = p\frac{1-\rho_L}{\rho_L}.$$

To summarise: We can estimate a and b from data by first estimating the default probability and the linear correlation from the data and then apply the formulas above.

The Poisson mixture model

The Poisson mixture model is different from the previous models in the sense that the X_i can attain infinitely many values, namely $X_i \in \{0,1,...\}$. We assume that the X_i are independent given $\mathbf{Z}$ and that they have a Poisson distribution conditionally on $\mathbf{Z}$, namely

$$P(X_i = k \mid \mathbf{Z}) = \frac{(\lambda_i(\mathbf{Z}))^k}{k!}e^{-\lambda_i(\mathbf{Z})}$$

for some function λ_i. Just like earlier, $\mathbf{Z}$ is a collection of m variables which we interpret as some underlying factors. We think of $X_i = 1$ as a default of the ith loan and $X_i = 0$ as no default. The other events, $X_i = k$ for $k \geq 2$, do not have a natural interpretation, but we choose λ_i small enough so that $X_i > 1$ happens with very low probability. We again consider $N_n = \sum_{i=1}^{n} X_i$, i.e. the number of defaults (at least for small λ_i). We will consider a special case of the Poisson mixture model which goes under the name CreditRisk$^+$. In CreditRisk$^+$, we assume the following:

- $Z_1, ..., Z_m$ are independent.
- $Z_j \sim \Gamma(\alpha_j, \beta_j^{-1})$ with $\alpha_j \beta_j = 1$.
- $\lambda_i(\mathbf{Z}) = \overline{\lambda}_i \sum_{j=1}^{m} a_{ij} Z_j$ with $a_{ij} \geq 0$ and $\sum_{j=1}^{m} a_{ij} = 1$. Here $\overline{\lambda}_i > 0$ is a constant for each i.

The assumption $\alpha_j \beta_j = 1$ implies that $E[Z_j] = 1$. With the above assumptions, we have $E[\lambda_i(\mathbf{Z})] = \overline{\lambda}_i$ which gives us control over the λ_i functions. We need to choose them small. With the above assumptions, it is possible to derive the distribution of N_n which is also a motivation for the model. In order to do so, we need to introduce some theory.

Definition 13.1. For a discrete random variable Y with values in $\{0, 1, ...\}$, the function

$$g_Y(t) = E[t^Y]$$

is called the *probability-generating function* of Y.

Remark 13.2. Notice the relation $g_Y(t) = E[e^{Y \log t}] = \kappa_Y(\log t)$ with κ_Y the moment generating function of Y. This relation implies that the probability-generating function (if it exists) determines the distribution of Y.

Example 13.3. Let N be Poisson distributed with parameter $\lambda > 0$. Then

$$\kappa_N(t) = e^{\lambda(e^t - 1)}, \quad t \in \mathbb{R}.$$

Hence the probability-generating function is

$$g_N(t) = e^{\lambda(t-1)}.$$

○

Example 13.4. Let N have a negative binomial distribution with parameters r and p i.e.

$$P(N = k) = \binom{k + r - 1}{k} (1 - p)^k p^r, \quad k = 0, 1, 2, ...$$

We leave it as an exercise for the reader to verify that the probability-generating function of N is given by

$$g_N(t) = \left(\frac{p}{1 - (1 - p)t} \right)^r \quad \text{for} \quad |t| < \frac{1}{p}.$$

○

Theorem 13.5. *With the assumptions of the CreditRisk$^+$ model, we have*

$$g_{N_n}(t) = \prod_{j=1}^{m} \left(\frac{1 - \delta_j}{1 - \delta_j t} \right)^{\alpha_j}, \quad \text{where} \quad \delta_j = \frac{\beta_j \sum_{i=1}^{n} \overline{\lambda}_i a_{ij}}{1 + \beta_j \sum_{i=1}^{n} \overline{\lambda}_i a_{ij}}.$$

Proof. The proof is essentially a computational exercise in the assumptions of the CreditRisk$^+$ model. We start by conditioning on $\mathbf{Z}$. The conditional probability-generating function for N_n given $\mathbf{Z}$ is

$$\begin{aligned} g_{N_n|\mathbf{Z}}(t) &= E\left[t^{N_n} \mid \mathbf{Z}\right] = E\left[t^{X_1+\cdots+X_n} \mid \mathbf{Z}\right] = \prod_{i=1}^{n} E\left[t^{X_i} \mid \mathbf{Z}\right] \\ &= \prod_{i=1}^{n} \kappa_{X_i|\mathbf{Z}}(\log t) = \prod_{i=1}^{n} e^{\lambda_i(\mathbf{Z})(t-1)} \end{aligned}$$

where we used the conditional independence and the above example for the Poisson distribution. By applying the tower property, we can remove the conditioning on $\mathbf{Z}$ by taking an expectation. Let f_j denote the density of Z_j. Then we have by independence of the Z_j that

$$\begin{aligned} g_{N_n}(t) &= E\left[E\left[t^{N_n} \mid \mathbf{Z}\right]\right] = E\left[\prod_{i=1}^{n} e^{\lambda_i(\mathbf{Z})(t-1)}\right] \\ &= \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^{n} e^{\lambda_i(\mathbf{z})(t-1)} f_1(z_1)\cdots f_m(z_m) dz_1 \cdots dz_m \\ &= \int_0^\infty \cdots \int_0^\infty e^{(t-1)\sum_{i=1}^{n} \overline{\lambda}_i \sum_{j=1}^{m} a_{ij} z_j} f_1(z_1)\cdots f_m(z_m) dz_1 \cdots dz_m \\ &= \int_0^\infty \cdots \int_0^\infty e^{(t-1)\sum_{j=1}^{m} \sum_{i=1}^{n} \overline{\lambda}_i a_{ij} z_j} f_1(z_1)\cdots f_m(z_m) dz_1 \cdots dz_m \end{aligned}$$

For the sake of simplicity, let $\mu_j = \sum_{i=1}^{n} \overline{\lambda}_i a_{ij}$. Then we can continue the computation as follows:

$$\begin{aligned} g_{N_n}(t) &= \int_0^\infty \cdots \int_0^\infty e^{(t-1)\sum_{j=1}^{m} \mu_j z_j} f_1(z_1)\cdots f_m(z_m) dz_1 \cdots dz_m \\ &= \int_0^\infty \cdots \int_0^\infty e^{(t-1)\mu_1 z_1} f(z_1) dz_1 \cdots e^{(t-1)\mu_m z_m} f(z_m) dz_m \\ &= \prod_{j=1}^{m} \int_0^\infty e^{(t-1)\mu_j z} f(z) dz = \prod_{j=1}^{m} \int_0^\infty e^{(t-1)\mu_j z} \frac{1}{\beta_j^{\alpha_j}\Gamma(\alpha_j)} z^{\alpha_j - 1} e^{-z/\beta_j} dz \\ &= \prod_{j=1}^{m} \int_0^\infty \frac{1}{\beta_j^{\alpha_j}\Gamma(\alpha_j)} z^{\alpha_j - 1} e^{-z(\beta_j^{-1} - (t-1)\mu_j)} dz. \end{aligned}$$

We now compute each integral (denoted by I_j) in the product:

$$\begin{aligned} I_j &= \int_0^\infty \frac{1}{\beta_j^{\alpha_j}\Gamma(\alpha_j)} z^{\alpha_j - 1} e^{-z(\beta_j^{-1} - (t-1)\mu_j)} dz \\ &= \frac{(\beta_j^{-1} - (t-1)\mu_j)^{-\alpha_j}\Gamma(\alpha_j)}{\beta_j^{\alpha_j}\Gamma(\alpha_j)} \int_0^\infty \frac{(\beta_j^{-1} - (t-1)\mu_j)^{\alpha_j}}{\Gamma(\alpha_j)} z^{\alpha_j - 1} e^{-z(\beta_j^{-1} - (t-1)\mu_j)} dz \\ &= \frac{1}{\beta_j^{\alpha_j}(\beta_j^{-1} - (t-1)\mu_j)^{\alpha_j}} = \frac{1}{(1 - (t-1)\beta_j\mu_j)^{\alpha_j}} = \left(\frac{1-\delta_j}{1-\delta_j t}\right)^{\alpha_j} \end{aligned}$$

by noting that δ_j as defined in the theorem is given by

$$\delta_j = \frac{\beta_j \mu_j}{1 + \beta_j \mu_j}.$$

Plugging this expression back into the one for $g_{N_n}(t)$ completes the proof. ■

Remark 13.6. Note that the terms

$$\left(\frac{1 - \delta_j}{1 - \delta_j t} \right)^{\alpha_j}$$

are probability-generating functions for negative binomial variables with parameters $p = 1 - \delta_j$ and $r = \alpha_j$.

In principle, one can invert g_{N_n}. There is a whole litterature dedicated to inverting moment and probability-generating functions. We will not pursue this here. We will only mention that one can make a very crude approximation that relies on Markov's inequality, namely

$$P(N_n > k) \leq \frac{E[t^{N_n}]}{t^k} = \frac{g_{N_n}(t)}{t^k}$$

for every $t > 0$. One can then minimise this expression over t.

14 Operational risk

Operational risk can be stated as ”loss from failed internal processes, people or systems or from external events”. To elaborate a bit on this, we can roughly divide such risks in categories. One category is *repetitive human errors* or *repetitive operational risks* (repetitive OR). These include IT failures, errors in settlements of transactions, litigation and the like. Other types of losses include fraud and external events such as flooding, fires, earthquakes and terrorism (although the latter is extremely hard to model). A difficulty in operational risk is that we often have little data available, and the data is often heavy tailed. The claim arrivals can also be hard to model since they often occur randomly in time (and often in clusters) and since the frequency changes over time. One can for example imagine that a large traded volume leads to a large number of back office errors.

Approaches in analyzing operational risk

We will now discuss the basics of two approaches in analyzing operational risk. These are

- The *basic indicator approach.*
- The *advanced measurement approach.*

Under the basic indicator (BI) approach, the capital requirement to cover OR (operational risk) losses at time n is given by

$$\mathrm{RC}^n_{BI}(\mathrm{OR}) = \frac{1}{Z_n} \sum_{i=1}^{3} \alpha \max(\mathrm{GI}^{n-i}, 0),$$

where $\alpha \approx 0.15$ is a constant,

$$Z_n = \#\text{years where } \mathrm{GI}^{n-i} > 0, i = 1, 2, 3,$$

and GI^s denotes the "gross income" at time s. Under the advanced measurement (AM) approach, we divide into K lines of business (typically $K = 8$ and lines include for example corporate finance, trading and sales). The capital requirement to cover OR losses is given by

$$\mathrm{RC}^n_{AM}(\mathrm{OR}) = \sum_{b=1}^{K} \rho_\alpha(L^{n,b})$$

where $0.99 \le \alpha \le 0.999$ and ρ_α is a risk measure such as VaR_α or ES_α.

Mathematical estimates

In this subsection, we will investigate methods to analyze the loss via a stochastic process approach. Let us start by recalling the definition of a Poisson process.

Definition 14.1. A stochastic process $\{N_t\}$ is called a *Poisson process* with *intensity* $\lambda > 0$ if N_t takes values in $\{0, 1, 2, ...\}$ and

(i) $P(N_h \ge 1) = \lambda h + o(h)$,

(ii) $P(N_h \ge 2) = o(h)$ and

(iii) $\{N_t\}$ has stationary and independent increments.

Recall that stationarity means that for every $s \le t$, $X_t - X_s \overset{\mathrm{d}}{=} X_{t-s}$. By independent increments, we mean that for every finite partition $0 < t_1 < t_2 < \cdots < t_k$, the variables $\{X_{t_{i+1}} - X_{t_i}\}_{i=1}^{k-1}$ are independent. Intuitively, we think of a Poisson process as a claim number process which satisfies

$$P(1 \text{ claim in } [t, t+h]) = \lambda h + o(h), \quad P(\ge 2 \text{ claims in } [t, t+h]) = o(h).$$

We will now model the loss of the company at time t by the process

$$L_t = \sum_{i=1}^{N_t} X_i$$

with $\{N_t\}$ a Poisson process with intensity λ independent of the iid sequence $\{X_i\}$. We will discuss the following ideas based on risk theory:

- Laplace transform method.
- Panjer recursion.
- A sophisticated large deviation approach based on the "Arwedson approximation" from risk theory.
- Time-dependent intensity.
- Stochastic processes for market risk.

Let us first discuss the Laplace transform method. Recall that the Laplace transform of a random variable Y is given by $\psi_Y(s) = E[e^{-sY}]$. We can compute the Laplace transform of the loss using the tower property as follows:

$$\psi_{L_t}(s) = E\left[e^{-sL_t}\right] = E\left[E\left[e^{-s\sum_{i=1}^{N_t} X_i} \mid N_t\right]\right] = E\left[\psi_X(s)^{N_t}\right]$$

where we have defined $\psi_X(s) = E\left[e^{-sX_1}\right]$. We continue the computation and get

$$\psi_{L_t}(s) = \sum_{n=0}^{\infty}(\psi_X(s))^n \frac{\lambda^n}{n!}e^{-\lambda} = \sum_{n=0}^{\infty}\frac{(\psi_X(s)\lambda)^n}{n!}e^{-\lambda} = e^{\lambda(\psi_X(s)-1)}.$$

We should note that all the Laplace transforms exist since we are working with non-negative random variables. After obtaining the Laplace-transform, numerical inversion techniques can be applied. The method is more flexible than this however. To illustrate this, consider a Poisson intensity that changes over time. To make this concrete, assume we are considering the time interval $[0,2]$, and we have Poisson processes N_1 and N_2 on $[0,1]$ and $(1,2]$, respectively, along with two claim sequences $\{X_i^{(1)}\}$ and $\{X_i^{(2)}\}$ belonging to each interval. The total loss is obtained by summing losses from each interval i.e. $L = L_1 + L_2$, where L_1 and L_2 are assumed to be independent. Then

$$\begin{aligned}\psi_L(s) &= E\left[e^{-s(L_1+L_2)}\right] = E\left[e^{-sL_1}\right]E\left[e^{-sL_2}\right] = \psi_{L_1}(s)\psi_{L_2}(s)\\ &= e^{-\lambda_1(\psi_{X^{(1)}}(s)-1)}e^{-\lambda_2(\psi_{X^{(2)}}(s)-1)},\end{aligned}$$

and we can invert this function to obtain the distribution. For those familiar with mixture distributions, the above Laplace transform is one of a compound Poisson sum

$$\sum_{i=1}^{\tilde{N}_t} Y_i, \quad t \in [0,1]$$

with $\{\tilde{N}_t\}$ a Poisson process with intensity $\lambda_1 + \lambda_2$ and Y_i mixture distributed with distribution function

$$G(y) = \frac{\lambda_1}{\lambda_1+\lambda_2}G_1(y) + \frac{\lambda_2}{\lambda_1+\lambda_2}G_2(y)$$

where $X_1^{(1)} \sim G_1$ and $X_1^{(2)} \sim G_2$. In practice, one often observes $E[N_t] < \mathrm{Var}(N_t)$, a phenomenon called *overdispersion.* To remedy this, one can use a *mixed Poisson process* $\{N_t\}$ defined by $N_t = \tilde{N}_{\Lambda t}$ where $\{\tilde{N}_t\}$ is a Poisson process with intensity 1 and $\Lambda > 0$ is a random variable.

Example 14.2. Choosing $\Lambda \sim \Gamma(\alpha, \beta)$ leads to the so-called *negative binomial process.* We let the reader verify that $N_t = \tilde{N}_{\Lambda t}$ is indeed negative binomial distributed. ○

We now leave the world of Laplace transforms and discuss the next topic, namely *Panjer recursion.* For this technique to be applicable, assume N_t satisfies the recursion

$$q_n := P(N_t = n) = \left(a + \frac{b}{n}\right)q_{n-1}, \quad n = 1, 2, ...$$

for constants a and b. We also assume that $X_i \in \{1, 2, ...\}$. Panjer recursion yields the exact recursive formula for $p_n = P(L_t = n)$:

$$p_n = \sum_{i=1}^{n} \left(a + \frac{bi}{n} \right) P(X_1 = i) p_{n-i}, \quad p_0 = q_0.$$

The assumption $X_i \in \{1, 2, ...\}$ is not as restrictive as it may seem. One can always scale the values as necessary.

We now consider the *Arwedson approximation.* For the loss process

$$L_t = \sum_{i=1}^{N_t} X_i,$$

we consider a "small" $\delta \in (0, 1)$ and the probability of a large loss over the small time interval $[0, \delta u]$,

$$\varphi_\delta(u) := P(L_t > u, \text{ some } 0 \le t \le \delta u).$$

Ultimately, we will choose $\delta u = 1$. Those familiar with classical ruin theory will immediately see the connection. But one should note that this situation is slightly different since we have no premium payments, and we are working over a finite time interval $[0, \delta u]$ and not $[0, \infty)$. The Arwedson approximation was originally developed in the study of *finite-time* ruin theory. In ruin theory, one studies the *Cramér-Lundberg process*

$$C_t = u + ct - \sum_{i=1}^{N_t} X_i$$

with $u \ge 0$ the initial capital of the company, $c > 0$ a constant premium rate and $\{X_i\}$ the insurance claim sizes. Arwedson considered the finite-time ruin probabilities

$$\Psi_K(u) = P(C_t < 0, \text{ for some } 0 \le t \le Ku)$$

where $0 \le K < \infty$. Under classical Cramér-Lundberg assumptions, Arwedson showed that

$$\Psi_K(u) \sim \begin{cases} \frac{C_K}{\sqrt{u}} e^{-uI(K)}, & \text{if } K \le \rho \\ Ce^{-Ru}, & \text{if } K > \rho \end{cases}.$$

Note that the case $K > \rho$ corresponds to the ordinary Cramér-Lundberg estimate. $I(K)$ would nowadays be called the "large deviation rate function" which describes the exponential decay of a probability as $u \to \infty$. R solves the equation $\Lambda(R) := \log E\left[e^{R(C_1 - u)}\right] = 0$ (called the *adjustment coefficient* in ruin theory). Here, $I(K) > R$ for $K < \rho$ and $\rho = (\Lambda'(R))^{-1}$ (one can show ρu is the "most likely" time of ruin).

Returning to $\varphi_\delta(u)$, if X_i has exponential moments (i.e. is "light tailed"), one can show

$$\varphi_\delta(u) \sim \frac{\mathcal{C}(\delta)}{\sqrt{u}} e^{-uJ(\delta)} \quad \text{as} \quad u \to \infty$$

which has connections to Arwedson's original result as well as the exponentially shifted measure from our discussion of importance sampling. If X_1 is *subexponential* (think: heavy tails), for example if X_1 is regularly varying, one can prove that

$$\varphi_\delta(u) \sim Du\overline{F}_X(u(1 - \delta\mu)) \quad \text{as} \quad u \to \infty.$$

The proof relies on the concept of "one large jump" in heavy tailed ruin problems. A concept that should be familiar with someone who has studied ruin theory.

We now turn to the point of time-dependent intensity. It makes sense to assume that the intensity of N_t changes over time.

Example 14.3. Assume N_t has intensity λ_n in the interval $(n-1, n]$ where

$$\lambda_n = F(Z_n), \quad Z_n = cZ_{n-1} + \xi_n, \quad |c| < 1,$$

and $\{\xi_n\}$ is iid $\mathcal{N}(0,1)$ and Z_n is a so-called AR(1) process. Z_n could represent traded volume, for example, which is linked to increases in operational risk. ○

For time-dependent intensities of ARMA-type, similar "Arwedson" approximations can be derived. Namely, one can likewise show

$$\varphi_\delta(u) \sim Du\overline{F}_X(u(1-\delta\mu)) \quad \text{as} \quad u \to \infty.$$

Similar insurance-based methods are potentially useful for market risk. We end this week (and this course) with a brief discussion on stochastic processes for market risk. Throughout the course, we only considered one-period models, and we assumed iid returns. Real-life data is not iid! Hence stochastic processes (time-dependent models) are called for. This is very complicated because multiple stochastic processes are usually dependent, and this is difficult to model, so most current research either considers dependent processes in one dimension or coordinatwise dependence in one-period models (but not both). A very classical model for dependence is the ARMA(p,q) model:

$$X_n - \sum_{j=1}^{p} \phi_j X_{n-j} = Z_n + \sum_{i=1}^{q} \theta_i Z_{n-i}, \quad t = 1, 2, ...$$

where $\{Z_n\}$ is an iid $\mathcal{N}(0,1)$ sequence and ϕ_j, θ_i constants. This is an example of a *time series model.* For "sufficiently small" ϕ_j and θ_i, we have that

$$X_n \xrightarrow{\mathrm{d}} X$$

i.e. that X_n converges to a stationary distribution where X is normally distributed. In an evolution of log-returns of stocks, one typically observes the following:

- Log-returns contain many "large" values i.e. the data is heavy-tailed.
- Exceedances of high thresholds occur in clusters, i.e. we have dependence in the tails.
- While dependent, returns show little serial correlation.
- Absolute (or squared) returns show strong serial correlation.
- Volatility varies over time.

To address these issues, the GARCH models were introduced. The first of these models, the ARCH(1) model, was introduced by Engle, see [11]. In this model, the log-returns $\{R_n\}$ satisfy

$$R_n^2 = (\phi_0 + \phi_1 R_{n-1}^2) Z_n^2, \quad n = 1, 2, ...$$

where $\{Z_n\}$ is an iid $\mathcal{N}(0,1)$ sequence. A more complicated model is the GARCH(1,1) model introduced by Bollerslev, see [7]. Here the log-returns $\{R_n\}$ satisfy

$$R_n = \sigma_n Z_n, \quad n = 1, 2, ...$$

where $\{Z_n\}$ is iid $\mathcal{N}(0,1)$ and

$$\sigma_n^2 = \alpha_0 + \beta_1 \sigma_{n-1}^2 + \alpha_1 R_{n-1}^2 = \alpha_0 + \sigma_{n-1}^2(\beta_1 + \alpha_1 Z_{n-1}^2).$$

Both ARCH(1) and GARCH(1,1) are examples of *stochastic recursive sequences.* Namely,

$$V_n = A_n V_{n-1} + B_n, \quad n = 1, 2, ...$$

where

$$\begin{aligned} V_n &= R_n^2 \quad \text{for ARCH(1)}, \\ V_n &= \sigma_n^2 \quad \text{for GARCH(1, 1)}. \end{aligned}$$

Here, $\{(A_n, B_n) : n = 1, 2, ...\}$ is any iid sequence on $(0, \infty) \times \mathbb{R}$. Under certain reasonable conditions, $V_n \xrightarrow{\text{d}} V$, and it is natural to consider $P(V > u)$ for large u. One can apply renewal theoretic methods (such as those presented in the course SkadeStok) to obtain

$$P(V > u) \sim Cu^{-R} \quad \text{as } u \to \infty$$

where

$$\Lambda(R) = 0, \quad \Lambda(\xi) = \log E\left[e^{\xi \log A}\right].$$

This shows that Pareto tails characterise the decay rate. For more complex models (e.g. GARCH(p,q)), one needs to consider *matrix recursions.* This is currently an active research area.

Notes and comments

The computation in the proof of Theorem 13.5 is inspired by the one in section 12.2 of [14]. For more information about Panjer recursion, we refer to [18], section 3.3.3. In the final discussion on operational risk, many tools from ruin theory were discussed. Ruin theory and related tools such as renewal theory were discussed in the course SkadeStok. See [16] for lecture notes from the last run of the course.

Exercises

Exercise 7.1:

Verify that the probability-generating function for a negative binomial variable N with parameters r and p is given by

$$g_N(t) = \left(\frac{p}{1-(1-p)t}\right)^r \quad \text{for} \quad |t| < \frac{1}{p}.$$

Exercise 7.2:

Let $\{\tilde{N}_t\}$ be a Poisson process with intensity 1 and $\Lambda \sim \Gamma(\alpha, \beta)$. Verify that the mixed Poisson process $N_t = \tilde{N}_{\Lambda t}$ has a negative binomial distribution and determine the parameters.

Exercise 7.3:

Let N be a discrete random variable with $N \in \{0, 1, 2, ...\}$. Define $q_n := P(N = n)$ and consider the relation

$$q_n = \left(a + \frac{b}{n}\right) q_{n-1}, \quad n = 1, 2, ...$$

Prove that N satisfies this relation for proper choices of a and b in the following cases.

1)N Poisson distributed with intensity $\lambda > 0$.

2)N Binomial distributed with parameters k and p.

3)N negative binomial distributed with parameters p and r.

One can prove that these three distributions are the only distributions satisfying this recursive relation.

Appendix A

Preliminaries

A.1 Generalised inverses

Definition A.1.1. Let $h : \mathbb{R} \to \mathbb{R}$ be a non-decreasing function. We define the *generalised inverse* of h as

$$h^{\leftarrow}(t) = \inf\{x \in \mathbb{R} : h(x) \geq t\}.$$

We have the convention $\inf \emptyset = \infty$.

Proposition A.1.2. *For a non-decreasing function h, $h^{\leftarrow}$ is left-continuous.*

Proof. This proof is from [21]. Assume that $t_n \uparrow t$ but $H^{\leftarrow}(t-) := \lim_{t_n \uparrow t} H^{\leftarrow}(t_n) < H^{\leftarrow}(t)$. Then we can find $x \in \mathbb{R}$ and $\delta > 0$ such that for all n,

$$H^{\leftarrow}(t_n) < x < H^{\leftarrow}(t) - \delta.$$

As $x \in \{y \in \mathbb{R} : H(y) \geq t_n\}$, we have $H(x) \geq t_n$ for all n. Let $n \to \infty$, then $H(x) \geq t$ so by definition of $H^{\leftarrow}$, $H^{\leftarrow}(t) \leq x$. This is in contradiction to $x < H^{\leftarrow}(t) - \delta$. ■

The following properties of generalised inverses will be useful.

Proposition A.1.3. *Let h be non-decreasing.*

(i) $x \geq h^{\leftarrow}(t)$ if and only if $h(x) \geq t$.

(ii) h is continuous if and only if $h^{\leftarrow}$ is strictly increasing.

(iii) $h(h^{\leftarrow}(t)) = t$ for all t if and only if h is continuous.

(iv) h is strictly increasing if and only if $h^{\leftarrow}(h(x)) = x$ for all x.

Proof. Point (i) is left as an exercise. Consider (ii). h is non-decreasing, so any discontinuity is a (positive) jump. Since a jump of h corresponds to a flat region for $h^{\leftarrow}$ (make a drawing!), the claim follows. One can make similar arguments for (iii) and (iv). For complete proofs of these statements and many others concerning generalised inverses, consult [9]. ■

Proposition A.1.4. *Let F be the distribution function of the random variable X.*

(i) $F^{-1}(U) \overset{\mathrm{d}}{=} X$ for $U \sim$ Unif$(0,1)$.

(ii) If F is continuous, $F(X) \overset{\mathrm{d}}{=} U$ for $U \sim \mathrm{Unif}(0,1)$.

(iii) $P(X \leq x) = P(F(X) \leq F(x))$.

Proof. (i) and (ii) are left as exercises. As for (iii), we note that $X \leq x$ implies $F(X) \leq x$ since F is non-decreasing. Conversely, consider the event $\{F(X) \leq F(x), X > x\}$. If $X > x$, we have $F(X) \leq F(x)$ and hence $\{F(X) \leq F(x), X > x\} \subseteq \{F(X) = F(x), X > x\}$ so that F is flat on $[x, X]$. This implies $P(F(X) \leq F(x), X > x) = 0$, completing the proof. ■

Exercises

Exercise A.1:
Prove (i) in Proposition A.1.3.

Exercise A.2:
Prove (i) and (ii) in Proposition A.1.4. Give a counterexample which shows that (ii) need not hold for general distribution functions.

A.2 Probability theory

Distribution functions

Definition A.2.1. If $(\Omega, \mathcal{F}, P)$ is a probability space, a *random variable* is a measurable map $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ where $\mathcal{B}$ is the Borel sigma-algebra on $\mathbb{R}$.

In this course, measurability is not an issue. We will also rarely worry about the background space $(\Omega, \mathcal{F}, P)$. We now go through distribution functions in some detail since distribution functions and quantile functions play a central role in the course.

Definition A.2.2. For a random variable X, we define the *distribution function* F of X as

$$F(x) = P(X \leq x).$$

Similarly, if $\mathbf{X} = (X_1, ..., X_d)$ is $\mathbb{R}^d$-valued, the distribution function F is given by

$$F(x_1, ..., x_d) = P(X_1 \leq x_1, ..., X_d \leq x_d).$$

If F is a distribution function for X, we call $\overline{F} = 1 - F$ the *survival function* of X.

In the univariate case, we have a nice characterisation of distribution functions.

Proposition A.2.3. *A function $F : \mathbb{R} \to \mathbb{R}$ is the unique distribution function of a random variable if and only if the following properties hold:*

1. *F is right-continuous.*
2. *F is non-decreasing.*
3. $\lim_{x \to \infty} F(x) = 1$.
4. $\lim_{x \to -\infty} F(x) = 0$.

Proof. Assume that F is a distribution function for the random variable X. For $\varepsilon > 0$, we have $\{X \leq x+\varepsilon\} \downarrow \{X \leq x\}$ for $\varepsilon \downarrow 0$. By continuity from above for measures, $F(x+\varepsilon) \to F(x)$ for $\varepsilon \downarrow 0$, showing that F has property 1. If $x \leq y$, then $\{X \leq x\} \subseteq \{X \leq y\}$ so that $F(x) = P(X \leq x) \leq P(X \leq y) = F(y)$, proving 2. Properties 3 and 4 follow from the fact that X is real-valued. Conversely, suppose F satisfies properties 1 - 4. Let $X = F^{\leftarrow}(U)$ where U is uniformly distributed on $(0,1)$. Then

$$P(X \leq x) = P(F^{\leftarrow}(U) \leq x) = P(U \leq F(x)) = F(x)$$

by Proposition A.1.3. Hence X has distribution function F. ■

Recall that a right-continuous function with left-limits is called càdlàg (French: "continue à droite, limite à gauche").

Corollary A.2.4. *Every distribution function is càdlàg.*

Proof. A non-decreasing function has left-limits. As a distribution is right-continuous by the above result, the corollary follows. ■

The distribution function of a random variable determines its distribution. Note also the useful identity

$$P(a < X \leq b) = F(b) - F(a)$$

for every $a < b$. This generalises to the two-dimensional case as follows: If (X, Y) has distribution function F, then

$$P(a < X \leq b, c < Y \leq d) = F(a,c) + F(b,d) - F(a,d) - F(b,c). \tag{A.1}$$

This is best seen by making a drawing of the rectangle with coordinates $(a,c), (a,d), (b,c)$ and (b,d). Multivariate distribution functions have similar properties as in the univariate case.

Proposition A.2.5. *Any multivariate distribution function $F : \mathbb{R}^d \to \mathbb{R}$ satisfies the following properties:*

1. *F is non-decreasing in each variable.*

2. *F is right-continuous in each variable.*

3. $\lim_{x_1,...,x_d \to \infty} F(x_1, ..., x_d) = 1$.

4. $0 \leq F(x_1, ..., x_d) \leq 1$.

5. $\lim_{x_i \to -\infty} F(x_1, ..., x_d) = 0$ *for every* $i = 1, ..., d$.

Proof. Left as an exercise for the reader. ■

Note that the above proposition is not an if and only if statement as in the univariate case. A counterexample is given in the exercises. We end this subsection about distribution functions with two tables containing the most important examples for this course.

Distribution	Density	Distribution function	Parameters
Normal, $\mathcal{N}(\mu,\sigma^2)$	$\varphi(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2}$	$\Phi(x)=\int_{-\infty}^{x}\varphi(t)dt$	$(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)$
Exponential, Exp(λ)	$\lambda e^{-\lambda x}, x>0$	$1-e^{-\lambda x}, x>0$	$\lambda\in(0,\infty)$
Gamma, $\Gamma(\alpha,\beta)$	$f(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, x>0$	$\int_0^x f(t)dt, x>0$	$(\alpha,\beta)\in(0,\infty)^2$
Student t	$f(x)=\frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$	$t_\nu(x)=\int_{-\infty}^{x}f(t)dt$	$\nu\in(0,\infty)$
Lognormal(μ,σ^2)	$f(x)=\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-(\log x-\mu)^2/2\sigma^2}, x>0$	$\int_0^x f(t)dt, x>0$	$(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)$
Pareto	$\frac{\alpha\kappa^\alpha}{(\kappa+x)^{\kappa+1}}, x>0$	$1-\left(\frac{\kappa}{\kappa+x}\right)^\alpha, x>0$	$(\alpha,\kappa)\in(0,\infty)^2$

Table A.1: Densities and distribution functions of some common continuous distributions.

Distribution	Density $P(N=k)$	Distribution function	Parameters
Poisson(λ)	$\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,2,...$	$\sum_{i=0}^{k}\frac{\lambda^i}{i!}e^{-\lambda}, k=0,1,2,...$	$\lambda\in(0,\infty)$
Binomial(n,p)	$\binom{n}{k}p^k(1-p)^{n-k}, k=0,1,...,n$	$\sum_{i=0}^{k}\binom{n}{i}p^i(1-p)^{n-i}, k=0,1,...,n$	$n\in\mathbb{N}, p\in[0,1]$
Geometric(p)	$(1-p)^k p, k=0,1,2,...$	$1-(1-p)^{k+1}, k=0,1,2,...$	$p\in[0,1]$
Negative binomial	$\binom{k+r-1}{k}(1-p)^k p^r, k=0,1,2,...$	$\sum_{i=0}^{k}\binom{i+r-1}{i}(1-p)^i p^r, k=0,1,2,...$	$p\in[0,1]$, $r\in\mathbb{N}$

Table A.2: Densities and distribution functions of some common discrete distributions.

Characteristic functions and moment-generating functions

An alternative characterization of distributions is via moment-generating functions and characteristic functions.

Definition A.2.6. For a random variable X, the function

$$\Phi_X(t)=E[e^{itX}]$$

is called the *characteristic function* of X. If there exists a neighbourhood $(-a,a)$ of zero $(a>0)$ such that

$$\kappa_X(t)=E[e^{tX}],\quad t\in(-a,a)$$

is finite, we call κ_X the *moment-generating function* of X.

For the sake of brevity, we will often write cf for characteristic function and mgf for moment-generating function. Note that the characteristic function of a random variable is always defined. Indeed, the integrand is bounded by 1 in norm.

Example A.2.7. For the $\mathcal{N}(0,1)$ distribution, the characteristic function is given by

$$\Phi(t)=e^{-t^2/2}$$

The case for the general normal distribution $\mathcal{N}(\mu,\sigma^2)$ is left as an exercise, see also the lemma below. ○

The cf and mgf are easily generalised to a multivariate random variable $\mathbf{X} = (X_1, ..., X_d)$ as follows:

$$\Phi_{\mathbf{X}}(\mathbf{t}) = E\left[e^{i\mathbf{t}^T\mathbf{X}}\right], \quad \kappa_{\mathbf{X}}(\mathbf{t}) = E\left[e^{\mathbf{t}^T\mathbf{X}}\right], \quad \mathbf{t} \in \mathbb{R}^d$$

where the mgf is only defined in the neighbourhood of the origin where it is finite.

Lemma A.2.8. *Let $\mathbf{X}$ be a $\mathbb{R}^d$-valued random variable, $\mathbf{a} \in \mathbb{R}^n$ and B a $n \times d$ matrix. Then the $\mathbb{R}^n$-valued random variable $\mathbf{a} + B\mathbf{X}$ has cf*

$$\Phi_{\mathbf{a}+B\mathbf{X}}(\mathbf{t}) = e^{i\mathbf{a}^T\mathbf{t}}\Phi_{\mathbf{X}}(B^T\mathbf{t}), \quad \mathbf{t} \in \mathbb{R}^n.$$

Similarly, whenever the mgfs exist,

$$\kappa_{\mathbf{a}+B\mathbf{X}}(\mathbf{t}) = e^{\mathbf{a}^T\mathbf{t}}\kappa_{\mathbf{X}}(B^T\mathbf{t}).$$

Proof. The proof is left to the reader. ■

The cf has the following important properties.

Theorem A.2.9. *If $\mathbf{X}$ and $\mathbf{Y}$ are random variables with the same characteristic functions, $\Phi_{\mathbf{X}} = \Phi_{\mathbf{Y}}$, then $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution.*

Proof. See Theorem 14.1 in [15]. ■

Corollary A.2.10. *The variables $X_1, ..., X_d$ are independent if and only if*

$$\Phi_{\mathbf{X}}(t_1, ..., t_d) = \prod_{i=1}^{d} \Phi_{X_i}(t_i)$$

for all $t_1, ..., t_d$ where $\mathbf{X} = (X_1, ..., X_d)$.

Proof. Assume the $X_1, ..., X_d$ are independent. Then

$$\Phi_{\mathbf{X}}(t_1, ..., t_d) = E\left[e^{i(t_1X_1 + \cdots t_dX_d)}\right] = E\left[e^{it_1X_1}\right] \cdots E\left[e^{it_dX_d}\right] = \prod_{i=1}^{d} \Phi_{X_i}(t_i).$$

Conversely, if the cf factors, it follows immediately from the uniqueness theorem above that the $X_1, ..., X_d$ are independent.

■

Proposition A.2.11. *Let X be a one-dimensional random variable with cf Φ_X. If $E[|X|^k] < \infty$ for some $k \in \mathbb{N}$, then Φ_X is C^k (k times differentiable and the k'th derivative is continuous) and*

$$\Phi_X^{(m)}(0) = i^m E[X^m], \quad m = 1, ..., k.$$

Proof. See Theorem 6.34 in [13] and the paragraph following the theorem. ■

Maybe not surprisingly, these properties more or less carry over to the mgf. A discussion of the result below can be found in [4], chapter 30.

Theorem A.2.12. *If the mgfs of $\mathbf{X}$ and $\mathbf{Y}$ exist in a neighbourhood around zero and are equal, then $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution.*

Corollary A.2.13. *Let the variables $X_1, ..., X_d$ have moment-generating functions $\kappa_{X_1}, ..., \kappa_{X_d}$ that exist in a neighbourhood around zero, then $X_1, ..., X_d$ are independent if and only*

$$\kappa_{(X_1,...,X_d)}(t_1, ..., t_d) = \prod_{i=1}^{d} \kappa_{X_i}(t_i).$$

Proposition A.2.14. *Let X be a one-dimensional random variable with mgf κ_X that exists in a neighbourhood $(-c, c)$ around zero. Then X has moments of all orders and for $k \in \mathbb{N}$,*

$$\kappa_X^{(k)}(0) = E[X^k].$$

We end this subsection with tables containing the mgf and cf of the distributions from the tables of distributions above.

Distribution	cf	mgf	Constraint for mgf
Normal, $\mathcal{N}(\mu, \sigma^2)$	$e^{\mu t - \frac{1}{2}t^2\sigma^2}$	$e^{\mu t + \frac{1}{2}t^2\sigma^2}$	$t \in \mathbb{R}$
Exponential, Exp(λ)	$\frac{\lambda}{\lambda - it}$	$\frac{\lambda}{\lambda - t}$	$t \in (-\infty, \lambda)$
Gamma, $\Gamma(\alpha, \beta)$	$\left(\frac{\beta}{\beta - it}\right)^\alpha$	$\left(\frac{\beta}{\beta - t}\right)^\alpha$	$t \in (-\infty, \beta)$
Student t	No explicit form	Doesn't exist	-
Lognormal(μ, σ^2)	No explicit form	Doesn't exist	-
Pareto	No explicit form	Doesn't exist	-

Table A.3: Characteristic functions and moment-generating functions for the distributions in table A.1.

Distribution	cf	mgf	Constraint for mgf
Poisson(λ)	$e^{\lambda(e^{it}-1)}$	$e^{\lambda(e^t-1)}$	$t \in \mathbb{R}$
Binomial(n, p)	$(pe^{it} + 1 - p)^n$	$(pe^t + 1 - p)^n$	$t \in \mathbb{R}$
Geometric(p)	$\frac{p}{1-(1-p)e^{it}}$	$\frac{p}{1-(1-p)e^t}$	$t < -\log(1-p)$
Negative binomial	$\left(\frac{p}{1-(1-p)e^{it}}\right)^r$	$\left(\frac{p}{1-(1-p)e^t}\right)^r$	$t < -\log(1-p)$

Table A.4: Characteristic functions and moment-generating functions for the distributions in table A.2.

The multivariate normal distribution

Definition A.2.15. An $\mathbb{R}^d$-valued random variable $\mathbf{X} = (X_1, ..., X_d)$ is *multivariate normal* if for every $\mathbf{a} \in \mathbb{R}^d$, the real-valued random variable $\mathbf{a}^T\mathbf{X}$ has a normal distribution.

The definition does not say that being multivariate normal is the same as all marginal variables being normal. A counterexample is provided in the exercises. In order to prove results with the multivariate normal distribution, the following theorem is essential.

Theorem A.2.16. $\mathbf{X}$ *is multivariate normal of dimension d if and only if there exists a symmetric positive semi-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$ and a vector $\mu \in \mathbb{R}^d$ such that*

$$\Phi_{\mathbf{X}}(\mathbf{t}) = e^{i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}}, \quad \mathbf{t} \in \mathbb{R}^d.$$

In this case, Σ is the covariance matrix of $\mathbf{X}$ and μ is the mean vector i.e. $E[X_i] = \mu_i$ and $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ for all $i, j = 1, ..., d$.

Proof. See Theorem 16.1 in [15]. ■

The theorem allows us to define the following.

Definition A.2.17. For a multivariate normal vector $\mathbf{X}$, we write $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ is the mean vector and Σ is the covariance matrix. If Σ is invertible (i.e. $\det \Sigma \neq 0$), we say that $\mathbf{X}$ has a *regular* multivariate normal distribution. Otherwise, $\mathbf{X}$ is called *singular.*

Theorem A.2.18. *A regular multivariate normal variable $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ in $\mathbb{R}^d$ has density*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} e^{-(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \mathbb{R}^d$$

with respect to Lebesgue measure on $\mathbb{R}^d$.

Proof. See Corollary 16.2 in [15]. Note the error in equation (16.5). It should say $(2\pi)^{n/2}$ and not $2\pi^{n/2}$. ■

The following result will be used extensively in the discussion on spherical and elliptical distributions.

Proposition A.2.19. *Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ be d-dimensional, let $\mathbf{a} \in \mathbb{R}^n$ and let B be an $n \times d$-matrix. Then $\mathbf{Y} \sim \mathcal{N}(\mathbf{a} + B\boldsymbol{\mu}, B\Sigma B^T)$.*

Proof. Left as an exercise for the reader ■

Convergence concepts and results

In this subsection, we will briefly touch upon the convergence concepts that we will use in this course.

Definition A.2.20. Let $X, X_1, X_2, ...$ be random variables. We say that the sequence $\{X_n\}$ *converges almost surely* to X for $n \to \infty$ if the event $\{X_n \to X \text{ for } n \to \infty\}$ has probability one. We write

$$P(X_n \to X) = 1.$$

Example A.2.21. Let $X_1, X_2, ...$ be iid Bernoulli distributed with success probability $p \in (0, 1)$ i.e. $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$ for all i. Consider the product process $Y_n = X_1 \cdots X_n$. We claim that $Y_n \to 0$ a.s. Indeed, note that $Y_n \in \{0, 1\}$ a.s. and

$$P(Y_n \to 0) = P(Y_n = 0 \text{ for some } n \in \mathbb{N}) = 1 - P(Y_n = 1 \text{ for all } n \in \mathbb{N}).$$

By independence, we have for any $N \in \mathbb{N}$ that

$$P(Y_N = 1) = P(X_1 = 1, ..., X_N = 1) = p^N$$

and

$$P(Y_n = 1 \text{ for all } n \in \mathbb{N}) \leq P(Y_N = 1) = p^N.$$

As this equality holds for all N, we can take limits on both sides and obtain $P(Y_n = 1 \text{ for all } n \in \mathbb{N}) = 0$ which yields $P(Y_n \to 0) = 1$ as desired. ○

Almost sure convergence is a strong form of convergence. A weaker type of convergence is convergence in probability.

Definition A.2.22. Let $X, X_1, X_2, ...$ be random variables. We say that the sequence $\{X_n\}$ *converges in probability* to X for $n \to \infty$ if for every $\varepsilon > 0$, we have

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0.$$

We write $X_n \xrightarrow{\text{P}} X$.

Lemma A.2.23. *Almost sure convergence implies convergence in probability.*

Proof. This proof is from [13]. Let $\{X_n\}$ be a sequence of random variables converging almost surely to X, and let $\varepsilon > 0$ be given. Consider an ω such that $X_n(\omega) \to X(\omega)$. There exists an $N \in \mathbb{N}$ (depending on ω) such that

$$|X_n(\omega) - X(\omega)| \leq \varepsilon \quad \text{for} \quad n \geq N,$$

implying that $1_{\{|X_n - X| > \varepsilon\}}(\omega) = 0$ for $n \geq N$. It follows that $1_{\{|X_n - X| > \varepsilon\}}(\omega) \to 0$ for $n \to \infty$ and since this holds for almost every ω, we have $1_{\{|X_n - X| > \varepsilon\}} \to 0$ almost surely. As this function is bounded by 1, dominated convergence implies

$$P(|X_n - X| > \varepsilon) = \int 1_{\{|X_n - X| > \varepsilon\}} dP \to 0$$

as desired. ■

It is not immediately clear that almost sure convergence is strictly stronger than convergence in probability. The difference is of a very technical nature. However, counterexamples exist, and we encourage the reader to look them up. See for example [22]. It is useful to have some tools to prove convergence almost surely and in probability. Such tools include the Markov inequality and Chebyshev's inequality.

Lemma A.2.24 (Markov's inequality). *Let X be a random variable. Then for any $\varepsilon > 0$,*

$$P(|X| > \varepsilon) \leq \frac{E[|X|]}{\varepsilon}.$$

Proof. Trivially, $\varepsilon 1_{\{|X| > \varepsilon\}} \leq |X|$. Now take expectations on both sides and rearrange. ■

Corollary A.2.25 (Chebyshev's inequality). *Let X be a random variable with finite expectation, $E[|X|] < \infty$. Then for any $\varepsilon > 0$,*

$$P(|X - E[X]| > \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}$$

Proof. Left as an exercise for the reader. ■

Example A.2.26. Consider a sequence of non-negative variables $X_1, X_2, ...$ with finite expectation and $E[X_n] = 1/n$. For any $\varepsilon > 0$, we have by the Markov inequality that

$$P(|X_n - 0| > \varepsilon) \leq \frac{E[X_n]}{\varepsilon} = \frac{1}{n\varepsilon} \to 0$$

so $X_n \xrightarrow{\text{P}} 0$. This result should not be surprising, considering the fact that a non-negative random variable is zero almost surely if and only if it has mean zero. ○

The above example of an application of the Markov inequality is not exactly interesting, but we want to stress that the inequality, while a triviality, is extremely useful and flexible. The following result shows that almost sure convergence follows if the probability $P(|X_n - X| > \varepsilon)$ goes to zero fast enough.

Proposition A.2.27 (**Borel-Cantelli Criterion for Almost Sure Convergence**). *Let $X, X_1, X_2, ...$ be random variables. If for every $\varepsilon > 0$,*

$$\sum_{n=1}^{\infty} P(|X_n - X| > \varepsilon) < \infty,$$

then $X_n \to X$ almost surely.

Proof. The proof is an immediate consequence of the Borel-Cantelli lemma, see Lemma 2.26 and Theorem 2.27 in [13]. ■

The following result is a cornerstone of probability theory.

Theorem A.2.28 (**Strong Law of Large Numbers**). *Let $\{X_i\}$ be an iid sequence of random variables with $E[|X_1|] < \infty$. Then*

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to E[X_1] \quad a.s.$$

Proof. A proof can be found in [13], see Theorem 4.25. ■

An even weaker form of convergence than convergence in probability is convergence in distribution.

Definition A.2.29. A sequence of random variables $X_1, X_2, ...$ is said to *converge in distribution* to X if for every continuous and bounded function $f : \mathbb{R} \to \mathbb{R}$,

$$\int f(X_n) dP \to \int f(X) dP.$$

We write $X_n \xrightarrow{\mathrm{d}} X$.

Remark A.2.30. Convergence in distribution is also called *weak convergence*.

The above definition is difficult to check in practice. The following results provide much easier ways to check convergence in distribution.

Theorem A.2.31 (**Helly-Bray**). *Let $X, X_1, X_2, ...$ be random variables with distribution functions $F, F_1, F_2, ...$ $X_n \xrightarrow{\mathrm{d}} X$ if and only if there exists a dense subset $A \subseteq \mathbb{R}$ such that $F_n(x) \to F(x)$ for $x \in A$. In this case, A can be chosen to be the set of continuity points of F.*

Proof. See Theorem 18.4 in [15] or Theorem 6.18 in [13]. ■

Proposition A.2.32. *If $X_n \xrightarrow{\mathrm{P}} X$, then $X_n \xrightarrow{\mathrm{d}} X$.*

Proof. See Theorem 18.2 in [15] or Lemma 6.12 in [13]. ■

We now state a version of the Central Limit Theorem, often abbreviated CLT.

Theorem A.2.33 (**Central Limit Theorem**)**.** *Let $\{X_i\}$ be iid with $E[X_1^2] < \infty$, $\mu = E[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1)$. Let $S_n = \sum_{i=1}^n X_i$. Then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathrm{d}} Z \sim \mathcal{N}(0,1).$$

Proof. See Theorem 21.1 in [15]. ■

In this course, this version of the CLT suffices. It is, however, only a little part of the whole story. More perspectives and versions of the CLT can be found in chapter 7 of [13].

Conditional expectations

The presentation in this subsection follows chapter 9 of [13]. Conditional expectations are essential in performing computations in probability theory and statistics. The (measure theoretic) definition of a conditional expectation is somewhat strange at first, but the definition has the advantage that all the theoretic properties follow almost trivially.

Definition A.2.34. Let X be a random variable on $(\Omega, \mathcal{F}, P)$, $E[|X|] < \infty$ and $\mathcal{G} \subseteq \mathcal{F}$ a sub-sigma-algebra. The conditional expectation of X with respect to $\mathcal{G}$, denoted by $E[X \mid \mathcal{G}]$, is a random variable satisfying the following properties:

(i) $E[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable.

(ii) For every $A \in \mathcal{G}$,

$$\int_A X dP = \int_A E[X \mid \mathcal{G}] dP.$$

Intuitively, we think of the conditional expectation $E[X \mid \mathcal{G}]$ as our best guess of the value of X given the information in $\mathcal{G}$. Try to keep this intuition in mind when reading the following examples and theoretical properties.

It is by no means trivial that the conditional expectation exists. An elegant construction is via the Radon-Nikodym theorem, see [13] chapter 8 and Theorem 9.1. We also remark that $E[X \mid \mathcal{G}]$ is only unique almost surely. To verify theoretical statements concerning $E[X \mid \mathcal{G}]$, it suffices to verify the two properties above. If another variable Z satisfies the above assumptions, we have $E[X \mid \mathcal{G}] = Z$ a.s. In the following, we will omit writing a.s. when considering computations involving conditional expectations. Also, if $\mathcal{G} = \sigma(Y)$ is the smallest sigma-algebra making Y measurable (intuitively, the information Y contains), we will write $E[X \mid Y]$ instead of $E[X \mid \sigma(Y)]$.

Example A.2.35. Assume X is $\mathcal{G}$-measurable. We claim that $X = E[X \mid \mathcal{G}]$. X satisfies (i) by assumption and for $A \in \mathcal{G}$, we have

$$\int_A X dP = \int_A E[X \mid \mathcal{G}] dP$$

by definition of $E[X \mid \mathcal{G}]$, verifying (ii). ○

Example A.2.36. Assume X is independent of $\mathcal{G}$ i.e. $P(A \cap \{X \in B\}) = P(A)P(X \in B)$ for all Borel sets B and $A \in \mathcal{G}$. We claim that $E[X \mid \mathcal{G}] = E[X]$. $E[X]$ is constant and thus trivially $\mathcal{G}$-measurable. Also, we get for $A \in \mathcal{G}$ that

$$\int_A X dP = E[1_A X] = E[1_A]E[X] = P(A)E[X] = \int_A E[X] dP$$

so both (i) and (ii) are satisfied by $E[X]$, proving the claim. ○

The following proposition allows us to compute a plethora of interesting examples.

Proposition A.2.37. *If $D_1, D_2, ...$ are disjoint sets in $\mathcal{F}$ with $\cup_n D_n = \Omega$ (such a collection is called a* partition*), $P(D_i) > 0$ for all i, $\mathcal{G} = \sigma(D_1, D_2, ...)$ and X is an integrable random variable, then*

$$E[X \mid \mathcal{G}](\omega) = \begin{cases} \frac{1}{P(D_1)} \int_{D_1} X dP, & \omega \in D_1 \\ \frac{1}{P(D_2)} \int_{D_2} X dP, & \omega \in D_2 \\ \vdots \end{cases}$$

Proof. It is not hard to verify that the sigma-algebra $\mathcal{G}$ consists of the sets that are unions of the D_i. Since $E[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable, $E[X \mid \mathcal{G}]$ must be constant when restricted to one of the D_i i.e.

$$E[X \mid \mathcal{G}](\omega) = \begin{cases} c_1, & \omega \in D_1 \\ c_2, & \omega \in D_2 \\ \vdots \end{cases}.$$

Since

$$\int_{D_i} X dP = \int_{D_i} E[X \mid \mathcal{G}] dP = \int_{D_i} c_i dP = c_i P(D_i),$$

the claim follows. ■

Corollary A.2.38. *Let N be a random variable with $N \in \{0, 1, 2, ...\}$. If X is an integrable random variable, then*

$$E[X \mid N] = \begin{cases} \frac{1}{P(N=0)} \int_{\{N=0\}} X dP & \text{on} \quad \{N = 0\} \\ \frac{1}{P(N=1)} \int_{\{N=1\}} X dP & \text{on} \quad \{N = 1\} \\ \vdots \end{cases}$$

Proof. This follows immediately from the previous proposition by letting $\mathcal{G} = \sigma(N)$ and noting that $\sigma(N)$ is generated by the partition $\{\{N = 0\}, \{N = 1\}, ...\}$. ■

Before computing some interesting examples, we state the following list of properties of conditional expectations.

Theorem A.2.39. *Let X and Y be random variables with finite expectation, $\mathcal{G} \subseteq \mathcal{F}$ a sub-sigma-algebra.*

(i) For $a, b \in \mathbb{R}$, $E[aX + bY \mid \mathcal{G}] = aE[X \mid \mathcal{G}] + bE[Y \mid \mathcal{G}]$ (linearity).

(ii) $E[X] = E[E[X \mid \mathcal{G}]]$ (tower property).

(iii) If $X \leq Y$ then $E[X \mid \mathcal{G}] \leq E[Y \mid \mathcal{G}]$ (monotonicity).

(iv) $|E[X \mid \mathcal{G}]| \leq E[|X| \mid \mathcal{G}]$ (triangle inequality).

(v) If X is $\mathcal{G}$-measurable, $E[XY \mid \mathcal{G}] = X[Y \mid \mathcal{G}]$.

Proof. Consider (i). We have to show that $aE[X \mid \mathcal{G}]+bE[X \mid \mathcal{G}]$ satisfies the two properties of $[aX + bY \mid \mathcal{G}]$. Measurability is obvious. For $A \in \mathcal{G}$, we have

$$\begin{aligned}\int_A aE[X \mid \mathcal{G}] + bE[Y \mid \mathcal{G}]dP &= a\int_A E[X \mid \mathcal{G}]dP + b\int_A E[Y \mid \mathcal{G}]dP \\ &= a\int_A XdP + b\int_A YdP \\ &= \int_A aX + bYdP\end{aligned}$$

which proves the claim. As for (ii), simply choose $A = \Omega$ in the second property of conditional expectations to obtain

$$E[X] = \int_\Omega XdP = \int_\Omega E[X \mid \mathcal{G}]dP = E[E[X \mid \mathcal{G}]].$$

(iii) follows immediately from the monotonicity property of integrals. As for (iv), we obviously have $-|X| \leq X \leq |X|$ so from (iii), we get

$$-E[|X| \mid \mathcal{G}] \leq E[X \mid \mathcal{G}] \leq E[|X| \mid \mathcal{G}]$$

which is the desired result. See the exercises for an outline of the proof of (v). ■

The tower property, (ii), has many names. It is also known as the *law of iterated expectations* and the *law of total expectation* to name a few.

Example A.2.40. A typical situation in for example non-life insurance is to have a sum of the form

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

where $\{X_i\}$ is an iid sequence independent of N, a random variable taking values in $\{0, 1, 2, ...\}$. Assume both X_1 and N have finite expectation. What is the expectation of S? Using Corollary A.2.38, we have[1]

$$\begin{aligned}\frac{1}{P(N=n)}\int_{\{N=n\}} SdP &= \frac{1}{P(N=n)}\int_{\{N=n\}} \sum_{i=1}^{n} X_i dP \\ &= \frac{1}{P(N=n)} E[1_{\{N=n\}}]E\left[\sum_{i=1}^{n} X_i\right] \\ &= nE[X_1]\end{aligned}$$

[1] We should in principle first verify that $E[|S|] < \infty$.

using that N and $\{X_i\}$ are independent. It follows that $E[S \mid N] = NE[X_1]$. From the tower property, it follows that

$$E[S] = E[E[S \mid N]] = E[NE[X_1]] = E[N]E[X_1].$$

$\circ$

In practice, one does not proceed as formally as in the above example. The corollary that we applied essentially says that when we condition on a variable, that variable can be treated as a constant. If N was constant equal to n, we would say that $E[S] = nE[X_1]$. Then we just replace n by the random variable N to obtain $E[S \mid N]$. Let us make this more precise by first observing that $E[X \mid Y]$ is a function of Y.

Theorem A.2.41 (Doob-Dynkin lemma). *If a random variable Z is $\sigma(Y)$-measurable, then there exists a measurable function ϕ such that $Z = \phi(Y)$.*

Proof. See Theorem 9.23 in [13]. ■

$$\begin{tikzcd}
\Omega \arrow[r, "Y"] \arrow[dr, "Z"'] & \mathbb{R} \arrow[d, "\phi"] \\
& \mathbb{R}
\end{tikzcd}$$

By definition of a conditional expectation, $E[X \mid Y]$ is $\sigma(Y)$-measurable. Hence $E[X \mid Y] = \phi(Y)$ for some function ϕ. While the Doob-Dynkin lemma does not provide an explicit recipe for ϕ, it is possible to compute ϕ in many situations of interest. The following result shows how to compute $\phi(y) = E[X \mid Y = y]$ in the (quite typical) case where (X, Y) has a density.

Theorem A.2.42. *Let $E[|X|] < \infty$ and assume (X, Y) has density $f(x, y)$. Then*

$$E[X \mid Y = y] = \int_{\mathbb{R}} x \frac{f(x, y)}{g(y)} dx$$

where $g(y) = \int_{\mathbb{R}} f(x, y) dx$ is the density of Y.

Proof. This is Corollary 9.28 in [13]. ■

Remark A.2.43. By recalling that the conditional density of X given $Y = y$ is defined by

$$f_{X|Y=y}(x) = \frac{f(x, y)}{g(y)},$$

we could also write the above result as

$$E[X \mid Y = y] = \int_{\mathbb{R}} x f_{X|Y=y}(x) dx$$

which also makes sense intuitively. Given densities, the conditional expectation can be computed as an ordinary expectation but with a conditional density.

Stochastic processes (including Brownian motions)

In this subsection, we provide a very brief review of stochastic processes in continuous time.

Definition A.2.44. A *(continuous time) stochastic process* is a collection of random variables $\{X_t\}$ indexed by $t \in [0, \infty)$.

Definition A.2.45 (Brownian motion). A stochastic process $\{X_t\}$ which satisfies the properties

- $X_0 = 0$,
- $X_t - X_s \sim \mathcal{N}(0, t-s)$ for all $0 \leq s < t$ and
- $X_{t_1}, X_{t_2} - X_{t_1}, ..., X_{t_n} - X_{t_{n-1}}$ are independent for $0 < t_1 < t_2 < \cdots < t_n$

is called a *(standard) Brownian motion.*

To model a flow of information in continuous time, we need the notion of a filtration.

Definition A.2.46. A filtration is a sequence $\{\mathcal{F}_t\}$ of sigma-algebras indexed by $t \in [0, \infty)$ such that $s \leq t$ implies $\mathcal{F}_s \subseteq \mathcal{F}_t$. We have $\mathcal{F}_0 = \{\emptyset, \Omega\}$ by convention.

Definition A.2.47. A stochastic process $\{X_t\}$ is called *adapted* to the filtration $\{\mathcal{F}_t\}$ if X_t is $\mathcal{F}_t$-measurable for all t.

Example A.2.48. For any stochastic process $\{X_t\}$, we can create a filtration by letting $\mathcal{F}_t = \sigma(X_s : s \leq t)$. This is the smallest filtration such that $\{X_t\}$ is adapted. The filtration is sometimes called the *natural filtration.* ◦

A particular nice type of stochastic process is a martingale.

Definition A.2.49. A stochastic process $\{X_t\}$ is called a *martingale* with respect to the filtration $\{\mathcal{F}_t\}$ if the following hold:

- $\{X_t\}$ is adapted to $\{\mathcal{F}_t\}$.
- $E[|X_t|] < \infty$ for each t.
- For every $s \leq t$, $E[X_t \mid \mathcal{F}_s] = X_s$.

Martingales do not play a major role in this course. Nevertheless, some examples are provided in the exercises.

Exercises

Exercise A.3:
If $Y \sim \mathcal{N}(\mu, \sigma^2)$, we say that $X = \exp(Y)$ has a lognormal distribution with parameters μ and σ. Using the density

$$\varphi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for the normal distribution, derive the density of the lognormal distribution.

Exercise A.4:
Prove Proposition A.2.5.

Exercise A.5:
Consider the function $F : \mathbb{R}^2 \to \mathbb{R}$ given by

$$F(x,y) = \begin{cases} 0, & x < 0 \text{ or } y < 0 \text{ or } x + y < 1 \\ 1, & \text{else} \end{cases}.$$

1)Verify that F satisfies all the properties in Proposition A.2.5.

2)Show that F cannot be a distribution function for a pair of random variables (X, Y). Hint: Use equation (A.1). Now consider $a = c = 1/3, b = d = 1$.

Exercise A.6:
Let Y be $\mathcal{N}(0,1)$ and let Z be Bernoulli distributed with success parameter $1/2$. Assume Y and Z are independent. Define $X_1 := Y$ and $X_2 := 1_{\{Z=1\}}Y - 1_{\{Z=0\}}Y$.

1)Verify that X_1 and X_2 are both $\mathcal{N}(0,1)$ variabels.

2)Show that (X_1, X_2) is not multivariate normal.

Exercise A.7:
Compute the moment-generating function for the $\Gamma(\alpha, \beta)$ distribution.

Exercise A.8:

1)Compute the moment-generating function for the Bernoulli distribution i.e. $P(X = 1) = p$ and $P(X = 0) = 1 - p$ for $p \in [0,1]$.

2)Compute the moment-generating function for the Binomial(n, p) distribution. Hint: Use the previous exercise.

3)Compute the moment-generating function for the Geometric(p) distribution.

Exercise A.9:

1)Prove Lemma A.2.8.

We know that the standard normal distribution has the characteristic function

$$\Phi(t) = e^{-t^2/2}.$$

2)Compute the characteristic function for the $\mathcal{N}(\mu, \sigma^2)$ distribution.

Exercise A.10:
Compute all moments of the exponential distribution.

Exercise A.11:
The $\Gamma(\lambda, n)$ distribution for $n \in \mathbb{N}$ is called the *Erlang distribution.* Verify that if $X \sim \Gamma(\lambda, n)$ then

$$X \overset{\mathrm{d}}{=} Y_1 + \cdots + Y_n$$

with $Y_1, ..., Y_n$ iid exponential distributed with parameter λ.

Exercise A.12:
Prove Proposition A.2.19.

Exercise A.13:
Prove Chebyshev's inequality, Corollary A.2.25.

Exercise A.14:
Without using the Strong Law of Large Numbers, prove the *Weak Law of Large Numbers*: If $\{X_i\}$ is an iid sequence of random variables with $E[X_1^2] < \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{P}} E[X_1].$$

Exercise A.15:
Assume $X_n \to X$ a.s. and that f is a continous function. Prove that $f(X_n) \to f(X)$ a.s.

Exercise A.16:
Let $\{X_i\}$ be iid variables with $E[X_1^2] = 4$ and $E[X_1] = 1$. Show that

$$\lim_{n\to\infty} \frac{X_1^2 + \cdots X_n^2}{X_1 + \cdots X_n}$$

exists a.s. and determine the value.

Exercise A.17:
Let $p \geq 1$. A sequence of random variables $\{X_i\}$ with $E[|X_i|^p] < \infty$ for all i is said to converge to X in L^p if $E[|X|^p] < \infty$ and

$$\lim_{n\to\infty} E[|X_n - X|^p] = 0.$$

In that case, we write $X_n \xrightarrow{L^p} X$.

1)Prove that if $X_n \xrightarrow{L^p} X$, then $X_n \xrightarrow{\text{P}} X$.

2)Let $\{X_i\}$ be a sequence of random variables with $E[X_i] = 0$ and $E[X_i^2] < \infty$ for all i. Prove that $(X_1 + \cdots + X_n)/n$ converges to zero in L^2 and in probability.

3)Prove the following dominated convergence statement: If $X_n \to X$ a.s. and there exists some variable Y with $E[|Y|^p] < \infty$ $(p \geq 1)$ such that $|X_n| \leq |Y|$ a.s. for all n, then $X_n \xrightarrow{L^p} X$.

Exercise A.18:
Assume X is a random variable with finite moment-generating function κ in the neighbourhood $(-c, c)$. Prove *Chernoff's bound*

$$P(X > \varepsilon) \leq \inf_{\alpha\in[0,c]} \frac{\kappa(\alpha)}{e^{\alpha\varepsilon}}.$$

Exercise A.19:
Assume N is Poisson distributed with parameter $\lambda > 0$ and $\{X_i\}$ is an iid sequence independent of N where X_1 has moment-generating function κ. Compute the moment-generating function of

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

Hint: Tower property.

Exercise A.20:
In this exercise, we will prove (v) in Theorem A.2.39.

1)Prove the result when $X = 1_{A_0}$ is an indicator function.

2)Prove the result when X is a simple function, e.g. $X = \sum_{i=1}^{n} c_i 1_{A_i}$, $A_1, ..., A_n \in \mathcal{G}$.

3)One can show the following dominated convergence statement for conditional expectations: If $X_n \to X$ a.s. and $|X_n| \leq Y$ for a random variable Y with $E[|Y|] < \infty$, then $E[X_n \mid \mathcal{G}] \to E[X \mid \mathcal{G}]$. Using this result, prove (v) for a general $\mathcal{G}$-measurable X. Hint: Recall that there exists a sequence of simple functions $\{X_n\}$ with $X_n \uparrow X$.

Exercise A.21:
Prove the following extension of the tower property: If $\mathcal{G} \subseteq \mathcal{H} \subseteq \mathcal{F}$, then

$$E[E[X \mid \mathcal{H}] \mid \mathcal{G}] = E[X \mid \mathcal{G}].$$

Exercise A.22:
In this exercise, we introduce the *conditional variance.* Assume $E[X^2] < \infty$ and that $\mathcal{G}$ is a sub-sigma-algebra, then the conditional variance is defined by

$$\mathrm{Var}(X \mid \mathcal{G}) = E[X^2 \mid \mathcal{G}] - E[X \mid \mathcal{G}]^2.$$

1)Prove the *law of total variance*,

$$\mathrm{Var}(X) = E[\mathrm{Var}(X \mid \mathcal{G})] + \mathrm{Var}(E[X \mid \mathcal{G}]).$$

2)Assume X is $\mathcal{G}$-measurable. Prove that $\mathrm{Var}(X \mid \mathcal{G}) = 0$.

3)Assume Y is $\mathcal{G}$-measurable. Prove that $\mathrm{Var}(X + Y \mid \mathcal{G}) = \mathrm{Var}(X \mid \mathcal{G})$.

4)Assume X is independent of $\mathcal{G}$. Prove that $\mathrm{Var}(X \mid \mathcal{G}) = \mathrm{Var}(X)$.

5)Give an intuitive interpretation of the previous three subproblems.

Exercise A.23:
Consider Example A.2.40. Assume that the X_i have finite second moment. Show that

$$\mathrm{Var}(S) = E[N]\,\mathrm{Var}(X_1) + \mathrm{Var}(N)E[X_1]^2.$$

Hint: Use the law of total variance from the previous exercise.

Exercise A.24:
Let $\{X_t\}$ be a Brownian motion

1)Define $Y_t = -X_t$. Show that $\{Y_t\}$ is a Brownian motion.

2)Let $c > 0$ and define $Y_t = cX_{t/c^2}$. Show that $\{Y_t\}$ is a Brownian motion.

Exercise A.25:
A continuous time process $\{X_t\}$ satisfies *continuity in probability* if for every sequence $\{t_n\}$ of non-negative real numbers, we have

$$t_n \to t \quad \Rightarrow \quad X_{t_n} \xrightarrow{\text{P}} X_t.$$

Show that a Brownian motion satisfies continuity in probability.

Exercise A.26:
Let $\{X_t\}$ be a Brownian motion and $\mathcal{F}_t = \sigma(X_s : s \leq t)$ the natural filtration.

1)Show that $\{X_t\}$ is a Brownian motion with respect to $\{\mathcal{F}_t\}$.

2)Show that $\{X_t^2 - t\}$ is a Brownian motion with respect to $\{\mathcal{F}_t\}$.

Exercise A.27:
A stochastic process $\{N_t\}$ satisfying the properties

- $N_0 = 0$,
- $N_t - N_s$ is Poisson distributed with parameter $\lambda(t-s)$,
- $N_{t_1}, N_{t_2} - N_{t_1}, ..., N_{t_n} - N_{t_{n-1}}$ are independent for $0 < t_1 < t_2 < \cdots < t_n$,
- $\{N_t\}$ has right-continuous sample paths and limits from the left,

then $\{N_t\}$ is called a *Poisson process* with *intensity* $\lambda > 0$. Verify that $\{N_t - \lambda t\}$ is a martingale with respect to the natural filtration.

A.3 Calculus

During the course, we will occasionally integrate with respect to functions of bounded variation. Examples of such functions include functions that are monotone, in particular distribution functions. Here we introduce the basic theory, following the lines of [20].

Definition A.3.1. Let $f : [0, \infty) \to \mathbb{R}$ be a function. The *variation* of f on the interval $[0, t]$ is given by

$$V^f(t) = \sup\left\{\sum_{i=1}^{n} |f(t_i) - f(t_{i-1})| : 0 = t_0 < t_1 < \cdots < t_n = t\right\}$$

i.e. the supremum of sums of absolute differences over all finite partitions of $[0, t]$. If $V^f(t) < \infty$ for all $t \geq 0$, we call f a *function of finite variation.*

Example A.3.2. Let $f : [0, \infty) \to \mathbb{R}$ be monotone. We claim that f is of finite variation. If f is non-decreasing, this follows immediately from the fact that if $0 = t_0 < t_1 < \cdots < t_n = t$ is a partition of $[0, t]$, we have

$$\sum_{i=1}^{n} |f(t_i) - f(t_{i-1})| = \sum_{i=1}^{n} f(t_i) - f(t_{i-1}) = f(t) - f(0)$$

by telescoping. Hence $V^f(t) = f(t) - f(0)$ and f is of finite variation. An analogous argument works for the case where f is non-increasing. ◦

Recall that for $x \in \mathbb{R}$, $x^+ = \max\{x, 0\}$ and $x^- = -\min\{x, 0\}$. It is easily seen that $x = x^+ - x^-$.

Definition A.3.3. For a function $f : [0, \infty) \to \mathbb{R}$, we define the *positive variation* by

$$V_+^f(t) = \sup\left\{\sum_{i=1}^{n}(f(t_i) - f(t_{i-1}))^+ : 0 = t_0 < t_1 < \cdots < t_n = t\right\}$$

and the *negative variation* by

$$V_-^f(t) = \sup\left\{\sum_{i=1}^{n}(f(t_i) - f(t_{i-1}))^- : 0 = t_0 < t_1 < \cdots < t_n = t\right\}.$$

Proposition A.3.4. *Let $f, g : [0, \infty) \to \mathbb{R}$ be functions of finite variation.*

(i) V^f, V_+^f and V_-^f are non-decreasing.

(ii) $af + bg$ is a function of finite variation for any $a, b \in \mathbb{R}$.

Proof. Left as an exercise for the reader. ■

The motivation for introducing the negative and positive variation is the following central result.

Theorem A.3.5 (Jordan decomposition). *A function $f : [0, \infty) \to \mathbb{R}$ is of finite variation if and only if f can be written as the difference of two non-decreasing functions. A possible decomposition is $f(t) - f(0) = V_+^f(t) - V_-^f(t)$.*

Proof. Assume $f = g - h$ for non-decreasing functions g and h. g and h are of finite variation as shown in the example above, and the previous proposition now implies that f is of finite variation. Conversely, assume f is of finite variation. For any partition $0 = t_0 < t_1 < \cdots < t_n = t$ of $[0, t]$, we have

$$f(t) - f(0) = \sum_{i=1}^{n}(f(t_i) - f(t_{i-1}))^+ - \sum_{i=1}^{n}(f(t_i) - f(t_{i-1}))^-$$

i.e.

$$\sum_{i=1}^{n}(f(t_i) - f(t_{i-1}))^+ = \sum_{i=1}^{n}(f(t_i) - f(t_{i-1}))^- + f(t) - f(0)$$

and taking supremum on both sides, we obtain the decomposition $f(t) - f(0) = V_+^f(t) - V_-^f(t)$, which is a difference of two non-decreasing functions as desired. ■

Remark A.3.6. Note that V^f has the decomposition $V^f = V^f_+ + V^f_-$ since for any $x \in \mathbb{R}$, $|x| = x^+ + x^-$.

We are almost ready to introduce integration. However, finite variation is not quite enough. We also require the additional property of being càdlàg.

Definition A.3.7. A function f is called càdlàg (French: continue à droite, limite à gauche) if f is right-continuous and has left limits.

The Jordan decomposition tells us how to proceed from here. If we define integration with respect to an increasing function, we can use linearity of the integral to define integration with respect to general functions of bounded variation. Let f be a non-decreasing càdlàg function. The function μ^f defined on the intervals $(a, b]$ given by $\mu^f((a, b]) = f(b) - f(a)$ extends to a (positive) measure (called a *Lebesgue-Stieltjes* measure) on all Borel sets. Note how this resembles the Lebesgue measure where f is just the identity. We can now define integration in the same way as in basic measure theory.

Definition A.3.8. Let $f : [0, \infty) \to \mathbb{R}$ be a non-decreasing càdlàg function. The *Lebesgue-Stieltjes integral* of a measurable function g with respect to f is given by

$$\int_{(0,\infty)} g(t)df(t) := \int_0^\infty g(t)d\mu^f(t)$$

given that $\int_0^\infty |g(t)|d\mu^f(t) < \infty$. For any Borel set B, we define

$$\int_B g(t)df(t) := \int_{(0,\infty)} 1_B(t)g(t)df(t).$$

If f is a càdlàg function of bounded variation, we have the Jordan decomposition $f(t) - f(0) = V^f_+(t) - V^f_-(t)$, and we define the Lebesgue-Stieltjes integral of g by

$$\int_0^\infty g(t)df(t) = \int_0^\infty g(t)dV_+(t) - \int_0^\infty g(t)dV_-(t).$$

The integral is well-defined whenever

$$\int_0^\infty |g(t)|dV^f(t) = \int_0^\infty |g(t)|dV^f_+(t) + \int_0^\infty |g(t)|dV^f_-(t) < \infty.$$

Example A.3.9. Let F be the distribution function for the random variable X. If $B = (a, b]$, then $P(X \in B) = F(b) - F(a) = \int_B dF$. Since the intervals $(a, b]$ generate the Borel sigma-algebra, we have $P(X \in B) = \int_B dF$ for all Borel sets B. ∘

Example A.3.10. Consider a positive random variable X with finite expectation and distribution function F. We claim that

$$E[X] = \int_0^\infty x dF(x).$$

Since $P(X \in (a, b]) = F(b) - F(a)$ for all $a < b$, the image measure of X, P^X, coincides with the measure induced by F. Hence

$$E[X] = \int_0^\infty x dP^X(x) = \int x dF(x)$$

as desired. ∘

We now go through some important properties of the Lebesgue-Stieltjes integral.

Proposition A.3.11. *Let f be a càdlàg function of finite variation. Assume all integrals below are well-defined.*

(i)

$$\int_{(s,t]} df(u) = f(t) - f(s).$$

(ii)

$$\int_{\{t\}} g(u)df(u) = g(u)\Delta f(t), \quad \Delta f(t) := f(t) - f(t-)$$

with $f(t-) = \lim_{s\uparrow t} f(s)$ the limit from the left.

(iii)

$$\int_{(s,t]} g(u)df(u) = 0$$

if f is constant on $(s,t]$.

Proof. Left as an exercise for the reader. ■

These properties will allow us to compute integrals with respect to functions that are piecewise constant. An important example is the empirical distribution function as we shall see in the first weeks of the course. We end this subsection (and the appendix) with the following key result.

Theorem A.3.12 (Integration by parts). *Let f and g be càdlàg functions of finite variation. Then (assuming all integrals are well-defined)*

$$f(t)g(t) - f(0)g(0) = \int_{(0,t]} g(s)df(x) + \int_{(0,t]} f(s-)dg(s).$$

Proof. Note first that

$$f(t)g(t) - f(0)g(0) - f(0)(g(t) - g(0)) - g(0)(f(t) - f(0)) = (f(t) - f(0))(g(t) - g(0)).$$

The result is now a direct consequence of the following computation based on Fubini's theorem:

$$\begin{aligned}(f(t) - f(0))(g(t) - g(0)) &= \int_{(0,t]}\int_{(0,t]} df(u)dg(s) = \int_{(0,t]}\int_{(0,s]} df(u)dg(s) + \int_{(0,t]}\int_{(s,t]} df(u)dg(s) \\ &= \int_{(0,t]} (f(s) - f(0))dg(s) + \int_{(0,t]}\int_{(0,u)} dg(s)df(u) \\ &= \int_{(0,t]} f(s)dg(s) - f(0)(g(t) - g(0)) + \int_{(0,t]} g(u-) - g(0)df(u) \\ &= \int_{(0,t]} f(s)dg(s) - f(0)(g(t) - g(0)) + \int_{(0,t]} g(u-)df(u) - g(0)(f(t) - f(0)).\end{aligned}$$

■

Remark A.3.13. It is not difficult to see that the result also works for other intervals such as $(a,b]$, (t,∞) etc.

Exercises

Exercise A.28:
Prove Proposition A.3.4.

Exercise A.29:
Prove Proposition A.3.11.

Exercise A.30:
In this exercise, we will consider some classes of functions of bounded variation.

1)Let $f : [0, \infty) \to \mathbb{R}$ be a function which is Lipschitz on every compact interval i.e. for every $[a, b] \subseteq [0, \infty)$, there exists a constant C such that

$$|f(s) - f(u)| \leq C|s - u|, \quad s, u \in [a, b].$$

Prove that f is of finite variation.

2)Let $f : [0, \infty) \to \mathbb{R}$ be C^1. Prove that f is Lipschitz on every compact interval and conclude that f is of bounded variation.

Exercise A.31:
Compute the integral $\int_{(0,4]} t df(t)$ in the following cases:

1)$f(t) = k$ for $k - 1 \leq t \leq k$, $k = 1, 2, ...$

2)$f(t) = e^t$.

3)$f(t) = k + e^t$ for $k - 1 \leq t \leq k$, $k = 1, 2, ...$

Bibliography

[1] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999. doi: https://doi.org/10.1111/1467-9965.00068.

[2] Søren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Stochastic Modelling and Applied Probability. Springer New York, NY, 1 edition, 2007. ISBN 978-0-387-30679-7.

[3] Sheldon Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer Cham, 4 edition, 2023. ISBN 978-3-031-41025-3.

[4] Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., anniversary edition, 2012. ISBN 978-1-118-12237-2.

[5] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1987. doi: 10.1017/CBO9780511721434.

[6] Tomas Björk. *Arbitrage Theory in Continuous Time*. Oxford University Press, 4 edition, 2020. ISBN 978–0–19–885161–5.

[7] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, (31):307–327, 1986. URL `https://public.econ.duke.edu/~boller/Published_Papers/joe_86.pdf`.

[8] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. URL `http://www.jstor.org/stable/2958830`.

[9] Paul Embrechts and Marius Hofert. A note on generalized inverses. 2014. URL `https://people.math.ethz.ch/~embrecht/ftp/generalized_inverse.pdf`.

[10] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*. Stochastic Modelling and Applied Probability. Springer Berlin, Heidelberg, 1 edition, 1997. ISBN 978-3-540-60931-5.

[11] Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1912773`.

[12] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Stochastic Modelling and Applied Probability. Springer New York, NY, 1 edition, 2003. ISBN 978-0-387-00451-8.

[13] Ernst Hansen. *Stochastic Processes.* Institut for Matematiske Fag Københavns Universitet, 4 edition, 2023. ISBN 978-87-71252-59-0.

[14] Henrik Hult and Filip Lindskog. *Mathematical Modeling and Statistical Methods for Risk Management.* 2007.

[15] Jean Jacod and Philip Protter. *Probability Essentials.* Universitext. Springer Berlin, Heidelberg, 2 edition, 2002. ISBN 978-3-642-55682-1.

[16] Rasmus Frigaard Lemvig. *Stochastic Processes in Non-Life Insurance (SkadeStok) Lecture Notes.* URL `https://rasmusfl.github.io/projects.html`.

[17] Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management - Concepts, Techniques and Tools.* Princeton University Press, revised edition, 2015. ISBN 978-0-691-16627-8.

[18] Thomas Mikosch. *Non-Life Insurance Mathematics.* Universitext. Springer Berlin, Heidelberg, 2 edition, 2009. ISBN 978-3-540-88232-9.

[19] Roger B. Nelsen. *An Introduction to Copulas.* Springer Series in Statistics. Springer New York, NY, 2 edition, 2006. ISBN 978-0-387-28659-4.

[20] Jesper Lund Pedersen. *Stochastic Processes in Life Insurance: The Dynamic Approach.* Department of Mathematical Sciences University of Copenhagen.

[21] Sidney I. Resnick. *Extreme Values, Regular Variation and Point Processes.* Springer Series in Operations Research and Financial Engineering. Springer New York, NY, 1 edition, 2007. ISBN 978-0-387-75952-4.

[22] Jordan M. Stoyanov. *Counterexamples in Probability.* Dover Books on Mathematics. Dover Publications Inc., 3 edition, 2013. ISBN 978-0-486-49998-7.

Index