

Quantitative Risk Management

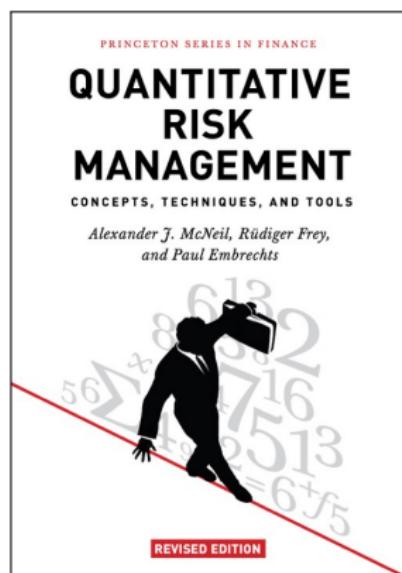
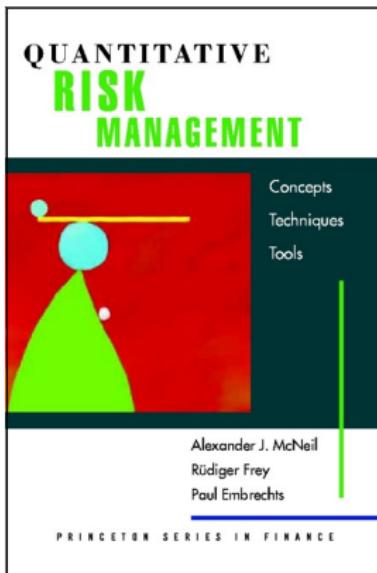
<http://www.qrmtutorial.org>

Last update: 2015-07-27

P. Embrechts, R. Frey, M. Hofert, A. J. McNeil

Course information

- Website: <http://www.qrmtutorial.org>
- Book: A. J. McNeil, R. Frey, P. Embrechts
Quantitative Risk Management (1st edition: 2005; revised edition: 2015)



Overview

- 1 Risk in perspective**
- 2 Basics concepts in risk management**
- 3 Empirical properties of financial data**
- 4 Financial time series**
- 5 Extreme value theory**
- 6 Multivariate models**
- 7 Copulas and dependence**

1 Risk in perspective

- 1.1 Risk
- 1.2 A brief history of risk management
- 1.3 The regulatory framework
- 1.4 Why manage financial risk?
- 1.5 Quantitative Risk Management

1.1 Risk

- The Concise Oxford English Dictionary: “hazard, a chance of bad consequences, loss or exposure to mischance”
- McNeil et al. (2005): “any event or action that may adversely affect an organization’s ability to achieve its objectives and execute its strategies”
- No single one-sentence definition captures all aspects of risk (risk means different things to different people).

For us: *risk = potential/chance for loss* \Rightarrow uncertainty \Rightarrow randomness

1.1.1 Risk and randomness

- To put this on solid ground, Kolmogorov (1933) introduced the notion of a *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$, where
 - ▶ Ω is a *sample space* and $\omega \in \Omega$ represents a realization of an experiment (“state of nature”);

- ▶ the σ -algebra \mathcal{F} contains all sets to which we can assign probabilities (“events”); and
 - ▶ $\mathbb{P}(\cdot)$ denotes a *probability measure*.
- We will mostly model situations in which an investor *holds today* an *asset* with an *uncertain future value*. To this end, we model the value of the asset/risky position as a *random variable* $X : \Omega \rightarrow \mathbb{R}$. Several risky positions are modeled by a *random vector* $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$.
 - Most of this modeling concerns the *distribution functions* $F_X(x) = \mathbb{P}(X \leq x)$ and $F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x})$ of X and \mathbf{X} , respectively.
 - If time matters, one can consider sequences of random variables $(\mathbf{X}_t)_{t \geq 0}$, so-called *stochastic processes*.
 - Our modeling tools will mainly come from *probability* and *statistics* (so *stochastics*; Greek “Stoхastikos” = art of guessing, skilled at aiming)

1.1.2 Financial Risk

There are various types of risks. We focus on (statistical methods for)

Market risk

The risk of losses in financial positions due to changes in the underlying components (e.g., stock and bond¹ prices, exchange rates, commodity prices)

Credit risk

The risk of a counterparty failing to meet its obligations (default), i.e., the risk of not receiving promised repayments (e.g., loans or bonds).

Operational risk (OpRisk) The risk of loss resulting from inadequate or failed internal processes, people and systems or from external events (e.g., fraud, fat-finger trades, earthquakes).

¹The bond issuer owes the bond holder a debt and is obliged to pay at maturity T the principal and a coupon (interest; typically paid at fixed time points).

There are many other types of risks ([not discussed](#) in detail here):

Liquidity risk

(Market) liquidity risk is the risk stemming from the lack of marketability of an investment that cannot be bought or sold quickly enough to prevent or minimize a loss. Liquidity is “[oxygen for a healthy market](#)”: one is not aware of its presence; its absence, however, is recognized immediately.

Funding liquidity risk refers to the [ease](#) with which institutions can [raise funding](#). The two often interact in stress periods.

Underwriting risk

In insurance, underwriting risk is the [risk inherent in insurance policies sold](#) (related, e.g., to natural catastrophes, political changes, changes in demographic tables).

Model risk

The risk of using a misspecified (inappropriate) model for measuring risk. This is always present to some degree! (e.g., heavily in OpRisk modeling)

Good risk management (RM) has to follow a *holistic approach*, i.e., all (possibly influential) types of risks and their interactions should be considered.

1.1.3 Measurement and management

Risk measurement

- Suppose we hold a portfolio of d investments with weights w_1, \dots, w_d . Let X_j denote the change in value of the j th investment. The *change in value (profit and loss (P&L))* of the portfolio over a given *holding period* is then $X = \sum_{j=1}^d w_j X_j$. Measuring the risk now consists of determining the *distribution function F* (or functionals of it, e.g., mean, variance, α -quantiles $F^-(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$).

- We need a properly calibrated joint model for $\boldsymbol{X} = (X_1, \dots, X_d)$ (statistical estimates of F or one of its functionals are obtained based on historical observations of this model). Not an easy (or unique) task!
- Good risk measurement is essential (for good RM). For any product sold, the underlying risks need to be properly quantified and openly communicated to a client to decide whether or not the product matches her/his risk appetite. The 2007–2009 crisis saw numerous violations of this principle (e.g., through collateralized debt obligations).

Risk management

What is RM? Kloman (1990) writes:

“To many analysts, politicians, and academics it is the management of environmental and nuclear risks, those technology-generated macro-risks that appear to threaten our existence. To bankers and financial officers it is the sophisticated use of such techniques as

currency hedging and interest-rate swaps. To insurance buyers or sellers it is coordination of insurable risks and the reduction of insurance costs. To hospital administrators it may mean “quality assurance”. To safety professionals it is reducing accidents and injuries. In summary, RM is a discipline for living with the possibility that future events may cause adverse effects.”

⇒ It is about ensuring resilience to future events.

- Note that financial firms are not passive/defensive towards risk, banks and insurers actively/willingly take risks because they seek a return. RM thus belongs to the core competence of an insurance company or bank.
- What does managing risks involve?
 - ▶ Determine the capital to hold to absorb losses, both for regulatory capital (to please regulators) and economic capital purposes (to survive as a company).

- ▶ Ensuring portfolios are well diversified.
- ▶ Optimizing portfolios according to risk-return considerations (e.g., via derivatives to hedge exposures to risks, or *securitization*, i.e., repackaging risks and selling them to investors).

1.2 A brief history of risk management

1.2.1 From Babylon to Wall Street

The ancient world to the the 20th century

- A *derivative* is a financial instrument derived from an underlying asset, e.g., stocks, bonds, commodities, currencies, interest rates etc. Examples are
 - ▶ *Options* (right, but not the obligation, to buy (*call*) or sell (*put*) an asset at an agreed-upon price (the *strike price K*) during a predetermined period (*American*) or date (*exercise date T* ; *European*);
 - ▶ *Futures* (obligation for the buyer (seller) to purchase (sell) an asset at a predetermined date and price);
 - ▶ *Swaps* (any exchange of an asset for another to change the maturity (e.g., of a bond) or because investment objectives have changed; include currency swaps, interest rate swaps).

- Babylon of 1800 BC: early evidence for options to provide financial cover against crop failure.

Academic innovation in the 20th century

- Before 1950: Desirability of an investment was measured in terms of its return.
- Markowitz (1952): Theory of portfolio selection; Desirability of an investment was decided upon a risk-return diagram (x-axis: risk, i.e., standard deviation; y-axis: expected return). An efficient frontier determined the optimal return for a given risk level.
- Late 20th century: Theory of valuation for derivatives (important milestone for quantifying and managing financial risk)
- Black and Scholes (1973): Black–Scholes–Merton formula for the price of a European call option (Nobel Prize 1997)

- Harrison and Kreps (1979), Harrison and Pliska (1981): Fundamental theorems of asset pricing
 - 1) A (model for) a market is arbitrage free if and only if there exists a (risk-neutral) probability measure Q equivalent to \mathbb{P} ;
 - 2) A market is complete (i.e., every contingent claim can be replicated) if and only if Q is unique.
- By 1995: Nominal values outstanding in derivatives markets: tens of trillions.

Disasters of the 1990s

- Growing volume of derivatives in banks' trading books (often not appearing as assets/liabilities in the balance sheet).
- 1995 Barings Bank: OpRisk losses + straddle position on the Nikkei (short in a call and put with the same strike; allows for a gain if Nikkei

does not move too far up or down) + Kobe earthquake = ruin due to loss of \$1.3 billion

- 1998 Long-Term Capital Management (LTCM): hedge fund; losses due to derivatives trading, required a \$3.5 billion payout to prevent collapse; M. Scholes and R. Merton were principles
- Life insurer Equitable Life: Prior to 1988 Equitable Life had sold pension products which offered the option of a guaranteed annuity rate of 7% at maturity. In 1993, current annuity rate fell below the guarantee rate and policyholders exercised their options. Equitable Life faced an enormous increase in their liabilities (not properly hedged). Around 2001, Equitable Life was underfunded by around £4.5 billion.

The turn of the century

- 1996–2000: dot-com bubble; Nasdaq index climbed from around 1000 to around 5400; many firms contributing to this rise belong to the internet

sector. Within one year, the Nasdaq falls by 50% (the bubble burst).

- During this time, financial engineers discovered *securitization* (bundling and repackaging of risks into securities with defined risk profiles that can be sold to investors).
- Different types of assets were transformed into *collateralized debt obligations (CDOs)*; boom to lend to borrowers with low credit ratings; CDO issuance volume by 2008 was around \$3 trillion; for *credit default swaps (CDS)*² around \$30 trillion. *Netting* considered (i.e., compensation of long versus short positions on the same underlying), the economic value of CDS and CDO markets was much smaller.
- CDSs were being increasingly used by investors to speculate on the changing credit outlook.

²Credit derivative which allows the (protection) buyer (pays premiums) to transfer credit risk inherent in a reference entity to a seller (investor; pays in case of default).

- The **consensus** was that all this activity was a **good thing!**
Alan Greenspan (Chairman Federal Reserve), November 2002:

“... instruments ... such as credit default swaps, collateralized debt obligations... have been developed and their use has grown rapidly in recent years. **The result? Improved credit RM** together with more and better risk-management tools appear to have significantly reduced ... stress on banks and other financial institutions. ... Obviously this market is still too new to have been tested in a widespread down-cycle for credit, but, to date, it appears to have functioned well.”

International Monetary Fund (IMF), April 2006:

“... dispersion of credit risk by banks to ... investors, rather than warehousing such risks on their balance sheets, **has helped to make** the banking and overall **financial system more resilient.**”

(but also warned about possible vulnerabilities)

CEO of AIG Financial Products, August 2007:

“It is hard for us, without being flippant, to even see a scenario within any kind of realm of reason that would see us losing one dollar in any of these transactions.”

- Not all of the risk from CDOs was dispersed, large banks held a lot of it themselves (see Acharya et al. (2009)):

“Starting in 2006, the CDO group at UBS noticed that their risk-management systems treated AAA securities as essentially riskless even though they yielded a premium (the proverbial free lunch). So they decided to hold onto them rather than sell them! After holding less than \$5 billion of them in 02/06, the CDO desk was warehousing a staggering \$50 billion in 09/07. . . . Similarly, by late summer of 2007, Citigroup had accumulated over \$55 billion of AAA-rated CDOs.”

The financial crisis of 2007–2009

- US house prices began to decline in 2006 and 2007.
- Subprime mortgage holders (having difficulties in refinancing their loans due to higher interest rates) defaulted on their payments. Starting in late 2007, this led to a rapid reassessment of the riskiness of securitization and losses in the value of CDOs. Banks were forced into write downs of the value of these assets on their balance sheets.
- The most serious crisis since the 1920s resulted:
 - ▶ March 2008: Bear Stearns collapsed and was sold to JP Morgan Chase
 - ▶ September 2008: Lehman Brothers filed for bankruptcy (⇒ worldwide panic, markets tumbled, liquidity vanished, many banks were near collapse)

- ▶ September 2008: AIG (insuring the default risk in securitized products by selling CDS protection) got into difficulty when many of the underlying securities defaulted ⇒ emergency loan of \$85 billion from the Federal Reserve Bank of New York.

Governments had to bail companies out by injecting capital or acquiring their distressed assets in arrangements such as the US TARP (Troubled Asset Relief Program).

- J. E. Stiglitz (Nobel Prize 2001) on securitization and the housing market in 1992:

“The question is, has the growth of securitization been a result of more efficient transaction technologies, or an unfounded reduction in concern about the importance of screening loan applicants? It is perhaps too early to tell, but we should at least entertain the possibility that it is the latter rather than the former. . . At the very least, the banks have demonstrated ignorance of two very

basic aspects of risk: (a) the importance of correlation ... (b) the possibility of price declines.”

- In parts, mathematicians/financial engineers were also blamed due to the failure of valuation/pricing models for complex securitized products, e.g., by F. Salmon (Wired Magazine, 2009-02-23, “Recipe for disaster: the formula that killed Wall Street”). The formula was the Gauss copula model and its application to credit risk was attributed to David Li.
- The reliance on mathematics was only one factor in the crisis, and certainly not the most important. Mathematicians had also warned well beforehand about securitization (see, e.g., Frey et al. (2001)). Political shortsightedness, the greed of market participants and the slow reaction of regulators had all contributed.

Recent developments and concerns

- The financial crisis led to recession and sovereign debt crises.

- High Frequency Trading (HFT) has raised concerns among regulators, triggered by such events as the Flash Crash of 2010-05-06 (United States trillion-dollar stock market crash; in about 36 minutes the S&P 500, the Nasdaq 100, and the Russell 2000 collapsed; the Dow Jones Industrial Average had its biggest intraday loss of about 9%; 2015-04-21 US Department of Justice: 22 criminal counts, including fraud and market manipulation).
- Trades are executed by computer (algorithms) in fractions of a second (no testing), computer centers are built near stock markets for faster trading. One notable casualty of algorithmic trading: Knight Capital lost \$460 million due to trading errors on 2012-08-01 (acquired by Getco LLC in December 2012).
- Ongoing concern: System risk, i.e., the risk of the collapse of the entire financial system due to the propagation of financial stress through a network of participants. The networks are complex. Besides banks

and insurance companies they contain largely unregulated hedge funds and structured investment vehicles (“shadow banking system”). One important theme is the identification of *systemically important financial institutions (SIFIs)* whose failure might cause a systemic crisis.

1.2.2 The road to regulation

- Main aim of regulation: Ensure that financial institutions have enough capital to remain solvent.
- Robert Jenkin (member of the Financial Policy Committee of the Bank of England, 2012-04-27):

“Capital is there to absorb losses from risks we understand and risks we may not understand. Evidence suggests that neither risk-takers nor their regulators fully understand the risks that banks sometimes take. That’s why banks need an appropriate level of loss absorbing equity.”

- *Basel Committee of Banking Supervision (BCBS)*: Committee established by the Central-Bank Governors of the Group of Ten (G-10) in 1974. The Basel Committee's secretariat is located at the [Bank for International Settlements \(BIS\)](#) in Basel (CH). The Basel Committee **does not have legal force** but it formulates broad supervisory standards, **guidelines** and statements of best practice, the *Basel Accords*, in the expectation that individual authorities will take steps to implement them.

The first Basel Accord (Basel I)

- Issued in [1988](#)
- Only addressed **credit risk**
- Fairly **coarse** (measured risk in an insufficiently differentiated way)
 - ▶ Claims were divided into **3 categories only**, counterparties being governments, regulated banks or others;

- ▶ Risk weighting identical for all corporate borrowers, independent of their credit rating;
- ▶ Unsatisfactory treatment of derivatives.

The birth of VaR

- 1993: G-30 published a seminal report addressing for the first time so-called off-balance-sheet products, e.g., derivatives. The banking industry saw the need for proper measurement of these risks.
- At JPMorgan the famous Weatherstone 4.15 report asked for a one-day, one-page summary of the bank's market risk to be delivered to the CEO in the late afternoon (hence the "4.15").
- Value-at-Risk (VaR) as a market risk measure was born and the JP-Morgan methodology (which became known as *RiskMetrics*), set an industry-wide standard.

- Banks pushed to be allowed to use *netting* effects (compensation of long versus short positions on the same underlying)
- Amendment to Basel 1 in 1996 ⇒ *standardized model* for market risk and *internal Value-at-Risk-based model* for more sophisticated banks
- Coarseness problem for *credit risk* remained (not enough incentives to diversify credit portfolios; regulatory capital rules too risk insensitive). Because of overcharging on the regulatory capital side, *banks started shifting business away from certain market segments.*

The second Basel Accord (Basel II)

- Initiated in 2001, document published in June 2004.
- Three pillar concept: 1) Quantification of *regulatory capital*; 2) *Regulatory review* of the modeling process; 3) *Disclosure requirements*.
- Important themes were:

- ▶ Under Pillar 1, banks are now allowed to use a more risk-sensitive approach for assessing credit risk of their portfolios (they could opt for an *internal ratings-based* approach which permitted the use of credit-rating systems).
- ▶ Operational risk was introduced as a new class of risk.
- Due to the financial crisis of 2007–2009, further amendments to the 2004 version were made (**criticism**: Basel II is *procyclical*, i.e., forcing firms to increase their capital ratios at exactly the wrong point in the business cycle with negative effects on liquidity), which delayed the implementation of the Basel II guidelines.

Basel 2.5

- CDOs had opened up opportunities for *regulatory arbitrage* (transferring credit risk from the capital-intensive banking book to the less-capitalized trading book).

- Some **enhancements to Basel II** were proposed in 2009 with the aim of addressing the build up of risk in the trading book. These enhancements, known as *Basel 2.5*, include a **stressed VaR** (calculating VaR from data for a 12-month period of market turmoil) and the **incremental risk charge** (to capture some of the default risk in trading book positions). There were also specific new rules for certain securitizations.

The third Basel Accord (Basel III)

- 2011: Five extensions of Basel II (and 2.5) were proposed:
 - 1) Measures to increase the quality and amount of capital by changing the definition of **key capital ratios** and allowing **countercyclical adjustments** to these ratios in crises;
 - 2) A strengthening of the framework for **counterparty credit risk** in **derivatives trading** with incentives to use central counterparties (exchanges);

- 3) Introduction of a leverage ratio to prevent excessive leverage;
 - 4) Introduction of various ratios that ensure that banks have sufficient funding liquidity;
 - 5) Measures to force systemically important banks (SIBs) to have even higher risk capital.
- Note: Basel III works alongside Basel II (and 2.5), not replacing it.
 - Targeted end date of implementation: 2019

Parallel developments in insurance regulation

- More fragmented, much less international coordination of efforts
- Exception: Solvency II framework in the European Union (EU)
- Overseen by EIOPA (European Insurance and Occupational Pensions Authority), but implementation is a matter for national regulators (e.g., the Prudential Regulatory Authority (PRA) in the UK).

- US: Insurance regulation is a matter for state governments. The National Association of Insurance Commissioners (NAIC) provides support to insurance regulators from the individual states (helps to promote best practices etc.; early 1990s: NAIC promoted the concept of risk-based capital (RBC), a rule-based (rather than model-based) method of measuring the minimum amount of capital appropriate for supporting overall business operations depending on size and profile).
- After the 2007–2009 crisis: 2010 Dodd–Frank Act (creation of a Federal Insurance Office to “monitor all aspects of the insurance sector” and the Financial Stability Oversight Council (FSOC) “charged with identifying risks to the financial stability of the United States”)
- The International Association of Insurance Supervisors (IASI) works on more international convergence for regulating the capital adequacy.
- There are also ongoing initiatives to bring about convergence of banking and insurance regulation.

From Solvency I to II

- Solvency I came into force in 2004: Rather coarse rules-based framework calling for companies to have a *minimum guarantee fund* (minimal capital) of €3 million and a solvency margin of 16–18% of non-life premiums together with 4% of the technical provisions for life ⇒ Single, robust system, **easy to understand, inexpensive to monitor**. However, on the negative side, it is **mainly volume based** and **not explicitly risk based**.
- Solvency II was initiated in 2001 (publication of the Sharma report): While the Solvency II Directive was adopted by the Council of the European Union and the European Parliament in November 2009, implementation of the framework is **not expected until 1 January 2016**.
- The process of refinement of the framework is managed by EIOPA (conducts a series of **quantitative impact studies (QIS)** in which companies

have tried out aspects of the proposals; information about the impact and practicability of the new regulations results).

- Solvency II goals: strengthen the capital adequacy by reducing the possibilities of consumer loss or market disruption in insurance (⇒ policyholder protection and financial stability motives)

Swiss Solvency Test (SST)

- Specific to Switzerland.
- Already developed and in force since 2011-01-01.
- Implements its own principles-based risk-capital regulation for insurers.
- Similar to Solvency II, but differs in its treatment of different types of risk. Also puts more emphasis on the development of internal models.
- The implementation of the SST belongs to the responsibilities of the Swiss Financial Markets Supervisory Authority (FINMA).

1.3 The regulatory framework

1.3.1 The Basel framework

The three-pillar concept

This concept is a [key feature](#) of the framework [starting from Basel II](#). From Basel Committee on Banking Supervision (2004):

“The Basel II Framework sets out the details for adopting more risk-sensitive minimum capital requirements [Pillar 1] for banking organizations. The new framework reinforces these risk-sensitive requirements by laying out principles for banks to assess the adequacy of their capital and for supervisors to review such assessments to ensure banks have adequate capital to support their risks [Pillar 2]. It also seeks to strengthen market discipline by enhancing transparency in banks' financial reporting [Pillar 3].”

Pillar 1 *Minimal capital charge*. Requirements for the calculation of the *regulatory capital* to ensure that a bank holds *sufficient capital* for its *market risk* in the trading book, *credit risk* in the banking book and *operational risk* (main quantifiable risks). For *market risk*, most banks use *internal models based on VaR methodology*; for *credit risk* and *operational risk* banks may *choose between several approaches* (see later). We will focus on *Pillar 1*.

Pillar 2 *Supervisory review process*. Local *regulators* review the checks and balances put in place for *capital adequacy assessments*, ensure that banks have adequate regulatory capital and *encourage them to use good techniques* for monitoring and managing risks. Interest-rate risk in the banking book must be considered and *stress tests* of a bank's capital adequacy performed. Aim: A bank should hold *capital in line with its true economic loss potential (economic capital)*.

Pillar 3 *Market discipline*. Addresses better public disclosure of risk measures and other RM relevant information (banks are required to provide better insight into the adequacy of their capitalization).

Credit and market risk; banking and trading book

- Banking activities are organized around the *banking book* (assets on the balance sheet held to maturity, at historic costs (*book value*); $\text{VaR}_{0.999}$ is calculated based on a one-year time horizon) and the *trading book* (assets held that are regularly traded; marked-to-market every day; $\text{VaR}_{0.99}$ is calculated based on a 10-day time horizon) reflecting the different accounting practices for different kinds of assets.
- Credit risk is mainly identified with the banking book; market risk with the trading book.
- The distinction is somewhat arbitrary and depends on “available to trade”. There can be incentives to move instruments from one book

to the other (often from the banking to the trading book) to benefit from a more favourable capital treatment. Basel Committee on Banking Supervision (2013):

“... the overall capital framework proved susceptible to arbitrage before and during the crisis ... To reduce the incentives for arbitrage, the Committee is seeking a less permeable boundary with strict limits on switching between books and measures to prevent “capital benefit” in instances where switching is permitted.”

The capital charge for the banking book

- The credit risk of the banking-book portfolio is assessed as the sum of *risk-weighted assets (RWAs)* (i.e., linear combination of notional exposures weighted by risk weights reflecting the creditworthiness of the counterparty)

- To calculate risk weights, banks use either the *standardized approach* or one of the more advanced *internal-ratings-based (IRB)* approaches (international banks have to follow the latter).
- The *capital charge* is determined as a *fraction (capital ratio)* of the sum of risk-weighted assets in the portfolio. The capital ratio was 8% under Basel II, but will be increased for Basel III in 2019.
- In the *standardized approach*, risk weights are prescribed by the regulator.
- Under the *IRB approaches* banks may make an *internal assessment* of the riskiness of a credit exposure, expressing this in terms of an estimated annualized *probability of default (PD)* and an estimated *loss-given-default (LGD)*, which are used as inputs in the calculation of risk-weighted assets. The total *sum of risk-weighted assets* is calculated using formulas specified by the Basel Committee, which also take positive correlation into account.

- IRB approaches allow for increased risk sensitivity in the capital charges compared with the standardized approach. Note, however, that the IRB approaches do not permit fully internal models of credit risk in the banking book (they only permit internal estimation of inputs to a model that has been specified by the regulator).

The capital charge for the trading book

- For market risk in the trading book there is also a standardized approach. However, most major banks use an *internal VaR model approach* ($\text{VaR}_{0.99}$ for a 10-day holding period; a 10-day $\text{VaR}_{0.99}$ of \$20 million means that our market portfolio is estimated to incur a loss of $\geq \$20 \text{ million}$ with probability 1% by the end of a 10-day holding period, if the portfolio composition remains fixed).
- For the conversion of VaR numbers into an actual capital charge, see later.

- VaR calculation is the main component of risk quantification, but **Basel 2.5 added**:
 - ▶ *Stressed VaR*: Banks are required to carry out VaR calculations based on their **models being calibrated to a historical 12-month period of financial stress**.
 - ▶ *Incremental Risk Charge (IRC)*: Banks must calculate an **additional charge** based on an estimate of the 99.9% quantile of the one-year loss distribution **due to defaults and rating changes** (since default and rating migration risk are not considered otherwise).
 - ▶ *Securitizations*: Exposures to securitizations in the trading book are subject to **new capital charges**.

The capital charge for OpRisk

There are **three options** of increasing sophistication. Under the **basic-indicator** and **standardized approaches** banks may calculate their OpRisk

charge using simple formulas based on gross annual income. Under the *advanced measurement approach* banks may develop internal models (most are based on internal and external historical data).

New elements of Basel III

The main changes will be (may change before final implementation):

- Banks will need to hold more and better quality capital (the latter is achieved through a more restrictive definition of eligible capital, the former relates to Basel II's 8% + a capital conservation buffer of 2.5% of risk-weighted assets + a countercyclical buffer of up to 2.5%)
- A leverage ratio will be imposed to put a floor under the build-up of excessive leverage (leverage will be measured through the ratio of Tier 1 capital to total assets; a minimum ratio of 3% is currently being tested).
- A charge for counterparty credit risk is included. When counterparty credit risk is taken into account in the valuation of an OTC derivative

contract, the default-risk-free value has to be adjusted by an amount known as the *credit valuation adjustment (CVA)*.

- Banks will become subject to *liquidity rules*; this is a completely new direction for the Basel framework which has previously only been concerned with capital adequacy. A *liquidity coverage ratio (LCR)* will be introduced to ensure that banks have enough highly liquid assets to withstand a period of net cash outflow lasting 30 days. A *net stable funding ratio (NSFR)* will ensure that sufficient funding is available in order to cover long-term commitments (\geq one year).

Risk quantification may change from VaR-based to being based on expected shortfall (ES); see later.

1.3.2 The Solvency II Framework

Main features

- As Basel II, Solvency II adopts a three-pillar system (Pillar 1: quantification of regulatory capital; Pillar 2: governance and supervision; Pillar 3: disclosure of information to the public)
- Under Pillar 1, a company calculates its *solvency capital requirement (SCR)* = amount of capital to ensure that the probability of insolvency over a one-year period is no more than 0.5% (referred to as a confidence level of 99.5%).
- The company also calculates a smaller *minimum capital requirement (MCR)* = minimum capital to continue operating without supervisory intervention.
- For calculating capital requirements, a *standard formula* or an *internal model* may be used. Either way, a *total balance sheet approach* is taken

(all risks and their interactions are considered).

- The insurer should have *own funds* (surplus of assets over liabilities) that exceed the SCR and the MCR.
- Under Pillar 2, the company must demonstrate that it has a RM system in place and that this system is integrated into decision making processes.
- An internal model must pass the “use test”: It must be an integral part of the RM system and be actively used in the running of the firm. Moreover, a firm must undertake an ORSA (*own risk and solvency assessment*) as described below.

Market-consistent valuation.

- Assets and liabilities of a firm must be valued in a *market-consistent* manner. Where possible, actual market values should be used (*marking-to-market*).

- When no market values exist, models (consistent with market information) have to be calibrated (a process known as *marking-to-model*).
- Market consistent valuation of the liabilities of an insurer is possible if cash flows to policyholders can be replicated by a replicating portfolio of matching assets.
- If this is not possible (e.g., for mortality risk), valuation is done by computing the sum of a *best estimate of the liabilities* (basically an expected value) plus a *risk margin*.

Standard formula approach

- Insurers calculate capital charges for different kinds of risk within a series of *modules* (e.g., for market risk, counterparty default risk, life underwriting risk, non-life underwriting risk and health insurance risk)
- Within each module, capital charges are calculated with respect to fundamental risk factors (e.g., within the market risk module, there

are interest-rate risk, equity risk or credit-spread risk). Capital charges are calculated by [considering stress scenarios](#) on the value of net assets (assets – liabilities). The stress scenarios are [intended to represent 1 in 200 year events](#) (i.e., events with annual probability 0.5%).

- The capital charges for each risk factor are [aggregated to obtain the module risk charge](#). Again a set of correlations is used to express the regulatory view of dependencies between the fundamental risk factors.
- The [risk charges arising from these modules](#) are aggregated to obtain [the SCR](#) using a formula that involves a set of prescribed correlation parameters.

Internal model approach.

- [On regulatory approval](#), firms can develop [an internal model](#) for the financial and underwriting risk factors.

- An internal model often takes the form of a so-called *economic scenario generator (ESG)* in which risk-factor scenarios for a one-year period are randomly generated and applied to the assets and liabilities to determine the SCR.

ORSA (Own risk and solvency assessment)

- *OSRA* = entirety of processes and procedures to identify, assess, monitor, manage, and report short and long term risks a (re)insurance company may face and to determine the own funds necessary to ensure the company's solvency at all times.
- OSRA (Pillar 2) is different from capital calculations (Pillar 1):
 - ▶ ORSA refers to a *process* (and not just an exercise in regulatory compliance);
 - ▶ Each firm's ORSA is its *own process* and likely to be *unique* (not bound by a common set of rules such as the standard-formula ap-

proach in Pillar 1; even firms using internal models under Pillar 1 are bound to similar constraints).

- ▶ ORSA goes beyond the one-year time horizon (which is a limitation of Pillar 1); e.g., for life insurance.

1.3.3 Criticism of regulatory frameworks

- Benefits of regulation: Customer protection, responsible corporate governance, fair and comparable accounting rules, transparent information on risk, capital and solvency for shareholders etc.
- The following aspects have raised criticism:
 - ▶ Costs and complexity for setting up and maintaining a sound risk management system compliant with present regulations (PRA: in the UK, Solvency II compliance costs at least £3 billion. Regulation becomes more and more complex.

- ▶ *Endogenous risk* (i.e., the risk generated within a system and amplified by the system due to feedback effects): Regulation may amplify shocks. It can lead to *risk-management herding* (institutions all run for the same exit by following the same (perhaps VaR-based) rules in times of crisis and thus further destabilize the whole system).
- ▶ **Market-consistent valuation** (at the core of the Basel rules for the trading book and Solvency II) implies that capital requirements are closely coupled to volatile financial markets. An insurer may appear to have insufficient solvency funds. However, if assets and liabilities are matched and contractual obligations can be met, this may not be a problem (insurance is a long-term business; no short-term need to sell assets or offload liabilities; a loss of capital need not be realised unless some of the bonds actually default).
- ▶ Highly quantitative nature of regulation: Extensive use of mathematical and statistical methods. Lord Turner (2009) (Turner Review of

the global banking crisis):

“The very complexity of the mathematics used to measure and manage risk, moreover, made it increasingly difficult for top management and boards to assess and exercise judgement over the risk being taken. Mathematical sophistication ended up not containing risk, but providing false assurances that other *prima facie* indicators of increasing risk (e.g. rapid credit extension and balance sheet growth) could be safely ignored.”

Overconfidence in the quality of risk measure estimates is a weakness. Quantitative modeling of OpRisk has been controversial: How can we measure human risk (e.g., incompetence, fraud), process risk (e.g., model risk, transaction risk), technology risk (e.g., system failure, programming and numerical errors) or legal risk?

- ▶ Can tighter regulation prevent a crisis such as that of 2007–2009? Rules are constantly overtaken by financial innovation.

1.4 Why manage financial risk?

1.4.1 A societal view

- Society relies on the stability of the banking and insurance system. The regulatory process (from which Basel II and Solvency II resulted) was motivated by the desire to prevent insolvency of individual institutions and thus protect customers (*microprudential perspective*).
- However, the reduction of systemic risk has become an important secondary focus since the 2007–2009 crisis (*macroprudential perspective*).
- Most would agree that the protection of customers and the promotion of financial stability are vital, but it is not always clear whether the two aims are well aligned (e.g., might be good to let a company go bankrupt to teach other companies a lesson).
- This is related to *systemic importance* of the company in question (size and connectivity to other firms). Considering some firms as too big to

fail creates a moral hazard (should be avoided!) since the management of such a firm may take more risk knowing that it would be bailed out in a crisis. Also, some companies are too big to save.

- Before the crisis, it was initially believed that the growth in securitization was dispersing credit risk throughout the system and was beneficial to financial stability. But inadequately valued credit risk (through CDOs) in the trading book, combined with the interconnectedness of banks through derivatives and interbank lending activities, meant that quite the opposite was true.
- Society suffered next. The world economy went into recession, households defaulted on their debts, and savings and pensions were hit hard. The crisis moved from “Wall Street into Main Street”.
- It seemed that the government-sponsored bail-outs had allowed banks “to privatize the gains and socialize the losses”.

- The interests of society are served by enforcing the discipline of risk management in financial firms, through the use of regulation. Better risk management can reduce the risk of company failure and protect customers and policyholders. However, regulation must be designed with care and should not promote herding, procyclical behaviour or other forms of endogenous risk that could result in a systemic crisis. Individual firms need to be allowed to fail on occasion, provided customers can be shielded from the worst consequences through appropriate compensation schemes.

1.4.2 The shareholder's view

- It is widely believed that proper financial RM can increase the value of a corporation and hence shareholder value. Questions to be answered include:
 - ▶ When does RM increase the value of a firm, and which risks should be managed?

- ▶ How should RM concerns factor into investment policy and capital budgeting?
- While *individual* investors are typically risk averse and should therefore manage the risk in their portfolios, it is not clear that risk management at the *corporate level* (e.g., hedging a foreign-currency exposure or holding a certain amount of risk capital) increases the value of a corporation and thus enhances shareholder value. The rationale for this is simple: If investors have access to perfect capital markets, they can do the RM transactions via their own trading and diversification.
- The famous *Modigliani–Miller Theorem*, which marks the beginning of modern corporate finance theory, states that, in an ideal world without taxes, bankruptcy costs and informational asymmetries, and with frictionless and arbitrage-free capital markets, the financial structure of a firm (thus its RM decisions) is irrelevant for the firm's value. In order to find reasons for corporate RM, one has to “turn the Modigliani–Miller

Theorem upside down":

- ▶ RM can *reduce tax costs*.
- ▶ RM can be beneficial, since a company may have better access to capital markets than individual investors.
- ▶ RM can *increase the firm value* in the presence of *bankruptcy costs* (e.g., cost of lawsuits or liquidation costs), *as it makes bankruptcy less likely*; this often has a positive effect on key employees or businesses (e.g., few customers would want to enter into a life insurance contract with an insurance company which is known to be close to bankruptcy; or banks might be faced with a bank run if close to bankruptcy).
- ▶ RM can *reduce the impact of costly external financing*, as it helps achieving optimal investment (*external funds are more costly* to obtain *than internal funds*); without RM the increased variability of a company's cash flow will be translated either into an increased

variability of external funds or to an increased variability in the amount of investment, both lead to decreasing (expected) profits).

1.5 Quantitative Risk Management

1.5.1 The Q in QRM

- In what follows, we adopt a somewhat narrower view and treat QRM as a quantitative science using the language of mathematics in general, and probability and statistics in particular.
- Mathematics and statistics provide us with with a suitable language and with appropriate concepts for describing financial risks which could otherwise not be done.
- A main theme is also to point out limitations of current methodology used. Furthermore, mathematicians are very well aware that a mathematical result not only has a conclusion, but equally importantly, has assumptions under which it holds. Statisticians are well aware that inductive reasoning on the basis of models relies on the assumption that these assumptions hold in the real world.

- By starting with questionable assumptions, models can be used (or manipulated) to deliver bad answers. The implication is that quantitative risk managers must become more worldly about the ways in which models are used. But equally, the regulatory system needs to be more vigilant about the ways in which models can be gamed.
- The Q in QRM is an essential part of the process. We believe it remains (if applied correctly and honestly) a part of the solution to managing risk (not the problem). See also Shreve (2008):

“Don’t blame the quants. Hire good ones instead and listen to them.”

1.5.2 The nature of the challenge

- Our approach to QRM has two main strands:
 - ▶ Put current practice onto a firmer mathematical ground;

- ▶ Put together techniques and tools which go beyond current practice and address some of the deficiencies.
- In particular, some of the challenges of QRM are:
 - ▶ Extremes matter. There is the need to address unexpected, abnormal or extreme outcomes (in contrast to many classical/statistical applications); see also A. Greenspan (Joint Central Bank Research Conference, 1995):

“... inappropriate use of the normal distribution can lead to an understatement of risk, which must be balanced against the significant advantage of simplification. From the central bank’s corner, the consequences are even more serious because we often need to concentrate on the left tail of the distribution in formulating lender-of-last-resort policies. Improving the characterization of the distribution of extreme values is of paramount importance.”

Or Lord Turner (2009):

"Price movements during the crisis have often been of a size whose probability was calculated by models (even using longer term inputs) to be almost infinitesimally small. This suggests that the **models systematically underestimated the chances of small probability high impact events** ... it is possible that financial market movements are inherently characterized by fat-tail distributions. **VaR models** need to be buttressed by the application of **stress test** techniques which consider the **impact of extreme movements** beyond those which the model suggests are at all probable."

- ▶ **Interdependence and concentration of risks.** **Risk is multivariate in nature**, we are **generally interested in** some form of **aggregate risk** that depends on **high-dimensional** vectors of underlying risk factors. A particular concern is the **dependence between extreme outcomes**,

when many risk factors move against us simultaneously. In connection with the LTCM case we find the following quote in [Business Week](#), September 1998.

“Extreme, synchronized rises and falls in financial markets occur infrequently but they do occur. The problem with the models is that they did not assign a high enough chance of occurrence to the scenario in which many things go wrong at the same time—the “perfect storm” scenario.”

In a perfect storm scenario, portfolio diversification arguments break down and there is much more concentration of risk; this was very much the case with the 2007–2009 crisis when borrowing rates rose, bond markets fell sharply, liquidity disappeared and many other asset classes declined in value.

- ▶ [The problem of scale](#). A portfolio may represent the entire position in risky assets of a financial institution. Calibration of detailed

multivariate models for all risk factors is impossible and hence any sensible strategy involves dimension reduction (i.e., identification of key risk drivers/features to be modeled). We are forced to adopt a fairly “broad-brush” approach. E.g., in the context of portfolio credit risk, we are more concerned with finding suitable models for the default dependence of counterparties than with accurately describing the individual default mechanism, since it is our belief that the former is at least as important as the latter in determining the risk of a large diversified portfolio.

- ▶ Interdisciplinarity. Ideas and techniques from several existing quantitative disciplines are drawn together. A combined quantitative skillset should include concepts, techniques and tools from mathematical finance, statistics, financial econometrics, financial economics and actuarial mathematics.
- ▶ Communication and education. A quantitative risk manager operates

in an environment where additional non-quantitative skills are equally important (communication, market practice, institutional details, humility). A lesson from the 2007–2009 crisis is that improved education in QRM is essential; from the front office to the back office to the boardroom, users of models and their outputs need to be better trained to understand model assumptions and limitations. This is part of the role of a quantitative risk manager, who should ideally have (or develop) the pedagogical skills to explain methods and conclusions to audiences at different levels of mathematical sophistication.

1.5.3 QRM beyond finance

- Some of the earliest applications of QRM are to be found in the **manufacturing industry**, where similar concepts and tools exist under names like **reliability** or **total quality control**. Industrial companies have recognized the **risks associated with bringing faulty products to the market**.
- QRM techniques have been adopted in the **transport and energy industries** (cost of storage and transport of electricity).
- There is an interest in the **transfer of risks between industries**; this process is known as ***alternative risk transfer (ART)***, e.g., the risk transfer between the **insurance and banking industries**.
- QRM methodology also applies to **individuals**, e.g., via the **risk of unemployment, depreciation in the housing market** or the investment in the **education of children**.

2 Basics concepts in risk management

2.1 Risk management for a financial firm

2.2 Modeling value and value change

2.3 Risk measurement

2.1 Risk management for a financial firm

2.1.1 Assets, liabilities and the balance sheet

The risks of a financial firm can be understood from its *balance sheet* (financial statement showing *assets* (investments) and *liabilities* (how funds have been raised; obligations)). A stylized balance sheet for a **bank** is:

Assets	Liabilities
Investments of the firm	Obligations from fundraising
- Cash	Debt capital
- Securities	- Customer deposits
- Loans, mortgages	- Bonds issued
- Property	- Reserves for losses on loans from banks
Equity	

A stylized balance sheet for an **insurer** (sells contracts, collects premiums, raises funds by issuing bonds \Rightarrow Liabilities are thus obligations to policy holders (reserve against future claims; obligations to bondholders)) is:

Assets	Liabilities
<i>Investments of the firm</i>	<i>Obligations to policy holders</i>
- Investments (e.g., bonds, stocks)	Debt capital
- Investments for unit-linked contracts	- Reserves for policies written
- Property	- Bonds issued
	Equity

- Balance sheet equation: **Assets = Liabilities = Debt + Equity.**
If equity > 0 , the company is *solvent*, otherwise *insolvent*. Distinction to **default** (not able to pay): Note that a solvent company can default because of liquidity problems.

- Valuation of the items on the balance sheet is a non-trivial task.
 - ▶ *Amortized cost accounting* values a position at its inception and this is carried forward/progressively reduced over time.
 - ▶ (Similar to market consistent valuation (a variant of)) *fair-value accounting* values assets at prices they are sold and liabilities at prices that would have to be paid in the market. This can be challenging for non-traded or illiquid assets or liabilities.

There is a tendency in the financial industry to move towards fair-value accounting.

2.1.2 Risks faced by a financial firm

- Decrease in the value of the investments on the asset side of the balance sheet (e.g., losses from securities trading or credit risk)
- *Maturity mismatch* (large parts of the assets are relatively illiquid (long-term) whereas large parts of the liabilities are rather short-term obligations. This can lead to a default of a solvent bank and even a **bank run**).
- The prime risk of an insurer is *insolvency* (risk that claims of policy holders cannot be met). On the asset side, risks are similar to those of a bank. On the liability side, the main risk is that reserves are insufficient to cover future claim payments. Note that the liabilities of a life insurer are of a long-term nature and subject to multiple categories of risk (e.g., interest rate risk, inflation risk and longevity risk).
- So risk is found on *both sides* of the balance sheet and thus RM should not focus on the asset side alone.

2.1.3 Capital

- There are different notions of capital. One distinguishes:

- Equity capital*
 - Value of assets – debt;
 - Measures the firm's value to its shareholders;
 - Can be split into *shareholder capital* (initial capital invested in the firm) and *retained earnings* (accumulated earnings not paid out to shareholders).

- Regulatory capital*
 - Capital required according to regulatory rules;
 - For European insurance companies: MCR + SCR (see Solvency II);
 - A regulatory framework also specifies the capital quality. Here one distinguishes *Tier 1 capital* (i.e., shareholder capital + retained earnings;

can act in full as buffer) and *Tier 2 capital* (includes other positions on the balance sheet, e.g., subordinated debt).

Economic capital

- Capital required to control the probability of **becoming insolvent** (typically over a one-year horizon);
 - Internal assessment or risk capital;
 - Aims at a holistic view (assets and liabilities) and works with fair values of balance sheet items.
- All of these notions refer to items on the liability side that entail no (or very limited) obligations to outside creditors and that can thus serve as a buffer against losses.

2.2 Modeling value and value change

2.2.1 Mapping of risks

Prob. space $(\Omega, \mathcal{F}, \mathbb{P})$
 X random variable (rv)
 $x = X(\omega)$ a realization
(ω = state of nature)

We now set up a general mathematical model for value and changes in value caused by financial risks. For this we assume to work on a *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ and consider a risk or loss as a *random variable* $X : \Omega \rightarrow \mathbb{R}$ (or: L , a random vector \mathbf{X} , ...).

- Consider a *portfolio* of assets and possibly liabilities. The *value* of the portfolio at time t (*today*) is denoted by V_t (a random variable; assumed to be known at t ; its *df* is typically *not trivial to determine!*).
- We consider a given *time horizon* Δt (e.g., 1 d or 10 d for market risk; 1 y for credit risk; 20 y for pension funds) and *assume*:
 - 1) the *portfolio composition* remains *fixed* over Δt ;
 - 2) there are *no intermediate payments* during Δt

⇒ Fine for $\Delta t \in \{1 \text{ d}, 10 \text{ d}\}$ but unlikely to hold for $\Delta t \in \{1 \text{ y}, 20 \text{ y}\}$.

- The *change* in value of our portfolio is then given by

$$\Delta V_{t+1} = V_{t+1} - V_t$$

and we define the (random) *loss* as the *sign-adjusted* value change

$$L_{t+1} = -\Delta V_{t+1}$$

(as QRM is mainly concerned with losses).

Remark 2.1

- The distribution of L_{t+1} is called *loss distribution* (df F_L or simply F).
- Practitioners often consider the *profit-and-loss (P&L) distribution* which is the distribution of $-L_{t+1} = \Delta V_{t+1}$.
- For longer time intervals, $\Delta V_{t+1} = V_{t+1}/(1 + r) - V_t$ (r = *risk-free interest rate*) would be more adequate, but we will mostly neglect this issue.

- V_t is typically modeled as a function f of time t and a d -dimensional random vector $\mathbf{Z} = (Z_{t,1}, \dots, Z_{t,d})$ of *risk factors* (d typically large), that is,

$$V_t = f(t, \mathbf{Z}_t) \quad (\text{mapping of risks})$$

for some measurable $f : \mathbb{R}_+ \times \mathbb{R}^d \rightarrow \mathbb{R}$. The choice of f and \mathbf{Z}_t is problem-specific (but *typically known* to a bank).

- It is often convenient to work with the *risk-factor changes*

$$\mathbf{X}_t = \mathbf{Z}_t - \mathbf{Z}_{t-1}.$$

We can rewrite L_{t+1} in terms of \mathbf{X}_t via

$$\begin{aligned} L_{t+1} &= -(V_{t+1} - V_t) = -(f(t+1, \mathbf{Z}_{t+1}) - f(t, \mathbf{Z}_t)) \\ &= -(f(t+1, \mathbf{Z}_t + \mathbf{X}_{t+1}) - f(t, \mathbf{Z}_t)) =: L(\mathbf{X}_{t+1}); \end{aligned}$$

$L(\cdot)$ is known as *loss operator*. We see that the *loss df* is determined by the loss df of \mathbf{X}_{t+1} .

- If f is differentiable, its first-order (Taylor) approximation is

$$f(t+1, \mathbf{Z}_t + \mathbf{X}_{t+1}) \approx f(t, \mathbf{Z}_t) + f_t(t, \mathbf{Z}_t) \cdot 1 + \sum_{j=1}^d f_{z_j}(t, \mathbf{Z}_t) \cdot X_{t+1,j}$$

We can thus approximate L_{t+1} by the *linearized loss*

$$L_{t+1}^\Delta = - \left(f_t(t, \mathbf{Z}_t) + \sum_{j=1}^d f_{z_j}(t, \mathbf{Z}_t) X_{t+1,j} \right) = -(c_t + \mathbf{b}_t^\top \mathbf{X}_{t+1}),$$

a linear function of $X_{t+1,1}, \dots, X_{t+1,d}$ (indices denote partial derivatives).

The approximation is best if the $|X_{t+1,j}|$'s are small (typically if Δt is small; questionable for extreme market changes) and if V_{t+1} is almost linear in \mathbf{Z}_t (i.e., if mixed partial derivatives $|f_{z_i z_j}|$ are small in absolute value).

Example 2.2 (Stock portfolio)

Consider a portfolio \mathcal{P} of d stocks $S_{t,1}, \dots, S_{t,d}$ ($S_{t,j}$ = value of stock j at time t) and denote by λ_j the number of shares of stock j in \mathcal{P} . In finance and risk management, one typically uses logarithmic prices as risk factors, i.e., $Z_{t,j} = \log S_{t,j}$, $j \in \{1, \dots, d\}$. Then

$$V_t = f(t, \mathbf{Z}_t) = \sum_{j=1}^d \lambda_j S_{t,j} = \sum_{j=1}^d \lambda_j e^{Z_{t,j}}.$$

- The one-period ahead loss is then given by

$$\begin{aligned} L_{t+1} &= -(V_{t+1} - V_t) = - \sum_{j=1}^d \lambda_j (e^{Z_{t,j} + X_{t+1,j}} - e^{Z_{t,j}}) \\ &= - \sum_{j=1}^d \lambda_j e^{Z_{t,j}} (e^{X_{t+1,j}} - 1) = - \sum_{j=1}^d \lambda_j S_{t,j} (e^{X_{t+1,j}} - 1). \end{aligned} \quad (1)$$

- With $f_{z_j}(t, \mathbf{Z}_t) = \lambda_j e^{Z_{t,j}} = \lambda_j S_{t,j}$, the linearized loss is

$$\begin{aligned} L_{t+1}^{\Delta} &= -\left(0 + \sum_{j=1}^d f_{z_j}(t, \mathbf{Z}_t) X_{t+1,j}\right) = -\sum_{j=1}^d \lambda_j S_{t,j} X_{t+1,j} \\ &= -\sum_{j=1}^d \tilde{w}_{t,j} X_{t+1,j} = -V_t \sum_{j=1}^d w_{t,j} X_{t+1,j}, \end{aligned}$$

where $\tilde{w}_{t,j} = \lambda_j S_{t,j}$ and $w_{t,j} = \lambda_j S_{t,j}/V_t$ (proportion of V_t invested in stock j). Note that $c_t = 0$ and $\mathbf{b}_t = \tilde{\mathbf{w}}_t$ here.

- If $\mu = \mathbb{E} \mathbf{X}_{t+1}$ and $\Sigma = \text{Cov } \mathbf{X}_{t+1}$ are known, then expectation and variance of the (linearized) one-period ahead loss are

$$\begin{aligned} \mathbb{E} L_{t+1}^{\Delta} &= -\tilde{\mathbf{w}}_t^\top \mu = -V_t \mathbf{w}_t^\top \mu, \\ \text{Var } L_{t+1}^{\Delta} &= \tilde{\mathbf{w}}_t^\top \Sigma \tilde{\mathbf{w}}_t = V_t^2 \mathbf{w}_t^\top \Sigma \mathbf{w}_t. \end{aligned}$$

Example 2.3 (European call option)

Consider a portfolio consisting of a European call option on a non-dividend-paying³ stock S_t with maturity T and strike (exercise price) K . The Black–Scholes formula says that

$$V_t = C^{\text{BS}}(t, S_t; r, \sigma, K, T) = S_t \Phi(d_1) - K e^{-r(T-t)} \Phi(d_2), \quad (2)$$

where

- t is the time in years;
- Φ is the df of $N(0, 1)$;
- r is the continuously compounded risk-free interest rate;
- $d_1 = \frac{\log(S_t/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}$ and $d_2 = d_1 - \sigma\sqrt{T-t}$; and
- σ is the annualized volatility (standard deviation) of S_t .

³No distribution of profits to the shareholders (dividends).

While (2) assumes r, σ to be **constant**, this is often **not true in real markets**. Hence, **besides $\log S_t$, we consider r_t, σ_t as risk factors**, so

$$\mathbf{Z}_t = (\log S_t, r_t, \sigma_t) \Rightarrow \mathbf{X}_{t+1} = (\log(S_{t+1}/S_t), r_{t+1} - r_t, \sigma_{t+1} - \sigma_t).$$

This implies that the mapping f is given by

$$V_t = C^{\text{BS}}(t, e^{Z_{t,1}}; Z_{t,2}, Z_{t,3}, K, T) =: f(t, \mathbf{Z}_t)$$

and the linearized one-day ahead loss (omitting the arguments of C^{BS}) is

$$\begin{aligned} L_{t+1}^\Delta &= -\left(f_t(t, \mathbf{Z}_t) + \sum_{j=1}^3 f_{z_j}(t, \mathbf{Z}_t) X_{t+1,j}\right) \\ &= -(C_t^{\text{BS}} \Delta t + C_{S_t}^{\text{BS}} S_t X_{t+1,1} + C_{r_t}^{\text{BS}} X_{t+1,2} + C_{\sigma_t}^{\text{BS}} X_{t+1,3}). \end{aligned}$$

Here $\Delta t = 1/250$ (as our risk management horizon is 1 d here) and **the “Greeks” enter** (C_t^{BS} is the ***theta*** of the option; $C_{S_t}^{\text{BS}}$ the ***delta***; $C_{r_t}^{\text{BS}}$ the ***rho***; $C_{\sigma_t}^{\text{BS}}$ the ***vega***).

For portfolios of derivatives, L_{t+1}^Δ can be a rather poor approximation to $L_{t+1} \Rightarrow$ higher-order (Taylor) approximations such as the ***delta-gamma***-

approximation (second-order) have been used, but one loses the tractability/ellipticality.

2.2.2 Valuation methods

Fair value accounting

The *fair value* of an asset (liability) is an *estimate of the price* which would be received (or paid) on an *active market*. This valuation principle only applies to a minority of balance sheet positions. US/worldwide accounting rules thus distinguish the following levels (of determining a fair value):

Level 1 *Mark-to-market*. The *fair value* of an investment is determined from *quoted prices* in an active market for the *same instrument* (e.g., the stock portfolio in Example 2.2 above).

Level 2 *Mark-to-model with objective inputs*. The *fair value* of an instrument is determined *using quoted prices* in active markets for *similar instruments* or by *using valuation techniques/models* with

inputs based on observable market data (e.g., the European call option in Example 2.3 above)

Level 3 *Mark-to-model with subjective inputs.* The **fair value** of an instrument is determined using valuation techniques/models for which **some inputs** are **not observable** in the market (e.g., determining default risk of portfolios of loans to companies for which no CDS spreads⁴ are available).

Risk-neutral valuation

- . . . is **widely used** for pricing financial products, e.g., derivatives
- **value** of a financial instrument **today** = **expected discounted values** of future **cash flows**; the expectation is taken w.r.t. to the **risk-neutral pricing measure Q** (also called **equivalent martingale measure (EMM)**)

⁴Annual amount the protection buyer must pay the protection seller over $[0, T]$, expressed as a fraction (often in 1 basis point = 0.01%) of the notional amount.

as it turns discounted prices into martingales, i.e., fair bets) as opposed to the **real world/physical measure \mathbb{P}** .

Example 2.4 (\mathbb{P} vs Q ; one-period default model)

- Consider a **defaultable bond** with principal 1 and maturity $T = 1 \text{ y}$. In case of a default (real world probability $p = 0.01$), the recovery rate is $R = 60\%$. The risk-free interest rate is $r = 0.05$. Moreover, assume the bond's current price to be $V_0 = 0.941$ ($t = 0$).
- The **expected discounted value** of the bond is

$$\frac{1}{1+r}(1 \cdot (1-p) + R \cdot p) = \frac{1}{1.05}(0.99 + 0.6p) = 0.949$$

which is $> V_0$ since investors demand a **premium** for bearing the bond's **default risk**.

- An **risk-neutral pricing measure** is a **probability measure Q** such that the **expectation of the discounted payoff w.r.t. Q** equals V_0 (investing

becomes a fair bet). Here, Q is determined by specifying q such that

$$\frac{1}{1+r}(1 \cdot (1 - q) + R \cdot q) = V_0 \quad \Rightarrow \quad q = 0.03 > 0.01 = p$$

(the larger q reflects the risk premium).

- P is estimated from historical data whereas Q is calibrated to current market prices.
- Risk-neutral valuation at t of a claim H at T is done via the *risk-neutral pricing rule*

$$V_0^H = \mathbb{E}_{Q,t}[e^{-r(T-t)} H], \quad t < T,$$

where $\mathbb{E}_{Q,t}[\cdot]$ denotes expectation w.r.t. Q given the information up to and including time t .

- Risk-neutral pricing applied to non-traded financial products is a typical example of level 2 valuation: Prices of traded securities are used to calibrate model parameters under the risk-neutral measure Q ; this measure is then used to price the non-traded products.

- There are two theoretical justifications for risk-neutral pricing:
 - ▶ (*First*) *Fundamental Theorem of Asset Pricing*: A model for security prices is **arbitrage free if and only if it admits at least one EMM Q** .
 - ▶ In financial models it is often possible to replicate the pay-off of a product by trading in the assets, a practice known as (*dynamic hedging*), and it is well-known that **in a frictionless market the cost of hedging is given by the risk-neutral pricing rule**.

Example 2.5 (European call option continued)

- Suppose that options with our desired strike K and/or maturity time T are **not traded**, but that other options on the same stock are traded.
- Under P the **stock price (S_t)** is assumed to follow a **geometric Brownian motion (GBM)** (the so-called *Black–Scholes model*) with dynamics

$$dS_t = \mu S_t dt + \sigma S_t dW_t$$

for constants $\mu \in \mathbb{R}$ (the drift) and $\sigma > 0$ (the volatility), and a standard Brownian motion (W_t).

- It is well known that there is an EMM Q under which $(e^{-rt}S_t)$ is a martingale; under Q , S_t follows a GBM with drift r and volatility σ . The European call option payoff is $H = \max\{S_T - K, 0\}$ and the risk-neutral valuation formula may be shown to be

$$V_t = E_t^Q(e^{-r(T-t)}(S_T - K)^+) = C^{\text{BS}}(t, S_t; r, \sigma, K, T), \quad t < T; \quad (3)$$

where t, S_t, r, K, T are known.

- We would typically use quoted prices $C^{\text{BS}}(t, S_t; r, \sigma, K^*, T^*)$ for options on the stock with different K^*, T^* to infer the unknown σ and then plug this so-called *implied volatility* into (3).

2.2.3 Loss distributions

From Example 2.2 we can identify the following key tasks of QRM:

- 1) Find a statistical model for \mathbf{X}_{t+1} (typically an estimated projection model used to forecast \mathbf{X}_{t+1} ; can also be a valuation model, see, e.g., Black–Scholes formula);
- 2) Compute/Derive the df $F_{L_{t+1}}$ (requires the df of $f(t+1, \mathbf{Z}_t + \mathbf{X}_{t+1})$);
- 3) Compute a risk measure from $F_{L_{t+1}}$.

There are three general methods to approach the challenges 1) and 2).

1) Analytical method

Idea: Choose $F_{\mathbf{X}_{t+1}}$ and f such that $F_{L_{t+1}}$ can be determined explicitly.

The prime example is the variance-covariance method; see RiskMetrics (1996):

Assumption 1 $\boldsymbol{X}_{t+1} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (e.g., if (Z_t) is a Brownian motion, (S_t) a geometric Brownian motion)

Assumption 2 $F_{L_{t+1}^\Delta}$ is a good approximation to $F_{L_{t+1}}$.

$$L_{t+1}^\Delta = -(c_t + \boldsymbol{b}_t^\top \boldsymbol{X}_{t+1}) \stackrel{\text{Ass. 1}}{\Rightarrow} L_{t+1}^\Delta \sim N(-c_t - \boldsymbol{b}_t^\top \boldsymbol{\mu}, \boldsymbol{b}_t^\top \boldsymbol{\Sigma} \boldsymbol{b}_t).$$

Advantages:

- $F_{L_{t+1}}$ explicit (\Rightarrow typically explicit risk measures)
- (Typically) easy to implement

Drawbacks:

Assumptions. Especially Assumption 1 is unlikely to be realistic for daily (probably also weekly/monthly) data. **Stylized facts** about risk-factor changes (see later)) suggest that $F_{\boldsymbol{X}_{t+1}}$ is *leptokurtic*, i.e., thinner body and heavier tail than $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
 $\Rightarrow \boldsymbol{X}_{t+1} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ underestimates the tail of $F_{L_{t+1}}$ and thus risk measures such as VaR.

Remark 2.6

- We have not talked about how to estimate $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ yet.

- When dynamic models for \mathbf{X}_{t+1} are considered (e.g., time series models), different estimation methods are possible depending on whether we focus on conditional distributions $F_{\mathbf{X}_{t+1} | (\mathbf{X}_s)_{s \leq t}}$ or the equilibrium distribution $F_{\mathbf{X}}$ in a stationary model.

2) Historical simulation

Idea: Estimate $F_{L_{t+1}}$ by its *empirical distribution function* based on the past risk-factor changes $\mathbf{X}_{t-n+1}, \dots, \mathbf{X}_t$, so

$$F_{L_{t+1}}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\tilde{L}_{t-i+1} \leq x\}}, \quad x \in \mathbb{R},$$

where $\tilde{L}_k = L(\mathbf{X}_k) = -(f(t+1, \mathbf{Z}_t + \mathbf{X}_k) - f(t, \mathbf{Z}_t))$.

The values $\tilde{L}_{t-n+1}, \dots, \tilde{L}_t$ show what would happen to the current portfolio if the risk-factor changes in periods $k \in \{t-n+1, \dots, t\}$ were to recur.

Advantages:

- Easy to implement

- No estimation of the unknown distribution of X_{t+1} required

Drawbacks:

- Sufficient (synchronized) data for all risk-factor changes required
- Only considers past losses (“driving a car by looking in the back mirror”)

3) Monte Carlo method

Idea: Take any (adequate) model for X_{t+1} , simulate from it, compute the corresponding simulated losses and $F_{L_{t+1}}$ via the empirical df.

Advantages:

- Quite general (applicable for any model of X_{t+1} which is easy to sample)

Drawbacks:

- Unclear how to find an appropriate model for X_{t+1} (any result is only as good as the chosen $F_{X_{t+1}}$)

- **Computational cost** (every simulation requires to evaluate the portfolio; expensive, e.g., if the latter contains derivatives which are priced via Monte Carlo themselves
⇒ Nested Monte Carlo simulations)

Remark 2.7

- So-called *economic scenario generators* (i.e., economically motivated dynamic models for the evolution and interaction of different risk factors) used in insurance also fall under the heading of Monte Carlo methods.
- Furthermore, there are methods from *extreme value theory* based on approximations of the **tail of the loss df** $F_{L_{t+1}}$ (see later).

2.3 Risk measurement

Definition 2.8 (Risk measure)

A *risk measure* for a financial position with (random) loss L is a **real number** which measures the “riskiness of L ”. It can be interpreted as the **amount of capital** required (today) to **account for future actual losses** (realizations of L) in that position.

- Alternatively, ... the amount of **capital required to make a position with loss L acceptable** to an (internal/external) regulator (> 0 if and only if **not acceptable**; equivalent to the amount of money to put aside now).
- Some **reasons for using risk measures** in practice:
 - ▶ To determine the amount of **capital** to hold as a **buffer against unexpected future losses** on a portfolio (in order to satisfy a regulator/manager concerned with the institution's solvency).

- ▶ By management, as a tool for **limiting** the amount of **risk** of a **business unit** (e.g., by requiring that the daily 95% Value-at-Risk (i.e., the 95%-quantile) of a trader's position should not exceed a given bound).
- ▶ To determine the **riskiness** (and **thus fair premium**) of an **insurance contract**.

2.3.1 Approaches to risk measurement

Existing approaches to measuring risk can be grouped into **three categories**:

1) Notional-amount approach

- oldest approach
- “standardized approaches” of Basel II (e.g., OpRisk) still use it
- **risk of a portfolio** \mathcal{P} : $\sum_{\text{securities in } \mathcal{P}}$ “notional value of the security”
 - “riskiness factor of the corresponding asset class”

- Advantages:
 - ▶ simplicity
- Drawbacks:
 - ▶ No differentiation between long and short positions and no netting: the risk of a long position in corporate bonds hedged by an offsetting position in credit default swaps is counted as twice the risk of the unhedged bond position.
 - ▶ No diversification benefits: risk of a portfolio of loans to many companies = risk of a portfolio where the whole amount is lent to a single company.
 - ▶ Problems for portfolios of derivatives: notional amount of the underlying can widely differ from the economic value of the derivative position.

2) Risk measures based on loss distributions

- Most modern risk measures are **characteristics** of the underlying (conditional or unconditional) **loss distribution** over some predetermined time horizon Δt .
- Examples: variance, **Value-at-Risk**, **expected shortfall** (see later)
- **Advantages:**
 - ▶ The concept of a loss distribution **makes sense on all levels** of aggregation (from single portfolios to the overall position of a financial institution).
 - ▶ If estimated properly, loss distributions **reflect netting** and **diversification effects**.
- Drawbacks:**
 - ▶ **Estimates** of loss distributions **are typically based on past data**.
 - ▶ It is **difficult to estimate loss distributions** accurately (especially for large portfolios).

⇒ Risk measures should be complemented by information from scenarios (forward-looking).

3) Scenario-based risk measures

- This approach to risk measurement is typically considered in stress testing.
- One considers possible future risk-factor changes (*scenarios*; e.g., a 20% drop in a market index).
- *Risk of a portfolio* = maximum (weighted) loss of the portfolio under all scenarios.
- If $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ denote the risk-factor changes (*scenarios*) with corresponding weights $\boldsymbol{w} = (w_1, \dots, w_n)$, the risk is

$$\psi_{\mathcal{X}, \boldsymbol{w}} = \max_{1 \leq i \leq n} \{w_i L(\boldsymbol{x}_i)\}, \quad (4)$$

where $L(\cdot)$ is the loss operator. Many risk measures used in practice are of the form (4); see, e.g., *CME SPAN: Standard Portfolio Analysis of Risk* (2010).

- Mathematical interpretation of (4):
 - ▶ Assume $L(\mathbf{0}) = 0$ (\checkmark if Δt small) and $w_i \in [0, 1]$, $i \in \{1, \dots, n\}$.
 - ▶ $w_i L(\mathbf{x}_i) = \mathbb{E}_{\mathbb{P}_i}[L(\mathbf{X}_i)]$ where $\mathbf{X}_i \sim \mathbb{P}_i = w_i \delta_{\mathbf{x}_i} + (1 - w_i) \delta_{\mathbf{0}}$ (δ_x the Dirac measure at x) is a probability measure on \mathbb{R}^d .

Therefore, $\psi_{\mathcal{X}, \mathbf{w}} = \max\{\mathbb{E}_{\mathbb{P}}[L(\mathbf{X})] : \mathbf{X} \sim \mathbb{P} \in \{\mathbb{P}_1, \dots, \mathbb{P}_n\}\}$. Such a risk measure is known as *generalized scenario*; they play an important role in the theory of coherent risk measures.

- Advantages:
 - ▶ Useful for portfolios with **few risk factors**.
 - ▶ Useful **complementary information** to risk measures based on loss distributions (past data).

Drawbacks: ▶ **Determining scenarios and weights**.

2.3.2 Value-at-Risk

One possible risk measure is the **maximum loss** $\inf\{x \in \mathbb{R} : F_L(x) = 1\}$. However, this is ∞ for most distributions of interest and neglects any probabilistic information. Idea of Value-at-Risk: replace “maximum loss” by “**maximum loss not exceeded with a given high probability**”.

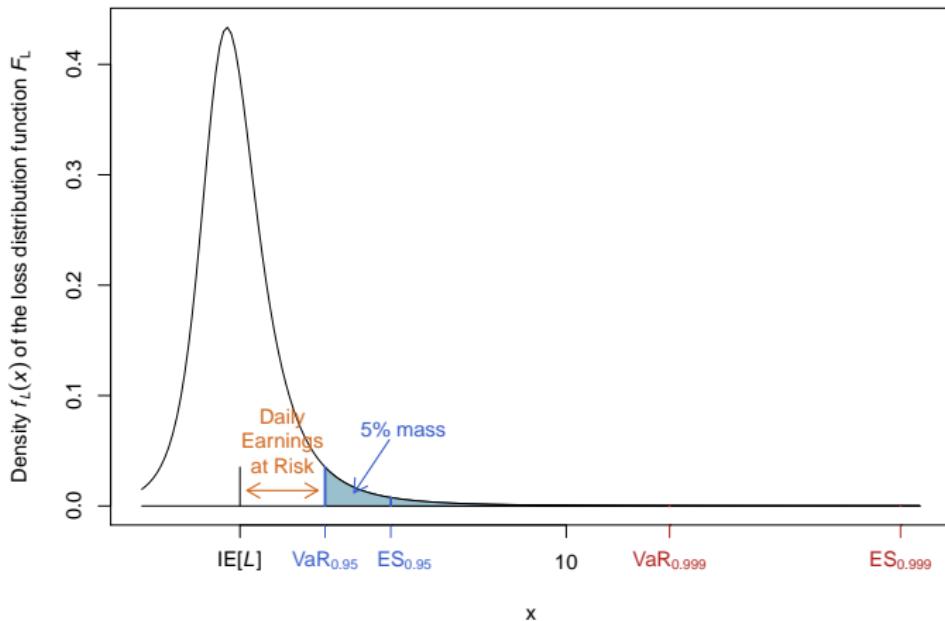
Definition 2.9 (Value-at-Risk)

For a loss $L \sim F_L$, **Value-at-Risk (VaR)** at confidence level $\alpha \in (0, 1)$ is defined by $\text{VaR}_\alpha = \text{VaR}_\alpha(L) = F_L^-(\alpha) = \inf\{x \in \mathbb{R} : F_L(x) \geq \alpha\}$.

- VaR_α is simply the **α -quantile of F_L** . As such, $F_L(x) < \alpha$ for all $x < \text{VaR}_\alpha(L)$ and $F_L(\text{VaR}_\alpha(L)) = F_L(F_L^-(\alpha)) \geq \alpha$.
- Known since 1994: Weatherstone 4¹⁵ report (J.P. Morgan; RiskMetrics)
- VaR is the **most widely used risk measure** (suggested by Basel II)
- $\text{VaR}_\alpha(L)$ also depends on the **estimator** of F_L and the **time horizon**.

- VaR is not a what if risk measure: $\text{VaR}_\alpha(L)$ does not provide information about the severity of losses which occur with probability $\leq 1 - \alpha$ (only about the loss frequency).

VaR and ES for a skew t_3 distribution



Example 2.10 (VaR for $N(\mu, \sigma^2)$, $t_\nu(\mu, \sigma^2)$, $\text{Par}(\theta)$)

1) Let $L \sim N(\mu, \sigma^2)$. Then $F_L(x) = \mathbb{P}(L \leq x) = \mathbb{P}((L - \mu)/\sigma \leq (x - \mu)/\sigma) = \Phi((x - \mu)/\sigma)$. This implies that

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

2) Let $L \sim t_\nu(\mu, \sigma^2)$, so $(L - \mu)/\sigma \sim t_\nu$ and thus, as above,

$$\text{VaR}_\alpha(L) = \mu + \sigma t_\nu^{-1}(\alpha).$$

Note that $X \sim t_\nu = t_\nu(0, 1)$ has density $f_X(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1 + x^2/\nu)^{-\frac{\nu+1}{2}}$, $\mathbb{E}X = 0$ (if $\nu > 1$) and $\text{Var } X = \frac{\nu}{\nu-2}$ (if $\nu > 2$).

3) Let $L \sim \text{Par}(\theta)$, $\theta > 0$, so $L \sim F_L(x) = 1 - x^{-\theta}$, $x \geq 1$. Then

$$\text{VaR}_\alpha(L) = (1 - \alpha)^{-1/\theta}.$$

Choices of parameters $\Delta t, \alpha$:

- Δt should reflect the time period over which the portfolio is held (unchanged) (e.g., insurance companies: $\Delta t = 1 \text{ y}$)
- Δt should be relatively small (more risk-factor change data is available).
- Typical choices:
 - ▶ For limiting traders: $\alpha = 0.95$, $\Delta t = 1 \text{ d}$
 - ▶ According to Basel II:
 - Market risk: $\alpha = 0.99$, $\Delta t = 10 \text{ d}$ (2 trading weeks)
 - Credit risk and operational risk: $\alpha = 0.999$, $\Delta t = 1 \text{ y}$
 - Economic capital: $\alpha = 0.9997$, $\Delta t = 1 \text{ y}$
 - ▶ According to Solvency II: $\alpha = 0.995$, $\Delta t = 1 \text{ y}$
- Backtesting often needs to be carried out at lower confidence levels in order to have sufficient statistical power to detect poor models.

- Be cautious with strict interpretations of $\text{VaR}_\alpha(L)$ and other risk measures, there is typically **considerable model/liquidity risk** behind.

Interlude: Generalized inverses

$T \nearrow$ means that T is *increasing*, i.e., $T(x) \leq T(y)$ for all $x < y$.

Definition 2.11 (Generalized inverse)

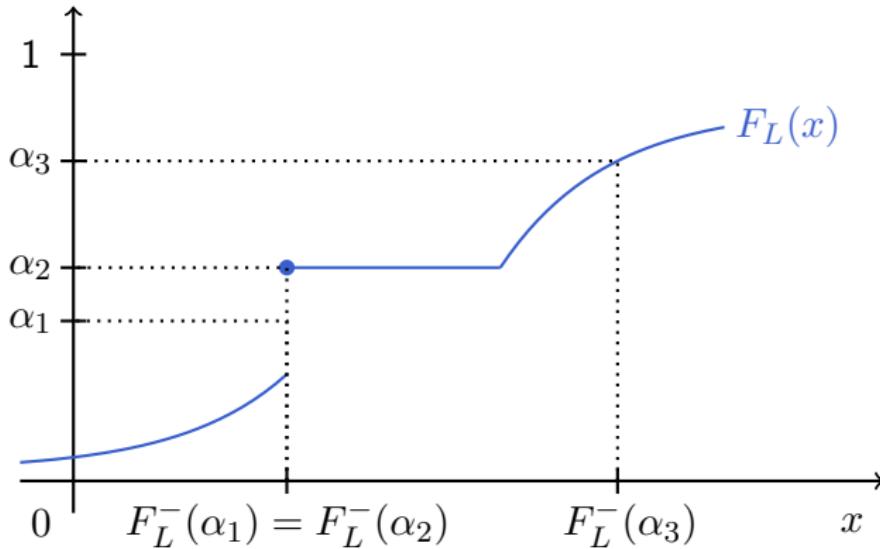
For any increasing function $T : \mathbb{R} \rightarrow \mathbb{R}$, with $T(-\infty) = \lim_{x \downarrow -\infty} T(x)$ and $T(\infty) = \lim_{x \uparrow \infty} T(x)$, the *generalized inverse* $T^- : \mathbb{R} \rightarrow \bar{\mathbb{R}} = [-\infty, \infty]$ of T is defined by

$$T^-(y) = \inf\{x \in \mathbb{R} : T(x) \geq y\}, \quad y \in \mathbb{R},$$

with the convention that $\inf \emptyset = \infty$. If T is a df, $T^- : [0, 1] \rightarrow \bar{\mathbb{R}}$ is the *quantile function* of T .

- If T is continuous and \uparrow , then $T^- \equiv T^{-1}$ (ordinary inverse).

- There are rules for working with T^- (similar to T^{-1}), see Embrechts and Hofert (2013a).
- F_L^- visualized (here: for a df F_L):



2.3.3 VaR in risk capital calculations

1) VaR in regulatory capital calculations for the trading book

For banks using the *internal model (IM)* approach for market risk in Basel II, the daily **risk capital formula** is

$$RC^t = \max \left\{ \text{VaR}_{0.99}^{t,10}, \frac{k}{60} \sum_{i=1}^{60} \text{VaR}_{0.99}^{t-i+1,10} \right\} + c.$$

- $\text{VaR}_{\alpha}^{s,10}$ denotes the 10-day VaR_{α} calculated at day s ($t = \text{today}$).
- $k \in [3, 4]$ is a multiplier (or *stress factor*).
- $c = \text{stressed VaR charge}$ (calculated from data from a volatile market period) + *incremental risk charge (IRC)*; $\text{VaR}_{0.999}$ -estimate of the annual distribution of losses due to defaults and downgrades) + *charges for specific risks*.

The averaging tends to lead to smooth changes in the capital charge over time unless $\text{VaR}_{0.99}^{t,10}$ is large.

2) The Solvency Capital Requirement in Solvency II

The *Solvency Capital Requirement (SCR)* is the amount of capital that enables the insurer to meet its obligations over $\Delta t = 1 \text{ y}$ with $\alpha = 0.995$.

Let $V_t = A_t - B_t$ (assets – liabilities; aka *own funds*) denote the equity capital. The insurer wants to determine the minimum amount of extra capital x_0 to put aside to be solvent in Δt with probability (\geq) α . So

$$\begin{aligned}x_0 &= \inf\{x \in \mathbb{R} : \mathbb{P}(V_{t+1} + x(1+r) \geq 0) \geq \alpha\} \\&= \inf\left\{x \in \mathbb{R} : \mathbb{P}\left(-\left(\frac{V_{t+1}}{1+r} - V_t\right) \leq x + V_t\right) \geq \alpha\right\} \\&= \inf\{x \in \mathbb{R} : \mathbb{P}(L_{t+1} \leq x + V_t) \geq \alpha\} \\&= \inf\{x \in \mathbb{R} : F_{L_{t+1}}(x + V_t) \geq \alpha\} \\&= \inf\{z - V_t \in \mathbb{R} : F_{L_{t+1}}(z) \geq \alpha\} = \text{VaR}_\alpha(L_{t+1}) - V_t\end{aligned}$$

and thus $\text{SCR} = V_t + x_0 = \text{VaR}_\alpha(L_{t+1})$ (available capital now + capital required to be solvent in Δt with probability (\geq) α). For a well-capitalized company ($x_0 \leq 0$), $-x_0$ (= own funds – SCR $\text{VaR}_\alpha(L_{t+1})$)

is called the *excess capital*.

3) Median shortfall

The more robust alternative to expected shortfall (see later) *median shortfall* ($\text{MS}_\alpha(L) = F_{L,\alpha}^-(1/2)$ where $F_{L,\alpha}(x) = \frac{F_L(x)-\alpha}{1-\alpha} \mathbb{1}_{\{x \geq F_L^-(\alpha)\}}$) is just $\text{VaR}_{\frac{1+\alpha}{2}}$.

Watch out for (badly defined) VaR

The “bible” on VaR is Jorion (2007). The following “definition” is very common:

“VaR is the *maximum* expected loss of a portfolio over a given time horizon with a certain confidence level.”

It is however mathematically *meaningless* and potentially *misleading*. In *no sense*, VaR is a *maximum loss*! We can lose more, sometimes much more, depending on the *heaviness of the tail* of the loss distribution.

2.3.4 Other risk measures based on loss distributions

1) Variance

- $\text{Var}[L]$ is historically the dominating risk measure in finance (due to Markowitz)
- Drawbacks:
 - ▶ $\mathbb{E}[L^2] < \infty$ required (not justifiable for non-life insurance or operational risk)
 - ▶ no distinction between positive/negative deviations from the mean (Var is only a good risk measure for F_L (approx.) symmetric around $\mathbb{E}L$, but F_L is typically skewed in credit and operational risk)

2) Upper partial moments

- Risk management is mainly concerned with the upper tail of F_L .

- Given an exponent $k \geq 0$ and a reference point q , the *upper partial moment* is defined by

$$\text{UPM}(k, q) = \int_q^\infty (x - q)^k dF_L(x).$$

- The larger k , the more conservative is this risk measure as more weight is put on large deviations from q .

3) Expected shortfall

Definition 2.12 (Expected shortfall)

For a loss $L \sim F_L$ with $\mathbb{E}|L| < \infty$, *expected shortfall (ES)* at confidence level $\alpha \in (0, 1)$ is defined by

$$\text{ES}_\alpha = \text{ES}_\alpha(L) = \frac{1}{1-\alpha} \int_\alpha^1 \text{VaR}_u(L) du. \quad (5)$$

- Besides VaR, ES is the *most important risk measure* in practice.

- ES_α is the average over VaR_u for all $u \geq \alpha$ (if F_L is continuous, ES_α is the average loss beyond VaR_α) $\Rightarrow \text{ES}_\alpha \geq \text{VaR}_\alpha$
- ES_α looks further into the tail of F_L , it is a what if risk measure (VaR_α is frequency-based; ES_α is severity-based).
- Due to considering the tail of F_L , ES_α is more difficult to estimate and backtest than VaR_α (larger sample size required).
- $\text{ES}_\alpha(L) < \infty$ requires $\mathbb{E}|L| < \infty$ (can be violated for OpRisk).
- Subadditivity and elicability
 - ▶ In contrast to VaR_α , ES_α is subadditive (see later)
 - ▶ In contrast to ES_α (see Gneiting (2011) or Kou and Peng (2014)), VaR_α is elicitable (and also exists if $\mathbb{E}|L| = \infty$)
 - ▶ Concerning going from VaR_α to ES_α , see BIS (2012, p. 41, Question 8).

- A risk measure ρ is *elicitable* w.r.t. a class of dfs \mathcal{F} if there exists a forecasting objective function $S : \mathbb{R}^2 \rightarrow \mathbb{R}$ such that

$$\rho(L) = \operatorname{arginf}_x \int_{\mathbb{R}} S(x, y) dF_L(y), \quad \forall F_L \in \mathcal{F} \quad (6)$$

(e.g., $S(x, y) = (x - y)^2 \Rightarrow \rho(L) = \mathbb{E}L$; $S(x, y) = |x - y| \Rightarrow \rho(L) = \operatorname{med}(L) = F_L^-(1/2)$). Not being *elicitable* implies that it is difficult/impossible to correctly compare models or optimize/minimize error functionals of type (6).

Proposition 2.13 (ES formulas)

Let $(x)_+ = \max\{x, 0\}$. For $\alpha \in (0, 1)$,

- 1) $\operatorname{ES}_\alpha(L) = \frac{\mathbb{E}[(L - F_L^-(\alpha))_+]}{1 - \alpha} + F_L^-(\alpha);$
- 2) $\operatorname{ES}_\alpha(L) = \frac{\mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}]}{1 - \alpha} + \frac{F_L^-(\alpha)(1 - \alpha - \bar{F}_L(F_L^-(\alpha)))}{1 - \alpha}.$

Proof.

1) $L \stackrel{d}{=} F_L^-(U)$, $U \sim U[0, 1]$, since $\mathbb{P}(F_L^-(U) \leq x) = \mathbb{P}(U \leq F_L(x)) = F_L(x)$. Therefore,

$$\begin{aligned}\frac{\mathbb{E}[(L - F_L^-(\alpha))_+]}{1 - \alpha} &= \frac{1}{1 - \alpha} \int_0^1 (F_L^-(u) - F_L^-(\alpha))_+ du \\ &= \frac{1}{1 - \alpha} \int_\alpha^1 (F_L^-(u) - F_L^-(\alpha)) du \\ &= \text{ES}_\alpha(L) - F_L^-(\alpha).\end{aligned}$$

2) First note that

$$\begin{aligned}\mathbb{E}[(L - F_L^-(\alpha))_+] &= \mathbb{E}[(L - F_L^-(\alpha)) \mathbb{1}_{\{L > F_L^-(\alpha)\}}] \\ &= \mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}] - F_L^-(\alpha) \mathbb{E}[\mathbb{1}_{\{L > F_L^-(\alpha)\}}] \\ &= \mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}] - F_L^-(\alpha) \bar{F}_L(F_L^-(\alpha)).\end{aligned}$$

Now apply 1), divide by $1 - \alpha$ and add $F_L^-(\alpha)$.

□

Corollary 2.14 (ES formulas under continuous F_L)

Let F_L be continuous. Then

$$1) \text{ ES}_\alpha(L) = \frac{\mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}]}{1-\alpha}$$

$$2) \text{ ES}_\alpha(L) = \mathbb{E}[L \mid L > F_L^-(\alpha)] \text{ (i.e., conditional VaR (CVaR))}$$

Proof.

- 1) Since $\bar{F}_L(F_L^-(\alpha)) = 1 - F_L(F_L^-(\alpha)) = 1 - \alpha$ for all $\alpha \in \text{ran } F_L \cup \{\inf F_L, \sup F_L\} \supseteq (0, 1)$, the claim follows from Proposition 2.13 2).
- 2) First note that

$$\begin{aligned} F_{L|L>F_L^-(\alpha)}(x) &= \mathbb{P}(L \leq x \mid L > F_L^-(\alpha)) = \frac{\mathbb{P}(F_L^-(\alpha) < L \leq x)}{\mathbb{P}(L > F_L^-(\alpha))} \\ &= \frac{F_L(x) - F_L(F_L^-(\alpha))}{1 - F_L(F_L^-(\alpha))} \mathbb{1}_{\{x > F_L^-(\alpha)\}} = \frac{F_L(x) - \alpha}{1 - \alpha} \mathbb{1}_{\{x > F_L^-(\alpha)\}}, \end{aligned}$$

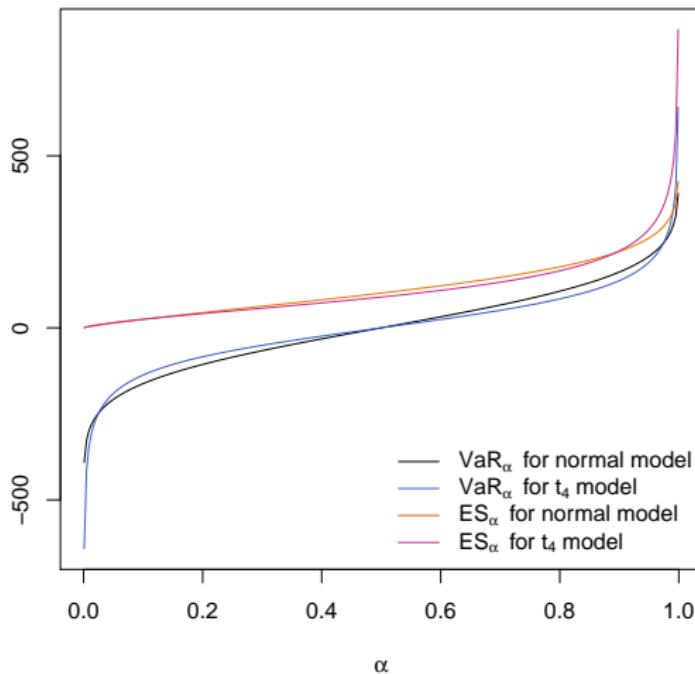
where the latter equality holds since $\alpha \in \text{ran } F_L$. This implies

$$\begin{aligned}\mathbb{E}[L \mid L > F_L^-(\alpha)] &= \int_{\mathbb{R}} x dF_{L|L>F_L^-(\alpha)}(x) = \int_{F_L^-(\alpha)}^{\infty} x \frac{dF_L(x)}{1-\alpha} \\ &= \frac{\mathbb{E}[L \mathbb{1}_{\{L>F_L^-(\alpha)\}}]}{1-\alpha} = \text{ES}_\alpha(L).\end{aligned}\quad \square$$

Example 2.15 (VaR and ES for stock returns)

- Consider a portfolio consisting of a single stock $V_t = S_t = 10\,000$. Example 2.2 implies that $L_{t+1}^\Delta = -V_t X_{t+1}$, where $X_{t+1} = \log(S_{t+1}/S_t)$.
- Let $\sigma = 0.2/\sqrt{250}$ (annualized volatility of 20%) and assume
 - 1) $X_{t+1} \sim N(0, \sigma^2) \Rightarrow L_{t+1}^\Delta \sim N(0, V_t^2 \sigma^2);$
 - 2) $X_{t+1} \sim t_4(0, \sigma^2 \frac{\nu-2}{\nu})$ ($\text{Var } X_{t+1} = \sigma^2$) or $X_{t+1} = \sqrt{\sigma^2 \frac{\nu-2}{\nu}} Y$ for $Y \sim t_4 \Rightarrow L_{t+1}^\Delta = -V_t \sqrt{\sigma^2 \frac{\nu-2}{\nu}} Y \sim t_4(0, V_t^2 \sigma^2 \frac{\nu-2}{\nu})$ ($\Rightarrow \text{Var}[L_{t+1}^\Delta] = V_t^2 \sigma^2$).

- Note that $ES_{\alpha}^{\text{normal}} \leq ES_{\alpha}^{t_4}$ for all α , but $VaR_{\alpha}^{\text{normal}} \leq VaR_{\alpha}^{t_4}$; in particular, the t_4 model is not always “riskier” than the normal model when VaR_{α} is used as a risk measures.



Example 2.16 (Example 2.10 continued)

1) Let $\tilde{L} \sim N(0, 1)$. Then $VaR_\alpha(\tilde{L}) = 0 + 1 \cdot \Phi^{-1}(\alpha)$ and thus

$$ES_\alpha(\tilde{L}) = \frac{1}{1 - \alpha} \int_{\alpha}^1 \Phi^{-1}(u) du \stackrel{x = \Phi^{-1}(u)}{=} \frac{1}{1 - \alpha} \int_{\Phi^{-1}(\alpha)}^{\infty} x \varphi(x) dx,$$

where $\varphi(x) = \Phi'(x) = \exp(-x^2/2)/\sqrt{2\pi}$. Note that $x\varphi(x) = -\varphi'(x)$, so that

$$ES_\alpha(\tilde{L}) = \frac{-[\varphi(x)]_{\Phi^{-1}(\alpha)}^{\infty}}{1 - \alpha} = \frac{-(0 - \varphi(\Phi^{-1}(\alpha)))}{1 - \alpha} = \frac{\varphi(\Phi^{-1}(\alpha))}{1 - \alpha}.$$

This implies that $L \sim N(\mu, \sigma^2)$ has expected shortfall

$$ES_\alpha(L) = \mu + \sigma ES_\alpha(\tilde{L}) = \mu + \sigma \frac{\varphi(\Phi^{-1}(\alpha))}{1 - \alpha}.$$

By l'Hôpital's Rule (case "0/0") and using $\varphi'(x) = -x\varphi(x)$, one can show that

$$1 \stackrel{\checkmark}{\leq} \lim_{\alpha \uparrow 1} \frac{ES_\alpha(L)}{VaR_\alpha(L)} = 1.$$

2) Benchmark model in finance

Let $L \sim t_\nu(\mu, \sigma^2)$, $\nu > 1$. Similarly as above, one obtains that

$$\text{ES}_\alpha(L) = \mu + \sigma \frac{f_{t_\nu}(t_\nu^{-1}(\alpha))(\nu + t_\nu^{-1}(\alpha)^2)}{(1 - \alpha)(\nu - 1)},$$

where f_{t_ν} denotes the density of t_ν (see Example 2.10). Again by l'Hôpital's Rule (case "0/0"), one can show that

$$1 \stackrel{\checkmark}{\leq} \lim_{\alpha \uparrow 1} \frac{\text{ES}_\alpha(L)}{\text{VaR}_\alpha(L)} = \frac{\nu}{\nu - 1} > 1 \quad (\text{and } \uparrow \infty \text{ for } \nu \downarrow 1).$$

In finance, often $\nu \in (3, 5)$. With $\nu = 3$, $\text{ES}_\alpha(L)$ is 50% larger than $\text{VaR}_\alpha(L)$ (in the limit for large α).

3) If $L \sim \text{Par}(\theta)$, $\theta > 1$, then $\text{VaR}_\alpha(L) = (1 - \alpha)^{-1/\theta}$, which implies

$$\text{ES}_\alpha(L) = \frac{\theta}{\theta - 1} \text{VaR}_\alpha(L)$$

and thus

$$1 \stackrel{\checkmark}{\leq} \lim_{\alpha \uparrow 1} \frac{\text{ES}_\alpha(L)}{\text{VaR}_\alpha(L)} = \frac{\theta}{\theta - 1} > 1 \quad (\text{and } \uparrow \infty \text{ for } \theta \downarrow 1).$$

Conclusion:

For losses with *heavy* (power-like) tails, the difference between using VaR and ES as risk measures for computing risk capital can be huge (for large α as required by Basel II).

2.3.5 Coherent and convex risk measures

- Artzner et al. (1999) (coherent risk measures) and Föllmer and Schied (2002) (convex risk measures) propose **axioms** a good risk measure should have.
- Here we assume that risk measures ρ are real-valued functions defined on a linear space of random variables \mathcal{M} (including constants).
- There are two possible interpretations of elements of \mathcal{M} :
 - 1) Future net asset values of portfolios/positions
Elements of \mathcal{M} are V_{t+1} ; a risk measure $\tilde{\rho}(V_{t+1})$ denotes the amount

of **additional capital** that needs to be added to a position with future net asset value V_{t+1} to make it acceptable to a regulator.

- 2) Losses L (related to 1) by $L = -(V_{t+1} - V_t)$) Elements of \mathcal{M} are losses L ; a risk measure $\rho(L)$ denotes the **total amount of equity capital** necessary to back a position with loss L .
- 1) and 2) are related via $\rho(L) = V_t + \tilde{\rho}(V_{t+1})$ (total capital = available capital + additional capital). In what follows, we focus on 2).

Axiom 1 (**monotonicity**) $L_1, L_2 \in \mathcal{M}, L_1 \leq L_2$ (a.s.) $\Rightarrow \rho(L_1) \leq \rho(L_2)$

Interpr.: Positions which lead to a higher loss in every state of the world require more risk capital.

Criticism: none

Axiom 2 (**translation invar.**) $\rho(L + l) = \rho(L) + l$ for all $L \in \mathcal{M}, l \in \mathbb{R}$

Interpr.: By adding $l \in \mathbb{R}$ to a position with loss L , we alter the capital requirements accordingly. If $\rho(L) > 0$, and $l = -\rho(L)$, then $\rho(L - \rho(L)) = \rho(L + l) = \rho(L) + l = 0$ so that adding $\rho(L)$ to a position with loss L makes it acceptable.

Criticism: Most people believe this to be reasonable; exception:
B. Rémillard (adding a constant value does not make a position riskier)

Axiom 3 (**subadditivity**) $\rho(L_1 + L_2) \leq \rho(L_1) + \rho(L_2)$ for all $L_1, L_2 \in \mathcal{M}$

Interpr.:

- Reflects the idea that risk can be reduced by **diversification**
- Using a non-subadditive ρ encourages institutions to legally break up into subsidiaries to reduce regulatory capital requirements.

- Subadditivity makes decentralization possible: if we want to bound the overall loss $L = L_1 + L_2$ of two positions by M , we can choose M_j such that $L_j \leq M_j$, $j \in \{1, 2\}$, with $M_1 + M_2 \leq M$ and require $\rho(L_j) \leq M_j$, $j \in \{1, 2\}$. Then $\rho(L) \leq_{\text{subadd.}} \rho(L_1) + \rho(L_2) \leq M_1 + M_2 \leq M$.

Criticism: VaR is ruled out in certain situations. Note that VaR is monotone ($L_1 \leq L_2$ (a.s.) $\Rightarrow F_{L_1}(x) \geq F_{L_2}(x)$, $x \in \mathbb{R} \Rightarrow F_{L_1}^-(u) \leq F_{L_2}^-(u)$, $u \in (0, 1)$), translation invariant ($F_{L+l}(x) = F_L(x - l) \Rightarrow F_{L+l}^-(u) = F_L^-(u) + l$, $u \in (0, 1)$) and positive homogeneous ($F_{\lambda L}(x) = F_L(x/\lambda) \Rightarrow F_{\lambda L}^-(u) = \lambda F_L^-(u)$), but in general not subadditive, especially not under one of the following scenarios (see below):

- 1) Independent, light-tailed L_1, L_2 and small α ;

- 2) L_1, L_2 have **skewed** distributions;
- 3) L_1, L_2 have **heavy tailed** distributions;
- 4) L_1, L_2 have **special dependence**.

Note that \mathcal{M} is important here. If it is sufficiently small (e.g., all multivariate elliptical distributions), VaR_α is subadditive (see later)!

Axiom 4 (**positive homogeneity**) $\rho(\lambda L) = \lambda\rho(L)$ for all $L \in \mathcal{M}$, $\lambda > 0$

Interpr.: $\lambda = n \in \mathbb{N}$, subadditivity $\Rightarrow \rho(nL) \leq n\rho(L)$. But n times the same loss L means no diversification, so equality should hold.

Criticism: If $\lambda > 0$ is **large**, **liquidity risk** plays a role and one should rather have $\rho(\lambda L) > \lambda\rho(L)$ (also to penalize concentration or risk), but this contradicts subadditivity. This has led to convex risk measures.

Definition 2.17 (Coherent risk measure)

A risk measure ρ is *coherent* if it satisfies Axioms 1–4 above.

Example 2.18 (Generalized scenario risk measures)

The generalized scenario risk measure $\psi_{\mathcal{X}, w}(L) = \max\{\mathbb{E}_{\mathbb{P}}[L(\mathbf{X})] : \mathbf{X} \sim \mathbb{P} \in \{\mathbb{P}_1, \dots, \mathbb{P}_n\}\}$ is *coherent*. Monotonicity, translation invariance, positive homogeneity are clear; for *subadditivity*, note that

$$\begin{aligned}\psi_{\mathcal{X}, w}(L_1 + L_2) &= \max\{\underbrace{\mathbb{E}_{\mathbb{P}}[L_1(\mathbf{X}) + L_2(\mathbf{X})]}_{=\mathbb{E}_{\mathbb{P}}[L_1(\mathbf{X})] + \mathbb{E}_{\mathbb{P}}[L_2(\mathbf{X})]}\} \\ &\leq \psi_{\mathcal{X}, w}(L_1) + \psi_{\mathcal{X}, w}(L_2),\end{aligned}$$

where $L_j(\mathbf{x})$ denotes the hypothetical loss of position j under scenario \mathbf{x} (risk-factor change). Note that *all coherent risk measures can be represented as generalized scenarios* via $\rho(L) = \sup\{\mathbb{E}_{\mathbb{P}}[L] : \mathbb{P} \in \mathcal{P}\}$ where \mathcal{P} is a set of probability measures; for a proof, see McNeil et al. (2005, Prop. 6.11 (ii)) for $|\Omega| < \infty$ and Delbaen (2000), Delbaen (2002) for the general case.

Example 2.19 (A coherent premium principle)

- Fischer (2003) proposed a class of coherent risk measures which are potentially useful for an insurance company that wants to compute premiums on a coherent basis without deviating too far from standard actuarial practice.
- Let $p > 1$, $\alpha \in [0, 1]$, $\mathcal{M} = L^p(\Omega, \mathcal{F}, \mathbb{P})$, $\|L\|_p = \mathbb{E}[|L|^p]^{1/p}$ and

$$\rho_{\alpha,p}(L) = \mathbb{E}[L] + \alpha \|\max\{L - \mathbb{E}[L], 0\}\|_p.$$

Risk = pure actuarial premium + *risk loading* (α -fraction of $(\int_{\mathbb{E}[L]}^{\infty} (x - \mathbb{E}[L])^p dF_L(x))^{1/p}$). The higher α or p , the more conservative is $\rho_{\alpha,p}(L)$.

- For *subadditivity* use $\max\{L_1 + L_2, 0\} \leq \max\{L_1, 0\} + \max\{L_2, 0\}$ and thus

$$\begin{aligned} & \|\max\{L_1 - \mathbb{E}[L_1] + L_2 - \mathbb{E}[L_2], 0\}\|_p \\ & \leq \|\max\{L_1 - \mathbb{E}[L_1], 0\} + \max\{L_2 - \mathbb{E}[L_2], 0\}\|_p \\ & \leq \|\max\{L_1 - \mathbb{E}[L_1], 0\}\|_p + \|\max\{L_2 - \mathbb{E}[L_2], 0\}\|_p. \end{aligned}$$

Minkowski

For **monotonicity**, let $L_1 \leq L_2$ a.s. and write $L = L_1 - L_2 (\leq 0)$. Thus, $\max\{L - \mathbb{E}[L], 0\} \leq \max\{0 - \mathbb{E}[L], 0\} = -\mathbb{E}[L]$ a.s., so $\|\max\{L - \mathbb{E}[L], 0\}\|_p \leq -\mathbb{E}[L]$. Since $\alpha \in [0, 1]$, $\rho_{\alpha,p}(L) \leq \mathbb{E}[L](1 - \alpha) \leq 0$. Using subadditivity, we obtain $\rho_{\alpha,p}(L_1) \leq \rho_{\alpha,p}(L) + \rho_{\alpha,p}(L_2) \leq \rho_{\alpha,p}(L_2)$. Translation invariance and positive homogeneity are trivial.

Definition 2.20 (Convex risk measure)

A risk measure ρ which is **monotone**, **translation invariant** and **convex** is called a ***convex risk measure***.

- Justification for their study is again diversification (but they don't have to be positive homogeneous).
- Let ρ be coherent. Then for all $\lambda \in [0, 1]$, $L_1, L_2 \in \mathcal{M}$, $\rho(\lambda L_1 + (1 - \lambda)L_2) \stackrel{\text{subadd.}}{\leq} \rho(\lambda L_1) + \rho((1 - \lambda)L_2) \stackrel{\text{pos.hom.}}{=} \lambda\rho(L_1) + (1 - \lambda)\rho(L_2)$ so ρ is convex.

The converse is not true in general, but for positive homogeneous risk measures, convexity and subadditivity are equivalent.

- Examples of convex but not positive homogeneous risk measures:
 - 1) Let $\rho'(L) = \rho(L) + 1$ for any coherent ρ .
 - 2) The *entropic risk measure* $\rho(L) = \mathbb{E}[e^{bL}]/b$, $b > 0$. To see that this is convex, use Young's inequality ($ab \leq a^p/p + b^q/q$ for all $a, b \geq 0$, $p, q \geq 1$ such that $1/p + 1/q = 1$) with $p = 1/\lambda$, $q = 1/(1 - \lambda)$, $a = e^{\lambda b L_1}$, $b = e^{(1-\lambda)b L_2}$.

Proposition 2.21 (Coherence of ES)

ES is a **coherent** risk measure.

Proof. Monotonicity, translation invariance and positive homogeneity follow from VaR. Subadditivity follows from Proposition 2.25 below. □

Proof of subadditivity of ES: A (mostly) analytic proof

We start with some auxiliary results.

Lemma 2.22

$\mathbb{P}(L = F_L^-(\alpha)) = 0$ implies $F_L(F_L^-(\alpha)) = \alpha$.

Proof. $F_L(F_L^-(\alpha)) - F_L(F_L^-(\alpha)-) = \mathbb{P}(L = F_L^-(\alpha)) = 0$, so F_L does not jump in $F_L^-(\alpha)$. By definition of F_L^- , $F_L(F_L^-(\alpha)) \geq \alpha$ and $F_L(F_L^-(\alpha)-) < \alpha$, which implies $F_L(F_L^-(\alpha)) = \alpha$. \square

For the following result let

$$\mathbb{1}_{\{L>q\}}^{(\alpha)} = \begin{cases} \mathbb{1}_{\{L>q\}}, & \text{if } \mathbb{P}(L = q) = 0, \\ \mathbb{1}_{\{L>q\}} + \frac{1-\alpha-\bar{F}_L(q)}{\mathbb{P}(L=q)} \mathbb{1}_{\{L=q\}}, & \text{if } \mathbb{P}(L = q) > 0. \end{cases}$$

Lemma 2.23 (Properties of $\mathbb{1}_{\{L>F_L^-(\alpha)\}}^{(\alpha)}$)

- 1) $\mathbb{1}_{\{L>F_L^-(\alpha)\}}^{(\alpha)} \in [0, 1]$
- 2) $\mathbb{E}[\mathbb{1}_{\{L>F_L^-(\alpha)\}}^{(\alpha)}] = 1 - \alpha$

Proof.

- 1) If $\mathbb{P}(L = F_L^-(\alpha)) = 0$ we are done, so consider $\mathbb{P}(L = F_L^-(\alpha)) > 0$. On the set of all $\omega \in \Omega$ such that $L(\omega) > F_L^-(\alpha)$, we are again done. Now consider all $\omega \in \Omega$ such that $L(\omega) = F_L^-(\alpha)$. Then $\mathbb{1}_{\{L>F_L^-(\alpha)\}}^{(\alpha)} = \frac{1-\alpha-\bar{F}_L(F_L^-(\alpha))}{\mathbb{P}(L=F_L^-(\alpha))}$. By definition, $F_L(F_L^-(\alpha)) \geq \alpha$, so $\bar{F}_L(F_L^-(\alpha)) \leq 1-\alpha$, thus $\mathbb{1}_{\{L>F_L^-(\alpha)\}}^{(\alpha)} \geq 0$. Also, $F_L(F_L^-(\alpha)-) < \alpha$, so $\mathbb{1}_{\{L>F_L^-(\alpha)\}}^{(\alpha)}$ equals $\frac{1-\alpha-(1-F_L(F_L^-(\alpha)))}{\mathbb{P}(L=F_L^-(\alpha))} = \frac{F_L(F_L^-(\alpha))- \alpha}{F_L(F_L^-(\alpha))-F_L(F_L^-(\alpha)-)} < 1$.

- 2) We have

$$\mathbb{E}[\mathbb{1}_{\{L>q\}}^{(\alpha)}] = \begin{cases} \bar{F}_L(q), & \text{if } \mathbb{P}(L = q) = 0, \\ \bar{F}_L(q) + \frac{1-\alpha-\bar{F}_L(q)}{\mathbb{P}(L=q)} \mathbb{P}(L = q) = 1 - \alpha, & \text{if } \mathbb{P}(L = q) > 0. \end{cases}$$

Consider $\mathbb{P}(L = q) = 0$. Since $q = F_L^-(\alpha)$, Lemma 2.22 implies that $\bar{F}_L(q) = 1 - F_L(F_L^-(\alpha)) = 1 - \alpha$. Thus $\mathbb{E}[\mathbb{1}_{\{L>q\}}^{(\alpha)}] = 1 - \alpha$. \square

Lemma 2.24 (Representation of ES_α in terms of $\mathbb{1}_{\{L > F_L^-(\alpha)\}}^{(\alpha)}$)

$$\text{ES}_\alpha(L) = \frac{\mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}^{(\alpha)}]}{1 - \alpha}$$

Proof.

- If $\mathbb{P}(L = F_L^-(\alpha)) = 0$, Lemma 2.22 implies that $\bar{F}_L(F_L^-(\alpha)) = 1 - \alpha$.
By Proposition 2.13 2) and since $\mathbb{P}(L = F_L^-(\alpha)) = 0$,

$$\begin{aligned}\text{ES}_\alpha(L) &= \frac{\mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}]}{1 - \alpha} + \frac{F_L^-(\alpha)(1 - \alpha - (1 - \alpha))}{1 - \alpha} \\ &= \frac{\mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}]}{1 - \alpha} = \frac{\mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}^{(\alpha)}]}{1 - \alpha}.\end{aligned}$$

- If $\mathbb{P}(L = F_L^-(\alpha)) > 0$, $\mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}^{(\alpha)}]$ equals

$$\mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}] + \frac{1 - \alpha - \bar{F}_L(F_L^-(\alpha))}{\mathbb{P}(L = F_L^-(\alpha))} \underbrace{\mathbb{E}[\textcolor{blue}{L} \mathbb{1}_{\{\textcolor{blue}{L} = \textcolor{blue}{F}_L^-(\alpha)\}}]}_{= \mathbb{E}[\textcolor{blue}{F}_L^-(\alpha) \mathbb{1}_{\{L = F_L^-(\alpha)\}}]} = \textcolor{blue}{F}_L^-(\alpha) \mathbb{P}(L = F_L^-(\alpha))$$

So, $\mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}^{(\alpha)}] = \mathbb{E}[L \mathbb{1}_{\{L > F_L^-(\alpha)\}}] + F_L^-(\alpha)(1 - \alpha - \bar{F}_L(F_L^-(\alpha)))$, which, by Proposition 2.13 2), equals $(1 - \alpha) \text{ES}_\alpha(L)$. \square

Proposition 2.25 (Subadditivity of ES)

ES_α is **subadditive** for all $\alpha \in (0, 1)$.

Proof. It suffices to show that

$$(1 - \alpha)(\text{ES}_\alpha(L_1) + \text{ES}_\alpha(L_2) - \text{ES}_\alpha(L_1 + L_2)) \geq 0.$$

Lemma 2.24 implies that

$$\begin{aligned} & \left(\sum_{j=1}^2 \mathbb{E}[L_j \mathbb{1}_{\{L_j > F_{L_j}^-(\alpha)\}}^{(\alpha)}] \right) - \mathbb{E}[(L_1 + L_2) \mathbb{1}_{\{L_1 + L_2 > F_{L_1 + L_2}^-(\alpha)\}}^{(\alpha)}] \\ & \stackrel{\text{Linearity}}{=} \sum_{j=1}^2 \mathbb{E}[L_j (\mathbb{1}_{\{L_j > F_{L_j}^-(\alpha)\}}^{(\alpha)} - \mathbb{1}_{\{L_1 + L_2 > F_{L_1 + L_2}^-(\alpha)\}}^{(\alpha)})]. \end{aligned} \quad (7)$$

- $L_j > F_{L_j}^-(\alpha) \Rightarrow \mathbb{1}_{\{L_j > F_{L_j}^-(\alpha)\}}^{(\alpha)} - \mathbb{1}_{\{L_1 + L_2 > F_{L_1 + L_2}^-(\alpha)\}}^{(\alpha)} = 1 - \dots \geq 0$
- $L_j < F_{L_j}^-(\alpha) \Rightarrow \mathbb{1}_{\{L_j > F_{L_j}^-(\alpha)\}}^{(\alpha)} - \mathbb{1}_{\{L_1 + L_2 > F_{L_1 + L_2}^-(\alpha)\}}^{(\alpha)} = 0 - \dots \leq 0$

In both cases, we make the expectations in (7) smaller by replacing L_j by $F_{L_j}^-(\alpha)$. Hence

$$(7) \geq \sum_{j=1}^2 F_{L_j}^-(\alpha) \underbrace{\mathbb{E}[\mathbb{1}_{\{L_j > F_{L_j}^-(\alpha)\}}^{(\alpha)} - \mathbb{1}_{\{L_1 + L_2 > F_{L_1 + L_2}^-(\alpha)\}}^{(\alpha)}]}_{\substack{= (1-\alpha) - (1-\alpha) = 0 \\ \text{Lem. 2.23 2)}}} \geq 0. \quad \square$$

Proof of subadditivity of ES: A (mostly) stochastic approach

Proposition 2.26 (Subadditivity of ES)

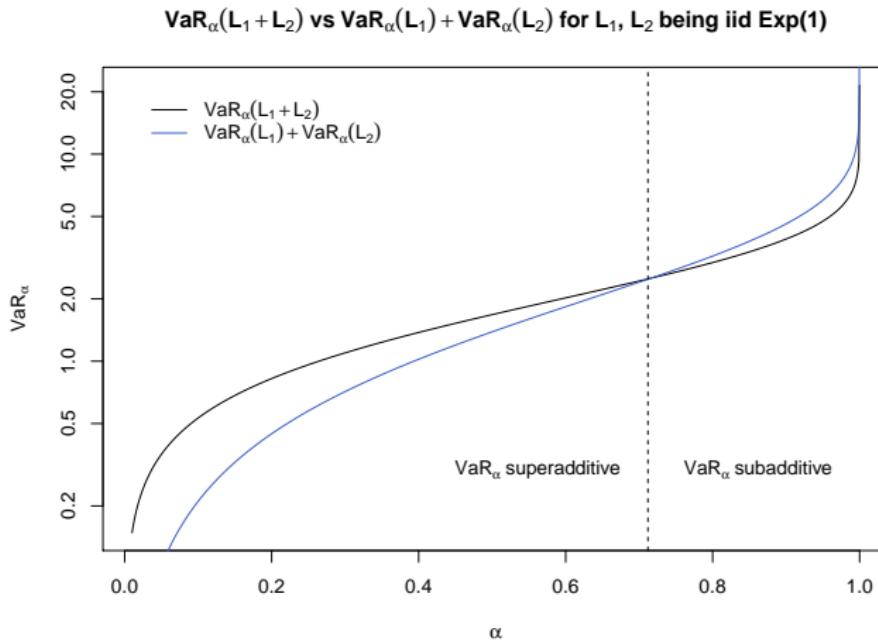
$$\text{ES}_\alpha(L) = \frac{\sup_{\{\tilde{L} \sim B(1,1-\alpha)\}} \mathbb{E}[L\tilde{L}]}{1-\alpha} \text{ (which, trivially, is subadditive).}$$

Proof (details become clear later). Let $L = F_L^-(U)$ for $U \sim U[0, 1]$ and $L' = \mathbb{1}_{\{U>\alpha\}} \sim B(1, 1 - \alpha)$. Then $\text{ES}_\alpha(L) = \frac{1}{1-\alpha} \int_\alpha^1 F_L^-(u) du = \frac{1}{1-\alpha} \int_0^1 F_L^-(u) \mathbb{1}_{\{u>\alpha\}} \cdot 1 du = \frac{1}{1-\alpha} \mathbb{E}[F_L^-(U) \mathbb{1}_{\{U>\alpha\}}] = \frac{1}{1-\alpha} \mathbb{E}[LL']$. Note that L and L' are comontone (see later), so that for any other $\tilde{L} \sim B(1, 1 - \alpha)$, Hoeffding's identity implies that $\mathbb{E}[L\tilde{L}] \leq \mathbb{E}[LL']$. Hence $\text{ES}_\alpha(L) = \sup_{\{\tilde{L} \sim B(1,1-\alpha)\}} \mathbb{E}[L\tilde{L}]/(1-\alpha)$. From this representation, ES_α is easily seen to be subadditive. □

Superadditivity scenarios for VaR

Exercise 2.27 (Independent L_1, L_2 and small α)

If $L_1, L_2 \stackrel{\text{ind.}}{\sim} \text{Exp}(1)$, VaR_α is superadditive $\iff \alpha < 0.71$.



Exercise 2.28 (Skewed loss distributions)

Consider a portfolio \mathcal{P} of two independent defaultable zero-coupon bonds with maturity $T = 1\text{y}$, nominal/face value 100, equal default probability $p = 0.009$, no recovery and interest rate 5%. Hence, for $j \in \{1, 2\}$, the loss of bond j (investor's/lender's perspective) is

$$L_j = \begin{cases} -5, & \text{with prob. } 1 - p = 0.991, \\ 100, & \text{with prob. } p = 0.009, \end{cases}$$

Set $\alpha = 0.99$. Since $\mathbb{P}(L_j < -5) = 0 < \alpha$ and $\mathbb{P}(L_j \leq -5) = 1 - p \geq \alpha$, $\text{VaR}_\alpha(L_j) = -5$, $j \in \{1, 2\}$. Since L_1, L_2 are independent, the loss $L = L_1 + L_2$ of \mathcal{P} is given by

$$L = \begin{cases} -10, & \text{with prob. } (1 - p)^2 = 0.982081, \\ 95, & \text{with prob. } 2p(1 - p) = 0.017838, \\ 200, & \text{with prob. } p^2 = 0.000081, \end{cases}$$

Since $\mathbb{P}(L < 95) < \alpha$ and $\mathbb{P}(L \leq 95) = 0.999919 \geq \alpha$, $\text{VaR}_\alpha(L) = 95 > -10 = \text{VaR}_\alpha(L_1) + \text{VaR}_\alpha(L_2)$. Hence VaR_α is superadditive. Note that

VaR_α punishes diversification since $\text{VaR}_\alpha(0.5L_1 + 0.5L_2) = \text{VaR}_\alpha(L)/2 = 47.5 > -5 = \text{VaR}_\alpha(L_1)$.

Another example of this type is the following.

Exercise 2.29 (Skewed loss distributions; extended example)

Consider d independent defaultable bonds with maturity $T = 1\text{y}$, nominal/face value $b > 0$, yearly coupon of $a/b > 0$, default probability $p \in [0, 1]$, and no recovery. Hence, for $j \in \{1, \dots, d\}$, the loss of bond j (investor's/lender's perspective) is

$$L_j = \begin{cases} -(b(1 + a/b) - b) = -a, & \text{with prob. } 1 - p, \\ b, & \text{with prob. } p. \end{cases}$$

Consider the two portfolios

$$\mathcal{P}_1 \text{ ("diversified") : } L = \sum_{j=1}^d L_j, \quad \mathcal{P}_2 \text{ ("concentrated") : } L = dL_1$$

and show that VaR_α is superadditive $\iff (1-p)^d < \alpha \leq 1-p$.

Solution. Let $\tilde{L}_j = (L_j + a)/(b + a) \in \{0, 1\}$. Then $\tilde{L}_j \sim \text{B}(1, p)$, $j \in \{1, \dots, d\}$, with

$$F_{\text{B}(1,p)}(x) = \begin{cases} 0 & \text{if } x \in (-\infty, 0), \\ 1-p & \text{if } x \in [0, 1), \\ 1 & \text{if } x \in [1, \infty), \end{cases}$$

and

$$F_{\text{B}(1,p)}^-(\alpha) = \begin{cases} -\infty, & \text{if } \alpha = 0, \\ 0, & \text{if } \alpha \in (0, 1-p], \\ 1, & \text{if } \alpha \in (1-p, 1]. \end{cases}$$

Furthermore, $\sum_{j=1}^d \tilde{L}_j \sim \text{B}(d, p)$ with distribution function $F_{\text{B}(d,p)}$.

- For \mathcal{P}_1 , translation invariance and positive homogeneity imply

$$\begin{aligned}\text{VaR}_\alpha \left(\sum_{j=1}^d L_j \right) &= \text{VaR}_\alpha \left(\sum_{j=1}^d ((b+a)\tilde{L}_j - a) \right) \\ &= (b+a) \text{VaR}_\alpha \left(\sum_{j=1}^d \tilde{L}_j \right) - da = (b+a) F_{B(d,p)}^-(\alpha) - da.\end{aligned}$$

- For \mathcal{P}_2 , $\text{VaR}_\alpha(L) = \text{VaR}_\alpha(dL_1) = d\text{VaR}_\alpha(L_1) = d\text{VaR}_\alpha((b+a)\tilde{L}_1 - a) = d(b+a)F_{B(1,p)}^-(\alpha) - da$.

Since VaR_α is superadditive if and only if $\text{VaR}_\alpha(\sum_{j=1}^d L_j) > \sum_{j=1}^d \text{VaR}_\alpha(L_j) = d\text{VaR}_\alpha(L_1)$, we obtain that

$$\begin{aligned}\text{VaR}_\alpha \text{ superadd.} &\iff (b+a)F_{B(d,p)}^-(\alpha) - da > d(b+a)F_{B(1,p)}^-(\alpha) - da \\ &\iff F_{B(d,p)}^-(\alpha) > dF_{B(1,p)}^-(\alpha) \\ &\iff F_{B(d,p)}(dF_{B(1,p)}^-(\alpha)) < \alpha.\end{aligned}$$

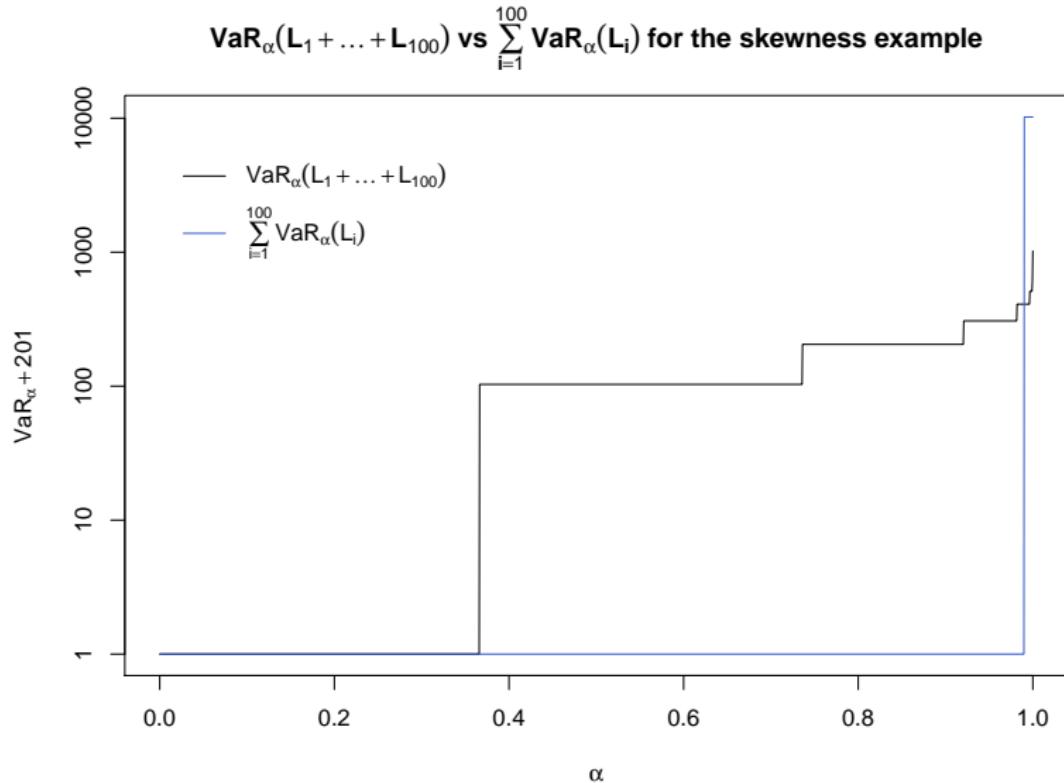
Since $dF_{B(1,p)}^-(\alpha) \in \{-\infty, 0, d\}$ we have that $F_{B(d,p)}(dF_{B(1,p)}^-(\alpha))$ equals

$$F_{B(d,p)}(dF_{B(1,p)}^-(\alpha)) = \begin{cases} 0, & \text{if } \alpha = 0, \\ F_{B(d,p)}(0) \underset{\mathbb{P}(\tilde{L}_j=0 \forall j)}{=} (1-p)^d, & \text{if } \alpha \in (0, 1-p], \\ F_{B(d,p)}(d) = 1, & \text{if } \alpha \in (1-p, 1]. \end{cases}$$

For $\alpha = 0$ or $\alpha \in (1-p, 1]$, $F_{B(d,p)}(dF_{B(1,p)}^-(\alpha)) < \alpha$ is not possible \Rightarrow VaR_α is superadditive if and only if $(1-p)^d < \alpha$ for $0 < \alpha \leq 1-p$. \square

- Note that the superadditivity range does not depend on a, b .
- For further generalizations of this result (e.g., to dependent bonds), see Hofert and McNeil (2014).

For $d = 100$, $T = 1$ y, nominal $b = 100$, coupon $a/b = 2\%$, $p = 1\%$:



Exercise 2.30 (Heavy tailed loss distributions)

$L_1, L_2 \stackrel{\text{ind.}}{\sim} \text{Par}(1/2)$ with df $F(x) = 1 - x^{-1/2}$, $x \in [1, \infty)$. Show that VaR_α is superadditive for all $\alpha \in (0, 1)$.

Solution. F has density $f(x) = \frac{1}{2x^{3/2}}$, $x \in [1, \infty)$. Since L_1, L_2 are independent, we can compute the density of $L_1 + L_2$ as the convolution

$$\begin{aligned} f_{L_1+L_2}(x) &= \int_{-\infty}^{\infty} f_{L_1}(t)f_{L_2}(x-t) dt = \frac{1}{4} \int_1^{x-1} \frac{1}{t^{3/2}} \frac{1}{(x-t)^{3/2}} dt \\ &= \frac{1}{4} \int_1^{x-1} \frac{1}{(xt - t^2)^{3/2}} dt. \end{aligned}$$

By completing the square and then substituting $s = t - x/2$, we obtain

$$f_{L_1+L_2}(x) = \frac{1}{4} \int_1^{x-1} \frac{1}{\left(\frac{x^2}{4} - (t - \frac{x}{2})\right)^{3/2}} dt = \frac{1}{4} \int_{1-x/2}^{x/2-1} \frac{1}{\left(\frac{x^2}{4} - s^2\right)^{3/2}} ds.$$

Substituting $s = \frac{x}{2} \sin t$ ($t = \arcsin(2s/x)$; $ds = \frac{x}{2} \cos t dt$) leads to

$$\begin{aligned} f_{L_1+L_2}(x) &= \frac{x}{8} \int_{\arcsin(2/x-1)}^{\arcsin(1-2/x)} \frac{\cos t}{\left(\frac{x^2}{4}(1-\sin^2 t)\right)^{3/2}} dt \\ &= \frac{1}{x^2} \int_{\arcsin(2/x-1)}^{\arcsin(1-2/x)} \frac{1}{\cos^2 t} dt = \frac{1}{x^2} [\tan t]_{\arcsin(2/x-1)}^{\arcsin(1-2/x)}. \end{aligned}$$

Note that $\tan \arcsin x = \frac{\sin \arcsin x}{\cos \arcsin x} = \frac{x}{\sqrt{1-x^2}}$. Hence,

$$\begin{aligned} f_{L_1+L_2}(x) &= \frac{1}{x^2} \left(\frac{1-2/x}{\sqrt{1-(1-2/x)^2}} - \frac{2/x-1}{\sqrt{1-(2/x-1)^2}} \right) \\ &= \frac{2}{x^2} \frac{x-2}{\sqrt{x^2-(x-2)^2}} = \frac{x-2}{x^2\sqrt{x-1}}, \quad x \in [2, \infty). \end{aligned}$$

The corresponding df equals $F_{L_1+L_2}(x) = 1 - 2\sqrt{x-1}/x$, $x \in [2, \infty)$.

To determine $\text{VaR}_\alpha(L_1 + L_2) = F_{L_1+L_2}^-(\alpha)$ we have to solve $F_{L_1+L_2}(x) = \alpha$ with respect to x . We obtain

$$F_{L_1+L_2}(x) = \alpha \iff \frac{\sqrt{x-1}}{x} = \frac{1-\alpha}{2} \iff \left(\frac{1-\alpha}{2}\right)^2 x^2 - x + 1 = 0,$$

with solutions $x_{1,2} = \frac{1 \pm \sqrt{1 - (1 - \alpha)^2}}{(1 - \alpha)^2 / 2}$. The solution has to satisfy $x_{1,2} \geq 2$, which happens if and only if $(1 - \alpha)^2 \leq 1 \pm \sqrt{1 - (1 - \alpha)^2}$. Note that $x < \sqrt{x}$ for all $x \in (0, 1)$, so this inequality is only valid for x_1 . Thus

$$\begin{aligned}\text{VaR}_\alpha(L_1 + L_2) &= \frac{1 + \sqrt{1 - (1 - \alpha)^2}}{(1 - \alpha)^2 / 2} = 2 \frac{1 + \sqrt{1 - (1 - \alpha)^2}}{(1 - \alpha)^2} \\ &> 2 \frac{1}{(1 - \alpha)^2} = 2 \text{VaR}_\alpha(L_1) = \text{VaR}_\alpha(L_1) + \text{VaR}_\alpha(L_2)\end{aligned}$$

for all $\alpha \in (0, 1)$.

□

Exercise 2.31 (Special dependence)

Let $\alpha \in (0, 1)$, $L_1 \sim U[0, 1]$ and

$$L_2 \stackrel{\text{a.s.}}{=} \begin{cases} L_1, & \text{if } L_1 < \alpha, \\ 1 + \alpha - L_1, & \text{if } L_1 \geq \alpha. \end{cases}$$

Let $\alpha \in (0, 1)$. Show that $VaR_{\alpha+\varepsilon}(L_1+L_2) > VaR_{\alpha+\varepsilon}(L_1) + VaR_{\alpha+\varepsilon}(L_2)$ for all $\varepsilon \in (0, (1 - \alpha)/2)$.

Solution. We first show that $L_2 \sim U[0, 1]$. By the law of total probability, $\mathbb{P}(L_2 \leq x) = \mathbb{P}(L_2 \leq x, L_1 < \alpha) + \mathbb{P}(L_2 \leq x, L_1 \geq \alpha)$. By continuity, the first summand equals $\mathbb{P}(L_1 \leq x, L_1 < \alpha) = \mathbb{P}(L_1 \leq \min\{x, \alpha\}) = \min\{x, \alpha\}$. For the second summand, note that $1 + \alpha - x \geq \alpha$ for all $x \in [0, 1]$, so that it equals

$$\begin{aligned} \mathbb{P}(1 + \alpha - L_1 \leq x, L_1 \geq \alpha) &= \mathbb{P}(L_1 \geq 1 + \alpha - x, L_1 \geq \alpha) \\ &= \mathbb{P}(L_1 \geq \max\{1 + \alpha - x, \alpha\}) = \mathbb{P}(L_1 \geq 1 + \alpha - x) \\ &= \mathbb{P}(1 + \alpha - x \leq L_1 \leq 1) = \max\{1 - (1 + \alpha - x), 0\} = \max\{x - \alpha, 0\}. \end{aligned}$$

Therefore, $\mathbb{P}(L_2 \leq x) = \min\{x, \alpha\} + \max\{x - \alpha, 0\} = x$, $x \in [0, 1]$.

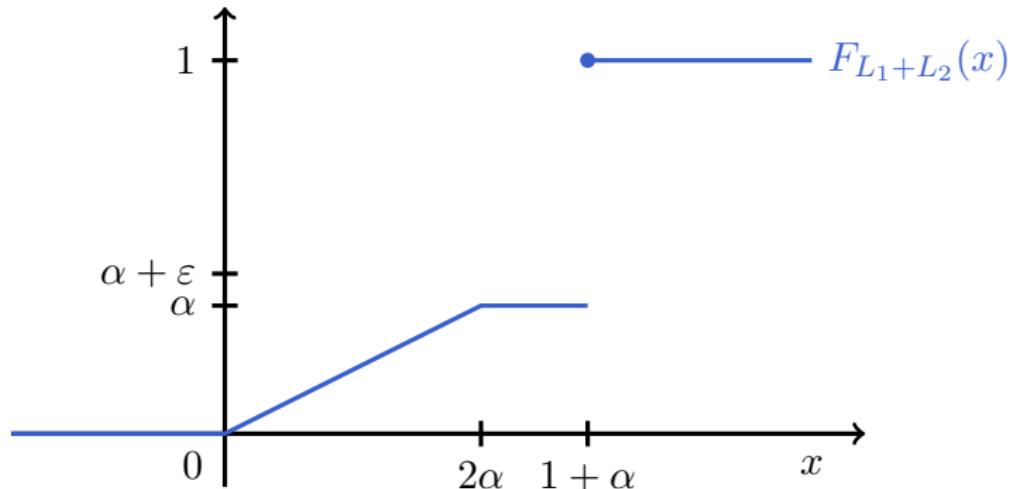
Note that

$$L_1 + L_2 = \begin{cases} 2L_1, & \text{if } L_1 < \alpha, \\ 1 + \alpha, & \text{if } L_1 \geq \alpha, \end{cases}$$

and $2\alpha = \alpha + \alpha < 1 + \alpha$ for all $\alpha \in (0, 1)$. Hence, $F_{L_1+L_2}(x)$ equals

$$\begin{aligned} \mathbb{P}(L_1 + L_2 \leq x) &= \mathbb{P}(2L_1 \leq x, L_1 < \alpha) + \mathbb{P}(1 + \alpha \leq x, L_1 \geq \alpha) \\ &= \min\{x/2, \alpha\} + \mathbb{1}_{\{1+\alpha \leq x\}}(1 - \alpha) \\ &= \begin{cases} 0, & \text{if } x < 0, \\ x/2, & \text{if } x \in [0, 2\alpha), \\ \alpha, & \text{if } x \in [2\alpha, 1 + \alpha), \\ 1, & \text{if } x \geq 1 + \alpha. \end{cases} \end{aligned}$$

A picture is worth a thousand words...



For all $\varepsilon \in (0, (1 - \alpha)/2)$, we thus obtain

$$\text{VaR}_{\alpha+\varepsilon}(L_1 + L_2) = 1 + \alpha > 2(\alpha + \varepsilon) = \text{VaR}_{\alpha+\varepsilon}(L_1) + \text{VaR}_{\alpha+\varepsilon}(L_2).$$

□

Remark 2.32 (Special case of comonotone risks; elliptical risks)

- If $L_1 \stackrel{\text{a.s.}}{=} L_2$ (special case of comonotone risks (strongest positive dependence; see later)) then positive homogeneity of VaR_α implies that $\text{VaR}_\alpha(L_1+L_2) = \text{VaR}_\alpha(2L_1) = 2 \text{VaR}_\alpha(L_1) = \text{VaR}_\alpha(L_1)+\text{VaR}_\alpha(L_2)$ for all $\alpha \in (0, 1)$, so VaR_α is additive (thus also subadditive). In comparison to Exercise 2.31, we see that the strongest positive dependence does not lead to the largest $\text{VaR}_\alpha(L_1 + L_2)$; if L_1 and L_2 have a special dependence structure, $\text{VaR}_\alpha(L_1 + L_2)$ can be larger than $\text{VaR}_\alpha(L_1) + \text{VaR}_\alpha(L_2)$.
- As we will see later, VaR_α is subadditive and thus coherent for all elliptical models (the “garden of eden of RM”) if $\alpha \in [1/2, 1]$. In the multivariate normal world, this can be seen as follows. Let $(L_1, L_2) \sim N(\boldsymbol{\mu}, \Sigma)$, $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. Then (see later)

$$L_1 + L_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2).$$

Since $|\rho| \leq 1$,

$$\begin{aligned}\text{VaR}_\alpha(L_1 + L_2) &= \mu_1 + \mu_2 + \sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2} \Phi^{-1}(\alpha) \\ &\leq \mu_1 + \mu_2 + \sqrt{(\sigma_1 + \sigma_2)^2} \Phi^{-1}(\alpha) = \text{VaR}_\alpha(L_1) + \text{VaR}_\alpha(L_2),\end{aligned}$$

so VaR_α is subadditive for all $\alpha \in [1/2, 1]$ for $(L_1, L_2) \sim N(\boldsymbol{\mu}, \Sigma)$.

- We also see that a statement like “VaR is coherent in the normal case” does not make sense unless we specify the joint distribution function of (L_1, L_2) (marginal dfs + dependence).
 - ▶ If the underlying copula is Gauss (see later), then (L_1, L_2) is multivariate normal and thus VaR_α , $\alpha \in [1/2, 1]$, is coherent.
 - ▶ If it is the copula underlying Exercise 2.31, then VaR_α is not coherent.

Furthermore, $\text{VaR}_\alpha(L_1 + L_2)$ for $L_i \sim N(\mu_i, \sigma_i^2)$, $i \in \{1, 2\}$, cannot even be computed unless we know the dependence between L_1, L_2 .

3 Empirical properties of financial data

3.1 Stylized facts of financial return series

3.2 Multivariate stylized facts

3.1 Stylized facts of financial return series

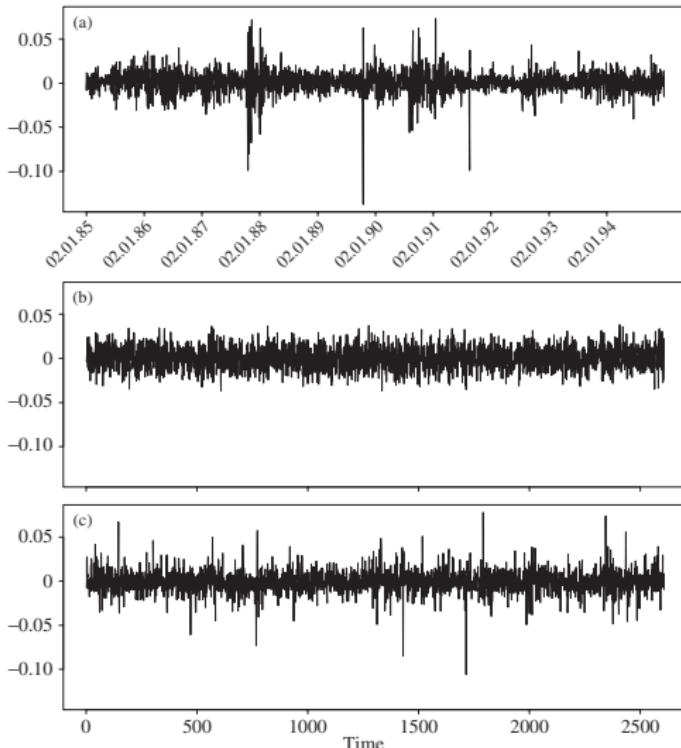
- Stylized facts are a collection of empirical observations and inferences drawn of such, which apply to many time series of risk-factor changes (e.g., log-returns on equities, indices, exchange rates, commodity prices).
- Consider discrete-time risk-factor changes $X_t = Z_t - Z_{t-1}$, e.g., $Z_t = \log S_t$, in which case

$$X_t = \log(S_t/S_{t-1}) \approx S_t/S_{t-1} - 1 = (S_t - S_{t-1})/S_{t-1},$$

is often called a *(log-)return*.

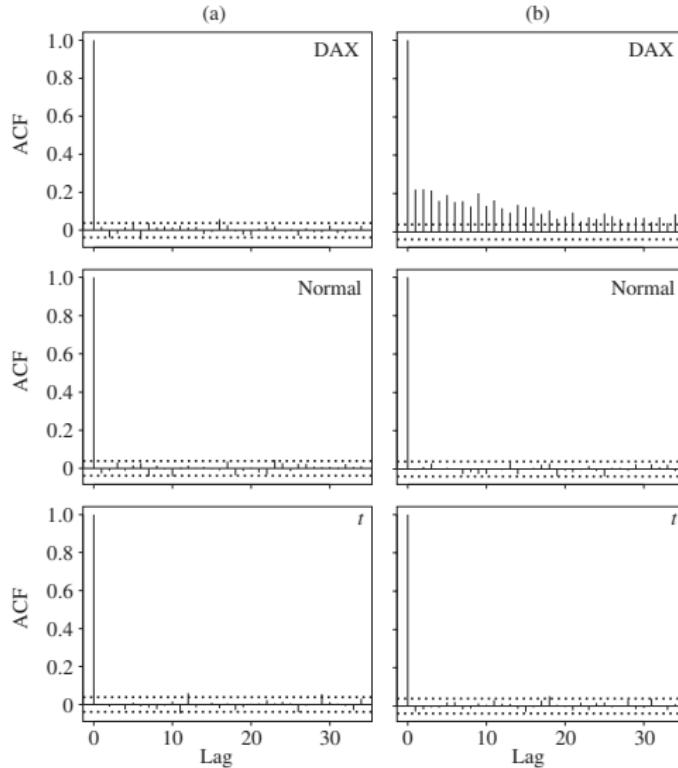
- Stylized facts often apply to daily log-returns (also to intra-daily, weekly, monthly). Tick-by-tick (high-frequency) data have their own stylized facts (not discussed here) and annual return (low-frequency) data are more difficult to investigate (data sparsity; non-stationarity).

3.1.1 Volatility Clustering



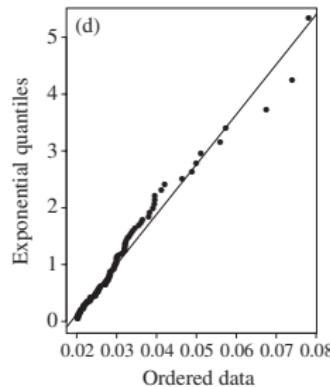
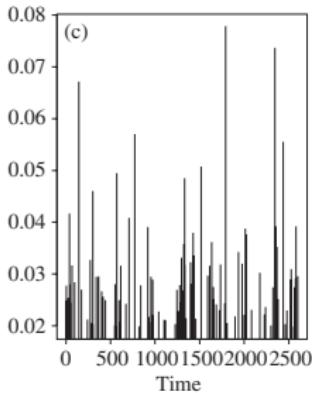
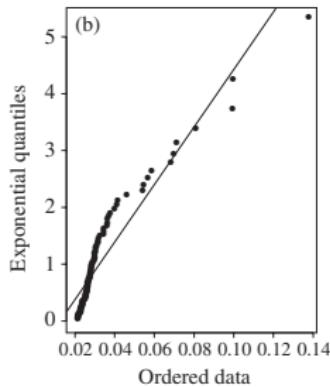
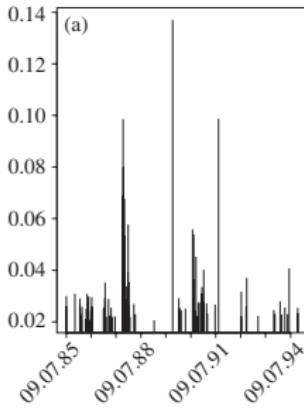
- (a) Log-returns for the DAX index from 1985-01-02 to 1994-12-30 ($n = 2608$)
(b) Simulated i.i.d. data from a fitted normal ($\hat{\mu} = \bar{X}_n$, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$;
⇒ range of extremes ↗)
(c) Simulated i.i.d. data from a fitted $t_{3.8}$ (num. max. of log-likelihood; still no volatility clustering = tendency for extreme returns to be followed by extreme returns)

Autocorrelation function (ACF) $\rho(h) = \text{Cor}[X_0, X_h]$, $h \in \mathbb{Z}$



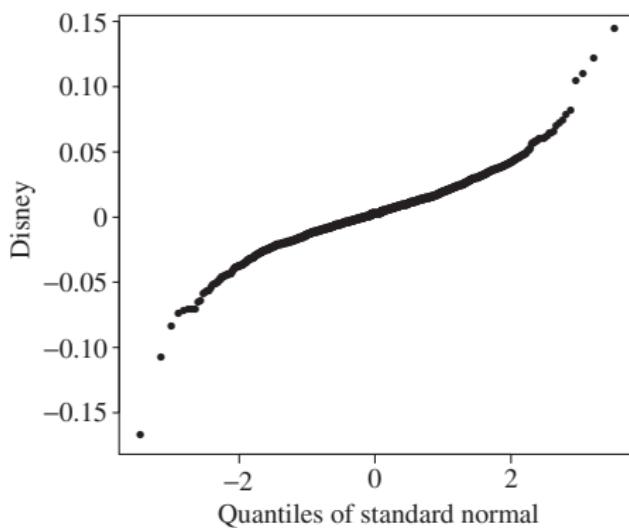
- (a) ACF of $(X_t)_{t \in \mathbb{Z}}$
- (b) ACF of $(|X_t|)_{t \in \mathbb{Z}}$
- non-zero ACF at lag 1 implies a tendency for a return to be followed by a return of equal sign; not the case here
⇒ predicted return = 0
- i.i.d. data $(X_t)_{t \in \mathbb{Z}}$ implies $\rho_X(h) = \rho_{|X|}(h) = \mathbb{1}_{\{h=0\}}$; not the case here; confirm with a Ljung–Box test (H_0 : ACF at first h lags = 0)
- Random-walk hypothesis $\not\approx$

100 largest negative log-returns (losses) of...



- (a) ... DAX index
- (c) ... simulated fitted $t_{3.8}$
- (b), (d) Q-Q plots of (theoretical) waiting times between extreme losses (should be $\text{Exp}(\lambda)$ for i.i.d. data; see EVT) against empirical ones.
- the DAX data shows shorter and longer waiting times than the i.i.d. data
⇒ clustering of extremes

3.1.2 Non-normality and heavy tails



Daily returns typically have $\text{kurt} > 3$ (*leptokurtic*; more narrow center, heavier tails). Typically **power-like tails** rather than exponential.

Non-normality can be detected via

- 1) **Q-Q plot** (an **S-shape** hints at **heavier tails**)
- 2) **Formal tests** (Jarque–Bera, Anderson–Darling, Shapiro–Wilk, D’Agostino)

Jarque–Bera test

- compares skew = $\frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3}$ and kurt = $\frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4}$ with sample versions
- test statistic $T = \frac{n}{6}(\widehat{\text{skew}}^2 + \frac{1}{4}(\widehat{\text{kurt}} - 3)^2)$ (asymptotically χ_2^2 under H_0 : data is normal)

3.1.3 Longer-interval return series

- By going from daily to weekly, monthly, quarterly and yearly data, these effects become less pronounced (returns look more i.i.d., less heavy-tailed).
- The (non-overlapping) h -period log-return at $t \in \{h, 2h, \dots, \lfloor \frac{n}{h} \rfloor h\}$ is

$$X_t^{(h)} = \log\left(\frac{S_t}{S_{t-h}}\right) = \log\left(\frac{S_t}{S_{t-1}} \cdots \frac{S_{t-h+1}}{S_{t-h}}\right) = \sum_{k=0}^{h-1} X_{t-k}$$

A CLT effect takes place (\Rightarrow less rejections with Ljung–Box, i.e., less evidence of serial correlation)

- Problem: the larger h , the less data is available
- Possible remedy: overlapping returns $\{X_t^{(h)} : t \in \{h, h+k, \dots, h + \lfloor \frac{n-h}{k} \rfloor k\}\}$ for $1 \leq k < h \Rightarrow$ more data but now serially dependent.

To summarize, we can infer the following **stylized facts** about **univariate financial return series**:

- (U1) Return series are **not i.i.d.** although they **show little serial correlation**;
- (U2) Series of **absolute** or **squared returns** show **profound serial correlation**;
- (U3) **Conditional expected returns** are **close to zero**;
- (U4) **Volatility** (conditional standard deviation) appears to **vary over time**;
- (U5) **Extreme returns** appear in **clusters**;
- (U6) Return series are **leptokurtic** or **heavy-tailed** (power-like tail).

3.2 Multivariate stylized facts

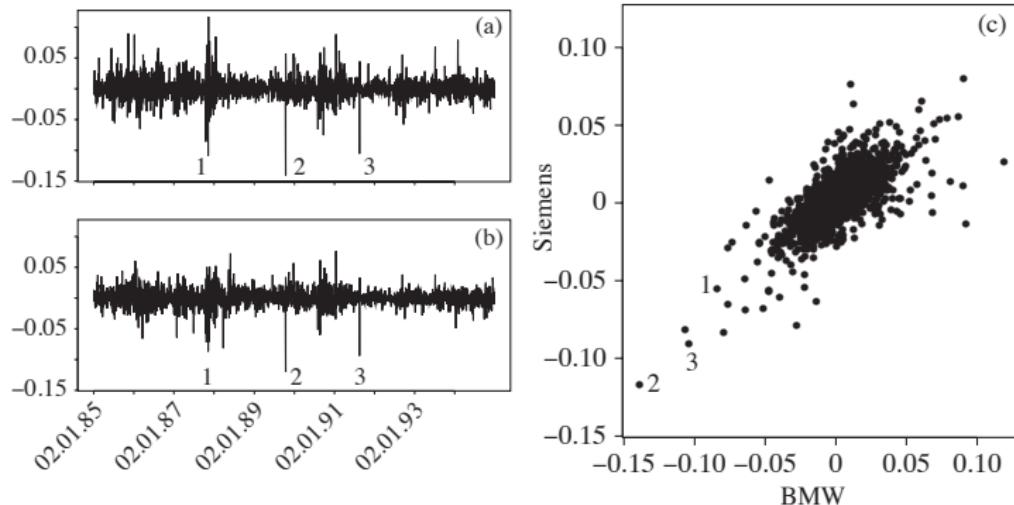
Consider multivariate log-return data $\mathbf{X}_1, \dots, \mathbf{X}_n$.

3.2.1 Correlation between series

- (U1) \Rightarrow Returns of stock A at t and $t + h$ show little correlation, so do the returns of stock A at t and stock B at $t + h$, $h > 0$. Stock A and stock B on day t may be correlated (due to factors that affect the whole market).
- Periods of high/low volatility are typically common to more than one stock \Rightarrow Returns of large magnitude in A at t may be followed by returns of large magnitude in A and B at $t + h$.
- Correlations of returns at t vary over time (difficult to detect whether changes are continual or constant within regimes; fit different models for changing correlation, then make a formal comparison).

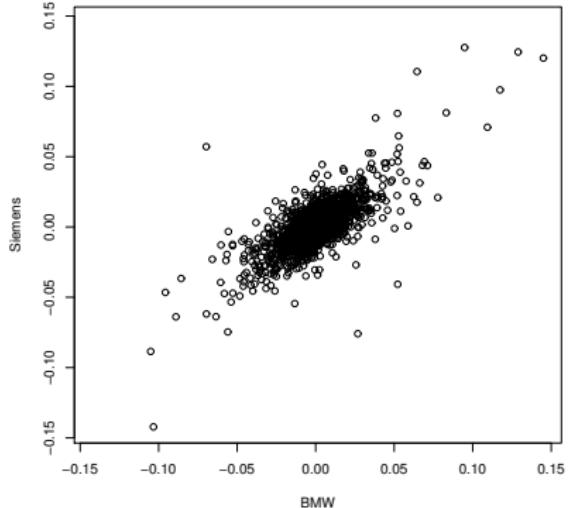
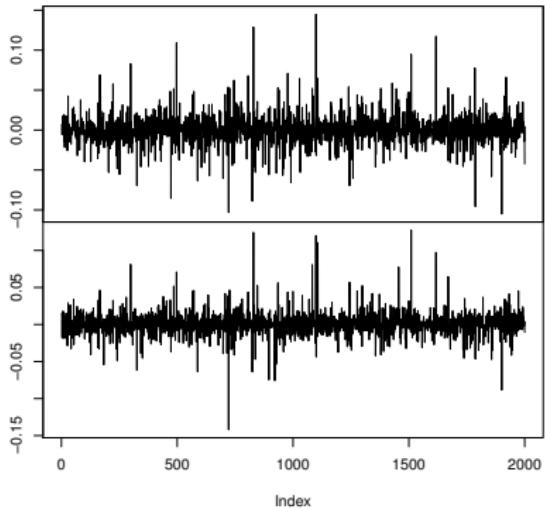
3.2.2 Tail dependence

(BMW, Siemens) log-returns from 1985-01-23 to 1994-09-22 ($n = 2000$)



In volatile/extreme periods, dependence seems stronger (1987-10-19 Black Monday (DJ drop by 22%, automatic trading, overvaluation, illiquidity, market psychology); 1989-10-16 Monday demonstrations in Leipzig (Wende); 1991-08-19 coup against soviet president Mikhail Gorbachev)

Simulated log-returns from a fitted bivariate t distribution ($n = 2000$; $\rho = 0.72$, $\nu = 2.8$ both fitted to (BMW, Siemens))



- The multivariate t distribution can replicate joint large gains/losses (but with the same probability)
- The multivariate normal distribution cannot replicate such a behavior, known as tail dependence (see later).

To summarize, we can infer the following **stylized facts** about multivariate financial return series:

- (M1) Multivariate return series show **little** evidence of **cross-correlation**, except for **contemporaneous returns**;
- (M2) Multivariate series of **absolute returns** show profound **cross-correlation**;
- (M3) **Correlations** between contemporaneous returns **vary over time** (not so easy to infer with empirical correlations due to considerable estimation error in small samples; most reliable way: fit various models and make a formal statistical comparison between the models);
- (M4) **Extreme returns** in one series often **coincide with extreme returns** in several **other series** (e.g., tail dependence).

4 Financial time series

4.1 Fundamentals of time series analysis

4.2 GARCH models for changing volatility

4.1 Fundamentals of time series analysis

For more details on time series, consider Brockwell and Davis (1991) and Brockwell and Davis (2002).

Interlude: Conditional expectations

Definition 4.1 (Conditional expectation, conditional probability)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathbf{X} \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ (i.e., $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$, \mathbf{X} is \mathcal{F} -measurable (i.e., $\mathbf{X}^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}(\mathbb{R}^d)$) and $\mathbb{E}|\mathbf{X}| < \infty$) and $\mathcal{G} \subseteq \mathcal{F}$ be a σ -algebra. Then any rv \mathbf{Y} such that

- 1) $\mathbf{Y} \in \mathcal{G}$ (\mathbf{Y} is \mathcal{G} -measurable);
 - 2) $\mathbb{E}|\mathbf{Y}| < \infty$; and
 - 3) $\mathbb{E}[\mathbf{Y} \mathbf{1}_G] = \int_G \mathbf{Y} d\mathbb{P} = \int_G \mathbf{X} d\mathbb{P} = \mathbb{E}[\mathbf{X} \mathbf{1}_G]$ for all $G \in \mathcal{G}$
- is called *conditional expectation of \mathbf{X} given \mathcal{G}* and denoted by $\mathbb{E}[\mathbf{X} | \mathcal{G}]$.
 $\mathbb{P}(A | \mathcal{G}) = \mathbb{E}[\mathbf{1}_A | \mathcal{G}]$ is called *conditional probability of A given \mathcal{G}* .

Example 4.2 (Do ordinary conditional expectations fit in this setup?)

Let $X \in \{x_1, \dots, x_n\}$, $Z \in \{z_1, \dots, z_m\}$ be rvs (w.l.o.g. all values are attained with non-zero probability).

- Let $\mathbb{P}(X = x_i | Z = z_j) = \frac{\mathbb{P}(X=x_i, Z=z_j)}{\mathbb{P}(Z=z_j)}$ and $\mathbb{E}[X | Z = z_j] = \sum_{i=1}^n x_i \mathbb{P}(X = x_i | Z = z_j)$. Then the *ordinary conditional expectation of X given Z* is defined by

$$\mathbb{E}[X | Z] = \sum_{j=1}^m \mathbb{E}[X | Z = z_j] \mathbb{1}_{\{Z=z_j\}}.$$

- To see that this definition coincides with our more general defintion, consider

$$\begin{aligned}\mathcal{G} &= \sigma(Z) = \{Z^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\} = \{\omega \in \Omega : Z(\omega) \in B \in \mathcal{B}(\mathbb{R})\} \\ &= \bigcup_{A \subseteq \{z_1, \dots, z_m\}} \{\omega \in \Omega : Z(\omega) \in A\} = \bigcup_{J \subseteq \{1, \dots, m\}} \{\omega \in \Omega : Z(\omega) \in (z_j)_{j \in J}\} \\ &= \bigcup_{J \subseteq \{1, \dots, m\}} \bigcup_{j \in J} G_j, \quad G_j = \{\omega \in \Omega : Z(\omega) = z_j\},\end{aligned}$$

and let $Y = \sum_{j=1}^m \mathbb{E}[X | Z = z_j] \mathbb{1}_{\{Z=z_j\}}$. Then Y is constant on G_j ($\Rightarrow Y \in \mathcal{G}$) and $\mathbb{E}|Y| < \infty$. Furthermore,

$$\begin{aligned}\int_{G_j} Y d\mathbb{P} &= \int_{G_j} \sum_{k=1}^m \mathbb{E}[X | Z = z_k] \mathbb{1}_{\{Z=z_k\}} d\mathbb{P} = \int_{G_j} \mathbb{E}[X | Z = z_j] d\mathbb{P} \\ &= \int_{G_j} \sum_{i=1}^n x_i \mathbb{P}(X = x_i | Z = z_j) d\mathbb{P} \\ &= \left(\sum_{i=1}^n x_i \mathbb{P}(X = x_i | Z = z_j) \right) \mathbb{P}(Z = z_j) \\ &= \sum_{i=1}^n x_i \mathbb{P}(X = x_i, Z = z_j) = \int_{G_j} X d\mathbb{P}.\end{aligned}$$

Since every $G \in \mathcal{G}$ is a disjoint union of G_j 's, we have that $\int_G Y d\mathbb{P} = \int_G X d\mathbb{P}$ for all $G \in \mathcal{G}$. So $\mathbb{E}[X | Z]$ corresponds to the ordinary conditional expectation here.

The following property of conditional expectations is used frequently and known as *tower property*.

Lemma 4.3 (Tower property; the smallest σ -algebra remains)

If $\mathcal{G} \subseteq \mathcal{F}$, then $\mathbb{E}[\mathbb{E}[X | \mathcal{G}] | \mathcal{F}] = \mathbb{E}[X | \mathcal{G}] = \mathbb{E}[\mathbb{E}[X | \mathcal{F}] | \mathcal{G}]$.

Idea of proof. Let $G \in \mathcal{G} \subseteq \mathcal{F}$. Applying Definition 4.1 Part 3) to $\mathbb{E}[\mathbb{E}[X | \mathcal{G}] | \mathcal{F}]$ and then to $\mathbb{E}[X | \mathcal{G}]$ implies that $\mathbb{E}[\mathbb{E}[\mathbb{E}[X | \mathcal{G}] | \mathcal{F}] \mathbf{1}_G] = \mathbb{E}[\mathbb{E}[X | \mathcal{G}] \mathbf{1}_G] = \mathbb{E}[X \mathbf{1}_G]$. \square

The next property of conditional expectations shows their importance.

Lemma 4.4 (Best \mathcal{G} -measurable L^2 approx./predictor of X)

Let $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subseteq \mathcal{F}$ be a sub- σ -algebra. Then

$$\min_{Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})} \mathbb{E}[(X - Y)^2] = \mathbb{E}[(X - \mathbb{E}[X | \mathcal{G}])^2]$$

Proof.

$$\begin{aligned}\mathbb{E}[(X - Y)^2] &= \underbrace{\mathbb{E}[(X - \mathbb{E}[X | \mathcal{G}])^2]}_{\text{independent of } Y} + \underbrace{\mathbb{E}[(\mathbb{E}[X | \mathcal{G}] - Y)^2]}_{= 0 \Leftrightarrow Y = \mathbb{E}[X | \mathcal{G}]} \\ &\quad + 2 \underbrace{\mathbb{E}[(X - \mathbb{E}[X | \mathcal{G}])(\mathbb{E}[X | \mathcal{G}] - Y)]}_{\substack{\text{tower} \\ \text{property}}} \\ &\stackrel{\text{take out}}{=} \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X | \mathcal{G}])(\mathbb{E}[X | \mathcal{G}] - Y) | \mathcal{G}]] \stackrel{\text{what is known}}{=} 0\end{aligned}$$

□

For more details on conditional expectations, see Williams (1991).

4.1.1 Basic definitions

Definition 4.5 (Mean function, autocovariance function)

A *stochastic process* is a family of rvs $(X_t)_{t \in I}$, $I \subseteq \mathbb{R}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A *time series* is a discrete-time ($I \subseteq \mathbb{Z}$) stochastic process. Assuming they exist, the *mean function* $\mu(t)$ and the *autocovariance function* $\gamma(t, s)$ of $(X_t)_{t \in \mathbb{Z}}$ are defined by

$$\mu(t) = \mathbb{E}[X_t], \quad t \in \mathbb{Z},$$

$$\gamma(t, s) = \text{Cov}[X_t, X_s] = \mathbb{E}[(X_t - \mathbb{E}X_t)(X_s - \mathbb{E}X_s)], \quad t, s \in \mathbb{Z}.$$

Definition 4.6 ((Weak/strict) stationarity)

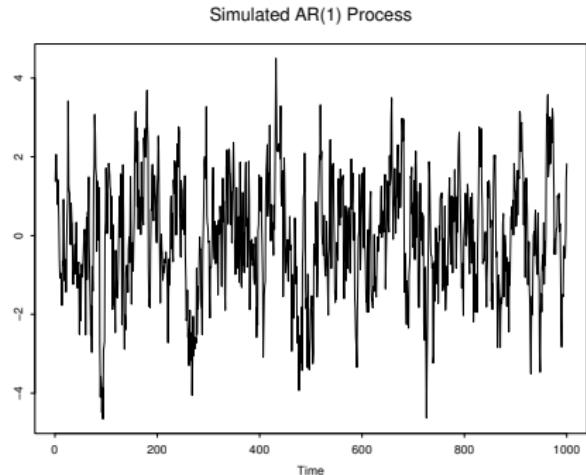
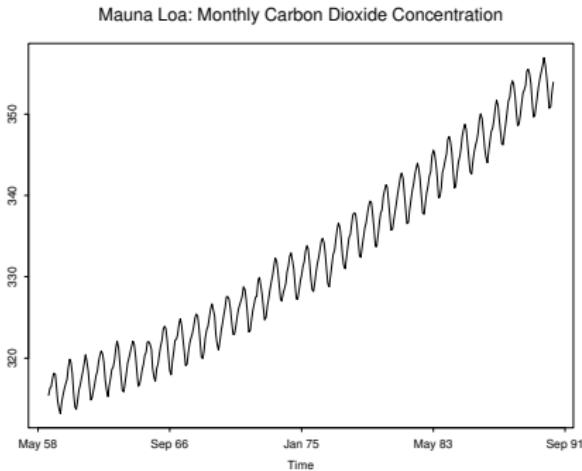
- 1) $(X_t)_{t \in \mathbb{Z}}$ is *(weakly/covariance) stationary* if $\mathbb{E}[X_t^2] < \infty$, $\mu(t) = \mu \in \mathbb{R}$ and $\gamma(t, s) = \gamma(t + h, s + h)$ for all $t, s, h \in \mathbb{Z}$.
- 2) $(X_t)_{t \in \mathbb{Z}}$ is *strictly stationary* if $(X_{t_1}, \dots, X_{t_n}) \stackrel{\text{d}}{=} (X_{t_1+h}, \dots, X_{t_n+h})$ for all $t_1, \dots, t_n, h \in \mathbb{Z}$, $n \in \mathbb{N}$.

Remark 4.7

- 1) Both types of stationarity formalize that $(X_t)_{t \in \mathbb{Z}}$ behaves similarly in any epoch.
- 2) Hölder's inequality: $\|XY\|_1 \leq \|X\|_p \|Y\|_q$ for $\|X\|_p = \mathbb{E}[|X|^p]^{1/p}$,
 $p, q \in [1, \infty] : \frac{1}{p} + \frac{1}{q} = 1$. Under stationarity, $|\gamma(t, s)| \stackrel{\text{triangle}}{\leq} \mathbb{E}[|X_t - EX_t| \cdot |X_s - EX_s|] \stackrel{\text{CSI}}{\leq} (\mathbb{E}[(X_t - EX_t)^2] \mathbb{E}[(X_s - EX_s)^2])^{1/2} < \infty$, so $\gamma(t, s)$ exists.
- 3) Strict stationarity $\not\equiv$ stationarity
 - i) $\mathbb{E}[X_t^2]$ doesn't have to exist (e.g., (G)ARCH processes). If it does, strict stationarity implies stationarity.
 - ii) $\mathbb{E}[|X_t|^p]$, $p > 2$, could be changing (e.g., Pearson type VII).
- 4) $(X_t)_{t \in \mathbb{Z}} \Rightarrow \gamma(t-s, 0) = \gamma(t, s) = \gamma(s, t) = \gamma(s-t, 0)$, so $\gamma(t, s)$ only depends on the emphlag $|t-s|$. We can thus consider $\gamma(h) := \gamma(h, 0)$,

$h \in \mathbb{Z}$. For $s = 0$, we obtain $\gamma(t) = \gamma(t - s, 0) = \gamma(s - t, 0) = \gamma(-t)$, $t \in \mathbb{Z}$, so it suffices to know $\gamma(h)$ for $h \in \mathbb{N}_0$ or $h \in -\mathbb{N}_0$. In particular, if we have shown that $\gamma(h) = f(h)$ for some f and all $h \in \mathbb{N}_0$, we also know $\gamma(-h) = \mathbb{E}[(X_t - \mu)(X_{t-h} - \mu)] = \mathbb{E}[(X_{t-h} - \mu)(X_t - \mu)] = \gamma(h)$, so that $\gamma(h) = f(|h|)$, $h \in \mathbb{Z}$.

Stationary?



(Partial) autocorrelation in stationary time series

Definition 4.8 (ACF)

The *autocorrelation function (ACF)* (or *serial correlation*) of a stationary time series $(X_t)_{t \in \mathbb{Z}}$ is defined by $\rho(h) = \text{Cor}[X_h, X_0] = \gamma(h)/\gamma(0)$, $h \in \mathbb{Z}$.

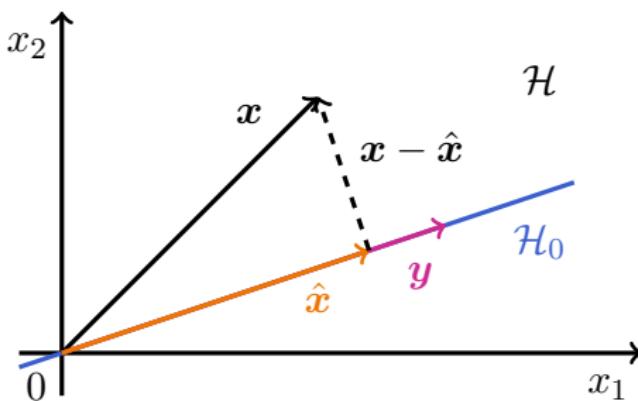
- The study of autocorrelation is known as *analysis in the time domain*. Another important quantity is the *partial autocorrelation function*. For introducing it, we need some tools.
- *Hilbert's Project Theorem* (see Brockwell and Davis (1991, p. 51)): If \mathcal{H}_0 is a closed subspace of the Hilbert space \mathcal{H} and $x \in \mathcal{H}$, then:
 - i) There exists a unique $\hat{x} \in \mathcal{H}_0$: $\|x - \hat{x}\| = \inf_{y \in \mathcal{H}_0} \|x - y\|$;
 - ii) $\hat{x} \in \mathcal{H}_0$, $\|x - \hat{x}\| = \inf_{y \in \mathcal{H}_0} \|x - y\|$ if and only if $\hat{x} \in \mathcal{H}_0$, $x - \hat{x} \in \mathcal{H}_0^\perp = \{x \in \mathcal{H} : \langle x, y \rangle = 0 \text{ for all } y \in \mathcal{H}_0\}$.

Note:

- \hat{x} is the (orthogonal) projection of x onto \mathcal{H}_0 , denoted by $P_{\mathcal{H}_0}x$.
- $\hat{x} = P_{\mathcal{H}_0}x$ is the unique element: $\langle x - \hat{x}, y \rangle = 0 \forall y \in \mathcal{H}_0$ (*prediction equations*; $P_{\mathcal{H}_0}x$ is the best approximation/prediction of x in \mathcal{H}_0).

Example 4.9

$x \in \mathcal{H} = \mathbb{R}^2$, $\mathcal{H}_0 = \text{span}\{\mathbf{y}\}$



- **Yule–Walker equations.** Let X_1, \dots, X_{n-1}, X_n be elements of a stationary time series $(X_t)_{t \in \mathbb{Z}}$ with $\mu(t) = 0$, $t \in \mathbb{Z}$. Suppose we would like to find $\hat{X}_n = \sum_{k=1}^{n-1} \phi_{n-1,k} X_{n-k}$ such that

$$\mathbb{E}[(X_n - \hat{X}_n)^2] \rightarrow \min_{(\phi_{n-1,k})_{k=1}^{n-1}} .$$

$\mathcal{H} = L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space with $\langle X, Y \rangle = \mathbb{E}[XY]$ and $\mathcal{H}_{n-1} = \text{span}\{X_1, \dots, X_{n-1}\} = \{\sum_{k=1}^{n-1} \alpha_k X_{n-k} : \alpha_1, \dots, \alpha_{n-1} \in \mathbb{R}\}$ is a subspace. Therefore, $\hat{X}_n = P_{\mathcal{H}_{n-1}} X_n$ satisfies the prediction equations

$$\langle X_n - \hat{X}_n, Y \rangle = 0, \quad \forall Y \in \mathcal{H}_{n-1}$$

$$\Leftrightarrow \underbrace{\langle X_n - \hat{X}_n, \sum_{k=1}^{n-1} \alpha_k X_{n-k} \rangle}_{= \sum_{k=1}^{n-1} \alpha_k \langle X_n - \hat{X}_n, X_{n-k} \rangle} = 0, \quad \forall \alpha_1, \dots, \alpha_{n-1} \in \mathbb{R}$$

$$\begin{aligned} & \Leftrightarrow \underbrace{\langle X_n - \hat{X}_n, X_l \rangle}_{= \mathbb{E}[(X_n - \sum_{k=1}^{n-1} \phi_{n-1,k} X_{n-k}) X_l]} = 0, \quad \forall l \in \{1, \dots, n-1\} \\ & \qquad \qquad \qquad = \mathbb{E}[X_n X_l] - \sum_{k=1}^{n-1} \phi_{n-1,k} \mathbb{E}[X_{n-k} X_l] \end{aligned}$$

$$\begin{aligned}
 &\Leftrightarrow \gamma(n-l) = \sum_{k=1}^{n-1} \gamma(n-k-l) \phi_{n-1,k} \\
 &\stackrel{\text{station.}}{\Leftrightarrow} \gamma(h) = \sum_{k=1}^{n-1} \gamma(h-k) \phi_{n-1,k}, \quad \forall h \in \{1, \dots, n-1\} \\
 &\Leftrightarrow \Gamma_{n-1} \phi_{n-1} = \gamma_{n-1}, \quad (\text{Yule-Walker equations})
 \end{aligned}$$

where

$$\begin{aligned}
 \phi_{n-1} &= (\phi_{n-1,1}, \dots, \phi_{n-1,n-1}), \\
 \gamma_{n-1} &= (\gamma(1), \dots, \gamma(n-1)), \\
 \Gamma_{n-1} &= (\gamma(|i-j|))_{i,j=1}^{n-1}.
 \end{aligned}$$

Hilbert's Projection Theorem ii) \Rightarrow there exists at least one solution ϕ_{n-1} and all of them lead to the same \hat{X}_n (unique by i)). If Γ_{n-1} is regular (invertible), ϕ_{n-1} is unique. This holds, e.g., if $\gamma(0) > 0$, $\gamma(h) \rightarrow 0$ ($h \rightarrow \infty$); see Brockwell and Davis (1991, p. 167).

- ϕ_n can be computed with the *Durbin–Levinson algorithm*: Let $(X_t)_{t \in \mathbb{Z}}$ be stationary with $\mu(t) = 0$, $t \in \mathbb{Z}$, $\gamma(0) > 0$, $\gamma(h) \rightarrow 0$ ($h \rightarrow \infty$). Then, for all $n \in \mathbb{N}$,

$$\phi_{n,n} = (*) \frac{\gamma(n) - \sum_{k=1}^{n-1} \gamma(n-k) \phi_{n-1,k}}{\gamma(0) - \sum_{k=1}^{n-1} \gamma(n-k) \phi_{n-1,n-k}}$$

$$= \frac{\rho(n) - \sum_{k=1}^{n-1} \rho(n-k) \phi_{n-1,k}}{1 - \sum_{k=1}^{n-1} \rho(n-k) \phi_{n-1,n-k}},$$

$$\begin{pmatrix} \phi_{n,1} \\ \vdots \\ \phi_{n,n-1} \end{pmatrix} \stackrel{(**)}{=} \begin{pmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{pmatrix} - \phi_{n,n} \begin{pmatrix} \phi_{n-1,n-1} \\ \vdots \\ \phi_{n-1,1} \end{pmatrix}.$$

Proof. The Yule–Walker equations hold if and only if

$$\begin{pmatrix} \gamma(0) & \cdots & \gamma(n-2) & \gamma(n-1) \\ \vdots & \ddots & \vdots & \vdots \\ \cdots & \cdots & \gamma(0) & \vdots \\ \cdots & \cdots & \cdots & \gamma(0) \end{pmatrix} \begin{pmatrix} \phi_{n,1} \\ \vdots \\ \phi_{n,n-1} \\ \phi_{n,n} \end{pmatrix} = \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(n-1) \\ \gamma(n) \end{pmatrix}$$

$$\Leftrightarrow \Gamma_{n-1} \begin{pmatrix} \phi_{n,1} \\ \vdots \\ \phi_{n,n-1} \end{pmatrix} + \phi_{n,n} \underbrace{\begin{pmatrix} \gamma(n-1) \\ \vdots \\ \gamma(1) \end{pmatrix}}_{\stackrel{\text{RW}}{=} \Gamma_{n-1} \begin{pmatrix} \phi_{n-1,n-1} \\ \vdots \\ \phi_{n-1,1} \end{pmatrix}} = \underbrace{\begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(n-1) \end{pmatrix}}_{\stackrel{\text{RW}}{=} \Gamma_{n-1} \begin{pmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{pmatrix}},$$

and $\sum_{k=1}^{n-1} \gamma(n-k) \phi_{n,k} + \phi_{n,n} \gamma(0) \stackrel{(***)}{=} \gamma(n)$. Multiplying with Γ_{n-1}^{-1}

leads (*). For (*), use the k th row $\phi_{n,k} = \phi_{n-1,k} - \phi_{n,n} \phi_{n-1,n-k}$ in (***) and solve w.r.t. $\phi_{n,n}$. □

Definition 4.10 (PACF)

The *partial autocorrelation function (PACF)* of a stationary time series $(X_t)_{t \in \mathbb{Z}}$ with $\mu(t) = 0$, $t \in \mathbb{Z}$, $\gamma(0) > 0$, $\gamma(h) \rightarrow 0$ ($h \rightarrow \infty$) is

$$\begin{aligned}\phi(h) &= \text{Cor}[X_0 - P_{\mathcal{H}_{h-1}} X_0, X_h - P_{\mathcal{H}_{h-1}} X_h] \\ &= \frac{\mathbb{E}[X_0(X_h - P_{\mathcal{H}_{h-1}} X_h)]}{\mathbb{E}[(X_h - P_{\mathcal{H}_{h-1}} X_h)(X_h - P_{\mathcal{H}_{h-1}} X_h)]} \\ &= \frac{\mathbb{E}[X_0 X_h] - \sum_{k=1}^{h-1} \phi_{h-1,k} \mathbb{E}[X_0 X_{h-k}]}{\mathbb{E}[X_h(X_h - P_{\mathcal{H}_{h-1}} X_h)]} \\ &= \frac{\gamma(h) - \sum_{k=1}^{h-1} \gamma(h-k) \phi_{h-1,k}}{\text{station. } \gamma(0) - \sum_{k=1}^{h-1} \gamma(k) \phi_{h-1,k}} \stackrel{\text{DL algo.}}{=} \phi_{h,h}, \quad h \in \mathbb{Z}.\end{aligned}$$

The PACF is . . .

- the Cor between X_0 and X_h with the linear dependence of X_1, \dots, X_{h-1} removed;

- $\phi_{h,h}$ (obtained from the Durbin-Levinson algorithm); and
- the coefficient of X_1 in the best linear L^2 -approximation of X_h by X_1, \dots, X_{h-1} .

Note that $\phi(1) = \phi_{1,1} = \gamma(1)/\gamma(0) = \rho(1)$.

White noise processes

Definition 4.11 ((Strict) white noise)

- 1) $(X_t)_{t \in \mathbb{Z}}$ is a *white noise* process if it is stationary with $\rho(h) = \mathbb{1}_{\{h=0\}}$ (no serial correlation). If $\mu(t) = 0$, $\gamma(0) = \sigma^2$, $(X_t)_{t \in \mathbb{Z}}$ is denoted by $\text{WN}(0, \sigma^2)$.
- 2) $(X_t)_{t \in \mathbb{Z}}$ is a *strict white noise* process if it is a sequence of i.i.d. rvs with $\gamma(0) = \sigma^2 < \infty$. If $\mu(t) = 0$, we write $\text{SWN}(0, \sigma^2)$.

One further noise concept is the following (see GARCH processes later).

Definition 4.12 (MGDS)

Let $(X_t)_{t \in \mathbb{Z}}$ be a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$. A sequence $(\mathcal{F}_{t \in \mathbb{Z}})$ (*accrual of information* over time) of σ -algebras is called *filtration* if $\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}$ for all $t \in \mathbb{Z}$. If $\mathcal{F}_t = \sigma(\{X_s : s \leq t\})$, we call $(\mathcal{F}_t)_{t \in \mathbb{Z}}$ the *natural filtration* of $(X_t)_{t \in \mathbb{Z}}$. $(X_t)_{t \in \mathbb{Z}}$ is *adapted* to $(\mathcal{F}_t)_{t \in \mathbb{Z}}$ if $X_t \in \mathcal{F}_t$ for all $t \in \mathbb{Z}$ (X_t is \mathcal{F}_t -measurable). $(X_t)_{t \in \mathbb{Z}}$ is a *martingale-difference sequence (MGDS)* w.r.t. $(\mathcal{F}_t)_{t \in \mathbb{Z}}$ if

- i) $\mathbb{E}|X_t| < \infty$ for all t ;
- ii) $(X_t)_{t \in \mathbb{Z}}$ is adapted to $(\mathcal{F}_t)_{t \in \mathbb{Z}}$; and
- iii) $\mathbb{E}[X_{t+1} | \mathcal{F}_t] = 0$ for all $t \in \mathbb{Z}$.

- If $\mathbb{E}[X_{t+1} | F_t] = X_t$ a.s., then (X_t) is a (discrete-time) *martingale* and $\varepsilon_t = X_t - X_{t-1}$ is a MGDS (winnings in rounds of a *fair game*).
- One can show that a MGDS $(\varepsilon_t)_{t \in \mathbb{Z}}$ with $\sigma^2 = \mathbb{E}[\varepsilon_t^2] < \infty$ satisfies $\rho(h) = 0$, $h \neq 0$, so $(\varepsilon_t)_{t \in \mathbb{Z}} \sim \text{WN}(0, \sigma^2)$.

4.1.2 ARMA processes

Definition 4.13 (ARMA(p, q))

Let $(\varepsilon_t)_{t \in \mathbb{Z}} \sim \text{WN}(0, \sigma^2)$. $(X_t)_{t \in \mathbb{Z}}$ is a *zero-mean ARMA(p, q) process* if it is stationary and satisfies, for all $t \in \mathbb{Z}$,

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}. \quad (8)$$

$(X_t)_{t \in \mathbb{Z}}$ is ARMA(p, q) with *mean μ* if $(X_t - \mu)_{t \in \mathbb{Z}}$ is a zero-mean ARMA(p, q).

Remark 4.14

- R, Brockwell and Davis (1991): ✓
S-Plus: θ_k 's have opposite signs.
- If the *innovations* $(\varepsilon_t)_{t \in \mathbb{Z}}$ are SWN($0, \sigma^2$), then $(X_t)_{t \in \mathbb{Z}}$ is strictly stationary (follows from the representation as a linear process below).

- The defining equation (8) can be written as

$$\phi(B)X_t = \theta(B)\varepsilon_t, \quad t \in \mathbb{Z},$$

where

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p,$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q,$$

$$B : B^k X_t = X_{t-k}, \quad k \in \mathbb{Z} \quad (\text{backshift operator})$$

Causal processes

For practical purposes, it suffices to consider *causal* ARMA processes, that is, $(X_t)_{t \in \mathbb{Z}}$ satisfying (8) which can be represented as

$$X_t = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k} \quad (\text{depends on the past/present, not the future})$$

for $\sum_{k=0}^{\infty} |\psi_k| < \infty$ (*absolute summability condition*).

This condition implies that

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k}\right|\right] &= \mathbb{E}\left[\lim_{n \rightarrow \infty} \left|\sum_{k=0}^n \psi_k \varepsilon_{t-k}\right|\right] \stackrel{\text{triangle}}{\leq} \mathbb{E}\left[\lim_{n \rightarrow \infty} \sum_{k=0}^n |\psi_k \varepsilon_{t-k}|\right] \\ &\stackrel{\substack{\text{monotone} \\ \text{convergence}}}{=} \lim_{n \rightarrow \infty} \mathbb{E}\left[\sum_{k=0}^n |\psi_k \varepsilon_{t-k}|\right] = \lim_{n \rightarrow \infty} \sum_{k=0}^n |\psi_k| \mathbb{E}|\varepsilon_{t-k}| \\ &\stackrel{\substack{\text{white} \\ \text{noise}}}{=} \mathbb{E}|\varepsilon_t| \cdot \lim_{n \rightarrow \infty} \sum_{k=0}^n |\psi_k| < \infty,\end{aligned}$$

so that $\sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k}$ converges (absolutely) a.s. It also converges in L^2 to the same limit (see Brockwell and Davis (2002, p. 83)).

Proposition 4.15 (ACF for causal processes)

Any process $X_t = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k}$ such that $\sum_{k=0}^{\infty} |\psi_k| < \infty$ is stationary with

$$\rho(h) = \frac{\sum_{k=0}^{\infty} \psi_k \psi_{k+|h|}}{\sum_{k=0}^{\infty} \psi_k^2}, \quad h \in \mathbb{Z}.$$

Proof. Since $\varepsilon_t \perp \varepsilon_{t+h}$ for all $h \neq 0$ and by the absolute summability condition, $\mathbb{E}[X_t^2] = \sigma^2 \sum_{k=0}^{\infty} \psi_k^2 < \infty$, $t \in \mathbb{Z}$. Furthermore, $\mu(t) = 0$, $t \in \mathbb{Z}$. Finally, $\gamma(h) = \mathbb{E}[X_t X_{t+h}] = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \psi_k \psi_l \mathbb{E}[\varepsilon_{t-k} \varepsilon_{t+h-l}] = \sigma^2 \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \psi_k \psi_l \mathbb{1}_{\{l=k+h\}} = \sigma^2 \sum_{k=0}^{\infty} \psi_k \psi_{k+h}$, $h \in \mathbb{N}_0$. Remark 4.7 4) implies that $\gamma(h) = \sigma^2 \sum_{k=0}^{\infty} \psi_k \psi_{k+|h|}$, $h \in \mathbb{Z}$, from which the claim follows. \square

Theorem 4.16 (Stationary and causal ARMA solutions)

Let $(X_t)_{t \in \mathbb{Z}}$ be an ARMA(p, q) process for which $\phi(z), \theta(z)$ have no roots in common. Then

$$(X_t)_{t \in \mathbb{Z}} \text{ is stationary and causal} \Leftrightarrow \phi(z) \neq 0 \quad \forall z \in \mathbb{C} : |z| \leq 1.$$

In this case, $X_t = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k}$ for $\sum_{k=0}^{\infty} \psi_k z^k = \theta(z)/\phi(z)$, $|z| \leq 1$.

Proof. Idea; see Brockwell and Davis (1991, p. 85) for more details.

" \Leftarrow " $\phi(z) \neq 0$, $|z| \leq 1 \Rightarrow 1/\phi(z)$ holomorphic on $|z| < 1 + \varepsilon$ for

some $\varepsilon > 0 \Rightarrow 1/\phi(z) = \sum_{k=0}^{\infty} a_k z^k$, $a_k(1 + \varepsilon/2)^k \rightarrow 0$ ($k \rightarrow \infty$)
 $\Rightarrow \exists c > 0 : |a_k| < c(1 + \varepsilon/2)^{-k}$, $k \in \mathbb{N}_0 \Rightarrow \sum_{k=0}^{\infty} |a_k| < \infty$.

Proposition 4.15 $\Rightarrow \varepsilon_t/\phi(B)$ is stationary. $\phi(B)X_t = \theta(B)\varepsilon_t \Rightarrow X_t = \frac{1}{\phi(B)}\phi(B)X_t = \theta(B)\varepsilon_t/\phi(B)$ is stationary (and causal).

“ \Rightarrow ” $X_t = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k} = \psi(B)\varepsilon_t$, $\sum_{k=0}^{\infty} |\psi_k| < \infty \Rightarrow \theta(B)\varepsilon_t = \phi(B)X_t = \eta(B)\varepsilon_t$ for $\eta(B) = \phi(B)\psi(B)$. Let $\eta(z) = \phi(z)\psi(z) = \sum_{k=0}^{\infty} \eta_k z^k$, $|z| \leq 1$. With $\theta_0 = 1$, it follows that $\sum_{k=0}^q \theta_k \varepsilon_{t-k} = \sum_{k=0}^{\infty} \eta_k \varepsilon_{t-k}$. Applying $\mathbb{E}[\cdot \varepsilon_{t-j}] (\langle \cdot, \varepsilon_{t-j} \rangle)$ and using that $(\varepsilon_t) \sim \text{WN}(0, \sigma^2)$, we obtain $\eta_k = \theta_k$, $k \in \{0, \dots, q\}$, and $\eta_k = 0$, $k > q$. This implies that $\theta(z) = \eta(z) = \phi(z)\psi(z)$ for all $|z| \leq 1$. Assume $\phi(z_0) = 0$ for some $|z_0| \leq 1$. Then $0 \neq \theta(z_0) = 0 \cdot \psi(z_0)$. Since $|\psi(z)| \leq \sum_{k=0}^{\infty} |\psi_k| < \infty$ for all $|z| \leq 1$, we obtain a contradiction. Thus $\phi(z) \neq 0$ for all $|z| \leq 1$. \square

Note that if $\theta(z) \neq 0$, $|z| \leq 1$ (known as *invertibility condition*), we can recover ε_t from $(X_s)_{s \leq t}$ via $\varepsilon_t = \phi(B)X_t/\theta(B)$.

- An ARMA(p, q) process with mean μ can be written as

$$X_t = \mu_t + \varepsilon_t$$

$$\mu_t = \mu + \sum_{k=1}^p \phi_k (X_{t-k} - \mu) + \sum_{k=1}^q \theta_k \varepsilon_{t-k}.$$

- If $(X_t)_{t \in \mathbb{Z}}$ is invertible then ε_{t-k} can be expressed in terms of $(X_s)_{s \leq t-k}$, hence μ_t can be expressed by $(X_s)_{s \leq t-1}$. It follows that μ_t is \mathcal{F}_{t-1} -measurable where $\mathcal{F}_{t-1} = \sigma(\{X_s : s \leq t-1\})$.
- If $(\varepsilon_t)_{t \in \mathbb{Z}}$ is a MGDS w.r.t. $(\mathcal{F}_t)_{t \in \mathbb{Z}}$, then $\mu_t = \mathbb{E}[X_t | \mathcal{F}_{t-1}]$.

⇒ An ARMA process puts a particular structure on the conditional mean μ_t given the past. As we will see, a (G)ARCH process puts a certain structure on the conditional variance $\sigma_t^2 = \text{Var}[X_t | \mathcal{F}_{t-1}]$.

Example 4.17

1) $\text{MA}(q) = \text{ARMA}(0, q)$: $X_t = \varepsilon_t + \sum_{k=1}^q \theta_k \varepsilon_{t-k} \stackrel{\theta_0=1}{=} \sum_{k=0}^q \theta_k \varepsilon_{t-k}$ (causal, absolute summability condition ✓).

- **ACF**: Proposition 4.15 $\Rightarrow \rho(h) = \frac{\sum_{k=0}^{q-|h|} \theta_k \theta_{k+|h|}}{\sum_{k=0}^q \theta_k^2}$, $|h| \in \{1, \dots, q\}$, and $\rho(h) = 0$ for all $|h| > q \Rightarrow$ ACF cuts off after lag q .
- **PACF** (for MA(1)): Let $\theta = \theta_1$. Yule–Walker equations $\Leftrightarrow P_h \phi_h = \rho_h$. One can show by induction (or the Durbin–Levinson algorithm) that

$$\phi_{h,h} = -\frac{(-\theta)^h (1 - \theta^2)}{1 - \theta^{2(h+1)}}, \quad h \in \mathbb{N},$$

$$\phi_{h,h-k} = (-\theta)^{-k} \left(\frac{1 - \theta^{2k}}{1 - \theta^2} \right) \phi_{h,h}, \quad k \in \{1, \dots, h-1\}.$$

In particular, $\phi(h) = \phi_{h,h} \searrow 0$ exponentially. One can show that for an $\text{MA}(q)$, $\phi(h)$ does not cut off but $|\phi(h)|$ is bounded by an exponentially decreasing function in h .

2) AR(p) = ARMA($p, 0$): $X_t - \sum_{k=1}^p \phi_k X_{t-k} = \varepsilon_t$.

- **ACF:** As for general ARMA processes, the ACF can be computed in several ways; see Brockwell and Davis (1991, Section 3.3), e.g., via $X_t = \theta(B)\varepsilon_t/\phi(B) = \psi(B)\varepsilon_t$ from $\rho(h)$ as in Proposition 4.15.

For AR(1), it is an exercise to show that $\rho(h) \searrow 0$ exponentially.

For AR(p), one can show from a general form of ψ_k (see Brockwell and Davis (1991, p. 92)) that $\rho(h) \searrow 0$ exponentially, possibly with damped sin waves.

- **PACF:** For $h > p$, let $Y \in \mathcal{H}_{h-1} = \text{span}\{X_1, \dots, X_{h-1}\}$. Since $(X_t)_{t \in \mathbb{Z}}$ is causal, $Y \in \text{span}\{\varepsilon_s : s \leq h-1\}$. Thus,

$$\left\langle X_h - \sum_{k=1}^p \phi_k X_{h-k}, Y \right\rangle = \langle \varepsilon_t, Y \rangle = 0.$$

Prediction equations $\Rightarrow \sum_{k=1}^p \phi_k X_{h-k}$ is the best linear approximation in the L^2 -sense to X_h from X_1, \dots, X_{h-1} , so $\sum_{k=1}^p \phi_k X_{h-k} =$

$P_{\mathcal{H}_{h-1}} X_h$. Hence,

$$\phi(h) = \text{Cor}[\underbrace{X_0 - P_{\mathcal{H}_{h-1}} X_0}_{\in \text{span}\{X_0, \dots, X_{h-1}\}}, \underbrace{X_h - P_{\mathcal{H}_{h-1}} X_h}_{=\varepsilon_h}] \underset{\text{causality}}{=} 0,$$

i.e., the **PACF** of an $\text{AR}(p)$ **cuts off after lag p** . For $1 \leq h \leq p$, one can use the Durbin–Levinson algorithm to compute $\phi(h)$.

- 3) **ARMA(1, 1)**: $X_t - \phi_1 X_{t-1} = \varepsilon_t + \theta_1 \varepsilon_{t-1}$, $|\phi_1| < 1$

(Theorem 4.16 \Rightarrow stationary and causal solution). For determining the **ACF**, we first rewrite the process as $X_t = \psi(B)\varepsilon_t$, where

$$\begin{aligned} \psi(z) &= \frac{\theta(z)}{\phi(z)} = \frac{1 + \theta_1 z}{1 - \phi_1 z} = (1 + \theta_1 z) \sum_{k=0}^{\infty} (\phi_1 z)^k \\ &= \sum_{k=0}^{\infty} \phi_1^k z^k + \sum_{k=1}^{\infty} \theta_1 \phi_1^{k-1} z^k = 1 + \sum_{k=1}^{\infty} \phi_1^{k-1} (\phi_1 + \theta_1) z^k, \end{aligned}$$

hence $\psi_0 = 1$ and $\psi_k = \phi_1^{k-1}(\phi_1 + \theta_1)$, $k \geq 1$. It follows that

$$\begin{aligned} \sum_{k=0}^{\infty} \psi_k \psi_{k+h} &\stackrel{h \geq 1}{=} \underbrace{\psi_0 \psi_h}_{=\phi_1^{h-1}(\phi_1 + \theta_1)} + \underbrace{\sum_{k=1}^{\infty} \phi_1^{k-1+k+h-1} (\phi_1 + \theta_1)^2}_{=(\phi_1 + \theta_1)^2 \phi_1^h \sum_{k=0}^{\infty} \phi_1^{2k}} \\ &= \phi_1^{h-1}(\phi_1 + \theta_1)(1 + (\phi_1 + \theta_1)\phi_1/(1 - \phi_1^2)) \\ &= \frac{\phi_1^{h-1}}{1 - \phi_1^2} (\phi_1 + \theta_1)(1 + \phi_1 \theta_1). \end{aligned}$$

Proposition 4.15 then implies that

$$\rho(h) = \phi_1^{h-1} \frac{(\phi_1 + \theta_1)(1 + \phi_1 \theta_1)}{1 + 2\phi_1 \theta_1 + \theta_1^2} = \phi_1^{h-1} \rho(1) \underset{(h \rightarrow \infty)}{\searrow} 0,$$

so that $\rho(h) = \phi_1^{|h|-1} \rho(1)$ for all $h \in \mathbb{Z} \setminus \{0\}$. The PACF can be computed from the Durbin–Levinson algorithm.

SARIMA models

$(X_t)_{t \in \mathbb{Z}}$ is a $\text{SARIMA}(p, d, q) \times (\tilde{p}, \tilde{d}, \tilde{q})_s$ (Seasonal; Integrated (i.e., may be made stationary by differencing)) process if

$$\underbrace{\phi(B)}_{\substack{\text{seasonal} \\ \text{order } p}} \underbrace{\tilde{\phi}(B^s)}_{\substack{\text{order } s\tilde{p}}} \underbrace{(1 - B)^d}_{\substack{\text{integrated part} \\ \text{order } d}} \underbrace{(1 - B^s)^{\tilde{d}}}_{\substack{\text{order } s\tilde{d}}} X_t = \underbrace{\theta(B)}_{\substack{\text{order } q}} \underbrace{\tilde{\theta}(B^s)}_{\substack{\text{order } s\tilde{q}}} \varepsilon_t, \quad t \in \mathbb{Z}.$$

We see that this is also an $\text{ARMA}(d + p + s(\tilde{d} + \tilde{p}), q + s\tilde{q})$ process. (Seasonal) “differences” are taken to get data from a **stationary** model.

4.1.3 Analysis in the time domain

Correlogram

A *correlogram* is a plot of $(h, \hat{\rho}(h))_{h \geq 0}$ for the sample ACF

$$\hat{\rho}(h) = \frac{\sum_{t=1}^n (X_{t+h} - \bar{X}_n)(X_t - \bar{X}_n)}{\sum_{t=1}^n (X_t - \bar{X}_n)^2}, \quad h \in \{0, \dots, n\}.$$

The sample PACF can be computed from $\hat{\rho}(h)$ via the DL algorithm.

Theorem 4.18

Let $X_t - \mu = \sum_{k=0}^{\infty} \psi_k Z_{t-k}$, $\sum_{k=0}^{\infty} |\psi_k| < \infty$ and $(Z_t) \sim \text{SWN}(0, \sigma^2)$. If either $\mathbb{E}[X_t^4] < \infty$ or $\sum_{k=0}^{\infty} k\psi_k^2 < \infty$, then

$$\sqrt{n} \left(\begin{pmatrix} \hat{\rho}(1) \\ \vdots \\ \hat{\rho}(h) \end{pmatrix} - \begin{pmatrix} \rho(1) \\ \vdots \\ \rho(h) \end{pmatrix} \right) \xrightarrow{(n \rightarrow \infty)}^{\text{d}} N_h(\mathbf{0}, W), \quad h \in \mathbb{N},$$

where $W_{ij} = \sum_{k=1}^{\infty} (\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k))(\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k))$, $i, j \in \{1, \dots, h\}$.

If the ARMA process is SWN itself, then $\sqrt{n} \begin{pmatrix} \hat{\rho}(1) \\ \vdots \\ \hat{\rho}(h) \end{pmatrix} \xrightarrow{(n \rightarrow \infty)}^{\text{d}} N_h(\mathbf{0}, I_h)$, $h \in \mathbb{N}$, so that with probability $1 - \alpha$,

$$\hat{\rho}(k) \underset{(n \text{ large})}{\in} \left[-\frac{q_{1-\alpha/2}}{\sqrt{n}}, \frac{q_{1-\alpha/2}}{\sqrt{n}} \right] = I_{\alpha, n}, \quad k \in \{1, \dots, h\},$$

where $q_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$. $I_{0.05, n}$ is typically displayed in the correlogram. If more than 5% of $\hat{\rho}(k)$, $k \in \{1, \dots, h\}$, lie outside $I_{0.05, n}$, this is evidence against the (i.i.d.) hypothesis of SWN \Rightarrow serial correlation.

Portmanteau tests

- As a [formal test](#) of this hypothesis (SWN), one can use the [Ljung–Box test](#) with test statistic

$$T = n(n+2) \sum_{k=1}^h \frac{\hat{\rho}(k)^2}{n-k} \underset{n \text{ large}}{\sim} \chi_h^2; \quad \text{reject if } T > \chi_h^{2-1}(1-\alpha).$$

- If $(X_t)_{t \in \mathbb{Z}}$ is SWN, $(|X_t|)_{t \in \mathbb{Z}}$ is also i.i.d. It is a good idea to also apply the correlogram and Ljung–Box tests to $(|X_t|)_{t \in \mathbb{Z}}$ as a further test.

4.1.4 Statistical analysis of time series

The Box–Jenkins approach

Approach for the statistical analysis of $(X_t)_{t \in \mathbb{Z}}$:

1) Preliminary analysis

- Plot the time series \Rightarrow Does it look stationary?
- If necessary, clean the (e.g., high-frequency) data and plot it again.

- iii) Make it stationary by removing trend and seasonality (regime switches etc.). A typical decomposition is

$$X_t = \underbrace{\mu_t}_{\text{trend}} + \underbrace{s_t}_{\text{seasonal component}} + \underbrace{\varepsilon_t}_{\text{residual process}}.$$

- A trend μ_t can be detected via smoothing with local averages:

$$\begin{aligned}\tilde{X}_t &= \frac{1}{2h+1} \sum_{k=-h}^h X_{t+k} \\ &= \underbrace{\sum_{k=-h}^h \frac{\mu_{t+k}}{2h+1}}_{\approx \mu_t} + \underbrace{\sum_{k=-h}^h \frac{s_{t+k}}{2h+1}}_{\approx 0} + \underbrace{\sum_{k=-h}^h \frac{\varepsilon_{t+k}}{2h+1}}_{=\tilde{\varepsilon}_t}.\end{aligned}$$

- A seasonal component s_t (from X_1 to X_S ; e.g., Jan–Dec) can be detected similarly, simply consider $(\tilde{X}_s)_{s=1}^S$ with

$$\tilde{X}_s = \frac{1}{N} \sum_{k=0}^{N-1} X_{s+kS}, \quad t \in \{1, \dots, S\}, \quad N = \left\lfloor \frac{n}{S} \right\rfloor.$$

Removing μ_t, s_t can be done non-parametrically (see R's `stl()`) or via regression (e.g., $X_t = a_0 + a_1 t + a_2 t^2 + a_3 t \sin(\frac{2\pi}{12}t + 4) + \varepsilon_t$ for the “airline” dataset) or by taking differences (SARIMA).

2) Analysis in the time domain

- i) Plot ACF, PACF and use the Ljung–Box test for $(X_t)_{t \in \mathbb{Z}}$ (hints at an ARMA) and $(|X_t|)_{t \in \mathbb{Z}}$ (hints at an (G)ARCH). If the SWN hypothesis cannot be rejected, fit a (static) distribution.
- ii) Do ACF (MA) or PACF (AR) cut off? (determines the order(s))

3) Model fitting

- i) Identify the order (if possible; see above);
- ii) Fit various (low-order) ARMA models (various ways; often (conditional) MLE);
- iii) Model-selection criterion (e.g., AIC, BIC) \Rightarrow select “best” model; see also the automatic procedure by Tsay and Tiao (1984).

4) Residual analysis

- i) Consider the residuals

$$\hat{\varepsilon}_t = X_t - \hat{\mu}_t, \quad \hat{\mu}_t = \hat{\mu} + \sum_{k=1}^p \hat{\phi}_k (X_{t-k} - \hat{\mu}) + \sum_{k=1}^q \hat{\theta}_k \hat{\varepsilon}_{t-k},$$

typically recursively computed (e.g., by letting the first q $\hat{\varepsilon}$'s be 0 and the first p X 's be \bar{X}_n)

- ii) Check (plots, ACF, Ljung–Box, ...) the model assumptions.
- iii) In a multivariate setting, determine a multivariate model for the residuals (\Rightarrow dependence between time series); see, e.g., `demo(copula_garch)` in the R package copula.

4.1.5 Prediction

Let X_{t-n+1}, \dots, X_t denote the data available at time t and suppose we want to compute $P_t X_{t+1}$. Assume we have the history $\mathcal{F}_t = \sigma(\{X_s : s \leq t\})$ of the underlying ARMA model available (including today t).

Exponential smoothing

- Used for prediction and trend estimation;
- Assume there is no deterministic seasonal component;
- Typically directly applied to price series;
- Prediction

$$P_t X_{t+1} = \sum_{k=0}^{n-1} \alpha(1-\alpha)^k X_{t-k} = \alpha X_t + (1-\alpha) P_{t-1} X_t.$$

$\alpha \in (0, 1) \uparrow \Rightarrow$ more weight is put on the last observation.

Conditional expectation

Let the ARMA $(X_t)_{t \in \mathbb{Z}}$ be invertible and $(\varepsilon_t)_{t \in \mathbb{Z}}$ be a MGDS w.r.t. $(\mathcal{F}_t)_{t \in \mathbb{Z}}$. By Lemma 4.4, $P_t X_{t+h} = \mathbb{E}[X_{t+h} | \mathcal{F}_t]$ ($\mathbb{E}[X_{t+h} | \mathcal{F}_t]$ minimizes $\mathbb{E}[(X_{t+h} - \cdot)^2]$) \Rightarrow For $h \in \mathbb{N}$, compute $\mathbb{E}[X_{t+h} | \mathcal{F}_t]$ recursively in terms of $\mathbb{E}[X_{t+h-1} | \mathcal{F}_t]$. Use that $\mathbb{E}[\varepsilon_{t+h} | \mathcal{F}_t] = 0$ and that $(X_s)_{s \leq t}$, $(\varepsilon_s)_{s \leq t}$ are

“known” at time t (invertibility insures that ε_t can be written as a function of $(X_s)_{s \leq t}$).

Example 4.19 (Prediction in the ARMA(1, 1) model)

ARMA(1, 1): $X_t - \mu = \phi_1(X_{t-1} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1}$. Then

$$\mathbb{E}[X_{t+1} | \mathcal{F}_t] = \mu + \phi_1(X_t - \mu) + \theta_1\varepsilon_t + \underbrace{\mathbb{E}[\varepsilon_{t+1} | \mathcal{F}_t]}_{=0};$$

$$\begin{aligned}\mathbb{E}[X_{t+2} | \mathcal{F}_t] &= \mu + \phi_1\mathbb{E}[X_{t+1} | \mathcal{F}_t] - \phi_1\mu \stackrel{\text{MGDS}}{=} 0 \\ &\quad + \theta_1\underbrace{\mathbb{E}[\varepsilon_{t+1} | \mathcal{F}_t]}_{=0} + \underbrace{\mathbb{E}[\varepsilon_{t+2} | \mathcal{F}_t]}_{\substack{\text{tower} \\ \text{property}}} \\ &= \mu + \phi_1(\mathbb{E}[X_{t+1} | \mathcal{F}_t] - \mu) = \mu + \phi_1^2(X_t - \mu) + \phi_1\theta_1\varepsilon_t;\end{aligned}$$

$$\mathbb{E}[X_{t+h} | \mathcal{F}_t] = \dots = \mu + \phi_1^h(X_t - \mu) + \phi_1^{h-1}\theta_1\varepsilon_t \underset{(h \rightarrow \infty)}{\rightarrow} \mu.$$

4.2 GARCH models for changing volatility

- (G)ARCH = (generalized) autoregressive conditionally heteroscedastic
- They are the most important models for daily risk-factor returns.

4.2.1 ARCH processes

Definition 4.20 (ARCH(p))

Let $(Z_t)_{t \in \mathbb{Z}} \sim \text{SWN}(0, 1)$ (often: $Z_t \stackrel{\text{ind.}}{\sim} N(0, 1)$ or $t_\nu(0, (\nu - 1)/\nu)$).
 $(X_t)_{t \in \mathbb{Z}}$ is an **ARCH(p) process** if it is strictly stationary and satisfies

$$X_t = \sigma_t Z_t,$$

$$\sigma_t^2 = \alpha_0 + \sum_{k=1}^p \alpha_k X_{t-k}^2,$$

where $\alpha_0 > 0$, $\alpha_k \geq 0$, $k \in \{1, \dots, p\}$.

Remark 4.21

- 1) σ_{t+1} is \mathcal{F}_t -measurable $\Rightarrow \mathbb{E}[X_{t+1} | \mathcal{F}_t] = \sigma_{t+1} \mathbb{E}[Z_{t+1} | \mathcal{F}_t] = \sigma_{t+1} \mathbb{E}[Z_{t+1}] = 0$. Thus, ARCH(p) processes are MGDSs w.r.t. the natural filtration $(\mathcal{F}_t)_{t \in \mathbb{Z}}$. If they are stationary, they are white noise since

$$\begin{aligned}\gamma(h) &= \mathbb{E}[X_t X_{t+h}] \stackrel{\substack{\text{tower} \\ \text{property}}}{=} \mathbb{E}[\mathbb{E}[X_t X_{t+h} | \mathcal{F}_{t+h-1}]] \\ &= \mathbb{E}[X_t \mathbb{E}[X_{t+h} | \mathcal{F}_{t+h-1}]] = 0, \quad h \in \mathbb{N}.\end{aligned}$$

This also applies to GARCH processes; see below.

- 2) If $(X_t)_{t \in \mathbb{Z}}$ is stationary, then $\text{Var}[X_{t+1} | \mathcal{F}_t] = \mathbb{E}[(\sigma_{t+1} Z_{t+1})^2 | \mathcal{F}_t] = \sigma_{t+1}^2 \mathbb{E}[Z_{t+1}^2 | \mathcal{F}_t] = \sigma_{t+1}^2 \mathbb{E}[Z_{t+1}^2] = \sigma_{t+1}^2$.
- \Rightarrow volatility ($= \sigma_t$, the conditional standard deviation) is changing in time, depending on past values of the process. This is where “autoregressive conditionally heteroscedastic” comes from. ARCH models can thus capture volatility clustering (if one of $|X_{t-1}|, \dots, |X_{t-p}|$ is large, X_t is drawn from a distribution with large variance).

Example 4.22 (ARCH(1))

- One can show (via stoch. recurrence relations) that an ARCH(1) process $(X_t)_{t \in \mathbb{Z}}$ is strictly stationary $\Leftrightarrow \mathbb{E}[\log(\alpha_1 Z_t^2)] < 0$. In this case,

$$X_t^2 = \alpha_0 + \alpha_1 \sum_{k=1}^{\infty} \alpha_1^k Z_{t-k}^2.$$

- $(X_t)_{t \in \mathbb{Z}}$ is stationary $\Leftrightarrow \alpha_1 < 1$. In this case, $\text{Var}[X_t] = \frac{\alpha_0}{1-\alpha_1}$.

Proof.

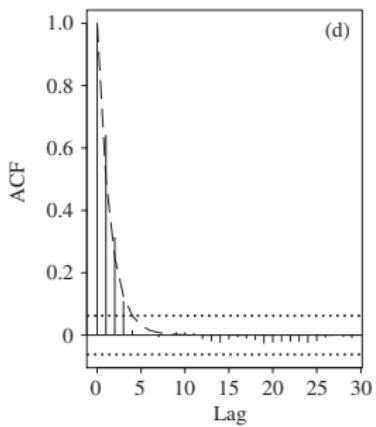
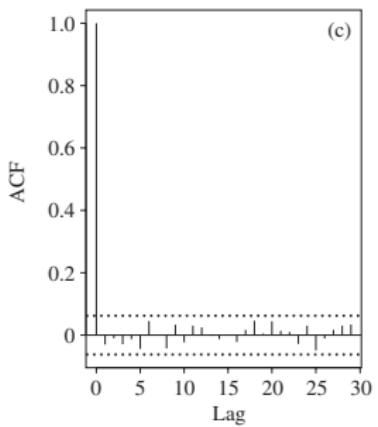
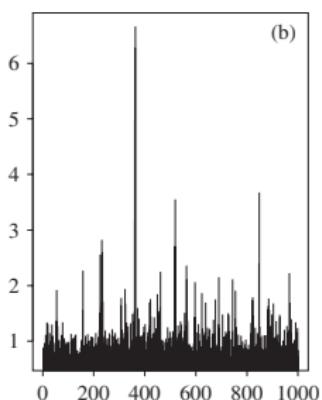
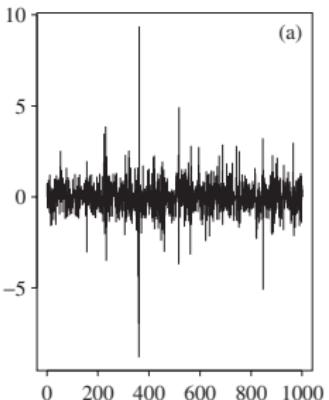
$$\begin{aligned} \Rightarrow X_t^2 &= \sigma_X^2 Z_t^2 = (\alpha_0 + \alpha_1 X_{t-1}^2) Z_t^2 \Rightarrow \mathbb{E}[X_t^2] = \alpha_0 + \alpha_1 \underbrace{\mathbb{E}[X_{t-1}^2 Z_t^2]}_{=\mathbb{E}[X_{t-1}^2] = \sigma_X^2} \\ &\Rightarrow \sigma_X^2 = \alpha_0 + \alpha_1 \sigma_X^2 \Rightarrow \sigma_X^2 = \frac{\alpha_0}{1-\alpha_1}, \quad \alpha_1 < 1. \end{aligned}$$

$$\begin{aligned} \Leftarrow \mathbb{E}[\log(\alpha_1 Z_t^2)] &\stackrel{\text{Jensen}}{\leq} \log \mathbb{E}[Z_t^2] = \log(\alpha_1 \mathbb{E}[Z_t^2]) = \log \alpha_1 < 0. \end{aligned}$$

$$\Rightarrow \mathbb{E}[X_t^2] \stackrel{\substack{\text{see above} \\ \text{i.i.d.}}}{=} \alpha_0 \sum_{k=1}^{\infty} \alpha_1^k \mathbb{E}[Z_{t-k}^2] = \frac{\alpha_0}{1-\alpha_1}$$

$\Rightarrow (X_t)_{t \in \mathbb{Z}}$ is a MGDS with finite, constant variance, hence a white noise process (see above). □

- One can show that for $\beta \geq 1$, the strictly stationary ARCH(1) process $(X_t)_{t \in \mathbb{Z}}$ satisfies $\mathbb{E}[X_t^{2\beta}] < \infty$ if and only if $\mathbb{E}[Z_t^{2\beta}] < \infty$ and $\alpha_1 < (\mathbb{E}[Z_t^{2\beta}])^{-1/\beta}$. From this result one can show that $\text{kurt}(X_t) = \frac{\mathbb{E}[X_t^4]}{\mathbb{E}[X_t^2]^2} = \frac{\text{kurt}(Z_t)(1-\alpha_1^2)}{(1-\alpha_1^2 \text{kurt}(Z_t))}$. If $\text{kurt}(Z_t) > 1$, $\text{kurt}(X_t) > \text{kurt}(Z_t)$. For Gaussian or t innovations, $\text{kurt}(X_t) > 3$ ([leptokurtic](#)).
- Parallels with the AR(1) process: Let $\alpha_1 < 1$ and $\varepsilon_t = \sigma_t^2(Z_t^2 - 1)$. Then $X_t^2 = \sigma_t^2 Z_t^2 = \sigma_t^2 + \varepsilon_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \varepsilon_t$, where $(\varepsilon_t)_{t \in \mathbb{Z}}$ is a MGDS since
 - ▶ $\mathbb{E}|\sigma_t^2(Z_t^2 - 1)| \leq \mathbb{E}[\sigma_t^2]\mathbb{E}[Z_t^2] = \mathbb{E}[\sigma_t^2] < \infty$;
 - ▶ $\sigma_t^2(Z_t^2 - 1) \in \mathcal{F}_t$ for all t ;
 - ▶ $\mathbb{E}[\sigma_{t+1}^2(Z_{t+1}^2 - 1) | \mathcal{F}_t] = \sigma_{t+1}^2(\mathbb{E}[Z_{t+1}^2] - 1) = 0$.
 If $\mathbb{E}[X_t^4] < \infty$, then $\mathbb{E}[\varepsilon_t^2] = \mathbb{E}[(X_t^2 - \alpha_0 - \alpha_1 X_{t-1}^2)^2] < \infty$. Hence, $(X_t^2)_{t \in \mathbb{Z}}$ is an AR(1) of the form $X_t^2 - \frac{\alpha_0}{1-\alpha_1} = \alpha_1 \left(X_{t-1}^2 - \frac{\alpha_0}{1-\alpha_1} \right) + \varepsilon_t$.



- a) Realization ($n = 1000$) of an **ARCH(1)** process with $\alpha_0 = 0.5$, $\alpha_1 = 0.5$ and **Gaussian innovations**;
- b) Realization of the **volatility** $(\sigma_t)_{t \in \mathbb{Z}}$;
- c) Correlogram of $(X_t)_{t \in \mathbb{Z}}$, compare with Remark 4.21 1);
- d) Correlogram of $(X_t^2)_{t \in \mathbb{Z}}$ (AR(1)); dashed line = true ACF

4.2.2 GARCH processes

Definition 4.23 (GARCH(p, q))

Let $(Z_t)_{t \in \mathbb{Z}} \sim \text{SWN}(0, 1)$. $(X_t)_{t \in \mathbb{Z}}$ is a **GARCH(p, q) process** if it is strictly stationary and satisfies

$$X_t = \sigma_t Z_t,$$

$$\sigma_t^2 = \alpha_0 + \sum_{k=1}^p \alpha_k X_{t-k}^2 + \sum_{k=1}^q \beta_k \sigma_{t-k}^2,$$

where $\alpha_0 > 0$, $\alpha_k \geq 0$, $k \in \{1, \dots, p\}$, $\beta_k \geq 0$, $k \in \{1, \dots, q\}$.

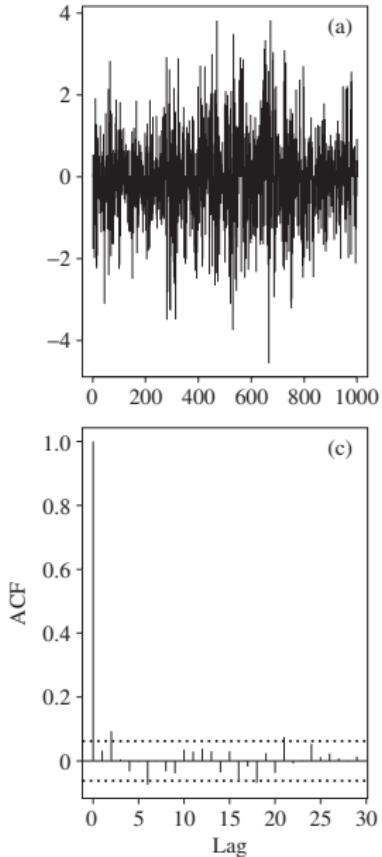
If one of $|X_{t-1}|, \dots, |X_{t-p}|$ or $\sigma_{t-1}, \dots, \sigma_{t-q}$ is large, X_t is drawn from a distribution with (persistently) large variance. Periods of high volatility tend to be more persistent.

Example 4.24 (GARCH(1, 1))

- One can show (via stoch. recurrence relations) that a GARCH(1, 1)

process $(X_t)_{t \in \mathbb{Z}}$ is strictly stationary if $\mathbb{E}[\log(\alpha_1 Z_t^2 + \beta_1)] < \infty$. In this case, $X_t = Z_t \sqrt{\alpha_0(1 + \sum_{k=1}^{\infty} \prod_{j=1}^k (\alpha_1 Z_{t-j}^2 + \beta_1))}$.

- $(X_t)_{t \in \mathbb{Z}}$ is stationary $\Leftrightarrow \alpha_1 + \beta_1 < 1$. In this case, $\text{Var}[X_t] = \frac{\alpha_0}{1-\alpha_1-\beta_1}$.
- One can show that $\mathbb{E}[X_t^4] < \infty$ if and only if $\mathbb{E}[(\alpha_1 Z_t^2 + \beta_1)^2] < 1$ (or $(\alpha_1 + \beta_1)^2 < 1 - (\text{kurt}(Z_t) - 1)\alpha_1^2$). From this result one can show that $\text{kurt}(X_t) = \frac{\text{kurt}(Z_t)(1-(\alpha_1+\beta_1)^2)}{1-(\alpha_1+\beta_1)^2-(\text{kurt}(Z_t)-1)\alpha_1^2}$. If $\text{kurt}(Z_t) > 1$ (Gaussian, scaled t innovations), $\text{kurt}(X_t) > \text{kurt}(Z_t)$.
- Parallels with the ARMA(1,1) process: Let $\alpha_1 + \beta_1 < 1$ and $\varepsilon_t = \sigma_t^2(Z_t^2 - 1)$ (MGDS), and assume $\mathbb{E}[X_t^4] < \infty$. Then $\sigma_{t-1}^2 = X_{t-1}^2 - \varepsilon_{t-1}$ implies that $X_t^2 = \sigma_t^2 Z_t^2 = \sigma_t^2 + \varepsilon_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \varepsilon_t = \alpha_0 + (\alpha_1 + \beta_1) X_{t-1}^2 + \varepsilon_t - \beta_1 \varepsilon_{t-1}$ which can be rewritten as $X_t^2 - \frac{\alpha_0}{1-\alpha_1-\beta_1} = (\alpha_1 + \beta_1)(X_{t-1}^2 - \frac{\alpha_0}{1-\alpha_1-\beta_1}) + \varepsilon_t - \beta_1 \varepsilon_{t-1}$, i.e., a GARCH(1,1) is an ARMA(1,1) for (X_t^2) .



- a) Realization ($n = 1000$) of a GARCH(1,1) process with $\alpha_0 = 0.5$, $\alpha_1 = 0.1$, $\beta_1 = 0.85$ and Gaussian innovations;
- b) Realization of the volatility $(\sigma_t)_{t \in \mathbb{Z}}$;
- c) Correlogram of $(X_t)_{t \in \mathbb{Z}}$, compare with Remark 4.21 1);
- d) Correlogram of $(X_t^2)_{t \in \mathbb{Z}}$ (ARMA(1,1)); dashed line = true ACF

Prediction of GARCH(1,1)

Assume $(X_t)_{t \in \mathbb{Z}}$ is a stationary GARCH(1,1) with $\mathbb{E}[X_t^4] < \infty$.

- $X_t = \sigma_t Z_t \Rightarrow \mathbb{E}[X_t | \mathcal{F}_{t-1}] = \sigma_t \mathbb{E}[Z_t] = 0$, so $(X_t)_{t \in \mathbb{Z}}$ is MGDS and thus, by the tower property, $\mathbb{E}[X_{t+h} | \mathcal{F}_t] = 0$, $h \in \mathbb{N}$.
- $\mathbb{E}[X_{t+1}^2 | \mathcal{F}_t] = \sigma_{t+1}^2 \mathbb{E}[Z_{t+1}] = \alpha_0 + \alpha_1 X_t^2 + \beta_1 \sigma_t^2$.
For $h \geq 2$, X_{t+h}^2 and σ_{t+h}^2 are rvs, and

$$\begin{aligned}\mathbb{E}[X_{t+h}^2 | \mathcal{F}_t] &\stackrel{(*)}{=} \mathbb{E}[\sigma_{t+h}^2 | \mathcal{F}_t] \mathbb{E}[Z_t^2] = \alpha_0 + \alpha_1 \mathbb{E}[X_{t+h-1}^2 | \mathcal{F}_t] \\ &\quad + \beta_1 \underbrace{\mathbb{E}[\sigma_{t+h-1}^2 | \mathcal{F}_t]}_{\stackrel{(*)}{=} \mathbb{E}[X_{t+h-1}^2 | \mathcal{F}_t]} = \alpha_0 + (\alpha_1 + \beta_1) \mathbb{E}[X_{t+h-1}^2 | \mathcal{F}_t] \\ &= \dots = \alpha_0 \sum_{k=0}^{h-1} (\alpha_1 + \beta_1)^k + (\alpha_1 + \beta_1)^{h-1} (\alpha_1 X_t^2 + \beta_1 \sigma_t^2).\end{aligned}$$
$$\Rightarrow \mathbb{E}[\sigma_{t+h}^2 | \mathcal{F}_t] = \mathbb{E}[X_{t+h}^2 | \mathcal{F}_t] \xrightarrow[(h \rightarrow \infty)]{\text{a.s.}} \frac{\alpha_0}{1 - \alpha_1 - \beta_1} = \text{Var}[X_t].$$

The GARCH(p,q) model

- Higher-order (G)ARCH models have the same general behavior as ARCH(1) and GARCH(1,1) models, but their mathematical analysis becomes more tedious.
- One can show that $(X_t)_{t \in \mathbb{Z}}$ is stationary $\Leftrightarrow \sum_{k=1}^p \alpha_k + \sum_{k=1}^q \beta_k < 1$.
- A squared GARCH(p, q) process has the structure

$$X_t^2 = \alpha_0 + \sum_{k=1}^{\max(p,q)} (\alpha_k + \beta_k) X_{t-k}^2 + \varepsilon_t - \sum_{k=1}^q \beta_k \varepsilon_{t-k},$$

where $\varepsilon_t = \sigma_t^2(Z_t^2 - 1)$, $\alpha_k = 0$, $k \in \{p+1, \dots, q\}$ if $q > p$, or $\beta_k = 0$ for $k \in \{q+1, \dots, p\}$ if $p > q$. This resembles the ARMA($\max(p, q), q$) process and is formally such a process provided $\mathbb{E}[X_t^4] < \infty$.

- There are also *IGARCH models* (i.e., non-stationary GARCH(p, q) models with $\sum_{k=1}^p \alpha_k + \sum_{k=1}^q \beta_k = 1$; infinite variance). A squared IGARCH(1,1) resembles an ARIMA(0,1,1) model.

4.2.3 Simple extensions of the GARCH model

Consider stationary GARCH processes as white noise for ARMA processes.

Definition 4.25 (ARMA(p_1, q_1) with GARCH(p_2, q_2) errors)

Let $(Z_t)_{t \in \mathbb{Z}} \sim \text{SWN}(0, 1)$. $(X_t)_{t \in \mathbb{Z}}$ is an ARMA(p_1, q_1) process with GARCH(p_2, q_2) errors if it is stationary and satisfies

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

$$\mu_t = \mu + \sum_{k=1}^{p_1} \phi_k (X_{t-k} - \mu) + \sum_{k=1}^{q_1} \theta_k (X_{t-k} - \mu_{t-k}),$$

$$\sigma_t^2 = \alpha_0 + \sum_{k=1}^{p_2} \alpha_k (X_{t-k} - \mu_{t-k})^2 + \sum_{k=1}^{q_2} \beta_k \sigma_{t-k}^2,$$

where $\alpha_0 > 0$, $\alpha_k \geq 0$, $k \in \{1, \dots, p_2\}$, $\beta_k \geq 0$, $k \in \{1, \dots, q_2\}$,
 $\sum_{k=1}^{p_2} \alpha_k + \sum_{k=1}^{q_2} \beta_k < 1$.

- For the ARMA process to be a causal and invertible linear process, as before, the polynomials $\tilde{\phi}(z) = 1 - \phi_1 z - \cdots - \phi_{p_1} z^{p_1}$ and $\tilde{\theta}(z) = 1 + \theta_1 z + \cdots + \theta_{q_1} z^{q_1}$ should have no common roots and no roots in $\{z \in \mathbb{C} : |z| \leq 1\}$.
- ARMA models with GARCH errors are quite flexible models. It is easy to see that the conditional mean of $(X_t)_{t \in \mathbb{Z}}$ is $\mu_t = \mathbb{E}[X_t | \mathcal{F}_{t-1}]$ and that the conditional variance of $(X_t)_{t \in \mathbb{Z}}$ is $\sigma_t^2 = \text{Var}[X_t | \mathcal{F}_{t-1}]$.
- Other extensions not further discussed here:
 - ▶ *GARCH with leverage*. These models introduce a parameter in the volatility equation in order for the volatility to react asymmetrically to recent returns (bad news leading to a fall in the equity value of a company tends to increase volatility, the so-called *leverage effect*).
 - ▶ *Threshold GARCH (TGARCH)*. More general models than GARCH with leverage in which the dynamics at time t depend on whether

X_{t-1} (or Z_{t-1} ; sometimes even a coefficient) was below/above a threshold.

- ▶ Note that one could also use an asymmetric innovation distribution with mean 0 and variance 1, e.g., from the generalized hyperbolic family.

4.2.4 Fitting (G)ARCH models to data

Building the likelihood

- The most widely used approach is maximum likelihood. We first consider ARCH(1) and GARCH(1, 1) models, the general case easily follows.
- ARCH(1). Suppose we have data X_0, X_1, \dots, X_n . The joint density can be written as

$$f_{X_0, \dots, X_n}(X_0, \dots, X_n) = f_{X_0}(X_0) \prod_{t=1}^n f_{X_t | X_{t-1}, \dots, X_0}(X_t | X_{t-1}, \dots, X_0)$$

$$\begin{aligned}
&= f_{X_0}(X_0) \prod_{t=1}^n f_{X_t|X_{t-1}}(X_t | X_{t-1}) \\
&= f_{X_0}(X_0) \prod_{t=1}^n \frac{1}{\sigma_t} f_Z\left(\frac{X_t}{\sigma_t}\right),
\end{aligned}$$

where $\sigma_t = \sqrt{\alpha_0 + \alpha_1 X_{t-1}^2}$ and f_Z denotes the density of the innovations $(Z_t)_{t \in \mathbb{Z}}$ (mean 0, variance 1; typically $N(0, 1)$ or $t_\nu(0, \frac{\nu-2}{\nu})$). The problem is that f_{X_0} is not known in tractable form. One thus typically considers the conditional likelihood given X_0

$$\begin{aligned}
L(\alpha_0, \alpha_1; X_0, \dots, X_n) &= f_{X_1, \dots, X_n | X_0}(X_1, \dots, X_n | X_0) \\
&= \frac{f_{X_0, \dots, X_n}(X_0, \dots, X_n)}{f_{X_0}(X_0)} = \prod_{t=1}^n \frac{1}{\sigma_t} f_Z\left(\frac{X_t}{\sigma_t}\right).
\end{aligned}$$

Similarly for ARCH(p) models, one considers the likelihood conditional the first p values.

- GARCH(1,1). Here we construct the joint density of X_1, \dots, X_n

conditional on both X_0 and σ_0 , so

$$\begin{aligned} L(\alpha_0, \alpha_1, \beta_1; X_0, \dots, X_n) &= f_{X_1, \dots, X_n | X_0, \sigma_0}(X_1, \dots, X_n | X_0, \sigma_0) \\ &= \prod_{t=1}^n f_{X_t | X_{t-1}, \dots, X_0, \sigma_0}(X_t | X_{t-1}, \dots, X_0, \sigma_0) = \prod_{t=1}^n f_{X_t | \sigma_t}(X_t | \sigma_t) \\ &= \prod_{t=1}^n \frac{1}{\sigma_t} f_Z\left(\frac{X_t}{\sigma_t}\right), \quad \text{where } \sigma_t = \sqrt{\alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2}. \end{aligned}$$

Note that σ_0^2 is not observed. One typically chooses the sample variance of X_1, \dots, X_n (or 0) as starting values.

- GARCH(p, q). Suppose we have data $X_{-p+1}, \dots, X_0, X_1, \dots, X_n$. Evaluate the likelihood conditional on the (observed) X_{-p+1}, \dots, X_0 as well as the (unobserved) $\sigma_{-q+1}, \dots, \sigma_0$ (choose starting values as above).

- Similarly for ARMA models with GARCH errors. In this case,

$$L(\boldsymbol{\theta}; X_0, \dots, X_n) = \prod_{t=1}^n \frac{1}{\sigma_t} f_Z\left(\frac{X_t - \mu_t}{\sigma_t}\right)$$

for the ARMA specification for μ_t and the GARCH specification for σ_t ; all parameters are collected in $\boldsymbol{\theta}$, including unknown parameters of the innovation distribution. The *log-likelihood* is thus given by

$$\ell(\boldsymbol{\theta}; X_0, \dots, X_n) = \sum_{t=1}^n \ell_t(\boldsymbol{\theta}) = \sum_{t=1}^n \log\left(\frac{1}{\sigma_t} f_Z\left(\frac{X_t - \mu_t}{\sigma_t}\right)\right).$$

- Extensions to models with leverage or threshold effects are also possible.
- The log-likelihood ℓ is typically maximized numerically to obtain $\hat{\boldsymbol{\theta}}_n$.

Properties of (Q)MLEs

- We consider two situations: The model which has been fitted...
 - ... has been correctly specified;

2) ... has the correct dynamics but the innovation distribution is erroneously assumed to be Gaussian (in this case the MLE is known as *quasi-maximum likelihood estimator (QMLE)*).

- The asymptotic results for GARCH models are similar to the results in the i.i.d. case; they have been derived in a series of papers. We only treat pure GARCH models, the form of the results will apply more generally (e.g., to ARMA models with GARCH errors).
- Under 1), one can show that for a GARCH(p, q) model with Gaussian innovations,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow[(n \rightarrow \infty)]{d} N_{p+q+1}(\mathbf{0}, I(\boldsymbol{\theta})^{-1}),$$

where

$$I(\boldsymbol{\theta}) := \mathbb{E}\left[\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{\top}\right] = -\mathbb{E}\left(\frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right) =: J(\boldsymbol{\theta})$$

is the *Fisher (or: expected) information* matrix. Thus we have a consistent and asymptotically normal estimator.

- In practice, the $I(\boldsymbol{\theta})$ is often approximated by an *observed information matrix*. Two candidates are

$$\bar{I}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^n \left(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \right) \quad \text{and} \quad \bar{J}(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{t=1}^n \frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2},$$

where the former has *outer-product* and the latter has *Hessian* form. Evaluating them at the MLEs leads to $\bar{I}(\hat{\boldsymbol{\theta}}_n)$ or $\bar{J}(\hat{\boldsymbol{\theta}}_n)$; in practice, the derivatives are often approximated using first and second-order differences. Under 1), $\bar{I}(\hat{\boldsymbol{\theta}}_n) \approx \bar{J}(\hat{\boldsymbol{\theta}}_n)$. One could also take the *sandwich estimator* $\bar{J}(\hat{\boldsymbol{\theta}}_n) \bar{I}(\hat{\boldsymbol{\theta}}_n)^{-1} \bar{J}(\hat{\boldsymbol{\theta}}_n)$.

- Under 2), one still obtains a consistent estimator. If the true innovation distribution has finite fourth moment, then

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow[(n \rightarrow \infty)]{d} N_{p+q+1}(\mathbf{0}, J(\boldsymbol{\theta})^{-1} I(\boldsymbol{\theta}) J(\boldsymbol{\theta})^{-1}),$$

Note that $I(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ typically differ in this case. $J(\boldsymbol{\theta})^{-1} I(\boldsymbol{\theta}) J(\boldsymbol{\theta})^{-1}$ can be estimated by the sandwich estimator.

- If model checking suggests that the **dynamics** have been adequately described by the GARCH model, but **the Gaussian assumption seems doubtful**, then **standard errors for parameter estimates** should be computed based on this covariance matrix estimate.

Model checking

- After model fitting, **check its residuals**. We consider an **ARMA** model with **GARCH errors** of the form $X_t - \mu_t = \varepsilon_t = \sigma_t Z_t$; see Definition 4.25.
- We distinguish **two kinds of residuals**:
 - 1) ***Unstandardized residuals***. These are the **residuals** $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ from the ARMA part of the model, calculated as in Section 4.1.4. Under the hypothesized model they **should behave like a realization of a GARCH process**.
 - 2) ***Standardized residuals***. These are reconstructed realizations of the **SWN** which drives the **GARCH process**. They are **calculated from**

the unstandardized residuals via

$$\hat{Z}_t = \hat{\varepsilon}_t / \hat{\sigma}_t, \quad \hat{\sigma}_t^2 = \hat{\alpha}_0 + \sum_{k=1}^{p_2} \hat{\alpha}_k \hat{\varepsilon}_{t-k}^2 + \sum_{k=1}^{q_2} \hat{\beta}_k \hat{\sigma}_{t-k}^2; \quad (9)$$

starting values for $\hat{\varepsilon}_t$ are taken as 0 and starting values for $\hat{\sigma}_t$ are taken as the sample variance (or 0); ignore the first few values then.

- The **standardized residuals should behave like SWN**. Check this via **correlograms of (\hat{Z}_t) and $(|\hat{Z}_t|)$** and by applying the **Ljung–Box test** of strict white noise. In case of no rejection (the dynamics have been satisfactorily captured), the **validity of the innovation distribution** can also be assessed (e.g., via **Q-Q plots or goodness-of-fit tests**).
⇒ **Two-stage analysis** possible: First estimate the dynamics via QMLE (known as **pre-whitening** of the data), then model the innovation distribution using the standardized residuals.
- Advantages: ▶ More **transparency in model building**;

- ▶ Separating of volatility modeling and modeling of shocks that drive the process;
- ▶ Practical in higher dimensions.

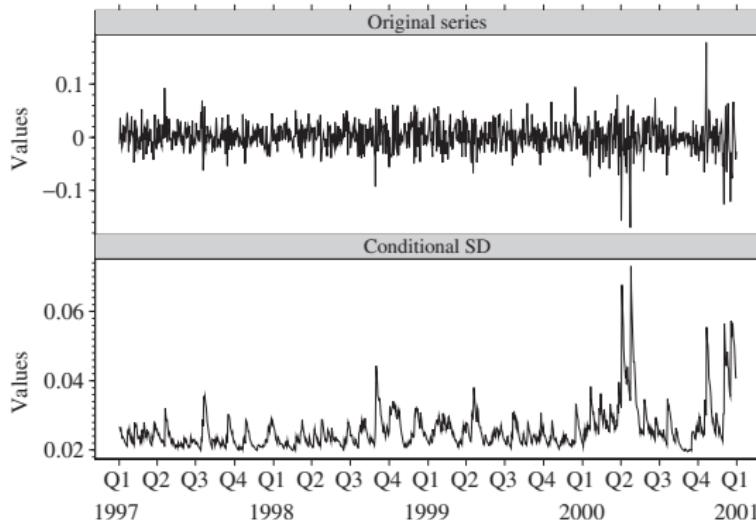
Drawbacks: ARMA fitting errors propagate through to the fitting of innovations (overall error hard to quantify).

Example 4.26 (GARCH model for Microsoft log-returns)

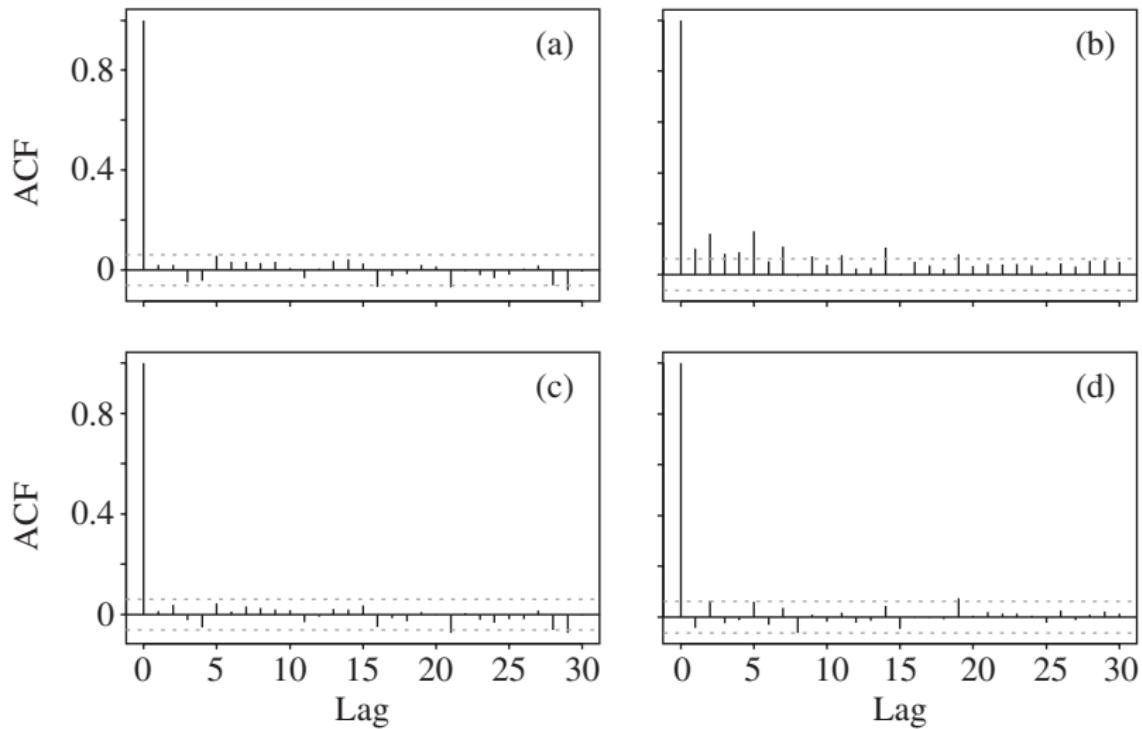
- Consider Microsoft daily log-returns from 1997–2000 (1009 values). The raw returns show no evidence of serial correlation, the absolute values do (Ljung–Box test based on the first 10 estimated correlations fails at the 5% level).
- Various models with t innovations are fitted via MLE: GARCH(1, 1), AR(1)–GARCH(1, 1), MA(1)–GARCH(1, 1), ARMA(1, 1)–GARCH(1, 1). The basic GARCH(1, 1) is favored according to Akaike's information criterion.

- A model with leverage effect further improves the fit (both raw and absolute standardized residuals show no serieal correlation; Ljung–Box does not reject).

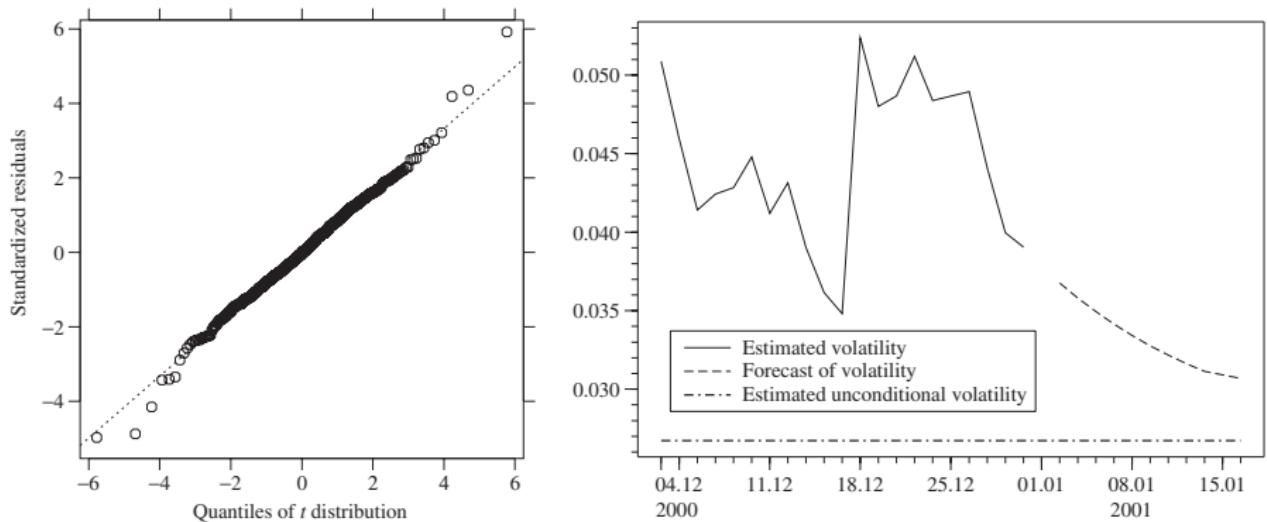
Microsoft log-returns 1997–2000: Data (top) and estimated volatility (bottom) from a GARCH(1, 1) with leverage term.



Correlograms of a) (X_t) ; b) $(|X_t|)$; c) (\hat{Z}_t) ; and d) $(|\hat{Z}_t|)$



Q-Q plot of the standardized residuals (left); Estimated and predicted volatility (right) for the first 10 days of 2001 based on a GARCH(1,1) model (here: without leverage effect).



4.2.5 Volatility forecasting and risk measure estimation

- Consider a weakly and strictly stationary time series $(X_t)_{t \in \mathbb{Z}}$ of the form

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

adapted to a filtration $(\mathcal{F}_t)_{t \in \mathbb{Z}}$, where $\mu_t, \sigma_t \in \mathcal{F}_{t-1}$ and $\mathbb{E} Z_t = 0$, $\text{Var } Z_t = 1$, independent of \mathcal{F}_{t-1} (e.g., $(X_t)_{t \in \mathbb{Z}}$ could be a (G)ARCH model or ARMA model with GARCH errors).

- Assume we know X_{t-n+1}, \dots, X_t and want to forecast σ_{t+h} , $h \geq 1$.
- Since $\mathbb{E}[\sigma_{t+h}^2 | \mathcal{F}_t] = \mathbb{E}[(X_{t+h} - \mu_{t+h})^2 | \mathcal{F}_t]$ our forecasting problem is related to the problem of predicting $(X_{t+h} - \mu_{t+h})^2$.
- Two possible approaches: Exponential smoothing and conditional expectations.

Exponential smoothing

- A one-period ahead forecast $P_t X_{t+1}$ of X_{t+1} based on \mathcal{F}_t is given by

$$P_t X_{t+1} = \alpha X_t + (1 - \alpha) P_{t-1} X_t. \quad (10)$$

Applied to $(X_{t+1} - \mu_{t+1})^2$ leads to

$$P_t (X_{t+1} - \mu_{t+1})^2 = \alpha (X_t - \mu_t)^2 + (1 - \alpha) P_{t-1} (X_t - \mu_t)^2. \quad (11)$$

- Since $\sigma_{t+1}^2 = \mathbb{E}[(X_{t+1} - \mu_{t+1})^2 | \mathcal{F}_t]$, we can use (11) as exponential smoothing scheme for the unobserved squared volatility σ_{t+1}^2 . This yields a recursive scheme for the one-step-ahead volatility forecast given by

$$\hat{\sigma}_{t+1}^2 = \alpha (X_t - \hat{\mu}_t)^2 + (1 - \alpha) \hat{\sigma}_t^2,$$

which is then iterated.

- α is typically chosen small (e.g., RiskMetrics: $\alpha = 0.06$); $\hat{\mu}_t$ is often chosen as 0 (see Section 3). Alternatively, apply exponential smoothing to μ_t via $P_{t-1} X_t$ in (10).

Conditional expectation

The general procedure becomes clear from the following two examples

Example 4.27 (Prediction in the GARCH(1,1) model)

- A GARCH(1,1) model is of type $X_t = \mu_t + \sigma_t Z_t$ for $\mu_t = 0$. Since $\mathbb{E}[X_{t+h} | \mathcal{F}_t] = 0$, $\hat{\mu}_{t+h} = P_t X_{t+h} = 0$ for all $h \in \mathbb{N}$.
- A natural prediction of X_{t+1}^2 based on \mathcal{F}_t is its conditional mean

$$\mathbb{E}[X_{t+1}^2 | \mathcal{F}_t] = \sigma_{t+1}^2 = \alpha_0 + \alpha_1 X_t^2 + \beta_1 \sigma_t^2.$$

If $\mathbb{E}[X_t^4] < \infty$, this is the optimal squared error prediction.

- We thus obtain the one-step-ahead forecast

$$\hat{\sigma}_{t+1}^2 = \widehat{\mathbb{E}[X_{t+1}^2 | \mathcal{F}_t]} = \alpha_0 + \alpha_1 X_t^2 + \beta_1 \hat{\sigma}_t^2.$$

- If $h > 1$, σ_{t+h}^2 and X_{t+h}^2 are rvs. Their predictions (coincide and) are

$$\mathbb{E}[\sigma_{t+h}^2 | \mathcal{F}_t] = \alpha_0 + \alpha_1 \mathbb{E}[X_{t+h-1}^2 | \mathcal{F}_t] + \beta_1 \mathbb{E}[\sigma_{t+h-1}^2 | \mathcal{F}_t]$$

$$\begin{aligned}
 &= \alpha_0 + (\alpha_1 + \beta_1) \mathbb{E}[X_{t+h-1}^2 | \mathcal{F}_t] \\
 &= \alpha_0 + (\alpha_1 + \beta_1) \mathbb{E}[\sigma_{t+h-1}^2 | \mathcal{F}_t]
 \end{aligned}$$

so that a general formula is

$$\mathbb{E}[\sigma_{t+h}^2 | \mathcal{F}_t] = \alpha_0 \sum_{k=0}^{h-1} (\alpha_1 + \beta_1)^k + (\alpha_1 + \beta_1)^{h-1} (\alpha_1 X_t^2 + \beta_1 \sigma_t^2).$$

Note that for $h \rightarrow \infty$, $\mathbb{E}[\sigma_{t+h}^2 | \mathcal{F}_t] \xrightarrow{\text{a.s.}} \frac{\alpha_0}{1-\alpha_1-\beta_1}$, so the prediction of squared volatility converges to the unconditional variance of the process.

Example 4.28 (Prediction in the ARMA(1, 1)–GARCH(1, 1) model)

Let $X_t - \mu_t = \sigma_t Z_t =: \varepsilon_t$ as before. It follows from Examples 4.19 and 4.27 that

$$\mathbb{E}[X_{t+h} | \mathcal{F}_t] = \mu + \phi_1^h (X_t - \mu) + \phi_1^{h-1} \theta_1 \varepsilon_t,$$

$$\text{Var}[X_{t+h} | \mathcal{F}_t] = \alpha_0 \sum_{k=0}^{h-1} (\alpha_1 + \beta_1)^k + (\alpha_1 + \beta_1)^{h-1} (\alpha_1 \varepsilon_t^2 + \beta_1 \sigma_t^2).$$

For ε_t, σ_t , substitute values obtained from (9).

Estimators of VaR_α and ES_α

- Suppose we have losses X_{t-n+1}, \dots, X_t and we would like to estimate VaR_α , ES_α for $F_{X_{t+1}|\mathcal{F}_t}$. Writing F_Z for the df of the innovations (Z_t), the \mathcal{F}_t -measurability of μ_{t+1} and σ_{t+1} implies that

$$F_{X_{t+1}|\mathcal{F}_t}(x) = \mathbb{P}(\mu_{t+1} + \sigma_{t+1}Z_{t+1} \leq x | \mathcal{F}_t) = F_Z\left(\frac{x - \mu_{t+1}}{\sigma_{t+1}}\right).$$

- Let $\text{VaR}_\alpha^t = F_{X_{t+1}|\mathcal{F}_t}^-(\alpha)$ and let ES_α^t denote the corresponding time-dynamic expected shortfall. We then have

$$\text{VaR}_\alpha^t = \mu_{t+1} + \sigma_{t+1}F_Z^-(\alpha), \quad \text{ES}_\alpha^t = \mu_{t+1} + \sigma_{t+1}\text{ES}_\alpha(Z).$$

- If we can estimate μ_{t+1} , σ_{t+1} (parametrically/non-parametrically/semi-parametrically), we only have left to estimate $F_Z^-(\alpha)$ and $\text{ES}_\alpha(Z)$.
- For GARCH-type models it is easy to calculate $F_Z^-(\alpha)$ and $\text{ES}_\alpha(Z)$. And if we use exponential smoothing or QMLE to estimate μ_{t+1} , σ_{t+1} , we can use the residuals $\hat{Z}_s = (X_s - \hat{\mu}_s)/\hat{\sigma}_s$, $s \in \{t-n+1, \dots, n\}$ to estimate $F_Z^-(\alpha)$ and $\text{ES}_\alpha(Z)$.

5 Extreme value theory

5.1 Maxima

5.2 Threshold exceedances

5.3 Point process models

For more theoretical details, see Embrechts et al. (1997), especially Chapters 2 and 3.

5.1 Maxima

Consider losses $(X_k)_{k \in \mathbb{N}}$ (e.g., negative log-returns).

5.1.1 Generalized extreme value distribution

Convergence of sums

Let $(X_k)_{k \in \mathbb{N}}$ be i.i.d. with $\mathbb{E}[X_1^2] < \infty$ (mean μ , variance σ^2) and

$$S_n = \sum_{k=1}^n X_k.$$

Note that $\bar{X}_n \xrightarrow[n \uparrow \infty]{\text{a.s.}} \mu$ by the Strong Law of Large Numbers (SLLN), so $(\bar{X}_n - \mu)/\sigma \xrightarrow[n \uparrow \infty]{\text{a.s.}} 0$. By the CLT,

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow[n \uparrow \infty]{d} N(0, 1),$$

that is,

$$\lim_{n \rightarrow \infty} \mathbb{P}((S_n - d_n)/c_n \leq x) = \Phi(x), \quad x \in \mathbb{R},$$

i.e., $c_n = \sqrt{n}\sigma$, $d_n = n\mu$, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-z^2/2} dz$. More generally ($\sigma^2 = \infty$), the **limiting distributions for appropriately normalized sums** are the class of **α -stable distributions** ($\alpha \in (0, 2]$; $\alpha = 2$: normal distribution).

Convergence of maxima

QRM is concerned with **maximal losses** (worst-case losses). Let $(X_i)_{i \in \mathbb{N}} \stackrel{\text{ind.}}{\sim} F$ (can be relaxed to a strictly stationary time series) and F continuous. Then the **block maxima** is given by

$$M_n = \max\{X_1, \dots, X_n\}.$$

Clearly, $\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, \dots, X_n \leq x) = F^n(x) \xrightarrow[n \uparrow \infty]{} \mathbb{1}_{\{x \geq x_F\}}$,

where $x_F := \sup\{x \in \mathbb{R} : F(x) < 1\} = F^-(1) \leq \infty$ denotes the *right endpoint of F* ; typically $x_F = \infty$ in finance/insurance. Thus, $M_n \xrightarrow[n \uparrow \infty]{d} x_F$.

One can show that, since x_F is a constant, $M_n \xrightarrow[n \uparrow \infty]{P} x_F$, and, since $(M_n) \nearrow$, $M_n \xrightarrow[n \uparrow \infty]{a.s.} x_F$ (similar as in the SLLN). Is there a “CLT” for block maxima?

Idea CLT: What about linear transformations (the simplest possible)?

Definition 5.1 (Maximum domain of attraction)

Suppose we find *normalizing sequences* of real numbers $(c_n) > 0$ and (d_n) such that $(M_n - d_n)/c_n$ converges in distribution, i.e.,

$$\mathbb{P}((M_n - d_n)/c_n \leq x) = \mathbb{P}(M_n \leq c_n x + d_n) = F^n(c_n x + d_n) \xrightarrow[n \uparrow \infty]{} H(x),$$

for some *non-degenerate (n.d.) df H* (not a unit jump). Then F is in the *maximum domain of attraction of H* ($F \in \text{MDA}(H)$).

H is determined up to location/scale, i.e., H specifies a unique *type* of distribution. In particular, we can always choose $(c_n), (d_n)$ such that the limit of $\frac{M_n - d_n}{c_n}$ appears in a location-scale transformed way.

Theorem 5.2 (Convergence to Types)

Suppose $(M_n)_n$ is a sequence of rvs such that $\frac{M_n - d_n}{c_n} \xrightarrow{d} Y$ for a rv Y and $d_n \in \mathbb{R}, c_n > 0$. Then

$$\frac{M_n - \delta_n}{\gamma_n} \xrightarrow{d} Z$$

for a rv Z and $\delta_n \in \mathbb{R}, \gamma_n > 0$ if and only if

$$(c_n/\gamma_n) \rightarrow c \in [0, \infty), \quad (d_n - \delta_n)/\gamma_n \rightarrow d \in \mathbb{R},$$

in which case $Z \stackrel{d}{=} cY + d$ (i.e., Y and Z are of the same type) and c, d are the unique such constants.

Proof. See Embrechts et al. (1997, p. 554). □

How does H look like? (Fisher and Tippett (1928), Gnedenko (1943))

Theorem 5.3 (Fisher–Tippett–Gnedenko)

If $F \in \text{MDA}(H)$ for some n.d. H , then H must be of GEV type, i.e., $H = H_\xi$ for some $\xi \in \mathbb{R}$ (see later).

Proof. Non-trivial. For a sketch, see Embrechts et al. (1997, p. 122). \square

- **Interpretation:** If location-scale transformed maxima converge in distribution to a n.d. limit, the limit distribution must be a GEV distribution.
- We can always choose normalizing sequences $(c_n) > 0$, (d_n) such that H_ξ appears in standard form.
- All commonly encountered continuous distributions are in the MDA of a GEV distribution.
- The following is often a useful result:

$$\lim_{n \uparrow \infty} F^n(c_n x + d_n) = H(x) \quad \underset{-\log x \approx 1-x}{\iff} \quad \lim_{n \uparrow \infty} n \bar{F}(c_n x + d_n) = -\log H(x).$$

Definition 5.4 (Generalized extreme value (GEV) distribution)

The (standard) *generalized extreme value (GEV) distribution* is given by

$$H_\xi(x) = \begin{cases} \exp(-(1 + \xi x)^{-1/\xi}), & \text{if } \xi \neq 0, \\ \exp(-e^{-x}), & \text{if } \xi = 0, \end{cases}$$

where $1 + \xi x > 0$ (MLE!). A three-parameter family is obtained by a location-scale transform $H_{\xi,\mu,\sigma}(x) = H_\xi((x - \mu)/\sigma)$, $\mu \in \mathbb{R}$, $\sigma > 0$.

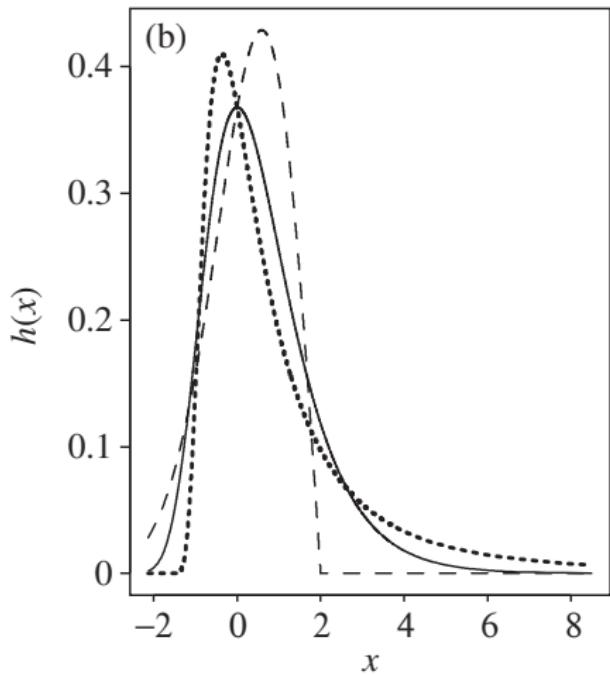
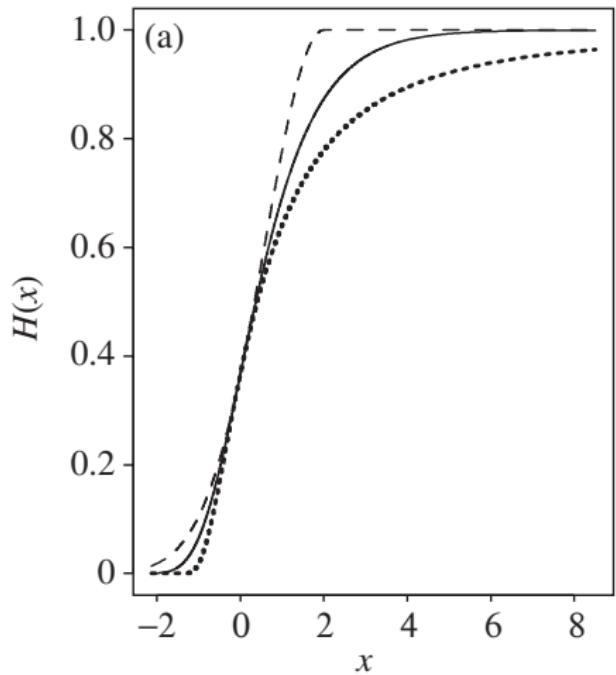
- The parameterization is *continuous in ξ* (simplifies statistical modeling).
- The *larger ξ* , the *heavier tailed H_ξ* (if $\xi > 0$, $\mathbb{E}[X^k] = \infty$ iff $k \geq \frac{1}{\xi}$).
- ξ is the *shape* (*determines moments, tail*). Special cases:
 - 1) $\xi < 0$: the *Weibull df*, short-tailed, $x_{H_\xi} < \infty$;
 - 2) $\xi = 0$: the *Gumbel df*, $x_{H_0} = \infty$, decays exponentially;
 - 3) $\xi > 0$: the *Fréchet df*, $x_{H_\xi} = \infty$, *heavy-tailed* ($\bar{H}_\xi(x) \approx (\xi x)^{-1/\xi}$), most important case for practice (typically, $\xi \in (1/5, 1/2)$).

- For $1 + \xi x > 0$, the density h_ξ of H_ξ is given by

$$h_\xi(x) = \begin{cases} (1 + \xi x)^{-1/\xi - 1} H_\xi(x), & \text{if } \xi \neq 0, \\ e^{-x} H_0(x), & \text{if } \xi = 0. \end{cases}$$

- One can show that *tail equivalence* $\lim_{x \uparrow x_F = x_G} \frac{\bar{F}(x)}{\bar{G}(x)} = c \in (0, \infty)$ implies $F \in \text{MDA}(H) \Leftrightarrow G \in \text{MDA}(H)$ with the same normalizing sequences, i.e., tail equivalent distributions belong to $\text{MDA}(H_\xi)$ for the same ξ .
- Minima:** $-X_1, \dots, -X_n \stackrel{\text{ind.}}{\sim} \bar{F}(-x) = 1 - F(-x)$. If $\bar{F}(-x) \in \text{MDA}(H_\xi)$, then, properly normalized, the limiting distribution of $\min\{X_1, \dots, X_n\} = -\max\{-X_1, \dots, -X_n\}$ is a type of $1 - H_\xi(-x)$.

(a): H_ξ ; (b): density h_ξ for $\xi \in \{-0.5, 0, 0.5\}$ (dashed, solid, dotted)



Example 5.5 (Exponential distribution)

For $(X_i)_{i \in \mathbb{N}} \stackrel{\text{ind.}}{\sim} \text{Exp}(\lambda)$, choosing $c_n = 1/\lambda$, $d_n = \log(n)/\lambda$, one obtains

$$\begin{aligned} F^n(c_n x + d_n) &= (1 - \exp(-\lambda((1/\lambda)x + \log(n)/\lambda)))^n \\ &= (1 - \exp(-x)/n)^n \underset{n \uparrow \infty}{\rightarrow} \exp(-e^{-x}) = H_0(x). \end{aligned}$$

Therefore, $F \in \text{MDA}(H_0)$ (Gumbel).

Example 5.6 (Pareto distribution)

For $(X_i)_{i \in \mathbb{N}} \stackrel{\text{ind.}}{\sim} \text{Par}(\theta, \kappa)$ with $F(x) = 1 - (\frac{\kappa}{\kappa+x})^\theta$, $x \geq 0$, $\theta, \kappa > 0$, choosing $c_n = \kappa n^{1/\theta}/\theta$, $d_n = \kappa(n^{1/\theta} - 1)$, one obtains

$$\begin{aligned} F^n(c_n x + d_n) &= \left(1 - \left(\frac{\kappa}{\kappa + x \kappa n^{1/\theta}/\theta + \kappa(n^{1/\theta} - 1)}\right)^\theta\right)^n \\ &= \left(1 - \left(\frac{1}{1 + xn^{1/\theta}/\theta + n^{1/\theta} - 1}\right)^\theta\right)^n \\ &= \left(1 - \frac{(1/(x/\theta))^\theta}{n}\right)^n = \left(1 - \frac{(\theta/x)^\theta}{n}\right)^n \underset{n \uparrow \infty}{\rightarrow} \exp(-(\theta/x)^\theta), \end{aligned}$$

which equals $H_{1/\theta, \theta, 1}$ and hence $F \in \text{MDA}(H_{1/\theta})$ (Fréchet).

We could have equally well chosen $c_n = \kappa(n^{1/\theta} - 1)$ and $d_n = 0$, since

$$\begin{aligned} F^n(c_n x + d_n) &= \left(1 - \left(\frac{\kappa}{\kappa + \kappa(n^{1/\theta} - 1)x}\right)^\theta\right)^n \\ &= \left(1 - \left(\frac{1}{1 - x + n^{1/\theta}x}\right)^\theta\right)^n = \left(1 - \frac{\left(\frac{1}{(1-x)/n^{1/\theta}+x}\right)^\theta}{n}\right)^n \\ &\xrightarrow[n \uparrow \infty]{} \exp(-(1/x)^\theta), \end{aligned}$$

which equals $H_{1/\theta, 1, 1/\theta}$.

Lurking in the background: $(a_n) \in \mathbb{C}$, $a_n \rightarrow a \in \mathbb{C} \Rightarrow (1 + a_n/n)^n \rightarrow e^a$.

5.1.2 Maximum domains of attraction

All commonly applied continuous F belong to $\text{MDA}(H_\xi)$ for some $\xi \in \mathbb{R}$. μ, σ can be estimated, but how can we characterize/determine ξ ? All $F \in \text{MDA}(H_\xi)$ for $\xi > 0$ have an elegant characterization involving the following notions.

Definition 5.7 (Slowly/regularly varying functions)

- 1) A positive, Lebesgue-measurable function L on $(0, \infty)$ is *slowly varying at ∞* if $\lim_{x \rightarrow \infty} \frac{L(tx)}{L(x)} = 1$, $t > 0$. The class of all such functions is denoted by SV ; e.g., $c, \log \in \text{SV}$.
- 2) A positive, Lebesgue-measurable function h on $(0, \infty)$ is *regularly varying at ∞ with index $\alpha \in \mathbb{R}$* if $\lim_{x \rightarrow \infty} \frac{h(tx)}{h(x)} = t^\alpha$, $t > 0$. The class of all such functions is denoted by RV_α ; $x^\alpha L(x) \in \text{RV}_\alpha$.

The Fréchet case

Theorem 5.8 (Fréchet MDA, Gnedenko (1943))

For $\xi > 0$, $F \in \text{MDA}(H_\xi)$ if and only if $\bar{F}(x) = x^{-1/\xi} L(x)$ for some $L \in \text{SV}$. If $F \in \text{MDA}(H_\xi)$, $\xi > 0$, the normalizing sequences can be chosen as $c_n = F^-(1 - 1/n)$ and $d_n = 0$, $n \in \mathbb{N}$.

Proof. Non-trivial. For a sketch, see Embrechts et al. (1997, p. 131). \square

- **Interpretation:** If $\bar{F} \in \text{RV}_{-1/\xi}$ (decay like a power function; Pareto like), then $F \in \text{MDA}(H_\xi)$ for $\xi > 0$; $\alpha = 1/\xi$ is known as *tail index*.
- *L can destroy the power*, but not too much (in (statistical) practice it matters, though)
- If $X \sim F$, $X \geq 0$, $\bar{F} \in \text{RV}_{-\alpha}$, $\alpha > 0$ (equivalently, $F \in \text{MDA}(H_\xi)$, $\xi > 0$), then $\mathbb{E}[X^k] < \infty$ if $k < \alpha = 1/\xi$, $\mathbb{E}[X^k] = \infty$ if $k > \alpha = 1/\xi$; see Embrechts et al. (1997, p. 568).

- One can show the *von Mises condition*:
If F has density f with $\lim_{x \uparrow \infty} \frac{xf(x)}{F(x)} = \alpha > 0$, then $F \in \text{MDA}(H_\xi)$, $\xi > 0$.
- Examples in $\text{MDA}(H_\xi)$, $\xi > 0$: inverse gamma, Student t , log-gamma, F , Cauchy, α -stable with $0 < \alpha < 2$, Burr and Pareto

Example 5.9 (Pareto distribution)

For $F = \text{Par}(\theta, \kappa)$, $\bar{F}(x) = (\kappa/(\kappa + x))^\theta = (1 + x/\kappa)^{-\theta} = x^{-\theta} L(x)$, $x \geq 0$, $\theta, \kappa > 0$, where $L(x) = (\kappa^{-1} + x^{-1})^{-\theta} \in \text{SV}$. We see (again) that $F \in \text{MDA}(H_\xi)$, $\xi > 0$.

The Gumbel case

The characterization of this class is more complicated. One can show the following result non-trivial; see Embrechts et al. (1997, p. 142).

Theorem 5.10 (Gumbel MDA)

$F \in \text{MDA}(H_0)$ if and only if there exists $z < x_F \leq \infty$ such that

$$\bar{F}(x) = c(x) \exp\left(-\int_z^x \frac{g(t)}{a(t)} dt\right), \quad x \in (z, x_F),$$

where c and g are measurable functions satisfying $c(x) \rightarrow c > 0$, $g(x) \rightarrow 1$ for $x \uparrow x_F$ and $a(x) > 0$ with density a' satisfying $\lim_{x \uparrow x_F} a'(x) = 0$.

If $F \in \text{MDA}(H_0)$, the normalizing sequences can be chosen as $c_n = a(d_n)$ for $a(x) = \int_x^{x_F} \bar{F}(t) dt / \bar{F}(x)$, $x < x_F$, (the mean excess function), and $d_n = F^{-}(1 - 1/n)$, $n \in \mathbb{N}$.

- Essentially $\text{MDA}(H_0)$ contains dfs whose tails decay roughly exponentially (*light-tailed*), but the tails can be quite different (up to moderately heavy). All moments exist for distributions in the Gumbel class, but both $x_F < \infty$ and $x_F = \infty$ are possible.

- **Examples in** $\text{MDA}(H_0)$: normal, log-normal, exponential, gamma (exponential, Erlang, χ^2), standard Weibull, Benktander type I and II, generalized hyperbolic (not: Student t).

The Weibull case

Theorem 5.11 (Weibull MDA)

For $\xi < 0$, $F \in \text{MDA}(H_\xi)$ if and only if $x_F < \infty$ and $\bar{F}(x_F - 1/x) = x^{1/\xi} L(x)$ for some $L \in \text{SV}$. If $F \in \text{MDA}(H_\xi)$, $\xi < 0$, the normalizing sequences can be chosen as $c_n = x_F - F^{-}(1 - 1/n)$ and $d_n = x_F$, $n \in \mathbb{N}$.

Proof. Non-trivial. For a sketch, see Embrechts et al. (1997, p. 135). □

- One can show the *von Mises condition*:

If F has density f which is positive on some finite interval (z, x_F) and if $\lim_{x \uparrow x_F} \frac{(x_F - x)f(x)}{\bar{F}(x)} = \alpha > 0$, then $F \in \text{MDA}(H_\xi)$, $\xi < 0$.

- **Examples in** $\text{MDA}(H_\xi)$, $\xi < 0$: beta (uniform). All $F \in \text{MDA}(H_\xi)$, $\xi < 0$, share $x_F < \infty$.

5.1.3 Maxima of strictly stationary time series

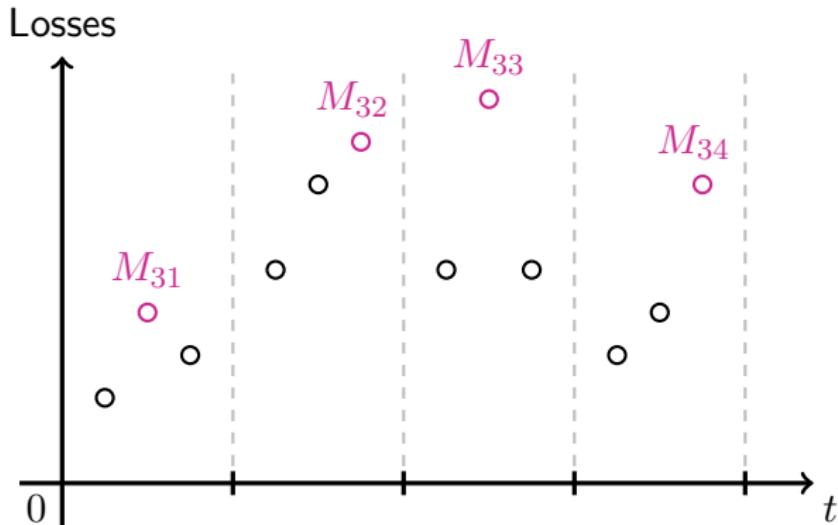
What about maxima of strictly stationary time series?

- Let $(X_k)_{k \in \mathbb{Z}}$ denote a strictly stationary time series with stationary distribution $X_k \sim F$, $k \in \mathbb{Z}$.
- Let $\tilde{X}_k \stackrel{\text{ind.}}{\sim} F$, $k \in \mathbb{Z}$, and $\tilde{M}_n = \max\{\tilde{X}_1, \dots, \tilde{X}_n\}$. For many processes one can show that there exists a real number $\theta \in (0, 1]$ such that $\lim_{n \uparrow \infty} \mathbb{P}((M_n - d_n)/c_n \leq x) = H^\theta(x)$ if and only if $\lim_{n \uparrow \infty} \mathbb{P}((\tilde{M}_n - d_n)/c_n \leq x) = H(x)$ (n.d.); θ is known as *extremal index*.
- If $F \in \text{MDA}(H_\xi)$ for some ξ $\Rightarrow \tilde{M}_n$ converges in distribution to H_ξ $\Rightarrow M_n$ converges in distribution to H_ξ^θ . Since H_ξ^θ is of the same type as H_ξ , the limiting distribution of the block maxima of the dependent series is the same as in the i.i.d. case (only location/scale may change).

- For large n , $\mathbb{P}((M_n - d_n)/c_n \leq x) \approx H^\theta(x) \approx F^{n\theta}(c_n x + d_n)$, so the distribution of M_n from a time series with extremal index θ can be approximated by the distribution $\tilde{M}_{n\theta}$ of the maximum of $n\theta < n$ observations from the associated i.i.d. series. $\Rightarrow n\theta$ counts the number of roughly independent clusters in n observations (θ is often interpreted as “1/mean cluster size”).
- If $\theta = 1$, large sample maxima behave as in the i.i.d. case; if $\theta \in (0, 1)$, large sample maxima tend to cluster.
- **Examples** (see Embrechts et al. (1997, pp. 216, pp. 415, pp. 476))
 - ▶ Strict white noise (iid rvs): $\theta = 1$;
 - ▶ ARMA processes with (ε_t) strict white noise: $\theta = 1$ (Gaussian); $\theta \in (0, 1)$ (if df of ε_t is in $MDA(H_\xi)$, $\xi > 0$);
 - ▶ (G)ARCH processes: $\theta \in (0, 1)$.

5.1.4 The block maxima method (BMM)

The basic idea in a picture based on losses X_1, \dots, X_{12} :



Consider the maximal loss from each block and fit $H_{\xi, \mu, \sigma}$ to them.

Fitting the GEV distribution

- Suppose $(x_i)_{i \in \mathbb{N}}$ are realizations of $(X_i)_{i \in \mathbb{N}} \stackrel{\text{ind.}}{\sim} F \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$, or of a process with an extremal index such as GARCH. The Fisher–Tippett–Gnedenko Theorem implies that

$$\mathbb{P}(M_n \leq x) = \mathbb{P}((M_n - d_n)/c_n \leq (x - d_n)/c_n) \underset{n \text{ large}}{\approx} H_{\xi, \mu=d_n, \sigma=c_n}(x).$$

- For fitting $\theta = (\xi, \mu, \sigma)$, we assume our realizations can be divided into m blocks of size n denoted by M_{n1}, \dots, M_{nm} (often naturally the case, e.g., in hydrology: daily water levels \Rightarrow yearly maxima; in finance: daily log-returns \Rightarrow monthly/quarterly/yearly maxima).
- Assume the block size n to be sufficiently large so that (regardless of whether the underlying data are dependent or not), the block maxima can be considered independent.

- The log-likelihood is

$$\ell(\boldsymbol{\theta}; M_{n1}, \dots, M_{nm}) = \sum_{i=1}^m \log\left(\frac{1}{\sigma} h_\xi\left(\frac{M_{ni} - \mu}{\sigma}\right) \mathbb{1}_{\{1+\xi(M_{ni}-\mu)/\sigma > 0\}}\right).$$

Maximize w.r.t. $\boldsymbol{\theta} = (\xi, \mu, \sigma)$ to get $\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\mu}, \hat{\sigma})$.

Remark 5.12

- 1) Sufficiently many/large blocks require large amounts of data.
- 2) Bias and variance must be traded off (*bias-variance tradeoff*):
 - Block size $n \uparrow \Rightarrow$ GEV approximation more accurate \Rightarrow bias \downarrow
 - Number of blocks $m \uparrow \Rightarrow$ more data for MLE \Rightarrow variance \downarrow
- 3) There is no general best strategy known to find the optimal block size.
- 4) The support of the density depends on the parameters \Rightarrow not differentiable; classical MLE regularity conditions for consistency and asymptotic efficiency do not apply. For $\xi > -1/2$ (fine for practice), Smith (1985) showed that the MLE is regular.

Example 5.13 (Block maxima analysis of S&P500)

Suppose it is Friday 1987-10-16. The S&P 500 index fell by 9.12% this week. On that Friday alone the index is down 5.16% on the previous day (largest one-day fall since 1962). We fit a GEV distribution to annual maxima of daily negative returns $X_t = S_t/S_{t-1} - 1$ since 1960.

Analysis 1: Based on annual maxima ($m = 28$; including the latest from the incomplete year 1987): $\hat{\theta} = (0.29, 2.03, 0.72) \Rightarrow$ heavy-tailed Fréchet distribution (infinite third moment). The corresponding standard errors are $(0.21, 0.0016, 0.0014) \Rightarrow$ High uncertainty (m small) for estimating ξ .

Analysis 2: Based on biannual maxima ($m = 56$): $\hat{\theta} = (0.33, 1.68, 0.55)$ with standard errors $(0.14, 0.0009, 0.0007) \Rightarrow$ Hints at even heavier tails.

Return levels and stress losses (exceedances)

The fitted GEV model can be used to estimate:

- 1) The size of an event with prescribed frequency (*return-level problem*)
- 2) The frequency of an event with prescribed size (*return-period problem*)

Definition 5.14 (Return level, return period)

Let $M_n \sim H$ (exact). The *k n-block return level* is $r_{n,k} = H^-(1 - 1/k)$.

The *return period* of the event $\{M_n > u\}$ is $k_{n,u} = 1/\bar{H}(u)$.

- $r_{n,k}$ is the level which is exceeded (on average) in one out of every k *n-blocks*, so $r_{n,k}$ solves $\mathbb{P}(M_n > r_{n,k}) = 1/k$ (e.g., 10-year return level $r_{260,10}$ = level exceeded in one out of every 10 years; 260d \approx 1 year).
- $k_{n,u}$ is the number of *n-blocks* for which we expect to see a single *n-block* exceeding u , so $k_{n,u}$ solves $r_{n,k_{n,u}} = H^-(1 - 1/k_{n,u}) = u$.

- Parametric estimators are given by

$$\hat{r}_{n,k} = H_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}^-(1 - 1/k) = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}((- \log(1 - 1/k))^{-\hat{\xi}} - 1),$$

$$\hat{k}_{n,u} = 1/\bar{H}_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}(u).$$

Confidence intervals for $r_{n,k}$, $k_{n,u}$ can be constructed via profile-likelihoods; see Davison (2003, pp. 126) and McNeil et al. (2005, p. 274).

Example 5.15 (Block maxima analysis of S&P500 (continued))

- The 10-year return level $r_{260,10}$ based on data up to and including Friday 1987-10-16 is estimated as $\hat{r}_{260,10} = 4.32\%$. The next trading day is Black Monday (1987-10-19), the event of an index drop of 20.47% is far beyond $\hat{r}_{260,10}$. One can show that 20.47% is in the 95% confidence interval of $r_{260,50}$ (estimated as $\hat{r}_{260,50} = 7.23\%$), but the 28 maxima are too few to get a reliable estimate of a once-in-50-years event.
- If we estimate the return period $k_{260,0.2047}$ of a loss of 20.47%, the point estimate is $\hat{k}_{260,0.2047} = 1629$ years. One can show that the 95%

confidence interval encompasses everything from 45 years to essentially never! \Rightarrow Very high uncertainty involved in estimating $k_{260,0.2047}$.

- In summary, on 1987-10-16 we simply did not have enough data to say anything meaningful about an event of this magnitude. This illustrates the difficulties of attempting to quantify events beyond our empirical experience.

5.2 Threshold exceedances

The BMM is wasteful of data (only the maxima of large blocks are used). It has been largely superseded in practice by methods based on threshold exceedances (*peaks-over-threshold (POT) approach*), where all data above a designated high threshold u are used.

5.2.1 Generalized Pareto distribution

Definition 5.16 (Generalized Pareto distribution (GPD))

The *generalized Pareto distribution (GPD)* is given by

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

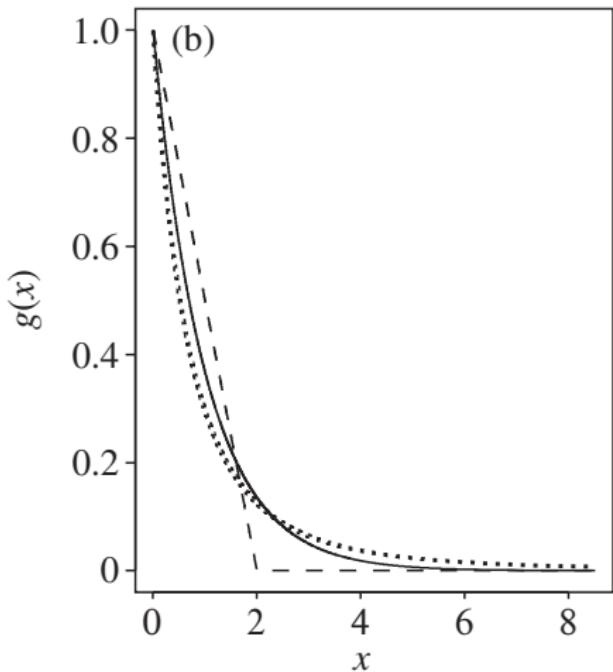
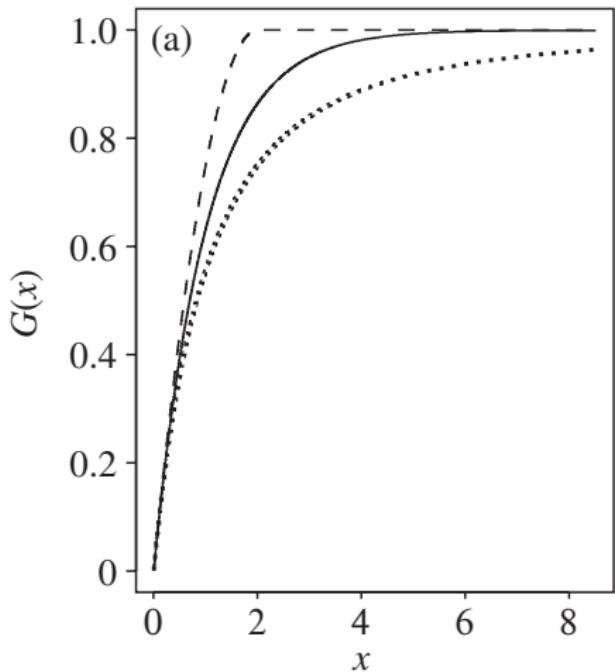
where $\beta > 0$, and the support is $x \geq 0$ when $\xi \geq 0$ and $x \in [0, -\beta/\xi]$ when $\xi < 0$.

- The parameterization is continuous in ξ .
- The larger ξ , the heavier tailed $G_{\xi,\beta}$ (if $\xi > 0$, $\mathbb{E}[X^k] = \infty$ iff $k \geq \frac{1}{\xi}$; if $\xi < 1$, then $\mathbb{E}[X] = \beta/(1 - \xi)$).
- ξ is known as *shape*; β as *scale*. Special cases:
 - 1) $\xi > 0$: Par($1/\xi, \beta/\xi$)
 - 2) $\xi = 0$: Exp($1/\beta$)
 - 3) $\xi < 0$: short-tailed Pareto type II distribution
- The density $g_{\xi,\beta}$ of $G_{\xi,\beta}$ is given by

$$g_{\xi,\beta}(x) = \begin{cases} \frac{1}{\beta}(1 + \xi x/\beta)^{-1/\xi-1}, & \text{if } \xi \neq 0, \\ \frac{1}{\beta} \exp(-x/\beta), & \text{if } \xi = 0, \end{cases}$$

where $x \geq 0$ when $\xi \geq 0$ and $x \in [0, -\beta/\xi)$ when $\xi < 0$ (MLE!).
- $G_{\xi,\beta} \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$ (follows from Theorems 5.8, 5.10 and 5.11)

(a): $G_{\xi,1}$; (b): density $g_{\xi,1}$ for $\xi \in \{-0.5, 0, 0.5\}$ (dashed, solid, dotted)



Definition 5.17 (Excess distribution over u , mean excess function)

Let $X \sim F$. The *excess distribution over the threshold u* is defined by

$$F_u(x) = \mathbb{P}(X - u \leq x \mid X > u) = \frac{F(x + u) - F(u)}{1 - F(u)}, \quad x \in [0, x_F - u].$$

If $\mathbb{E}|X| < \infty$, the *mean excess function* is defined by

$$e(u) = \mathbb{E}[X - u \mid X > u] \quad (\text{mean w.r.t. } F_u)$$

Interpretation

F_u describes the distribution of the loss over u (excess), given that u is exceeded. $e(u)$ is the mean of F_u as a function in u .

- One can show the useful formula $e(u) = \frac{1}{\bar{F}(u)} \int_u^{x_F} \bar{F}(x) dx$.
- For continuous $X \sim F$ with $\mathbb{E}|X| < \infty$, the following formula holds:

$$\text{ES}_\alpha(X) = e(\text{VaR}_\alpha(X)) + \text{VaR}_\alpha(X), \quad \alpha \in (0, 1); \quad (12)$$

- The results of the following example are easy to check.

Example 5.18 (F_u , $e(u)$ for $\text{Exp}(\lambda)$, $G_{\xi,\beta}$)

- 1) If F is $\text{Exp}(\lambda)$, then $F_u(x) = 1 - e^{-\lambda x}$, $x \geq 0$ (so again $\text{Exp}(\lambda)$; lack-of-memory property). The mean excess function is $e(u) = 1/\lambda = \mathbb{E}X$.
 - 2) If F is $G_{\xi,\beta}$, then $F_u(x) = G_{\xi,\beta+\xi u}(x)$, $x \geq 0$ (so again GPD, with the same shape, only the scale grows linearly in u) \Rightarrow Important for (re)insurance (u denotes the threshold determined by an insurance contract; everything above needs to be covered by reinsurance). This will also allow us to conduct estimation of risk measures lower in the tail and then scale up (see later; one of the core applications of EVT).
- The mean excess function of $G_{\xi,\beta}$ is

$$e(u) = \frac{\beta + \xi u}{1 - \xi}, \quad \text{for all } u : \beta + \xi u > 0,$$

which is linear in u (this is a characterizing property of the GPD and used to determine u). Note that ξ determines the slope of $e(u)$.

Theorem 5.19 (Pickands–Balkema–de Haan (1974/75))

There exists a positive, measurable function $\beta(u)$, such that

$$\lim_{u \uparrow x_F} \sup_{0 \leq x < x_F - u} |F_u(x) - G_{\xi, \beta(u)}(x)| = 0.$$

if and only if $F \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$.

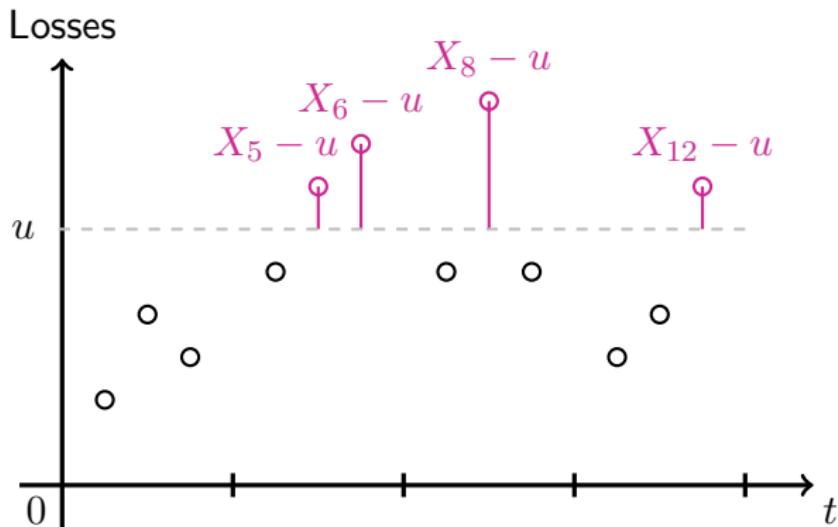
Proof. Non-trivial; see, e.g., Pickands (1975) and Balkema and de Haan (1974). \square

Interpretation

- GPD = Canoncial df for modeling excess losses over high u .
- The result is also a characterization of $\text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$. All $F \in \text{MDA}(H_\xi)$ form a set of df for which the excess distribution converges to the GPD $G_{\xi, \beta}$ with the same ξ as in H_ξ as the threshold u is raised.

5.2.2 Modeling excess losses

The basic idea in a picture based on losses X_1, \dots, X_{12} .



Consider all excesses over u and fit $G_{\xi,\beta}$ to them.

The method

- Given losses $X_1, \dots, X_n \sim F \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$, let
 - ▶ $N_u = |\{i \in \{1, \dots, n\} : X_i > u\}|$ denote the *number of exceedances* over the (given; see later) threshold u ;
 - ▶ $\tilde{X}_1 < \dots < \tilde{X}_{N_u}$ denote the ordered *exceedances*; and
 - ▶ $Y_k = \tilde{X}_k - u$, $k \in \{1, \dots, N_u\}$, the corresponding *excesses*.
- If Y_1, \dots, Y_{N_u} are *i.i.d.* and (roughly) distributed as $G_{\xi, \beta}$, the *log-likelihood* is given by

$$\begin{aligned}\ell(\xi, \beta; Y_1, \dots, Y_{N_u}) &= \sum_{k=1}^{N_u} \log g_{\xi, \beta}(Y_k) \\ &= -N_u \log(\beta) - (1 + 1/\xi) \sum_{k=1}^{N_u} \log(1 + \xi Y_k / \beta)\end{aligned}$$

⇒ Maximize w.r.t. $\beta > 0$ and $1 + \xi Y_k / \beta > 0$ for all $k \in \{1, \dots, N_u\}$.

Non-i.i.d. data

- If X_1, \dots, X_n are serially dependent and show no tendency of clusters of extreme values (extremal index $\theta = 1$), asymptotic theory of point processes suggests a limiting model for high-level threshold exceedances, in which exceedances occur according to a Poisson process and the excess losses are i.i.d. generalized Pareto distributed.
- If extremal clustering is present ($\theta < 1$; e.g., (G)ARCH processes), the assumption of independent excess losses is less satisfactory. Easiest approach: neglect the problem, simply apply MLE which is then a quasi-MLE (QMLE) (likelihood misspecified); point estimates should still be reasonable, standard errors may be too small.
- See Section 5.3 for more details on threshold exceedances.

Excesses over higher thresholds

Once a model is fitted to F_u , we can infer a model for F_v , $v \geq u$.

Lemma 5.20

Assume, for some u , $F_u(x) = G_{\xi,\beta}(x)$ for $0 \leq x < x_F - u$. Then $F_v(x) = G_{\xi,\beta+\xi(v-u)}(x)$ for all $v \geq u$.

Proof. Recall that $F_u(x) = \mathbb{P}(X - u \leq x | X > u) = \frac{F(u+x) - F(u)}{\bar{F}(u)}$, so $\bar{F}_u(x) = \bar{F}(u+x)/\bar{F}(u)$. For $v \geq u$, we have

$$\begin{aligned}\bar{F}_v(x) &= \frac{\bar{F}(v+x)}{\bar{F}(v)} = \frac{\bar{F}(u + (v+x-u))}{\bar{F}(u)} \frac{\bar{F}(u)}{\bar{F}(u + (v-u))} \\ &= \frac{\bar{F}_u(v+x-u)}{\bar{F}_u(v-u)} = \frac{\bar{G}_{\xi,\beta}(x+v-u)}{\bar{G}_{\xi,\beta}(v-u)} \stackrel{\text{check}}{=} \bar{G}_{\xi,\beta+\xi(v-u)}(x) \quad \square\end{aligned}$$

⇒ The excess distribution over $v \geq u$ remains GPD with the same ξ (and β growing linearly in v); makes sense for a limiting distribution for $u \uparrow$.

If $\xi < 1$, the mean excess function is given by

$$e(v) = \frac{\xi}{1-\xi} v + \frac{\beta - \xi u}{1-\xi}, \quad v \in [u, \infty) \text{ if } \xi \in [0, 1), \quad (13)$$

and $v \in [u, u - \beta/\xi]$ if $\xi < 0$. This forms the bases for a graphical method for choosing u .

Sample mean excess plot and choice of the threshold

Definition 5.21 (Sample mean excess function, mean excess plot)

Based on positive loss data X_1, \dots, X_n , the sample mean excess function is defined by

$$e_n(v) = \frac{\sum_{i=1}^n (X_i - v) \mathbb{1}_{\{X_i > v\}}}{\sum_{i=1}^n \mathbb{1}_{\{X_i > v\}}}, \quad X_{(n)} > v.$$

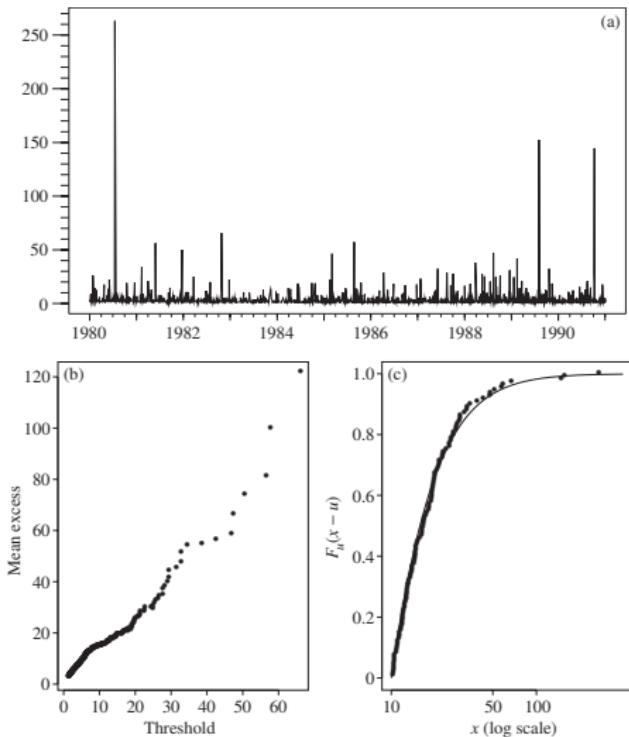
The mean excess plot is the plot of $\{(X_{(i)}, e_n(X_{(i)})) : 1 \leq i \leq n-1\}$, where $X_{(i)}$ denotes the i th order statistic.

- If the GPD model over u , the $e_n(v)$ should become increasingly “linear” for higher values of u . An upward/zero/downward trend indicates $\xi > 0/\xi = 0/\xi < 0$.
- The sample mean excess plot is rarely perfectly linear (particularly for large u where one averages over a small number of excesses).
- The choice of a good threshold u is as difficult as finding an adequate block size for the Block Maxima method. There are data-driven tools (e.g., sample mean excess plot), but there is no general method to determine an optimal threshold (without second-order assumptions on $L \in \text{SV}$).
- Typically, select u as the smallest point where $e_n(v)$, $v \geq u$, becomes linear. Rule-of-thumb: One needs a couple of thousand data points and can often take u around the 0.9-quantile.
- One should always analyze the data for several u and check the sensitivity of the choice of u .

Example 5.22 (Danish fire loss data)

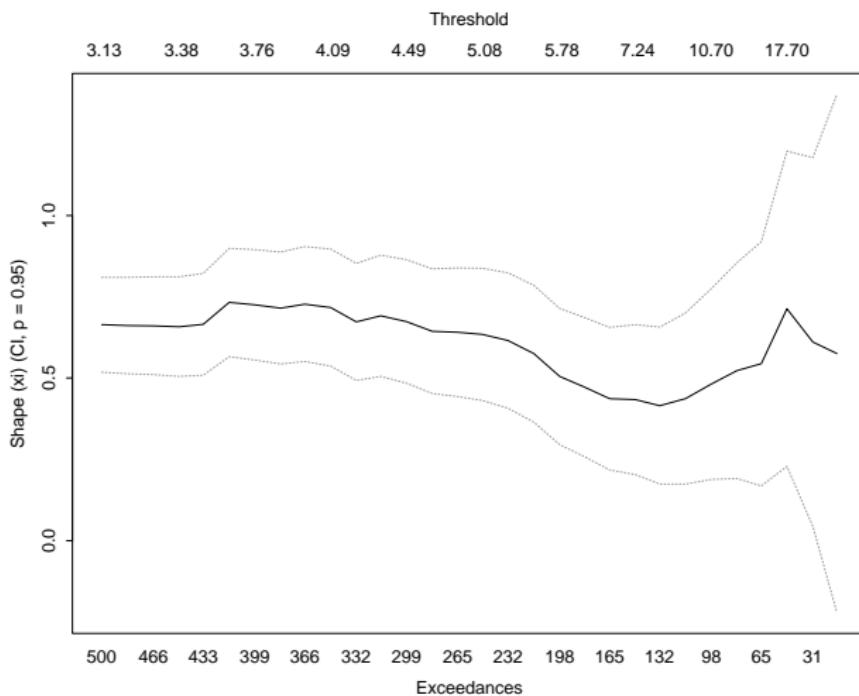
- 2156 fire insurance losses over 1M Danish kroner from 1980-01-03 to 1990-12-31; combined loss for a building and its contents, in some cases also a loss of business earnings. The losses are inflation adjusted to reflect values as of 1985.
- The mean excess function shows a “kink” below 10; “straightening out” above 10 \Rightarrow Our choice is $u = 10$ (so 10M Danish kroner).
- MLE $(\hat{\xi}, \hat{\beta}) = (0.50, 7.0)$ (with standard errors $(0.14, 1.1)$)
 \Rightarrow very heavy-tailed, infinite-variance model
- We can then estimate the expected loss given exceedance of 10M kroner or any higher threshold (via $e(v)$ in (13) based on $\hat{\xi}, \hat{\beta}$ and the chosen u), even beyond the data.
 \Rightarrow EVT allows us to estimate “in the data” and then “scale up”.

(a): Losses ($> 1M$; in M); (b): $e_n(u)$ (\uparrow); (c) empirical $F_u(x - u)$, $G_{\hat{\xi}, \hat{\beta}}$



⇒ Choose the threshold $u = 10$

Sensitivity of the estimated shape parameter $\hat{\xi}$ to changes in u :



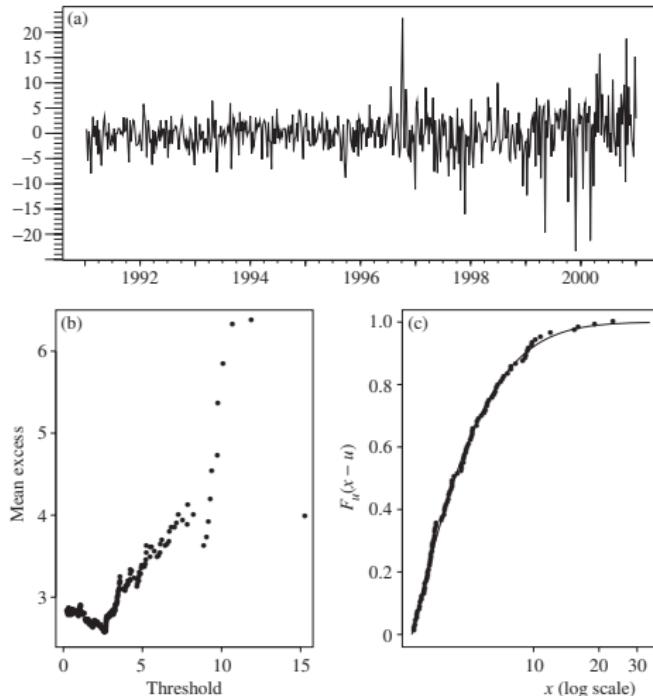
Example 5.23 (AT&T weekly loss data)

- Let (X_t) denote weekly log-returns and consider the percentage one-week loss as a fraction of S_t , given by

$$100L_{t+1}/S_t \stackrel{(1)}{=} 100(-S_t(\exp(X_{t+1}) - 1))/S_t = 100(1 - \exp(X_{t+1})).$$

- We have 521 such losses (period 1991–2000).
- The estimated GPD parameters are $\hat{\xi} = 0.22$ and $\hat{\beta} = 2.1$ (MLEs) with standard errors 0.13 and 0.34, respectively. The fitted model is thus close to having an infinite fourth moment.
- Note that we ignored here that monthly AT&T data over 1993–2000 is actually not consistent with the i.i.d. assumption (absolute values of log-returns reject the hypothesis of serial uncorrelatedness via the Ljung–Box test).

(a): % losses (1991–2000); (b): $e_n(u)$; (c): empirical $F_u(x - u)$, $G_{\hat{\xi}, \hat{\beta}}$.



⇒ Choose the threshold $u = 2.75\%$ (102 exceedances)

5.2.3 Modeling tails and measures of tail risk

- How can the fitted GPD model be used to estimate the tail of the loss distribution F and associated risk measures?
- Assume $F_u(x) = G_{\xi,\beta}(x)$ for $0 \leq x < x_F - u$, $\xi \neq 0$ and some u .
- We obtain the following GPD-based formula for tail probabilities:

$$\begin{aligned}\bar{F}(x) &= \mathbb{P}(X > u)\mathbb{P}(X > x | X > u) \\ &= \bar{F}(u)\mathbb{P}(X - u > x - u | X > u) = \bar{F}(u)\bar{F}_u(x - u) \\ &= \bar{F}(u)\left(1 + \xi \frac{x - u}{\beta}\right)^{-1/\xi}, \quad x \geq u.\end{aligned}\tag{14}$$

- Assuming we know $\bar{F}(u)$, inverting this formula for $\alpha \geq F(u)$ leads to

$$\text{VaR}_\alpha = F^-(\alpha) = u + \frac{\beta}{\xi} \left(\left(\frac{1 - \alpha}{\bar{F}(u)} \right)^{-\xi} - 1 \right),\tag{15}$$

$$\text{ES}_\alpha = \frac{\text{VaR}_\alpha}{1 - \xi} + \frac{\beta - \xi u}{1 - \xi}, \quad \xi < 1.\tag{16}$$

The formula for ES_α can also be obtained from $e(\cdot)$ via (12) and (13).

- $\bar{F}(x)$, VaR_α and ES_α are all of the form $g(\xi, \beta, \bar{F}(u))$. If we have sufficient samples above u , we obtain semi-parametric plug-in estimators via $g(\hat{\xi}, \hat{\beta}, N_u/n)$.
- We hope to gain over empirical estimators by using a kind of extrapolation based on the GPD for more extreme tail probabilities and risk measures.
- For example, based on (14), Smith (1987) proposed the semi-parametric tail estimator

$$\hat{\bar{F}}(x) = \frac{N_u}{n} \left(1 + \hat{\xi} \frac{x - u}{\hat{\beta}} \right)^{-1/\hat{\xi}}, \quad x \geq u;$$

also known as the *Smith estimator* (note that it is only valid for $x \geq u$).

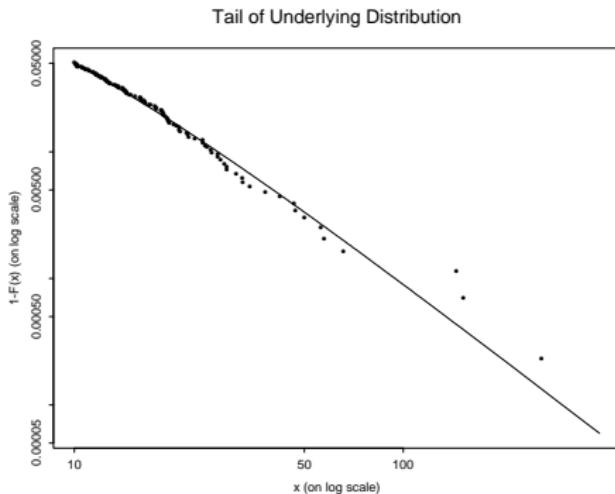
\Rightarrow Bias-variance tradeoff: $u \uparrow \Rightarrow$ bias of parametrically estimating $\bar{F}_u(x - u) \downarrow$, but variance of non-parametrically estimating $\bar{F}(u) \uparrow$

- GPD-based $\widehat{\text{VaR}}_\alpha$, $\widehat{\text{ES}}_\alpha$ for $\alpha \geq 1 - N_u/n$ can be obtained similarly from (15), (16).

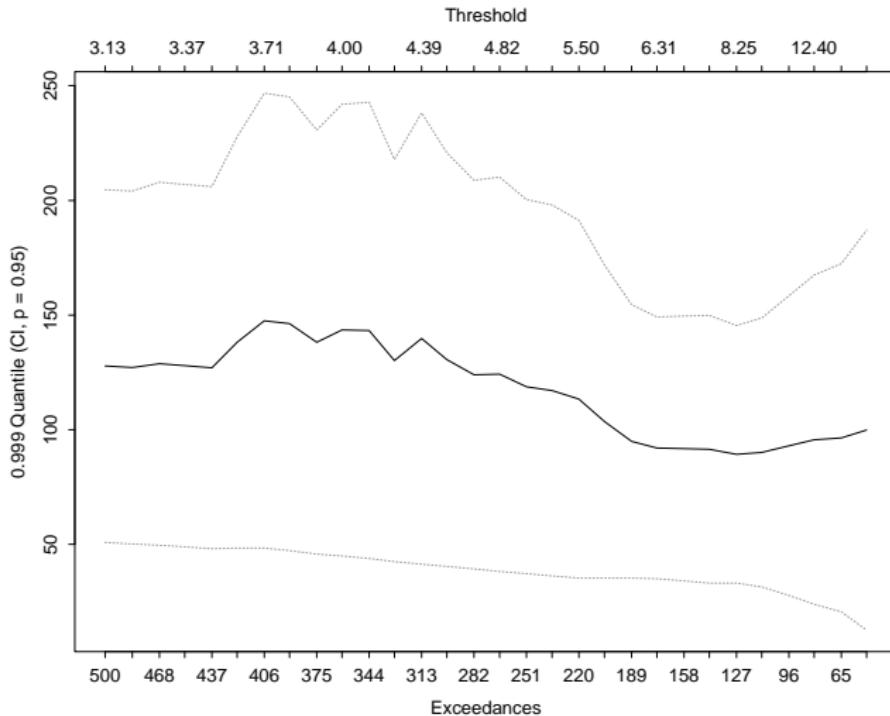
- Confidence intervals for $\bar{F}(x)$, $x \geq u$, VaR_α , ES_α can be obtained likelihood-based (neglecting the uncertainty in N_u/n): Reparametrize the GPD model in terms of $\phi = g(\xi, \beta, N_u/n)$ and construct a confidence interval for ϕ based on the likelihood ratio test.

Example 5.24 (Danish fire loss data (continued))

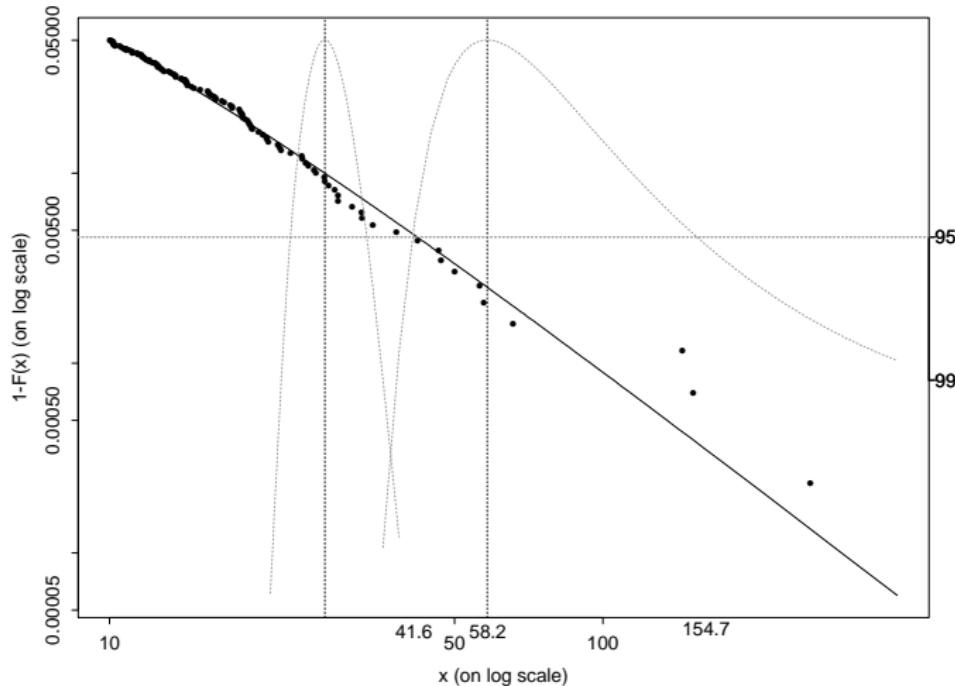
The semi-parametric Smith/tail estimator $\hat{\bar{F}}(x)$, $x \geq u$ is given by:



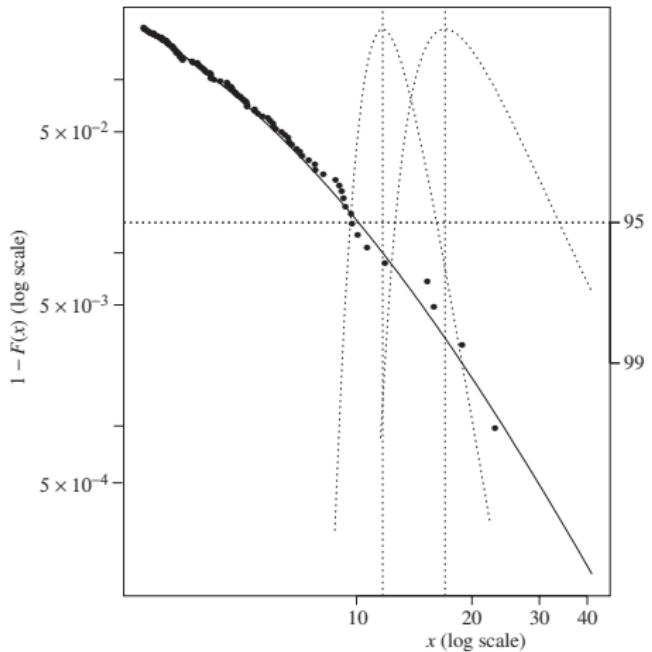
It is important to check the sensitivity of \hat{F} (or $\widehat{\text{VaR}}_\alpha$, $\widehat{\text{ES}}_\alpha$) w.r.t. u .



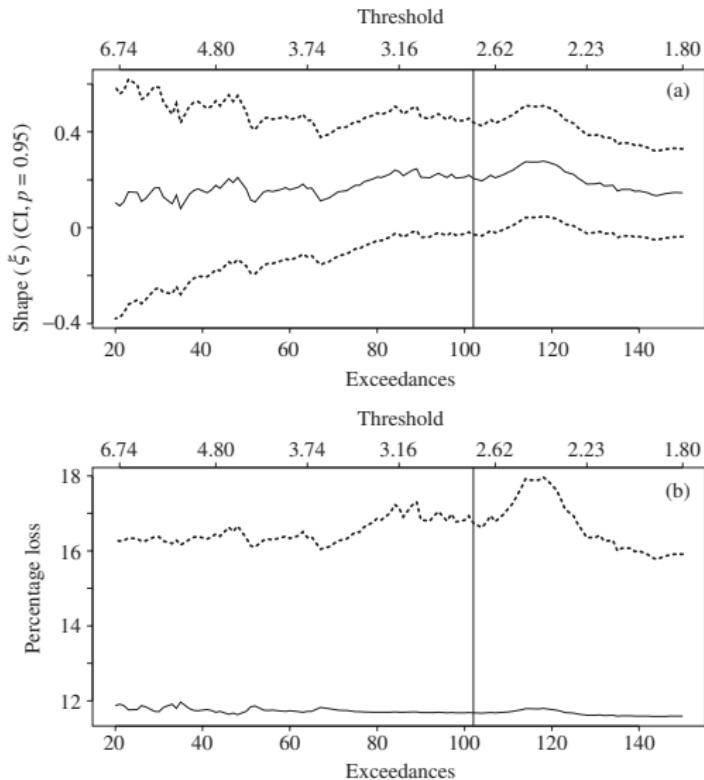
Here are $\hat{F}(x)$, $x \geq u$, $\widehat{\text{VaR}}_{0.99}$, $\widehat{\text{ES}}_{0.99}$ including confidence intervals.



Example 5.25 (AT&T weekly loss data (continued))



- Fitted GPD model as in Example 5.23.
- Plot of $\hat{F}(x)$.
- Vertical lines: $\widehat{\text{VaR}}_{0.99}$, $\widehat{\text{ES}}_{0.99}$
- **log-log scale often good:** $\bar{F}(x) = x^{-\alpha} L(x)$ and therefore
 $\log \bar{F}(x) = -\alpha \log(x) + \log L(x)$
 \approx linear in $\log(x)$



- Sensitivity w.r.t. u
- Top: $\hat{\xi}$ for different u or N_u , including a 95% CI based on standard error
- Bottom: Corresponding $\widehat{\text{VaR}}_{0.99}$ (solid line), $\widehat{\text{ES}}_{0.99}$ (dotted line)

5.2.4 The Hill estimator

- Assume $F \in \text{MDA}(H_\xi)$, $\xi > 0$, so that $\bar{F}(x) = x^{-\alpha} L(x)$, $\alpha > 0$.
- Let e^* be the mean excess function for $\log X$. Using partial integration ($\int H dG = [HG] - \int G dH$), we obtain

$$e^*(\log u) = \mathbb{E}(\log X - \log u \mid \log X > \log u)$$

$$\begin{aligned} &= \frac{1}{\bar{F}(u)} \int_u^\infty (\log x - \log u) dF(x) = -\frac{1}{\bar{F}(u)} \int_u^\infty \log\left(\frac{x}{u}\right) d\bar{F}(x) \\ &= -\frac{1}{\bar{F}(u)} \left(\underbrace{\left[\log\left(\frac{x}{u}\right) \bar{F}(x) \right]_u^\infty}_{=0} - \int_u^\infty \bar{F}(x) \frac{1}{x} dx \right) \\ &= \frac{1}{\bar{F}(u)} \int_u^\infty \frac{\bar{F}(x)}{x} dx = \frac{1}{\bar{F}(u)} \int_u^\infty x^{-\alpha-1} L(x) dx. \end{aligned}$$

For u sufficiently large, $L(x) \approx L(u)$, $x \geq u$ (Karamata's Theorem), so

$$e^*(\log u) \underset{u \text{ large}}{\approx} \frac{L(u)u^{-\alpha}/\alpha}{\bar{F}(u)} = \frac{1}{\alpha}.$$

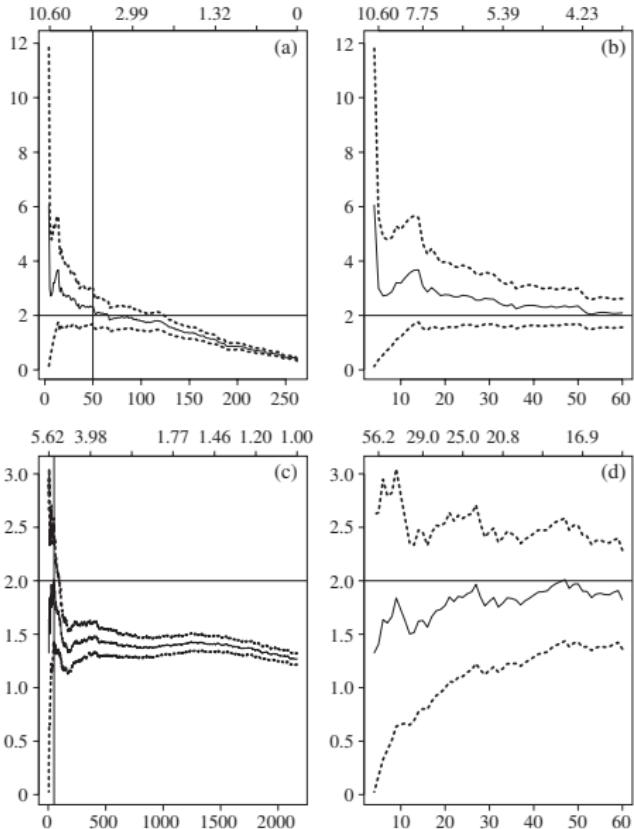
- For n large, k sufficiently small, use $u = X_{[k]}$, so

$$\begin{aligned}\frac{1}{\alpha} \approx e_n^*(\log X_{[k]}) &= \frac{\sum_{i=1}^n (\log X_i - \log X_{[k]}) \mathbb{1}_{\{\log X_i > \log X_{[k]}\}}}{\sum_{i=1}^n \mathbb{1}_{\{\log X_i > \log X_{[k]}\}}} \\ &= \frac{\sum_{i=1}^{k-1} (\log X_{[i]} - \log X_{[k]})}{k-1} = \frac{1}{k-1} \sum_{i=1}^{k-1} \log X_{[i]} - \log X_{[k]}\end{aligned}$$

- The standard form of the *Hill estimator* of the tail index α is

$$\hat{\alpha}_{k,n}^{(\text{H})} = \left(\frac{1}{k} \sum_{i=1}^k \log X_{[i]} - \log X_{[k]} \right)^{-1}, \quad 2 \leq k \leq n, \quad k \text{ sufficiently small.}$$

- Choosing k : Find a small k where the *Hill plot* $\{(k, \hat{\alpha}_{k,n}^{(\text{H})}) : 2 \leq k \leq n\}$ stabilizes (typically, $k = \lceil \beta n \rceil$, $\beta \in [0.01, 0.05]$).
- Interpreting Hill plots can be difficult. If F does not have a regularly varying tail (or if it has serial dependence), Hill plots can be very misleading.



- Hill plots showing estimates of $\alpha = 1/\xi$ for (a), (b) the AT&T data and (c),(d) the Danish fire loss data (rhs = expanded version of the lhs).
- (a),(b): suggests estimates of $\alpha \in [1.5, 2]$ ($\xi \in [1/2, 2/3]$; close to the estimated $\hat{\xi} = 0.50$, see Example 5.22); (c),(d): suggests estimates of $\alpha \in [2, 4]$ ($\xi \in [1/4, 1/2]$; larger than the estimated $\hat{\xi} = 0.22$, see Example 5.23)

Hill-based tail and risk measure estimates

- Assume $\bar{F}(x) = cx^{-\alpha}$, $x \geq u > 0$ (replacing L by a constant). Estimate α by $\hat{\alpha}_{k,n}^{(H)}$ and u by $X_{[k]}$ (for k sufficiently small).
- Note that $c = u^\alpha \bar{F}(u)$ so $\hat{c} = X_{[k]}^{\hat{\alpha}_{k,n}^{(H)}} \hat{F}_n(X_{[k]}) \approx X_{[k]}^{\hat{\alpha}_{k,n}^{(H)}} \frac{k}{n}$. We thus obtain the semi-parametric *Hill tail estimator*

$$\hat{F}(x) = \frac{k}{n} \left(\frac{x}{X_{[k]}} \right)^{-\hat{\alpha}_{k,n}^{(H)}}, \quad x \geq X_{[k]}.$$

- From this result we obtain the semi-parametric *Hill VaR estimator*

$$\widehat{\text{VaR}}_\alpha(X) = \left(\frac{n}{k} (1 - \alpha) \right)^{-\frac{1}{\hat{\alpha}_{k,n}^{(H)}}} X_{[k]}, \quad \alpha \geq F(u) \approx 1 - \frac{k}{n},$$

and, for $\hat{\alpha}_{k,n}^{(H)} > 1$, $\alpha \geq F(u) \approx 1 - \frac{k}{n}$, the semi-param. *Hill ES estimator*

$$\widehat{\text{ES}}_\alpha(X) = \frac{\left(\frac{n}{k} \right)^{\frac{1}{\hat{\alpha}_{k,n}^{(H)}}} X_{[k]}}{1 - \alpha} \int_\alpha^1 (1 - z)^{-\frac{1}{\hat{\alpha}_{k,n}^{(H)}}} dz = \frac{\hat{\alpha}_{k,n}^{(H)}}{\hat{\alpha}_{k,n}^{(H)} - 1} \widehat{\text{VaR}}_\alpha(X).$$

Interlude: Scaling of the risk measures $\text{VaR}_\alpha, \text{ES}_\alpha$

- Again assume $\bar{F}(x) = cx^{-\alpha}$, $x \geq u > 0$, and let $\hat{\alpha}$ denote a tail index estimator.
- As $\frac{\bar{F}(u)}{\bar{F}(x)} = (\frac{x}{u})^\alpha$, using $x := \text{VaR}_\beta(X)$ and $u := \text{VaR}_{\beta_u}(X)$ implies

$$\text{VaR}_\beta(X) = \left(\frac{1 - \beta_u}{1 - \beta} \right)^{\frac{1}{\alpha}} \text{VaR}_{\beta_u}(X). \quad (17)$$

This allows one to estimate VaR_β at $\beta_u \leq \beta$ (for $\beta_u \geq F(u)$):

$$\widehat{\text{VaR}}_\beta(X) = \left(\frac{1 - \beta_u}{1 - \beta} \right)^{\frac{1}{\hat{\alpha}}} \widehat{\text{VaR}}_{\beta_u}(X).$$

- For $\alpha > 1$, $\beta \geq \beta_u \geq F(u)$, a similar scaling for $\text{ES}_\beta(X)$ is

$$\begin{aligned} \text{ES}_\beta(X) &\stackrel{(17)}{=} \frac{(1 - \beta_u)^{\frac{1}{\alpha}} \text{VaR}_{\beta_u}(X)}{1 - \beta} \underbrace{\int_\beta^1 (1 - \tilde{\beta})^{-\frac{1}{\alpha}} d\tilde{\beta}}_{= \frac{\alpha}{\alpha-1} (1-\beta)^{1-\frac{1}{\alpha}}} \stackrel{(17)}{=} \frac{\alpha}{\alpha - 1} \text{VaR}_\beta(X) \end{aligned}$$

5.2.5 Simulation study of EVT quantile estimators

We compare estimators for ξ (Study 1) and $\text{VaR}_{0.99}$ (Study 2) based on

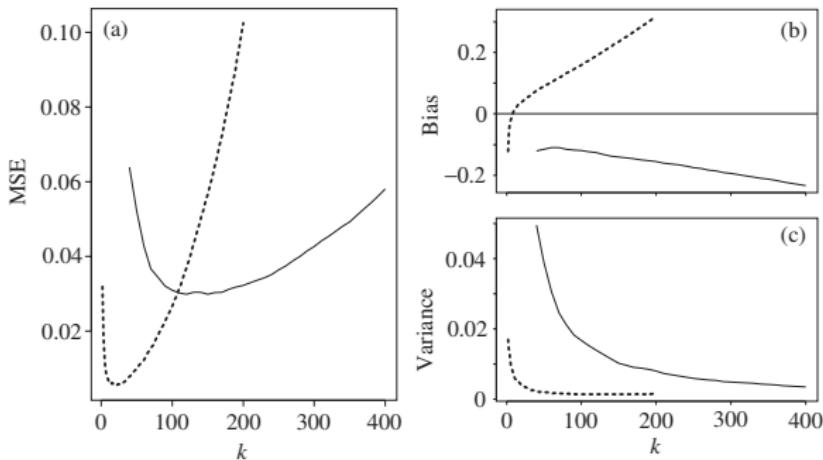
$$\begin{aligned}\text{MSE}[\hat{\theta}] &= \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + \mathbb{E}[2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)] + \mathbb{E}[(\mathbb{E}[\hat{\theta}] - \theta)^2] \\ &= (\mathbb{E}[\hat{\theta}] - \theta)^2 + \text{Var}[\hat{\theta}] = \text{bias}[\hat{\theta}]^2 + \text{Var}[\hat{\theta}]\end{aligned}$$

with a Monte Carlo study (Sample size $N = 1000$; drawn from a t_4 distribution with corresponding true $\xi = 1/4$); analytical evaluation of bias and variance is not possible.

Study 1: Estimating ξ

We estimate ξ with a fitted GPD (via MLE; $k \in \{30, 40, \dots, 400\}$) and with the Hill estimator ($\hat{\xi} = 1/\hat{\alpha}_{k,n}^{(H)}$; $k \in \{2, 3, \dots, 200\}$). Note that the t_4 distribution has a well-behaved regularly varying tail.

(a): $\widehat{\text{MSE}}[\hat{\xi}]$; (b): $\widehat{\text{bias}}[\hat{\xi}]$; (c): $\widehat{\text{Var}}[\hat{\xi}]$ (solid: GPD; dotted: Hill)

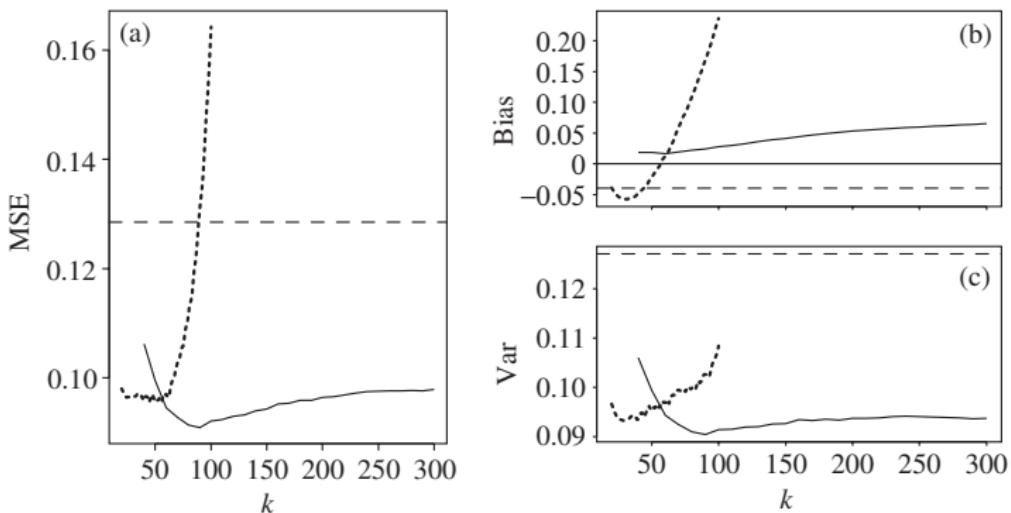


- The Hill estimator outperforms the GPD estimator (optimal k around 20–30) according to the variance for small k (number of order statistics)
- The biases are closer, with the Hill (GPD) estimator tending to overestimate (underestimate) ξ .
- For the GPD method, the optimal u is around 100–150 exceedances.

Study 2: Estimating VaR_{0.99}

Estimate VaR_{0.99} based on a fitted GPD, with the Hill VaR estimator and with the empirical quantile estimator. Here the situation changes.

(a): $\widehat{\text{MSE}}[\widehat{\text{VaR}}_{0.99}]$; (b): $\widehat{\text{bias}}[\widehat{\text{VaR}}_{0.99}]$; (c): $\widehat{\text{Var}}[\widehat{\text{VaR}}_{0.99}]$ (solid: GPD; dotted: Hill; dashed: empirical quantile estimator)



- The empirical VaR_{0.99} estimator has a negative bias.
- The Hill VaR_{0.99} estimator has a negative bias for small k but a rapidly growing positive bias for larger k .
- The GPD VaR_{0.99} estimator has a positive bias which grows much more slowly.
- The GPD VaR_{0.99} estimator attains lowest MSE for a value of k around 100, but the MSE is very robust to the choice of k (because of the slow growth of the bias) ⇒ Choice of u less critical!
- The Hill VaR_{0.99} estimator performs well for $20 \leq k \leq 75$ (we only use k values that lead to a quantile estimate beyond the effective threshold $X_{[k]}$) but then deteriorates rapidly.
- Both EVT methods outperform the empirical quantile estimator.

5.2.6 Conditional EVT for financial time series

- The GPD method is an unconditional approach for estimating \bar{F} and associated risk measures. A conditional (time-dependent) risk-measurement approach may be more appropriate.
- We now consider a simple adaptation of the GPD method to obtain conditional risk-measure estimates in a GARCH context.
- Assume X_{t-n+1}, \dots, X_t are negative log-returns generated by a strictly stationary time series process (X_t) of the form

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

where μ_t and σ_t are \mathcal{F}_{t-1} -measurable and $Z_t \stackrel{\text{ind.}}{\sim} F_Z$; e.g., ARMA model with GARCH errors. Furthermore, let $Z \sim F_Z$.

- VaR $_\alpha^t$ and ES $_\alpha^t$ based on $F_{X_{t+1}|\mathcal{F}_t}$ are given by

$$\text{VaR}_\alpha^t(X_{t+1}) = \mu_{t+1} + \sigma_{t+1} \text{VaR}_\alpha(Z),$$

$$\text{ES}_\alpha^t(X_{t+1}) = \mu_{t+1} + \sigma_{t+1} \text{ES}_\alpha(Z).$$

- To obtain estimates $\widehat{\text{VaR}}_{\alpha}^t(X_{t+1})$ and $\widehat{\text{ES}}_{\alpha}^t(X_{t+1})$, proceed as follows:
 - 1) Fit an ARMA-GARCH model (via exponential smoothing or QMLE based on normal innovations (since we do not assume a particular innovation distribution)) \Rightarrow Estimates of μ_{t+1} and σ_{t+1} .
 - 2) Fit a GPD to F_Z (treat the residuals from the GARCH fitting procedure as i.i.d. from F_Z) \Rightarrow GPD-based estimates of $\text{VaR}_{\alpha}(Z)$ (see (15)) and $\text{ES}_{\alpha}(Z)$ (see (16)).

5.3 Point process models

So far: loss size distribution. Now: loss frequency distribution

5.3.1 Threshold exceedances for strict white noise

- Consider a strict white noise $(X_i)_{i \in \mathbb{N}}$ (i.i.d. from $F \in \text{MDA}(H_\xi)$; can be extended to dependent processes with extremal index $\theta = 1$).
- Let $u_n(x) = c_n x + d_n$ (x fixed). We know $F^n(u_n(x)) \xrightarrow[n \uparrow \infty]{} H_\xi(x)$. Taking $-\log(\cdot)$ and using $-\log y \approx 1 - y$ for $y \rightarrow 1$, we obtain $n\bar{F}(u_n(x)) \approx -n \log F(u_n(x)) = -\log(F^n(u_n(x))) \xrightarrow[n \uparrow \infty]{} -\log H_\xi(x)$.
- $N_{u_n(x)}$ (exceedances among X_1, \dots, X_n) fulfills $N_{u_n(x)} \sim \text{B}(n, \bar{F}(u_n(x)))$
- The Poisson Limit Theorem ($n \rightarrow \infty$, $p = \bar{F}(u_n(x)) \rightarrow 0$, $np = n\bar{F}(u_n(x)) \rightarrow \lambda = -\log H_\xi(x)$) implies $N_{u_n(x)} \xrightarrow[n \uparrow \infty]{} \text{Poi}(-\log H_\xi(x))$.
- One can show: Not only is $N_{u_n(x)}$ asymptotically Poisson, but the exceedances occur according to a Poisson process.

On point processes

- Suppose Y_1, \dots, Y_n take values in some *state space* \mathcal{X} (e.g., \mathbb{R}, \mathbb{R}^2). Define for any $A \subseteq \mathcal{X}$, the counting rv

$$N(A) = \sum_{i=1}^n \mathbb{1}_{\{Y_i \in A\}}.$$

Under technical conditions, see Embrechts et al. (1997, pp. 220), $N(\cdot)$ defines a point process.

- $N(\cdot)$ is a *Poisson point process* on \mathcal{X} with *intensity measure* Λ if:

- For $A \subseteq \mathcal{X}$ and $k \geq 0$,

$$\mathbb{P}(N(A) = k) = \begin{cases} e^{-\Lambda(A)} \frac{\Lambda(A)^k}{k!}, & \text{if } \Lambda(A) < \infty, \\ 0, & \text{if } \Lambda(A) = \infty. \end{cases}$$

- $N(A_1), \dots, N(A_m)$ are independent for any mutually disjoint subsets A_1, \dots, A_m of \mathcal{X} .

- Note that $\mathbb{E}N(A) = \Lambda(A)$. Also, the *intensity (function)* is the function $\lambda(x)$ which satisfies $\Lambda(A) = \int_A \lambda(x) dx$.

Asymptotic behavior of the point process of exceedances

- For $n \in \mathbb{N}$ and $i \in \{1, \dots, n\}$ let $Y_{i,n} = \frac{i}{n} \mathbb{1}_{\{X_i > u_n(x)\}}$. The *point process of exceedances over u_n* is the process $N_n(\cdot)$ with state space $\mathcal{X} = (0, 1]$ given by

$$N_n(A) = \sum_{i=1}^n \mathbb{1}_{\{Y_{i,n} \in A\}}, \quad A \subseteq \mathcal{X}.$$

- N_n is an element of the sequence of point processes (N_n) . N_n counts the *exceedances with time of occurrence in A* and we are interested in the behaviour of N_n as $n \rightarrow \infty$.
- Embrechts et al. (1997, Theorem 5.3.2) show that $N_n(\cdot)$ converges in distribution on \mathcal{X} to a Poisson process $N(\cdot)$ with intensity $\Lambda(\cdot)$ satisfying $\Lambda(A) = (t_2 - t_1)\lambda(x)$ for $A = (t_1, t_2) \subseteq \mathcal{X}$, $\lambda(x) = -\log H_\xi(x)$.

- In particular, $\mathbb{E}N_n(A) \underset{n \uparrow \infty}{\rightarrow} \mathbb{E}N(A) = \Lambda(A) = (t_2 - t_1)\lambda(x)$. λ does not depend on time and takes the constant value $\lambda = \lambda(x)$.
- We refer to the limiting process as a *homogeneous Poisson process with intensity* (or rate) λ .

Application of the result in practice

- Fix a large n and $u = c_n x + d_n$ for some x .
- Approximate N_u by a Poisson rv and the point process of exceedances of u by a homogeneous Poisson process with rate $\lambda = -\log H_\xi(x) = -\log H_\xi((u - d_n)/c_n) = -\log H_{\xi, \mu=d_n, \sigma=c_n}(u)$.
 - ⇒ Relationship between the GEV model and a Poisson model for the occurrence in time of exceedances of u .
- We see that exceedances of i.i.d. data over u are separated by i.i.d. exponential waiting times.

5.3.2 The POT model

- Putting the pieces together, we obtain an asymptotic model for threshold exceedances in regularly spaced i.i.d. data (or data with $\theta = 1$).
- This so-called *peaks-over-threshold (POT) model* makes the following assumptions:
 - 1) Exceedances times occur according to a *homogeneous Poisson process*.
 - 2) Excesses above u are *i.i.d.* and independent of exceedance times.
 - 3) The *excess distribution is generalized Pareto*.
- This model can also be viewed as a *marked Poisson point process* (exceedance times = points; GPD-distributed excesses = marks) or a (non-homogeneous) *two-dimensional Poisson* point process (point (t, x) = (time, magnitude of exceedance))

Two-dimensional Poisson formulation of POT model

- Assume that, on the state space $\mathcal{X} = (0, 1] \times (u, \infty)$, the point process defined by $N(A) = \sum_{i=1}^n \mathbb{1}_{\{(i/n, X_i) \in A\}}$ is a Poisson process with intensity at (t, x) given by

$$\lambda(x) = \lambda(t, x) = \begin{cases} \frac{1}{\sigma} \left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-1/\xi-1}, & \text{if } (1 + \xi(x - \mu)/\sigma) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

- For $A = (t_1, t_2) \times (x, \infty) \subseteq \mathcal{X}$, the intensity measure is

$$\Lambda(A) = \int_{t_1}^{t_2} \int_x^\infty \lambda(y) dy dt = -(t_2 - t_1) \log H_{\xi, \mu, \sigma}(x)$$

Thus, for any $x \geq u$, the one-dimensional process of exceedances of x is a homogeneous Poisson process with intensity $\tau(x) = -\log H_{\xi, \mu, \sigma}(x)$.

- $\bar{F}_u(x)$ can be calculated as the ratio of the rates of exceeding $u+x$ and u via

$$\bar{F}_u(x) = \frac{\tau(u+x)}{\tau(u)} = \left(1 + \frac{\xi x}{\sigma + \xi(u - \mu)}\right)^{-1/\xi} = \bar{G}_{\xi, \sigma + \xi(u - \mu)}(x)$$

This is precisely the POT model.

- The model also implies the GEV model. Consider $\{M_n \leq x\}$ for some $x \geq u$, i.e., the event that there are no points in $A = (0, 1] \times (x, \infty)$. Thus, $\mathbb{P}(M_n \leq x) = \mathbb{P}(N(A) = 0) = \exp(-\Lambda(A)) = H_{\xi, \mu, \sigma}(x)$, $x \geq u$, which is precisely the GEV model.

Statistical estimation of the POT model

- Given the exceedances $\tilde{X}_1 < \dots < \tilde{X}_{N_u}$, $A = (0, 1] \times (u, \infty)$ and $\Lambda(A) = \tau(u) =: \tau_u$, the likelihood $L(\xi, \sigma, \mu; \tilde{X}_1, \dots, \tilde{X}_{N_u})$ is

$$\underbrace{\frac{N_u!}{\text{ordered sample prob. of } N_u \text{ samples}} \underbrace{\frac{e^{-\Lambda(A)} \Lambda(A)^{N_u}}{N_u!}}_{\text{sample prob. of } N_u \text{ samples}} \prod_{i=1}^{N_u} \underbrace{\frac{\lambda(\tilde{X}_i)}{\Lambda(A)}}_{\text{density of } \tilde{X}_i} = e^{-\Lambda(A)} \prod_{i=1}^{N_u} \lambda(\tilde{X}_i) = e^{-\tau_u} \prod_{i=1}^{N_u} \lambda(\tilde{X}_i).$$

- Reparametrizing λ by $\tau_u = -\log H_{\xi, \mu, \sigma}(u) = (1 + \xi \frac{u-\mu}{\sigma})^{-1/\xi}$ and

$\beta = \sigma + \xi(u - \mu)$, we obtain

$$\begin{aligned}\lambda(x) &= \frac{1}{\sigma} \left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-\frac{1}{\xi}-1} = \frac{1}{\sigma} \left(\left(1 + \xi \frac{u - \mu}{\sigma}\right) \left(1 + \frac{\xi \frac{x-u}{\sigma}}{1 + \xi \frac{u-\mu}{\sigma}}\right)\right)^{-\frac{1}{\xi}-1} \\ &= \frac{\tau_u}{\sigma(1 + \xi \frac{u-\mu}{\sigma})} \left(1 + \frac{\xi \frac{x-u}{\sigma}}{1 + \xi \frac{u-\mu}{\sigma}}\right)^{-\frac{1}{\xi}-1} = \frac{\tau_u}{\beta} \left(1 + \frac{\xi(x-u)}{\sigma + \xi(u-\mu)}\right)^{-\frac{1}{\xi}-1} \\ &= \frac{\tau_u}{\beta} \left(1 + \frac{\xi(x-u)}{\beta}\right)^{-\frac{1}{\xi}-1} = \tau_u g_{\xi, \beta}(x-u),\end{aligned}$$

where $\xi \in \mathbb{R}$ and $\tau_u, \beta > 0$. Therefore, $\ell(\xi, \sigma, \mu; \tilde{X}_1, \dots, \tilde{X}_{N_u})$ equals

$$\begin{aligned}&= -\tau_u + \sum_{i=1}^{N_u} \log \lambda(\tilde{X}_i) = -\tau_u + N_u \log \tau_u + \overbrace{\sum_{i=1}^{N_u} (\log \lambda(\tilde{X}_i) - \log \tau_u)}^{= \log g_{\xi, \beta}(\tilde{X}_i - u)} \\ &= \ell_{\text{Poi}}(\tau_u; N_u) - N_u \log(T) + \ell_{\text{GPD}}(\xi, \beta; \tilde{X}_1 - u, \dots, \tilde{X}_{N_u} - u),\end{aligned}\tag{18}$$

where ℓ_{Poi} is the log-likelihood for a one-dimensional homogeneous Poisson process with rate τ_u and ℓ_{GPD} is the log-likelihood for fitting a GPD to the excesses $\tilde{X}_i - u$, $i \in \{1, \dots, N_u\}$.

- We can thus separate inferences about (ξ, β) and τ_u . Estimate (ξ, β) in a GPD analysis and then τ_u by its MLE N_u . Use these estimates to infer estimates of $\mu = u - \beta(1 - \tau_u^\xi)/\xi$ and $\sigma = \tau_u^\xi \beta$.

Advantages of the POT model formulation

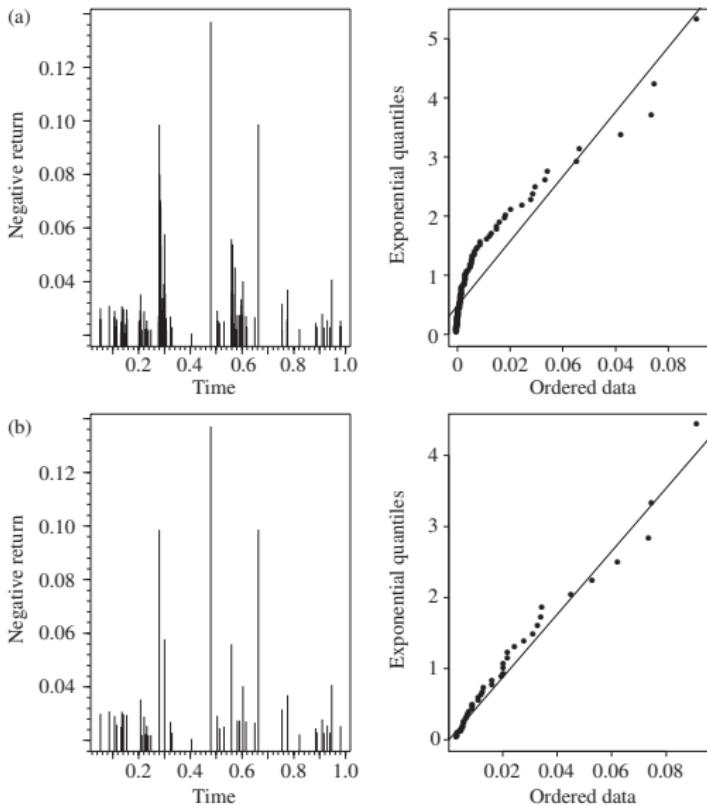
- One advantage of the two-dimensional Poisson point process model is that ξ , μ and σ do not depend on u (unlike β in the GPD model).
⇒ In practice, we would expect the estimated parameters of the Poisson model to be roughly stable over a range of high thresholds.
- The intensity λ is thus often used to introduce covariates to obtain Poisson processes which are non-homogeneous in time, e.g., by replacing μ and σ by parameters that vary over time as functions of covariates; see, e.g., Chavez-Demoulin et al. (2013).

Applicability of the POT model to return series data

- Returns do not really form genuine point events in time (in contrast to, e.g., water levels). They are discrete-time measurements that describe short-term changes (a day or a week). Nonetheless, assume that under a longer-term perspective, such data can be approximated by point events in time.
- Exceedances of u for daily financial return series do not necessarily occur according to a homogeneous Poisson process. They tend to cluster. Thus the standard POT model is not directly applicable.
- For stochastic processes with extremal index $\theta < 1$, e.g., GARCH processes, the extremal clusters themselves should occur according to a homogeneous Poisson process in time \Rightarrow Individual exceedances occur according to a Poisson cluster process; see Leadbetter (1991). Thus a suitable model for the occurrence and magnitude of exceedances in a

financial return series might be some form of marked Poisson cluster process.

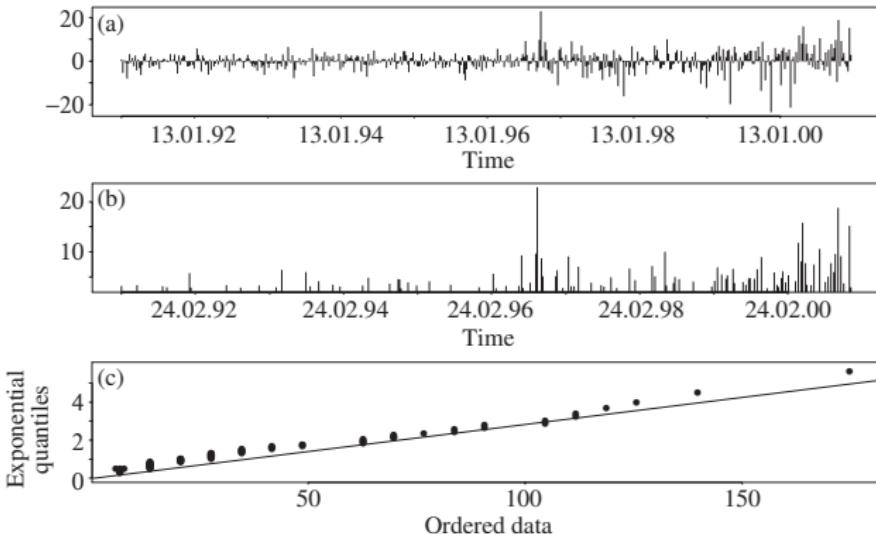
- *Declustering* may circumvent the problem. One identifies clusters (ad hoc; not easy) of exceedances and then applies the POT model to cluster maxima only.
- A possible declustering algorithm is the *runs method*. A run size r is fixed and two successive exceedances are said to belong to two different clusters if they are separated by a run of at least r values below u ; see Embrechts et al. (1997, pp. 422).
- In the following figure the DAX daily negative returns have been declustered with $r = 10$ trading days; this reduces the 100 exceedances to 42 cluster maxima.



- (a): DAX daily negative returns and a Q-Q plot of their spacings
- (b): Declustered data (runs method with $r = 10$ trading days \Rightarrow spacings are more consistent with a Poisson model)
- However, by neglecting the modeling of cluster formation, we cannot make more dynamic statements about the intensity of occurrence of exceedances.

Example 5.26 (POT analysis of AT&T weekly losses (continued))

Consider the 102 weekly percentage losses exceeding $u = 2.75\%$:



- Inter-exceedance times seem to follow an exponential distribution.
- But exceedances become more frequent over time ($\not\rightarrow$ homogeneous Poisson process \Rightarrow Possibly consider an inhomogeneous Poisson process).

- Using the log-likelihood (18), we fit a two-dimensional Poisson model to the 102 exceedances of $u = 2.75\%$. The parameter estimates are $\hat{\xi} = 0.22$, $\hat{\mu} = 19.9$ and $\hat{\sigma} = 5.95$.
- The implied GPD scale parameter is $\hat{\beta} = \hat{\sigma} + \hat{\xi}(u - \hat{\mu}) = 2.1 \Rightarrow$ The same $\hat{\xi}$ and $\hat{\beta}$ as in Example 5.23.
- The estimated exceedance rate over $u = 2.75$ is $\hat{\tau}(u) = -\log H_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}(u) = 102$ (= number of exceedances; as theory suggests).
- Higher thresholds, e.g., 15%: Since $\hat{\tau}(15) = 2.50$, losses exceeding 15% occur as a Poisson process with rate 2.5 losses per 10-year period (\approx a four-year event). \Rightarrow The Poisson model provides an alternative method of defining the return period of an event.
- Similarly, estimate return levels: If the 10-year return level is the level which is exceeded according to a Poisson process with rate one loss per 10 years, estimate the level by solving $\hat{\tau}(u) = 1$ w.r.t. u , so

$u = H_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}^{-1}(\exp(-1)) = 19.9$ so the 10-year event is a weekly loss of roughly 20%.

- Confidence intervals for such quantities can be constructed via profile likelihoods.

6 Multivariate models

- 6.1 Basics of multivariate modeling
- 6.2 Normal mixture distributions
- 6.3 Spherical and elliptical distributions
- 6.4 Dimension reduction techniques

6.1 Basics of multivariate modeling

6.1.1 Random vectors and their distributions

Joint and marginal distributions

- Let $\mathbf{X} = (X_1, \dots, X_d) : \Omega \rightarrow \mathbb{R}^d$ be a d -dimensional *random vector* (representing risk-factor changes, risks, etc.).
- The *(joint) distribution function (df) H of \mathbf{X}* is

$$H(\mathbf{x}) = F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}) = \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d), \quad \mathbf{x} \in \mathbb{R}^d.$$

- The *j th margin* or *marginal df F_j of \mathbf{X}* is

$$F_j(x_j) = \mathbb{P}(X_j \leq x_j)$$

$$\begin{aligned} &= \mathbb{P}(X_1 \leq \infty, \dots, X_{j-1} \leq \infty, X_j \leq x_j, X_{j+1} \leq \infty, \dots, X_d \leq \infty) \\ &= H(\infty, \dots, \infty, x_j, \infty, \dots, \infty), \quad x_j \in \mathbb{R}, \quad j \in \{1, \dots, d\}. \end{aligned}$$

(interpreted as a **limit**).

- Similarly for *k-dimensional margins*. Suppose we partition \mathbf{X} into $(\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$, where $\mathbf{X}_1 = (X_1, \dots, X_k)^\top$ and $\mathbf{X}_2 = (X_{k+1}, \dots, X_d)^\top$, then the marginal distribution function of \mathbf{X}_1 is

$$F_{\mathbf{X}_1}(\mathbf{x}_1) = \mathbb{P}(\mathbf{X}_1 \leq \mathbf{x}_1) = H(x_1, \dots, x_k, \infty, \dots, \infty).$$

- H is absolutely continuous* if

$$H(\mathbf{x}) = \int_{-\infty}^{x_d} \dots \int_{-\infty}^{x_1} h(z_1, \dots, z_d) dz_1 \dots dz_d = \int_{(-\infty, \mathbf{x}]} h(\mathbf{z}) d\mathbf{z}$$

for some $h \geq 0$ then known as the *(joint) density of \mathbf{X} (or H)*. Similarly, the *jth marginal df F_j is absolutely continuous* if $F_j(x) = \int_{-\infty}^x f_j(z) dz$ for some $f_j \geq 0$ then known as the *density of X_j (or F_j)*.

- In case h exists, $F_j(x_j) = \int_{-\infty}^{x_j} \int_{(-\infty, \infty)} h(\mathbf{z}) d\mathbf{z}_{-j} dz_j = \int_{-\infty}^{x_j} f_j(z_j) dz_j$, so that $f_j(z_j)$ can be recovered from h via

$$\underbrace{\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty}}_{d-1\text{-many}} h(z_1, \dots, z_{j-1}, z_j, z_{j+1}, \dots, z_d) dz_1 \dots dz_{j-1} dz_{j+1} \dots dz_d.$$

- Existence of a joint density \Rightarrow Existence of marginal densities for all k -dimensional marginals, $1 \leq k \leq d - 1$. The converse is false in general (counter-examples can be constructed with copulas; see later).
- By replacing integrals by sums, one obtains similar formulas for the discrete case, in which we call densities *probability mass functions*.
- Sometimes it's convenient to work with the survival function \bar{H} of \mathbf{X} , given by

$$\bar{H}(\mathbf{x}) = \bar{F}_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} > \mathbf{x}) = \mathbb{P}(X_1 > x_1, \dots, X_d > x_d), \quad \mathbf{x} \in \mathbb{R}^d,$$

with corresponding *jth marginal survival function* \bar{F}_j

$$\begin{aligned}\bar{F}_j(x_j) &= \mathbb{P}(X_j > x_j) \\ &= \bar{H}(-\infty, \dots, -\infty, x_j, -\infty, \dots, -\infty), \quad x_j \in \mathbb{R}, \quad j \in \{1, \dots, d\}.\end{aligned}$$

- Note that, unlike for $d = 1$, $\bar{H}(\mathbf{x}) \neq 1 - H(\mathbf{x})$ in general.

Conditional distributions and independence

- A multivariate model for risks in the form of a joint df, survival function or density, implicitly describes their *dependence structure*. We can then make statements about conditional probabilities.
- As before, consider $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top \sim H$. The *conditional df of \mathbf{X}_2 given $\mathbf{X}_1 = \mathbf{x}_1$* is $F_{\mathbf{X}_2|\mathbf{X}_1}(x_2 | \mathbf{x}_1) = \mathbb{P}(\mathbf{X}_2 \leq x_2 | \mathbf{X}_1 = \mathbf{x}_1) = \mathbb{E}[\mathbb{1}_{\{\mathbf{X}_2 \leq x_2\}} | \mathbf{X}_1 = \mathbf{x}_1]$, where $\mathbb{E}[\cdot | \cdot]$ denotes conditional expectation (not discussed here).
- A useful identity for conditional dfs is

$$\begin{aligned}& \int_{(-\infty, \mathbf{x}_1]} F_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{z}) dF_{\mathbf{X}_1}(\mathbf{z}) \\&= \int_{\mathbb{R}^d} \mathbb{1}_{\{\mathbf{z} \leq \mathbf{x}_1\}} \mathbb{E}[\mathbb{1}_{\{\mathbf{X}_2 \leq \mathbf{x}_2\}} | \mathbf{X}_1 = \mathbf{z}] dF_{\mathbf{X}_1}(\mathbf{z}) \\&= \mathbb{E}[\mathbb{1}_{\{\mathbf{X}_1 \leq \mathbf{x}_1\}} \mathbb{E}[\mathbb{1}_{\{\mathbf{X}_2 \leq \mathbf{x}_2\}} | \mathbf{X}_1]] = \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{\mathbf{X}_1 \leq \mathbf{x}_1, \mathbf{X}_2 \leq \mathbf{x}_2\}} | \mathbf{X}_1]] \\&\stackrel{\text{tower property}}{=} \mathbb{E}[\mathbb{1}_{\{\mathbf{X}_1 \leq \mathbf{x}_1, \mathbf{X}_2 \leq \mathbf{x}_2\}}] = H(\mathbf{x}),\end{aligned}$$

where the second-last equality holds by the [tower property](#) of conditional expectations.

- If $x_1 \rightarrow \infty$, then $F_{\mathbf{X}_2}(\mathbf{x}_2) = \int_{\mathbb{R}^d} F_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{z}) dF_{\mathbf{X}_1}(\mathbf{z})$. Furthermore, if H has a density h , then $f_{\mathbf{X}_2}(\mathbf{x}_2) = \int_{\mathbb{R}^d} f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{z}) dF_{\mathbf{X}_1}(\mathbf{z})$.
- If H has density h and $f_{\mathbf{X}_1}$ denotes the density of \mathbf{X}_1 , then

$$\begin{aligned} h(\mathbf{x}_1, \mathbf{x}_2) &= \frac{\partial^2}{\partial \mathbf{x}_2 \partial \mathbf{x}_1} H(\mathbf{x}_1, \mathbf{x}_2) = \frac{\partial}{\partial \mathbf{x}_2} F_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{x}_1) f_{\mathbf{X}_1}(\mathbf{x}_1) \\ &= f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{x}_1) f_{\mathbf{X}_1}(\mathbf{x}_1). \end{aligned}$$

We call

$$f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{x}_1) = \frac{h(\mathbf{x}_1, \mathbf{x}_2)}{f_{\mathbf{X}_1}(\mathbf{x}_1)} \quad (19)$$

the [conditional density of \$\mathbf{X}_2\$ given \$\mathbf{X}_1 = \mathbf{x}_1\$](#) . In this case, the conditional df $F_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{x}_1)$ is given by

$$F_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{x}_1) = \int_{-\infty}^{x_{k+1}} \cdots \int_{-\infty}^{x_d} \frac{h(\mathbf{x}_1, \mathbf{z})}{f_{\mathbf{X}_1}(\mathbf{x}_1)} dz_{k+1} \cdots dz_d.$$

- Inspired by (19), we call \boldsymbol{X}_1 and \boldsymbol{X}_2 *independent* if

$$H(\boldsymbol{x}_1, \boldsymbol{x}_2) = F_{\boldsymbol{X}_1}(\boldsymbol{x}_1)F_{\boldsymbol{X}_2}(\boldsymbol{x}_2), \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2.$$

- If H has density h , then \boldsymbol{X}_1 and \boldsymbol{X}_2 are independent if

$$h(\boldsymbol{x}_1, \boldsymbol{x}_2) = f_{\boldsymbol{X}_1}(\boldsymbol{x}_1)f_{\boldsymbol{X}_2}(\boldsymbol{x}_2), \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2.$$

In this case, $f_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 | \boldsymbol{x}_1) = h(\boldsymbol{x}_1, \boldsymbol{x}_2)/f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) = f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)$.

- The components $\boldsymbol{X}_1, \dots, \boldsymbol{X}_d$ of \boldsymbol{X} are (*mutually*) *independent* if

$$H(\boldsymbol{x}) = \prod_{j=1}^d F_j(x_j), \quad \forall \boldsymbol{x},$$

or, if H has density h ,

$$h(\boldsymbol{x}) = \prod_{j=1}^d f_j(x_j), \quad \forall \boldsymbol{x}.$$

Moments and characteristic function

- If $\mathbb{E}|X_j| < \infty$, $j \in \{1, \dots, d\}$, the *mean vector of \mathbf{X}* is defined by

$$\boldsymbol{\mu} = \mathbb{E}\mathbf{X} = (\mathbb{E}X_1, \dots, \mathbb{E}X_d).$$

One can show: X_1, \dots, X_d independent $\Rightarrow \mathbb{E}[X_1 \cdots X_d] = \prod_{j=1}^d \mathbb{E}[X_j]$

- If $\mathbb{E}[X_j^2] < \infty$ for all j , the *covariance matrix of \mathbf{X}* is defined as

$$\Sigma = \text{Cov}[\mathbf{X}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^\top].$$

Its (i, j) th element is

$$\begin{aligned}\sigma_{ij} = \Sigma_{ij} &= \text{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)] \\ &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j];\end{aligned}$$

the diagonal elements are $\sigma_{jj} = \text{Var}[X_j]$, $j \in \{1, \dots, d\}$.

- The *cross covariance matrix between* two (admissible) random vectors \mathbf{X}, \mathbf{Y} is defined as $\text{Cov}[\mathbf{X}, \mathbf{Y}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^\top]$. Note that $\text{Cov}[\mathbf{X}, \mathbf{X}] = \text{Cov}[\mathbf{X}]$.

- If $\mathbb{E}[X_j^2] < \infty$, $j \in \{1, \dots, d\}$, the *correlation matrix* of \mathbf{X} is defined as the matrix $P = \text{Cor}[\mathbf{X}]$ with (i, j) th element

$$\rho_{ij} = P_{ij} = \text{Cor}[X_i, X_j] = \frac{\text{Cov}[X_i, X_j]}{\sqrt{\text{Var}[X_i] \text{Var}[X_j]}}$$

which is in $[-1, 1]$ with $\rho_{ij} = \pm 1$ if and only if $X_j \stackrel{\text{a.s.}}{=} aX_i + b$ for some $a \geq 0$ and $b \in \mathbb{R}$. This follows from the Cauchy–Schwarz inequality $|\langle X_i, X_j \rangle| \leq \sqrt{\langle X_i, X_i \rangle \langle X_j, X_j \rangle}$ applied to the inner product $\langle X_i, X_j \rangle := \mathbb{E}[X_i X_j]$.

- X_i, X_j ($i \neq j$) independent $\not\Rightarrow \text{Cov}[X_i, X_j] = 0$. The only known distribution for which uncorrelatedness implies independence is the **multivariate normal distribution**.
- **Some properties:**

1) For all $A \in \mathbb{R}^{k \times d}$, $\mathbf{b} \in \mathbb{R}^k$:

- ▶ $\mathbb{E}[A\mathbf{X} + \mathbf{b}] = A\mathbb{E}\mathbf{X} + \mathbf{b} = A\boldsymbol{\mu} + \mathbf{b}$;

- $\text{Cov}[A\mathbf{X} + \mathbf{b}] = A \text{Cov}[\mathbf{X}]A^\top = A\Sigma A^\top$; if $k = 1$ ($A = \mathbf{a}^\top$),

$$\mathbf{a}^\top \Sigma \mathbf{a} = \text{Cov}[\mathbf{a}^\top \mathbf{X}] = \text{Var}[\mathbf{a}^\top \mathbf{X}] \geq 0, \quad \mathbf{a} \in \mathbb{R}^d, \quad (20)$$

i.e., covariance matrices are *positive semidefinite* (and, trivially, symmetric).

- $\text{Cov}[\mathbf{X}_1 + \mathbf{X}_2] = \text{Cov}[\mathbf{X}_1] + \text{Cov}[\mathbf{X}_2] + \text{Cov}[\mathbf{X}_1, \mathbf{X}_2] + \text{Cov}[\mathbf{X}_2, \mathbf{X}_1]$

- 2) If Σ is a *positive definite matrix* (i.e., $\mathbf{a}^\top \Sigma \mathbf{a} > 0$ for all $\mathbf{a} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$), then Σ is invertible (since pos. def. $\Rightarrow \Sigma \mathbf{a} \neq \mathbf{0} \Rightarrow \text{rank } \Sigma = d - \dim \ker \Sigma = d - \dim \{\mathbf{b} : \Sigma \mathbf{b} = \mathbf{0}\} = d$ (Rank-nullity Theorem)).
- 3) A *symmetric, positive definite (positive semidefinite) Σ* can be written as

$$\Sigma = AA^\top \quad (21)$$

for a lower triangular matrix A with $A_{jj} > 0$ ($A_{jj} \geq 0$) for all j . L is known as *Cholesky factor* (also denoted by $\Sigma^{1/2}$) of the *Cholesky decomposition* (21).

- Properties of \mathbf{X} can often be shown with the *characteristic function*

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[\exp(i\mathbf{t}^\top \mathbf{X})], \quad \mathbf{t} \in \mathbb{R}^d.$$

X_1, \dots, X_d are independent $\Leftrightarrow \phi_{\mathbf{X}}(\mathbf{t}) = \prod_{j=1}^d \phi_{X_j}(t_j)$ for all \mathbf{t} .

Proposition 6.1

A symmetric matrix Σ is a covariance matrix if and only if it is positive semidefinite.

Proof.

“ \Rightarrow ” As we have seen in (20), a covariance matrix Σ is positive semidefinite.

“ \Leftarrow ” Let Σ be positive semidefinite with Cholesky factor A . Let \mathbf{X} be a random vector with $\text{Cov } \mathbf{X} = I_d = \text{diag}(1, \dots, 1)$ (e.g., $X_j \stackrel{\text{ind.}}{\sim} N(0, 1)$). Then $\text{Cov}[A\mathbf{X}] = A \text{Cov}[\mathbf{X}] A^\top = A A^\top = \Sigma$, i.e., Σ is a covariance matrix (namely that of $A\mathbf{X}$). \square

6.1.2 Standard estimators of covariance and correlation

- Assume $\mathbf{X}_1, \dots, \mathbf{X}_n \sim H$ (daily/weekly/monthly/yearly risk-factor changes) to be serially uncorrelated (i.e., multivariate white noise) with $\mu = \mathbb{E}\mathbf{X}_1$, $\Sigma = \text{Cov } \mathbf{X}_1$, $P = \text{Cor } \mathbf{X}_1$.
- Non-parametric method-of-moments-like estimators of μ, Σ, P are

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \quad (\text{sample mean}; \text{unbiased}; \text{colMeans})$$

$$S = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top \quad (\text{sample cov. mat.}; \text{unbiased}; \text{var})$$

$$R = (R_{ij}) \text{ for } R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} \quad (\text{sample cor. matrix}; \text{unbiased}; \text{cor})$$

- $\bar{\mathbf{X}}$ and R are also MLEs; $\frac{n-1}{n}S$ is the MLE for Σ .

- Clearly, $\bar{\mathbf{X}}$ is unbiased. Since the \mathbf{X}_i 's are uncorrelated,

$$\text{Cov}[\bar{\mathbf{X}}] = \frac{1}{n^2} \sum_{i=1}^n \text{Cov}[\mathbf{X}_i] = \frac{1}{n} \text{Cov}[\mathbf{X}_1] = \frac{1}{n} \Sigma.$$

- S is unbiased since

$$\begin{aligned}\mathbb{E}S &= \frac{1}{n-1} \sum_{i=1}^n \mathbb{E}[(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top] \\ &= \frac{1}{n-1} \sum_{i=1}^n \mathbb{E}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^\top - (\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})^\top] \\ &= \frac{1}{n-1} \sum_{i=1}^n (\Sigma - \text{Cov } \bar{\mathbf{X}}) \underset{\text{Cov}[\bar{\mathbf{X}}] = \frac{\Sigma}{n}}{=} \frac{n}{n-1} (1 - 1/n) \Sigma = \Sigma.\end{aligned}$$

- Check that $S = \frac{1}{n-1} ABA^\top$ for $A = (\mathbf{X}_1, \dots, \mathbf{X}_n) \in \mathbb{R}^{d \times n}$ and $B = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \in \mathbb{R}^{n \times n}$ (where $\mathbf{1}_n = (1, \dots, 1) \in \mathbb{R}^n$). Since $\text{rank } A \leq \min\{n, d\}$, $\text{rank } B = n - \dim \ker B = n - 1$, and $\text{rank}(ABA^\top) \leq \min\{\text{rank } A, \text{rank } B\}$, it follows that $\text{rank } S \leq \min\{d, n-1\}$. If $n \leq d$,

S is not invertible. To obtain a positive definite (and thus invertible) estimator of Σ , see, e.g., `Matrix::nearPD()`.

- Further properties of \bar{X}, S, R depend on H .

6.1.3 The multivariate normal distribution

Definition 6.2 (Multivariate normal distribution)

$\mathbf{X} = (X_1, \dots, X_d)$ has a *multivariate normal* (or *Gaussian*) *distribution* if

$$\mathbf{X} \stackrel{\text{d}}{=} \boldsymbol{\mu} + A\mathbf{Z}, \quad (22)$$

where $\mathbf{Z} = (Z_1, \dots, Z_k)$, $Z_j \stackrel{\text{ind.}}{\sim} N(0, 1)$, $A \in \mathbb{R}^{d \times k}$, $\boldsymbol{\mu} \in \mathbb{R}^d$.

- $\mathbb{E}\mathbf{X} = \boldsymbol{\mu} + A\mathbb{E}\mathbf{Z} = \boldsymbol{\mu}$
- $\text{Cov}[\mathbf{X}] = \text{Cov}[\boldsymbol{\mu} + A\mathbf{Z}] = A \text{Cov}[\mathbf{Z}]A^\top = AA^\top =: \Sigma$

Proposition 6.3 (Characteristic function)

Let \mathbf{X} be as in (22) and $\Sigma = AA^\top$. Then the cf of \mathbf{X} is

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[\exp(i\mathbf{t}^\top \mathbf{X})] = \exp\left(i\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t}\right), \quad \mathbf{t} \in \mathbb{R}^d.$$

Proof.

- $Z_1 \sim N(0, 1)$ has density $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ which satisfies
 - i) $\varphi(x) = \varphi(-x);$
 - ii) $\varphi'(x) = -x\varphi(x).$

By Euler's Formula, the characteristic function $\phi_{Z_1}(t)$ of Z_1 is given by

$$\phi_{Z_1}(t) = \int_{-\infty}^{\infty} (\cos(tx) + i \sin(tx)) \varphi(x) dx \stackrel{\text{i)}}{=} \int_{-\infty}^{\infty} \cos(tx) \varphi(x) dx.$$

Hence,

$$\phi'_{Z_1}(t) = \int_{-\infty}^{\infty} \sin(tx)(-x)\varphi(x) dx \stackrel{\text{ii)}}{=} \int_{-\infty}^{\infty} \sin(tx)\varphi'(x) dx \stackrel{\text{by parts}}{=} -t\phi_{Z_1}(t).$$

We also know that $\phi_{Z_1}(0) = 1$. This initial value problem has the unique solution $\phi_{Z_1}(t) = \exp(-t^2/2)$.

- Now let $\tilde{\mathbf{t}}^\top = \mathbf{t}^\top A$. Then

$$\begin{aligned}\phi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}[\exp(i\mathbf{t}^\top(\boldsymbol{\mu} + A\mathbf{Z}))] = \exp(i\mathbf{t}^\top\boldsymbol{\mu})\mathbb{E}[\exp(i\tilde{\mathbf{t}}^\top\mathbf{Z})] \\ &\stackrel{\text{ind.}}{=} \exp(i\mathbf{t}^\top\boldsymbol{\mu}) \prod_{j=1}^d \mathbb{E}[\exp(i(\tilde{t}_j Z_j))] = \exp\left(i\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2} \sum_{j=1}^d \tilde{t}_j^2\right) \\ &= \exp\left(i\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\tilde{\mathbf{t}}^\top\tilde{\mathbf{t}}\right) = \exp\left(i\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top A A^\top \mathbf{t}\right) \\ &= \exp\left(i\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top\Sigma\mathbf{t}\right)\end{aligned}$$

□

- We see that the multivariate normal distribution is characterized by $\boldsymbol{\mu}$ and Σ , hence the notation $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$.
- $\mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$ can be characterized by univariate normal distributions. To see this we first need the following theoretical result.

Theorem 6.4 (Cramér–Wold)

Let \mathbf{X}, \mathbf{X}_n , $n \in \mathbb{N}$, be random vectors. Then

$$\mathbf{X}_n \xrightarrow[n \uparrow \infty]{\text{d}} \mathbf{X} \iff \mathbf{a}^\top \mathbf{X}_n \xrightarrow[n \uparrow \infty]{\text{d}} \mathbf{a}^\top \mathbf{X} \quad \forall \mathbf{a} \in \mathbb{R}^d$$

Proof.

“ \Rightarrow ” This follows directly from the Continuous Mapping Theorem with the continuous map being $g(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$.

“ \Leftarrow ” $\phi_{\mathbf{X}_n}(\mathbf{t}) = \mathbb{E}[\exp(i \cdot \mathbf{1} \cdot \mathbf{t}^\top \mathbf{X}_n)] = \phi_{\mathbf{t}^\top \mathbf{X}_n}(1) \xrightarrow[n \uparrow \infty]{} \phi_{\mathbf{t}^\top \mathbf{X}}(1) = \phi_{\mathbf{X}}(\mathbf{t})$ for all \mathbf{t} . The result then follows by the Lévy Continuity Theorem. \square

Corollary 6.5

Let \mathbf{X}, \mathbf{Y} be two random vectors. Then

$$\mathbf{X} \stackrel{\text{d}}{=} \mathbf{Y} \iff \mathbf{a}^\top \mathbf{X} \stackrel{\text{d}}{=} \mathbf{a}^\top \mathbf{Y} \quad \forall \mathbf{a} \in \mathbb{R}^d.$$

Proposition 6.6 (Characterization of $N_d(\mu, \Sigma)$)

$$\mathbf{X} \sim N_d(\mu, \Sigma) \iff \mathbf{a}^\top \mathbf{X} \sim N(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}) \quad \forall \mathbf{a} \in \mathbb{R}^d.$$

Proof.

$$\begin{aligned} \Rightarrow \quad \phi_{\mathbf{a}^\top \mathbf{X}}(t) &= \mathbb{E}[\exp(it\mathbf{a}^\top \mathbf{X})] \\ &= \phi_{\mathbf{X}}(t\mathbf{a}) = \exp\left(i(t\mathbf{a})^\top \boldsymbol{\mu} - \frac{1}{2}(t\mathbf{a})^\top \boldsymbol{\Sigma}(t\mathbf{a})\right) \\ &= \exp\left(it(\mathbf{a}^\top \boldsymbol{\mu}) - \frac{1}{2}t^2(\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a})\right). \end{aligned}$$

Uniqueness of characteristic functions $\Rightarrow \mathbf{a}^\top \mathbf{X} \sim N(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a})$.

\Leftarrow Let $\mathbf{Y} \sim N_d(\mu, \Sigma)$. We have just seen that $\mathbf{a}^\top \mathbf{Y} \sim N(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a})$ for all $\mathbf{a} \in \mathbb{R}^d$, so $\mathbf{a}^\top \mathbf{X} \stackrel{d}{=} \mathbf{a}^\top \mathbf{Y}$ for all $\mathbf{a} \in \mathbb{R}^d$. By Corollary 6.5, $\mathbf{X} \stackrel{d}{=} \mathbf{Y} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. □

Consequences:

- Margins: $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma) \stackrel{\substack{a=e_j \\ \text{see copulas}}}{\not\Rightarrow} X_j \sim N(\mu_j, \sigma_{jj}^2), \quad j \in \{1, \dots, d\}.$
- Sums: $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma) \stackrel{a=1}{\Rightarrow} \sum_{j=1}^d X_j \sim N(\sum_{j=1}^d \mu_j, \sum_{i,j} \sigma_{ij}).$

Proposition 6.7 (Density)

Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ with $d \leq k$, $\text{rank } A = d$ ($\Rightarrow \Sigma$ pos. definite, invertible). Then \mathbf{X} has density

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right), \quad \mathbf{x} \in \mathbb{R}^d.$$

Proof. Let $\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + A\mathbf{Z}$ with $\text{rank } A = d$, $\mathbf{Z} = (Z_1, \dots, Z_d)$, $Z_i \stackrel{\text{ind.}}{\sim} N(0, 1)$, $i \in \{1, \dots, d\}$. The density of \mathbf{Z} is

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{j=1}^d f_{Z_j}(z_j) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\mathbf{z}^\top \mathbf{z}\right), \quad \mathbf{z} \in \mathbb{R}^d.$$

By the Density Transformation Theorem,

$$f_{\mathbf{X}}(\mathbf{x}) = f_{T(\mathbf{Z})}(\mathbf{x}) = f_{\mathbf{Z}}(T^{-1}(\mathbf{x})) \left| \det \frac{d}{d\mathbf{x}} T^{-1}(\mathbf{x}) \right|.$$

With $T(\mathbf{z}) = \boldsymbol{\mu} + A\mathbf{z}$, we have $T^{-1}(\mathbf{x}) = A^{-1}(\mathbf{x} - \boldsymbol{\mu})$ and $\frac{d}{d\mathbf{x}} T^{-1}(\mathbf{x}) = A^{-1}$, and thus

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(A^{-1}(\mathbf{x} - \boldsymbol{\mu}))^\top (A^{-1}(\mathbf{x} - \boldsymbol{\mu}))\right) |\det(A^{-1})|.$$

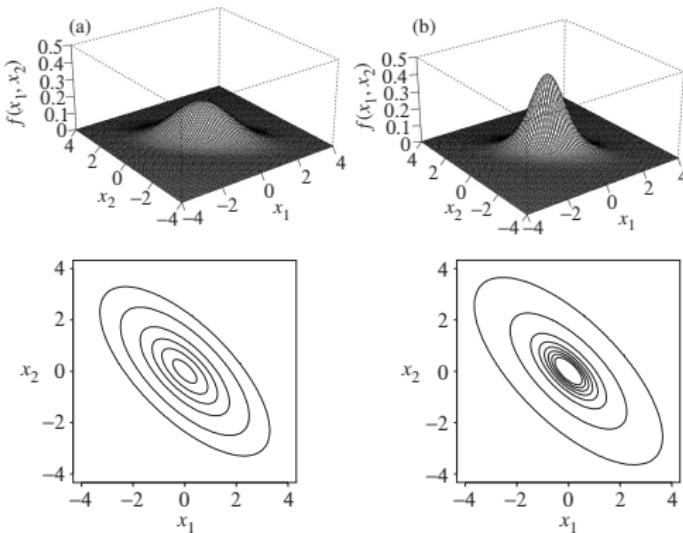
Now $(A^{-1})^\top A^{-1} = (A^\top)^{-1} A^{-1} = (AA^\top)^{-1} = \Sigma^{-1}$ and $\det(A^{-1}) = 1/\det(A) = 1/\sqrt{\det(A)\det(A^\top)} = 1/\sqrt{\det\Sigma}$ and thus the result follows. □

Consequences:

- Sets of the form $S_c = \{\mathbf{x} \in \mathbb{R}^d : (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c\}$, $c > 0$, describe points of equal density. Contours of equal density are thus ellipsoids. Whenever a multivariate density $f_{\mathbf{X}}(\mathbf{x})$ depends on \mathbf{x} only

through the quadratic form $(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})$, it is the density of an elliptical distribution (see later).

- The components of $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ are mutually independent if and only if Σ is diagonal, i.e., if and only if the components of \mathbf{X} are uncorrelated.



Left: $N_d(\boldsymbol{\mu} = (\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}), \Sigma = (\begin{smallmatrix} 1 & -0.7 \\ -0.7 & 1 \end{smallmatrix}))$; Right: $t_{\nu=4}(\boldsymbol{\mu}, \frac{\nu-2}{\nu}\Sigma)$ (same mean and covariance matrix as on the left-hand side)

The definition of $N_d(\mu, \Sigma)$ in terms of a stochastic representation ($\mathbf{X} \stackrel{d}{=} \mu + A\mathbf{Z}$) directly justifies the following sampling algorithm; see also `mvtnorm::rmvnorm(, method="chol")`.

Algorithm 6.8 (Sampling $N_d(\mu, \Sigma)$)

Let $\mathbf{X} \sim N_d(\mu, \Sigma)$ with Σ positive definite.

- 1) Compute the Cholesky factor A of Σ ; see, e.g., Press et al. (1992).
- 2) Generate $Z_j \stackrel{\text{ind.}}{\sim} N(0, 1)$, $j \in \{1, \dots, d\}$ (R: done with inversion!).
- 3) Return $\mathbf{X} = \mu + A\mathbf{Z}$, where $\mathbf{Z} = (Z_1, \dots, Z_d)$.

Further useful properties of multivariate normal distributions

■ Linear combinations

If $\mathbf{X} \sim N_d(\mu, \Sigma)$ and $B \in \mathbb{R}^{k \times d}$, $\mathbf{b} \in \mathbb{R}^k$, then

$$\begin{aligned} B\mathbf{X} + \mathbf{b} &= B(\mu + A\mathbf{Z}) + \mathbf{b} = (B\mu + \mathbf{b}) + BA\mathbf{Z} \\ &\sim N_k(B\mu + \mathbf{b}, BA(BA)^T) = N_k(B\mu + \mathbf{b}, B\Sigma B^T). \end{aligned}$$

Special case (see variance-covariance method; or Proposition 6.6):
 $\mathbf{b}^\top \mathbf{X} \sim N(\mathbf{b}^\top \boldsymbol{\mu}, \mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b})$

- **Marginal dfs**

Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and write $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$, where $\mathbf{X}_1 \in \mathbb{R}^k$, $\mathbf{X}_2 \in \mathbb{R}^{d-k}$, and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top$, $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$. Then

$$\mathbf{X}_1 \sim N_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \quad \text{and} \quad \mathbf{X}_2 \sim N_{d-k}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}).$$

Proof. Choose $B = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 0 & 0 \\ 0 & I_{d-k} \end{pmatrix}$, respectively.

- **Conditional distributions**

Let \mathbf{X} be as before and $\boldsymbol{\Sigma}$ be positive definite. Then

$$\mathbf{X}_2 | \mathbf{X}_1 = \mathbf{x}_1 \sim N_{d-k}(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}),$$

where $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)$ and $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$.

Proof. Via conditional densities or as follows: Consider $\mathbf{Z} = A\mathbf{X}_1 + \mathbf{X}_2$ with $A = -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$. Note that $(\mathbf{Z}, \mathbf{X}_1) = \begin{pmatrix} A & I_{d-k} \\ I_k & 0 \end{pmatrix} \mathbf{X}$ is jointly

normal. Since $\mathbf{Z} = \mathbf{X}_2 + A\mathbf{X}_1$, we know that $(\mathbf{X}_2 | \mathbf{X}_1 = \mathbf{x}_1) \stackrel{\text{d}}{=} \mathbf{Z} - Ax_1$ is multivariate normal (since \mathbf{Z} is). We have left to show that the formulas for $\mu_{2.1}$ and $\Sigma_{22.1}$ hold. Since $A = -\Sigma_{21}\Sigma_{11}^{-1}$,

$$\text{Cov}[\mathbf{Z}, \mathbf{X}_1] = \text{Cov}[\mathbf{X}_2, \mathbf{X}_1] + A\text{Cov}[\mathbf{X}_1] = \Sigma_{21} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{11} = 0,$$

hence \mathbf{Z} and \mathbf{X}_1 are **independent**. Therefore

$$\begin{aligned}\mu_{2.1} &= \mathbb{E}[\mathbf{X}_2 | \mathbf{X}_1 = \mathbf{x}_1] = \mathbb{E}_{\text{ind.}}[\mathbf{Z}] - \mathbb{E}[A\mathbf{X}_1 | \mathbf{X}_1 = \mathbf{x}_1] \\ &= \mathbb{E}\mathbf{X}_2 + A\mathbb{E}\mathbf{X}_1 - Ax_1 = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1),\end{aligned}$$

$$\begin{aligned}\Sigma_{22.1} &= \text{Cov}[\mathbf{X}_2 | \mathbf{X}_1 = \mathbf{x}_1] = \text{Cov}[\mathbf{Z} - A\mathbf{X}_1 | \mathbf{X}_1 = \mathbf{x}_1] \\ &= \text{Cov}_{\text{ind.}}[\mathbf{Z} - Ax_1] = \text{Cov} \mathbf{Z} = \text{Cov}[\mathbf{X}_2 + A\mathbf{X}_1] \\ &= \text{Cov}[\mathbf{X}_2] + A \text{Cov}[\mathbf{X}_1]A^\top + \text{Cov}[\mathbf{X}_2, \mathbf{X}_1]A^\top + A \text{Cov}[\mathbf{X}_1, \mathbf{X}_2] \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{11}(-\Sigma_{21}\Sigma_{11}^{-1})^\top + \Sigma_{21}(-\Sigma_{21}\Sigma_{11}^{-1})^\top - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.\end{aligned}$$

Noting that $(\Sigma_{21}\Sigma_{11}^{-1})^\top = (\Sigma_{11}^{-1})^\top\Sigma_{21}^\top = (\Sigma_{11}^\top)^{-1}\Sigma_{12} = \Sigma_{11}^{-1}\Sigma_{12}$, the form of $\Sigma_{22.1}$ easily follows. □

■ Quadratic forms

Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$, Σ positive definite with Cholesky factor A . Furthermore, let $\mathbf{Z} = A^{-1}(\mathbf{X} - \boldsymbol{\mu})$. Then $\mathbf{Z} \sim N_d(\mathbf{0}, I_d)$. Moreover,

$$(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Z}^\top \mathbf{Z} \sim \chi_d^2, \quad (23)$$

which is useful for (goodness-of-fit) testing of $N_d(\boldsymbol{\mu}, \Sigma)$; see later.

Proof. Clear via linearity and definition, and the definition of χ_d^2 .

■ Convolutions

Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ and $\mathbf{Y} \sim N_d(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$ be independent. Then

$$\mathbf{X} + \mathbf{Y} \sim N_d(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \Sigma + \tilde{\Sigma}).$$

Proof. By independence, $\phi_{\mathbf{X}+\mathbf{Y}}(\mathbf{t})$ factors into

$$\begin{aligned} &= \phi_{\mathbf{X}}(\mathbf{t})\phi_{\mathbf{Y}}(\mathbf{t}) = \exp\left(it^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t}\right) \exp\left(it^\top \tilde{\boldsymbol{\mu}} - \frac{1}{2}\mathbf{t}^\top \tilde{\Sigma} \mathbf{t}\right) \\ &= \exp\left(it^\top (\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}) - \frac{1}{2}\mathbf{t}^\top (\Sigma + \tilde{\Sigma}) \mathbf{t}\right), \end{aligned}$$

which is the cf of $N_d(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \Sigma + \tilde{\Sigma})$. □

Further properties:

- 1) **Univariate t_ν distribution:** $Z \sim N(0, 1)$, $W \sim \chi_\nu^2$ independent $\Rightarrow X = Z/\sqrt{W/\nu} \sim t_\nu$. With $Y = \mu + \sigma X$, one has $\mathbb{E}Y = \mu$ if $\nu > 1$ and $\text{Var}[Y] = \frac{\nu}{\nu-2}\sigma^2$ if $\nu > 2$.

Generalization of χ_ν^2 to $\nu > 0$: $\chi_\nu^2 = \Gamma(\nu/2, 1/2)$ where $\Gamma(\alpha, \beta)$ has density $f(x) = \beta^\alpha x^{\alpha-1} e^{-\beta x}/\Gamma(\alpha)$ (β is the rate; see also R).

- 2) If $\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{\text{ind.}}{\sim} N_d(\boldsymbol{\mu}, \Sigma)$ with rank $\Sigma = d$, then

$$\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top \sim W_d(\Sigma, n-1) \quad (\text{Wishart distr.}) \quad (24)$$

and $\bar{\mathbf{X}}$ and (24) are independent. For $X = (\mathbf{X}_1^\top, \dots, \mathbf{X}_n^\top)^\top$, one has $X^\top X = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top \sim W_d(\Sigma, n)$; special case: $W_1(1, n) = \chi_n^2$.

6.1.4 Testing multivariate normality

By Proposition 6.6,

$$\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{\text{ind.}}{\sim} N_d(\boldsymbol{\mu}, \Sigma) \Rightarrow \mathbf{a}^\top \mathbf{X}_1, \dots, \mathbf{a}^\top \mathbf{X}_n \stackrel{\text{ind.}}{\sim} N(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \Sigma \mathbf{a}).$$

This can be tested statistically (for some \mathbf{a}) with various goodness-of-fit tests (e.g., Q-Q plots) known for univariate normality (however, for $\mathbf{a} = \mathbf{e}_j$, $j \in \{1, \dots, d\}$, we would only test normality of the margins, not joint normality). Alternatively, (23) could be used to test joint normality.

Univariate tests

Formal statistical tests (see `fBasics::NormalityTests`)

- For general univariate df F :
 - ▶ Kolmogorov–Smirnov (`stats::ks.test()`)
 - ▶ Cramér–von Mises (for normal df: `nortest::cvm.test()`)

- ▶ Anderson–Darling (recommended by D'Agostino and Stephens (1986);
`ADGofTest::ad.test()`)
- For $N(\mu, \sigma^2)$:
 - ▶ D'Agostino (`fBasics::dagoTest()`, `moments::agostino.test()`)
 - ▶ Shapiro–Wilk (`stats::shapiro.test()`)
 - ▶ Jarque–Bera (`tseries::jarque.bera.test()`,
`moments::jarque.test()`)

Graphical tests

Let X_1, \dots, X_n be iid, $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}$ the corresponding empirical distribution function (edf). Suppose we want to graphically test whether $X_1, \dots, X_n \sim F$ for some df F based on given realizations x_1, \dots, x_n . Let $x_{(1)} \leq \dots \leq x_{(n)}$ denote the corresponding ordered statistics. Possible options are:

- **P-P plot:** Plot $\{(p_i, F(x_{(i)})) : i = 1, \dots, n\}$, where $p_i := \text{ppoints}(n)[i]$
 $\approx \frac{i-1/2}{n}.$
- **Q-Q plot:** Plot $\{(F^-(p_i), x_{(i)}) : i = 1, \dots, n\}$ (differences in tails better visible).

Justification:

- 1) Glivenko–Cantelli: $\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow[n \uparrow \infty]{\text{a.s.}} 0$
- 2) $\hat{F}_n(x) \xrightarrow{n \rightarrow \infty} F(x) \quad \forall x \in C(F) \Leftrightarrow \hat{F}_n^-(u) \xrightarrow{n \rightarrow \infty} F^-(u) \quad \forall u \in C(F^-);$
 see van der Vaart (2000, Lemma 21.2)

By 1), the first (and thus the 2nd) part of 2) holds. Hence, for the true underlying F , $x_{(i)} = \hat{F}_n^-(i/n) \approx \hat{F}_n^-(p_i) \approx F^-(p_i)$.

Interpretation: If F is (reasonably close to) the underlying unknown df, P-P and Q-Q plots resemble lines close to $y = x$ (possibly after standardization to mean 0 and variance 1).

Multivariate tests

Formal statistical tests

- Multivariate Shapiro–Wilk (`mvnormtest::mshapiro.test()`)
- Mardia's test (`dprep::mardia()`):
 - ▶ According to (23), if $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ with Σ positive definite, then $(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_d^2$.
 - ▶ Let $D_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})^\top S^{-1} (\mathbf{X}_i - \bar{\mathbf{X}})$ denote the *squared Mahalanobis distances* and $D_{ij} = (\mathbf{X}_i - \bar{\mathbf{X}})^\top S^{-1} (\mathbf{X}_j - \bar{\mathbf{X}})$ the *Mahalanobis angles*.
 - ▶ Let $b_d = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n D_{ij}^3$ and $k_d = \frac{1}{n} \sum_{i=1}^n D_i^4$. Under the null hypothesis one can show that asymptotically for $n \rightarrow \infty$,

$$\frac{n}{6} b_d \sim \chi_{d(d+1)(d+2)/6}^2, \quad \frac{k_d - d(d+2)}{\sqrt{8d(d+2)/n}} \sim N(0, 1),$$

which can be used for testing; see Joenssen and Vogel (2014).

Graphical test

- Due to \bar{X} and S , the D_i^2 's are not exactly following a χ_d^2 anymore. It turns out that $\frac{n}{(n-1)^2} D_i^2 \stackrel{H_0}{\sim} \text{Beta}(d/2, (n-d-1)/2)$; see Gnanadesikan and Kettenring (1972). Check this with a Q-Q plot. For large n , the approximate χ_d^2 distribution is fine.

Example 6.9 (Multivariate (non-)normality of 10 Dow Jones stocks)

- We apply Mardia's test (of multivariate skewness and kurtosis) to daily/weekly/monthly/quarterly log-returns of 10 (of the 30) Dow Jones stocks from 1993–2000.
- We also compare D_i^2 data to a χ_{10}^2 using a Q-Q plot.

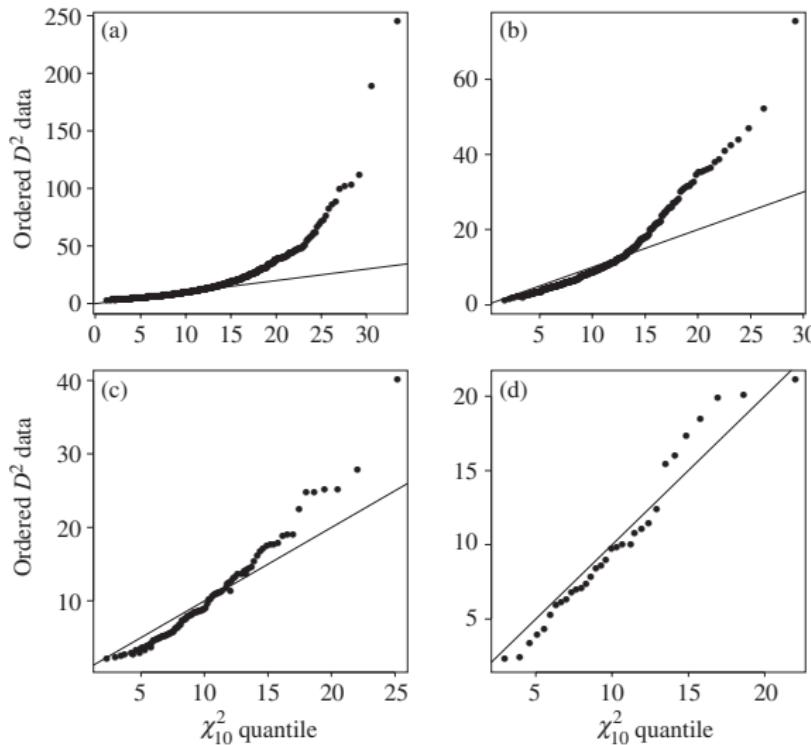
Mardia's (asymptotic) test based on the multivariate measures of skewness and kurtosis:

	Daily	Weekly	Monthly	Quarterly
n	2020	416	96	32
b_{10}	9.31	9.91	21.10	50.10
$p\text{-value}$	0.00	0.00	0.00	0.02
k_{10}	242.45	177.04	142.65	120.83
$p\text{-value}$	0.00	0.00	0.00	0.44

Conclusion: Daily/weekly/monthly data: Evidence against joint normality
Quarterly data: CLT effect seems to take place (but too little data to say more); still evidence against joint normality.

Q-Q plot of D_i^2 data against a χ_{10}^2 distribution:

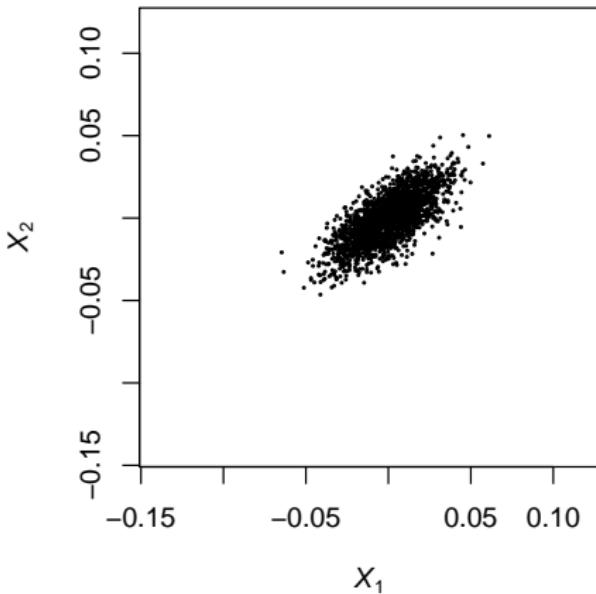
(a) daily data; (b) weekly data; (c) monthly data; and (d) quarterly data



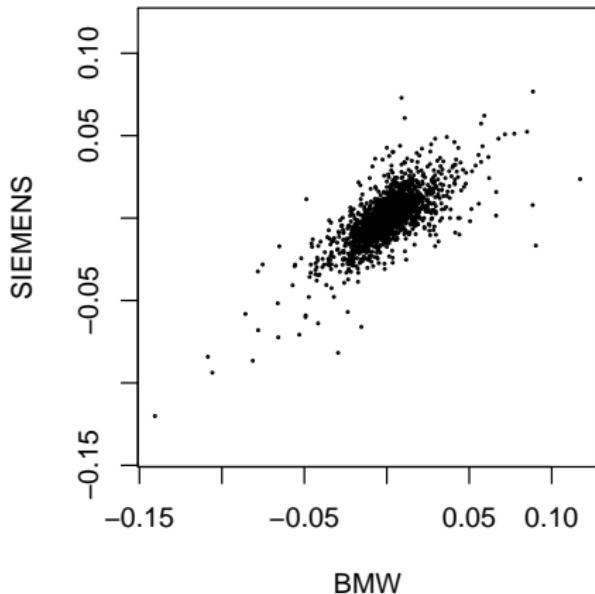
Example 6.10 (Simulated data vs BMW–Siemens)

Is the [BMW–Siemens data](#) (see Section 3.2.2) [jointly normal](#)?

Simulated data (fitted multivariate normal)

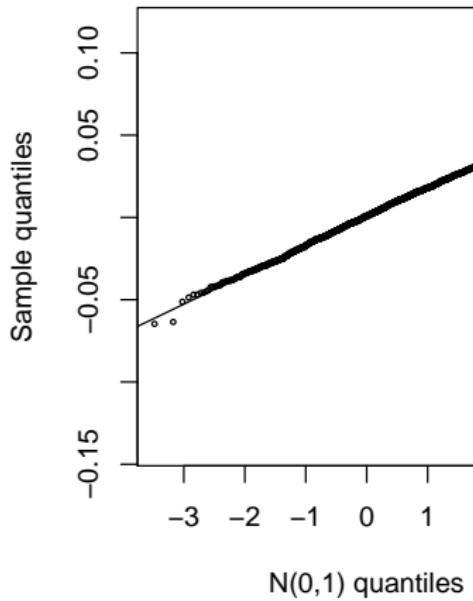


Real risk-factor changes

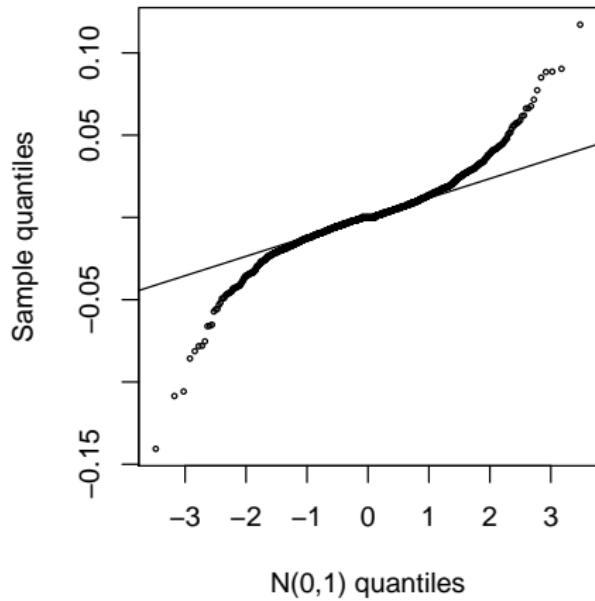


Considering the first margin only:

Q-Q plot for margin 1 (simulated data)

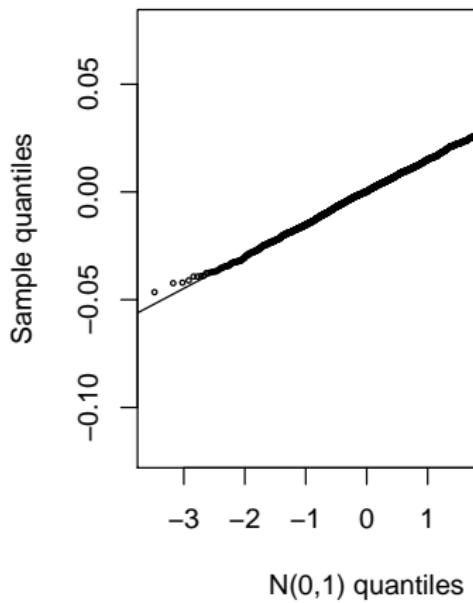


Q-Q plot for margin 1 (real data)

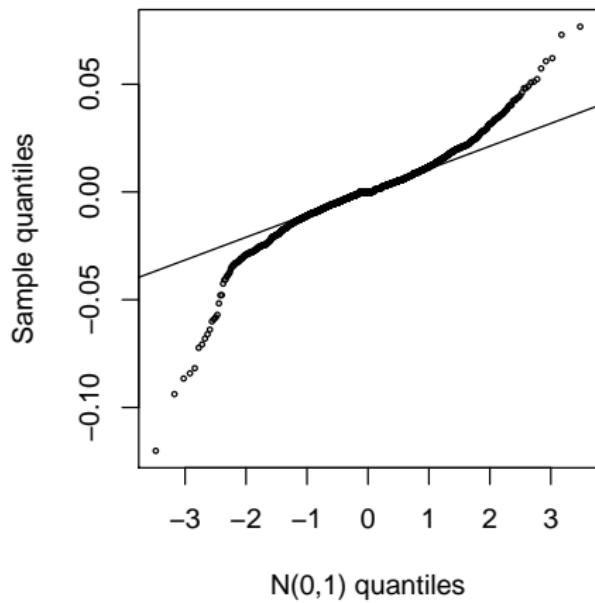


Considering the second margin only:

Q-Q plot for margin 2 (simulated data)

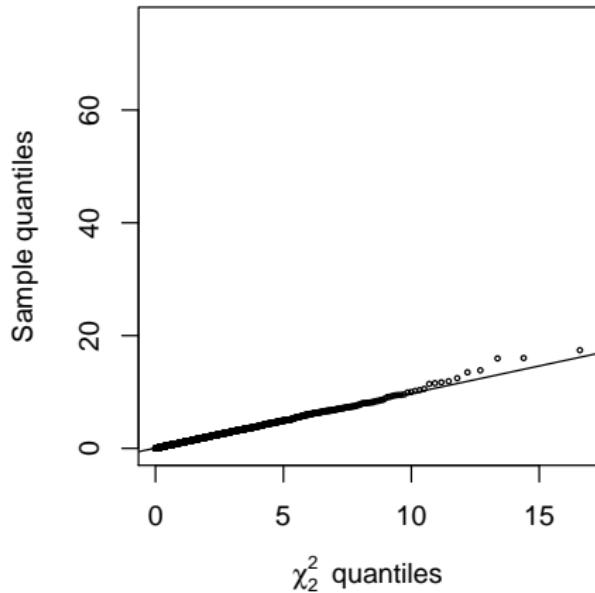


Q-Q plot for margin 2 (real data)

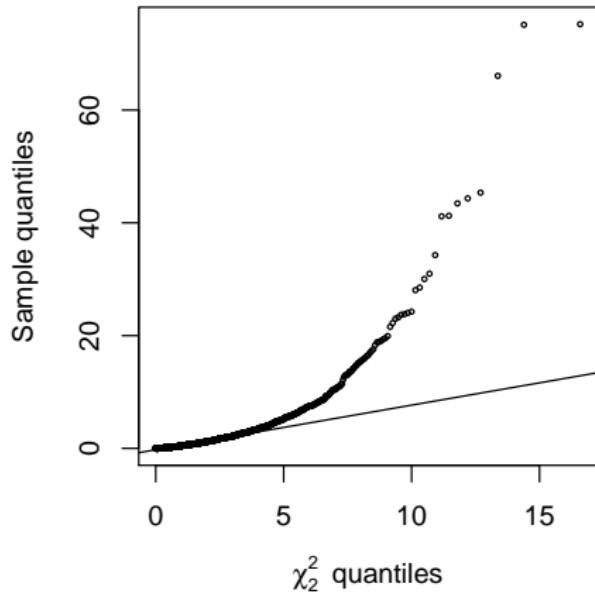


Q-Q plot of the simulated (left) or real (right) D_i^2 's against a χ_2^2 :

Q-Q plot of D_i^2 (simulated data)



Q-Q plot of D_i^2 (real data)



Advantages of $N_d(\mu, \Sigma)$

- Inference “easy”.
- Distribution is determined by μ and Σ .
- Linear combinations are normal (\Rightarrow VaR_α and ES_α calculations (for portfolios, for example) are easy).
- Marginal distributions are normal.
- Conditional distributions are normal.
- Quadratic forms are known.
- Convolutions are normal.
- Sampling is straightforward.
- Independence and uncorrelatedness are equivalent.

Drawbacks of $N_d(\mu, \Sigma)$ for modeling risk-factor changes

- 1) Tails of univariate (normal) margins are too thin (generate too few extreme events).
- 2) Joint tails are too thin (generate too few joint extreme events). $N_d(\mu, \Sigma)$ cannot capture the notion of tail dependence (see later).
- 3) Very strong symmetry known as radial symmetry: \mathbf{X} is called *radially symmetric about μ* if $\mathbf{X} - \mu \stackrel{d}{=} \mu - \mathbf{X}$. For $N_d(\mu, \Sigma)$: $\mathbf{X} - \mu \stackrel{d}{=} A\mathbf{Z} \stackrel{d}{=} A(-\mathbf{Z}) = -A\mathbf{Z} \stackrel{d}{=} -(\mathbf{X} - \mu) = \mu - \mathbf{X}$.

In short:

- Elliptical distributions (a generalization of normal mixture distributions) can address 1) and 2) while sharing many of the desirable properties of $N_d(\mu, \Sigma)$.
- Normal mean-variance mixture distribution can also address 3) (but at the expense of tractability in comparison to $N_d(\mu, \Sigma)$).

6.2 Normal mixture distributions

Idea: Randomize Σ (and μ) with a non-negative rv W .

6.2.1 Normal variance mixtures

Definition 6.11 (Multivariate normal variance mixtures)

The random vector \mathbf{X} has a (multivariate) *normal variance mixture distribution* if

$$\mathbf{X} \stackrel{\text{d}}{=} \boldsymbol{\mu} + \sqrt{W} A \mathbf{Z}, \quad (25)$$

where $\mathbf{Z} \sim N_k(\mathbf{0}, I_k)$, $W \geq 0$ is a rv independent of \mathbf{Z} , $A \in \mathbb{R}^{d \times k}$, and $\boldsymbol{\mu} \in \mathbb{R}^d$. $\boldsymbol{\mu}$ is called *location vector* and $\Sigma = AA^\top$ *scale* (or *dispersion matrix*).

Observe that $(\mathbf{X} | W = w) \stackrel{\text{d}}{=} \boldsymbol{\mu} + \sqrt{w} A \mathbf{Z} = N_d(\boldsymbol{\mu}, wAA^\top) = N_d(\boldsymbol{\mu}, w\Sigma)$; or $(\mathbf{X} | W) \stackrel{\text{d}}{=} N_d(\boldsymbol{\mu}, W\Sigma)$. W can be interpreted as a shock affecting the volatilities of all risk factors.

Properties of multivariate normal variance mixtures

Assume $\text{rank}(A) = d \leq k$ and that Σ is positive definite. Let $\mathbf{Y} = \boldsymbol{\mu} + A\mathbf{Z}$.

- If $\mathbb{E}\sqrt{W} < \infty$, then $\mathbb{E}[\mathbf{X}] \stackrel{\text{ind.}}{=} \boldsymbol{\mu} + \mathbb{E}[\sqrt{W}]A\mathbb{E}[\mathbf{Z}] = \boldsymbol{\mu} + \mathbf{0} = \boldsymbol{\mu}$ ($= \mathbb{E}\mathbf{Y}$)
- If $\mathbb{E}W < \infty$, then

$$\begin{aligned}\text{Cov}[\mathbf{X}] &= \text{Cov}[\sqrt{W}A\mathbf{Z}] = \mathbb{E}[(\sqrt{W}A\mathbf{Z})(\sqrt{W}A\mathbf{Z})^\top] \\ &\stackrel{\text{ind.}}{=} \mathbb{E}[W] \cdot \mathbb{E}[A\mathbf{Z}\mathbf{Z}^\top A^\top] = \mathbb{E}[W] \cdot A\mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]A^\top \\ &= \mathbb{E}[W]AI_kA^\top = \mathbb{E}[W]\Sigma \underset{\text{in general}}{\neq} \Sigma \quad (= \text{Cov}[\mathbf{Y}])\end{aligned}$$

- However, if they exist (i.e., if $\mathbb{E}W < \infty$), $\text{Cor}[\mathbf{X}]$ and $\text{Cor}[\mathbf{Y}]$ are equal:

Proof. $\text{Cov}[\mathbf{X}] = \mathbb{E}[W]\Sigma \Rightarrow \text{Cov}[X_i, X_j] = \mathbb{E}[W]\Sigma_{ij}$ and $\text{Var}[X_i] = \mathbb{E}[W]\Sigma_{ii}$. This implies that

$$\text{Cor}[X_i, X_j] = \frac{\text{Cov}[X_i, X_j]}{\sqrt{\text{Var}[X_i]\text{Var}[X_j]}} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}} = \text{Cor}[Y_i, Y_j]. \quad \square$$

Lemma 6.12 (Independence in normal variance mixtures)

Let $\mathbf{X} = \boldsymbol{\mu} + \sqrt{W}\mathbf{A}\mathbf{Z}$ with $\mathbb{E}W < \infty$ and $A = I_d$ ($\Rightarrow \text{Cov}[\mathbf{X}] = \mathbb{E}[W]\text{Cov}[\mathbf{Z}] = \mathbb{E}[W]I_d$ (uncorrelated)). Then

X_i and X_j are independent $\iff W$ is a.s. constant (i.e., $\mathbf{X} \sim N_d$).

Proof. W.l.o.g. assume $\boldsymbol{\mu} = \mathbf{0}$.

$$\begin{aligned} \Rightarrow \mathbb{E}|X_i| \mathbb{E}|X_j| &\stackrel{\text{ind.}}{=} \mathbb{E}[|X_i||X_j|] = \mathbb{E}[W|Z_i||Z_j|] \stackrel{\text{ind.}}{=} \mathbb{E}[W] \mathbb{E}|Z_i| \mathbb{E}|Z_j| \\ &\geq \mathbb{E}[\sqrt{W}]^2 \mathbb{E}|Z_i| \mathbb{E}|Z_j| \stackrel{\text{Jensen}}{=} \mathbb{E}|\sqrt{W}Z_i| \mathbb{E}|\sqrt{W}Z_j| = \mathbb{E}|X_i| \mathbb{E}|X_j| \end{aligned}$$

\Rightarrow We must have “=” in Jensen's inequality. This holds if and only if

W is constant a.s.; so $\mathbf{X} \sim N_d(\mathbf{0}, WI_d)$ in this case.

\Leftarrow W a.s. constant $\Rightarrow \mathbf{X} \sim N_d(\mathbf{0}, WI_d) \Rightarrow X_i, X_j$ independent. \square

Recall: If $\mathbf{X} \sim \text{N}_d(\boldsymbol{\mu}, \Sigma)$, then $\phi_{\mathbf{X}}(\mathbf{t}) = \exp(i\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t})$.

Furthermore, $\mathbf{X} | W = w \sim \text{N}_d(\boldsymbol{\mu}, w\Sigma)$ (or: $\mathbf{X} | W \sim \text{N}_d(\boldsymbol{\mu}, W\Sigma)$)

- **Characteristic function:** The cf of a multivariate normal variance mixtures is

$$\begin{aligned}\phi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}[\exp(i\mathbf{t}^\top \mathbf{X})] = \mathbb{E}[\mathbb{E}[\exp(i\mathbf{t}^\top \mathbf{X}) | W]] \\ &= \mathbb{E}[\exp(i\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}W\mathbf{t}^\top \Sigma \mathbf{t})] = \exp(i\mathbf{t}^\top \boldsymbol{\mu}) \mathbb{E}[\exp(-W\frac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t})].\end{aligned}$$

- **LS transform:** The *Laplace-Stieltjes transform* of F_W is

$$\hat{F}_W(\theta) := \mathcal{LS}[F_W](\theta) := \mathbb{E}[\exp(-\theta W)] = \int_0^\infty e^{-\theta w} dF_W(w).$$

Therefore, $\phi_{\mathbf{X}}(\mathbf{t}) = \exp(i\mathbf{t}^\top \boldsymbol{\mu}) \hat{F}_W(\frac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t})$. We thus introduce the notation $\mathbf{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{F}_W)$ for a d -dimensional multivariate normal variance mixture.

- **Density:** If Σ is positive definite, $\mathbb{P}(W = 0) = 0$, the density of \mathbf{X} is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \int_0^\infty f_{\mathbf{X}|W}(\mathbf{x} | w) dF_W(w) \\ &= \int_0^\infty \frac{1}{(2\pi)^{d/2} w^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2w}\right) dF_W(w). \end{aligned}$$

\Rightarrow Only depends on \mathbf{x} through $(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$.

\Rightarrow Multivariate normal variance mixtures are elliptical distributions.

If Σ is diagonal and $\mathbb{E}W < \infty$, \mathbf{X} is uncorrelated (as $\text{Cov}[\mathbf{X}] = \mathbb{E}[W]\Sigma$) but not independent unless W is constant (\sqrt{W} creates dependence).

- **Linear combinations:** For $\mathbf{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{F}_W)$ and $\mathbf{Y} = B\mathbf{X} + \mathbf{b}$, where $B \in \mathbb{R}^{k \times d}$ and $\mathbf{b} \in \mathbb{R}^k$, we have $\mathbf{Y} \sim M_k(B\boldsymbol{\mu} + \mathbf{b}, B\Sigma B^\top, \hat{F}_W)$.

Proof. Recall that $\phi_{\mathbf{X}}(\mathbf{t}) = \exp(i\mathbf{t}^\top \boldsymbol{\mu}) \hat{F}_W(\frac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t})$. Thus,

$$\begin{aligned} \phi_{\mathbf{Y}}(\mathbf{t}) &= \mathbb{E}[\exp(i\mathbf{t}^\top (B\mathbf{X} + \mathbf{b}))] = \exp(i\mathbf{t}^\top \mathbf{b}) \cdot \mathbb{E}[\exp(i(B^\top \mathbf{t})^\top \mathbf{X})] \\ &= \exp(i\mathbf{t}^\top \mathbf{b}) \phi_{\mathbf{X}}(B^\top \mathbf{t}) = \exp(i\mathbf{t}^\top (\mathbf{b} + B\boldsymbol{\mu})) \hat{F}_W(\frac{1}{2}\mathbf{t}^\top B\Sigma B^\top \mathbf{t}). \quad \square \end{aligned}$$

If $\mathbf{a} \in \mathbb{R}^d$ ($\mathbf{b} = \mathbf{0}$, $B = \mathbf{a}^\top \in \mathbb{R}^{1 \times d}$), $\mathbf{a}^\top \mathbf{X} \sim M_1(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \Sigma \mathbf{a}, \hat{F}_W)$.

- **Sampling:**

Algorithm 6.13 (Simulation of $\mathbf{X} = \boldsymbol{\mu} + \sqrt{W}\mathbf{A}\mathbf{Z} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{F}_W)$)

- 1) Generate $\mathbf{Z} \sim N_d(\mathbf{0}, I_d)$.
- 2) Generate $W \sim F_W$ (with LS transform \hat{F}_W), independent of \mathbf{Z} .
- 3) Compute the Cholesky factor A (such that $AA^\top = \Sigma$).
- 4) Return $\mathbf{X} = \boldsymbol{\mu} + \sqrt{W}\mathbf{A}\mathbf{Z}$.

Example 6.14 ($t_d(\nu, \boldsymbol{\mu}, \Sigma)$ distribution)

- 1) Generate $\mathbf{Z} \sim N_d(\mathbf{0}, I_d)$.
- 2) Generate $V \sim \chi^2_\nu$ and set $W = \frac{\nu}{V} \sim \text{Ig}(\nu/2, \nu/2)$.
Alternatively, $W = 1/V$ with $V \sim \Gamma(\nu/2, \text{rate} = \nu/2)$.
- 3) Compute the Cholesky factor A (such that $AA^\top = \Sigma$).
- 4) Return $\mathbf{X} = \boldsymbol{\mu} + \sqrt{W}\mathbf{A}\mathbf{Z}$.

Examples of multivariate normal variance mixtures

- **Multivariate normal distribution**

$W = 1$ a.s. (degenerate case)

- **Two point mixture**

$$W = \begin{cases} w_1 & \text{with probability } p, \\ w_2 & \text{with probability } 1 - p \end{cases} \quad w_1, w_2 > 0, w_1 \neq w_2.$$

Can be used to model [ordinary and stress regimes](#); extends to k regimes.

- **Symmetric generalised hyperbolic distribution**

W has a generalised inverse Gaussian distribution (GIG); see McNeil et al. (2015, p. 187)

- **Multivariate t distribution**

W has an inverse gamma distribution $W = 1/V$ for $V \sim \Gamma(\nu/2, \nu/2)$.

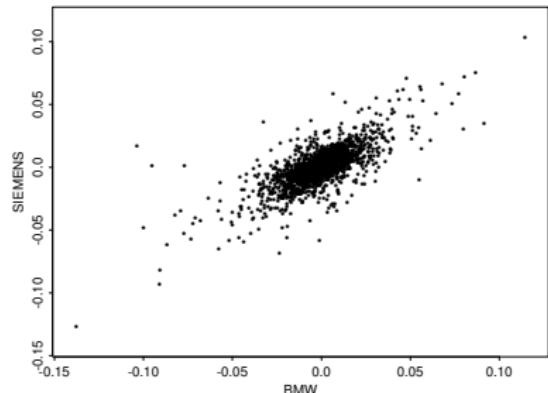
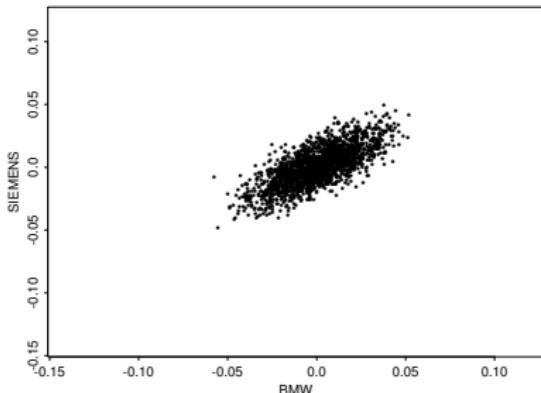
► $\mathbb{E}[W] = \frac{\nu}{\nu-2} \Rightarrow \text{Cov}[\mathbf{X}] = \frac{\nu}{\nu-2} \Sigma$. For finite variances/correlations, $\nu > 2$ is required. For finite mean, $\nu > 1$ is required.

- The (elliptical) density of the multivariate t distribution is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\Gamma((\nu + d)/2)}{\Gamma(\nu/2)(\nu\pi)^{d/2}|\Sigma|^{1/2}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+d}{2}},$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ is a positive definite matrix, and ν is the degrees of freedom. Notation: $\mathbf{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$.

- $t_d(\nu, \boldsymbol{\mu}, \Sigma)$ has heavier marginal and joint tails than $N_d(\boldsymbol{\mu}, \Sigma)$.
- BMW–Siemens data: Simulations from fitted $N_d(\boldsymbol{\mu}, \Sigma)$ and $t_d(3, \boldsymbol{\mu}, \Sigma)$:



6.2.2 Normal mean-variance mixtures

- Radial symmetry implies that all one-dimensional margins of normal variance mixtures are symmetric.
- Often visible in data: Joint losses have heavier tails than joint gains.

Idea: Introduce asymmetry by mixing normal distributions with different means and variances.

\mathbf{X} has a (multivariate) *normal mean-variance mixture distribution* if

$$\mathbf{X} \stackrel{\text{d}}{=} \mathbf{m}(W) + \sqrt{W} A \mathbf{Z}, \quad (26)$$

where

- $\mathbf{Z} \sim \mathcal{N}_k(\mathbf{0}, I_k)$;
- $W \geq 0$ is a scalar random variable which is independent of \mathbf{Z} ;
- $A \in \mathbb{R}^{d \times k}$ is a matrix of constants;
- $\mathbf{m} : [0, \infty) \rightarrow \mathbb{R}^d$ is a measurable function.

- Normal mean-variance mixtures add radial asymmetry: Let $\Sigma = AA^\top$ and observe that $\mathbf{X} | W = w \sim N_d(\boldsymbol{m}(w), w\Sigma)$. In general, they are no longer elliptical and $\text{Cor}(\mathbf{X}) \neq \text{Cor}(\mathbf{Y})$ (where $\mathbf{Y} = \boldsymbol{\mu} + A\mathbf{Z}$)

Example 6.15 (Generalized hyperbolic distribution)

- Here, $\boldsymbol{m}(W) = \boldsymbol{\mu} + W\boldsymbol{\gamma}$. Since

$$\mathbb{E}[\mathbf{X} | W] = \boldsymbol{\mu} + W\boldsymbol{\gamma},$$

$$\text{Cov}[\mathbf{X} | W] = W\Sigma$$

one has

$$\mathbb{E}\mathbf{X} = \mathbb{E}[\mathbb{E}[\mathbf{X} | W]] = \boldsymbol{\mu} + \mathbb{E}[W]\boldsymbol{\gamma} \quad \text{if } \mathbb{E}W < \infty,$$

$$\begin{aligned} \text{Cov}[\mathbf{X}] &= \mathbb{E}[\text{Cov}[\mathbf{X} | W]] + \text{Cov}[\mathbb{E}[\mathbf{X} | W]] \\ &= \mathbb{E}[W]\Sigma + \text{Var}[W]\boldsymbol{\gamma}\boldsymbol{\gamma}^\top \quad \text{if } \mathbb{E}[W^2] < \infty. \end{aligned}$$

- If W has a GIG distribution, then \mathbf{X} follows a generalised hyperbolic distribution. $\boldsymbol{\gamma} = \mathbf{0}$ leads to (elliptical) normal variance mixtures; see McNeil et al. (2015, Sections 6.2.3) for details.

6.3 Spherical and elliptical distributions

Empirical examples (see McNeil et al. (2015, Sections 6.2.4)) show that

- 1) $M_d(\mu, \Sigma, \hat{F}_W)$ (e.g., multivariate t , NIG) provide superior models to $N_d(\mu, \Sigma)$ for daily/weekly US stock-return data;
- 2) the more general radially asymmetric normal mean-variance mixture distributions did not seem to offer much of an improvement.

We soon study elliptical distributions, a generalization of $M_d(\mu, \Sigma, \hat{F}_W)$.

6.3.1 Spherical distributions

Definition 6.16 (Spherical distribution)

A random vector $\mathbf{Y} = (Y_1, \dots, Y_d)$ has a *spherical distribution* if for every orthogonal $U \in \mathbb{R}^{d \times d}$ (i.e., $U \in \mathbb{R}^{d \times d}$ with $UU^\top = U^\top U = I_d$)

$$\mathbf{Y} \stackrel{d}{=} U\mathbf{Y} \quad (\text{distributionally invariant under rotations and reflections})$$

Theorem 6.17 (Characterization of spherical distributions)

Let $\|\mathbf{t}\| = (t_1^2 + \cdots + t_d^2)^{1/2}$, $\mathbf{t} \in \mathbb{R}^d$. The following are equivalent:

- 1) \mathbf{Y} is spherical.
- 2) \exists a characteristic generator $\psi : [0, \infty) \rightarrow \mathbb{R}$, such that $\phi_{\mathbf{Y}}(\mathbf{t}) = \mathbb{E}[e^{i\mathbf{t}^\top \mathbf{Y}}] = \psi(\|\mathbf{t}\|^2)$, $\forall \mathbf{t} \in \mathbb{R}^d$.
- 3) For every $\mathbf{a} \in \mathbb{R}^d$, $\mathbf{a}^\top \mathbf{Y} \stackrel{d}{=} \|\mathbf{a}\| Y_1$ (linear combinations are of the same type \Rightarrow subadditivity of VaR_α for elliptically distr. losses).
see later

Proof. 1) \Rightarrow 2): $\phi_{\mathbf{Y}}(\mathbf{t}) = \phi_{U\mathbf{Y}}(\mathbf{t}) = \phi_{\mathbf{Y}}(U^\top \mathbf{t})$ for all $U \in \mathbb{R}^{d \times d}$ orthogonal. Since U can only change the direction of \mathbf{t} but not its length, $\phi_{\mathbf{Y}}(\mathbf{t})$ only depends on $\|\mathbf{t}\|$, i.e., the length of \mathbf{t} \Rightarrow we can define $\psi(\|\mathbf{t}\|^2) = \phi_{\mathbf{Y}}(\mathbf{t})$.

2) \Rightarrow 3): $\phi_{Y_1}(t) = \phi_{\mathbf{Y}}(te_1) \stackrel{2)}{=} \psi(t^2)$ (*). Now $\phi_{\mathbf{a}^\top \mathbf{Y}}(t) = \phi_{\mathbf{Y}}(t\mathbf{a}) \stackrel{2)}{=} \psi(t^2 \|\mathbf{a}\|^2) = \psi((t\|\mathbf{a}\|)^2) \stackrel{(*)}{=} \phi_{Y_1}(t\|\mathbf{a}\|) = \phi_{\|\mathbf{a}\| Y_1}(t)$

$$\begin{aligned}
 3) \Rightarrow 1): \phi_{U\mathbf{Y}}(\mathbf{t}) &= \mathbb{E}[\exp(i(U^\top \mathbf{t})^\top \mathbf{Y})] = \mathbb{E}[\exp(i\mathbf{a}^\top \mathbf{Y})] \stackrel{3)}{=} \mathbb{E}[\exp(i\|\mathbf{a}\|Y_1)] \\
 &= \mathbb{E}[\exp(i\|\mathbf{t}\|Y_1)] \stackrel{3)}{=} \mathbb{E}[\exp(it^\top \mathbf{Y})] = \phi_{\mathbf{Y}}(\mathbf{t})
 \end{aligned}
 \quad \square$$

Due to the above characterizations, we introduce the notation $\mathbf{Y} \sim S_d(\psi)$.

Theorem 6.18 (Stochastic representation)

$\mathbf{Y} \sim S_d(\psi)$ if and only if

$$\mathbf{Y} \stackrel{\text{d}}{=} R\mathbf{S}, \quad (27)$$

for independent *radial part* $R \geq 0$ and $\mathbf{S} \sim U(\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\})$.

Proof. Let Ω_d be the characteristic generator of \mathbf{S} .

“ \Rightarrow ” $\mathbf{Y} \sim S_d(\psi) \Rightarrow \phi_{\mathbf{Y}}(\|\mathbf{t}\|\mathbf{u}) \stackrel{2)}{=} \psi(\|\mathbf{t}\|^2 \mathbf{u}^\top \mathbf{u}) = \psi(\|\mathbf{t}\|^2)$ for all $\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1$. Replacing \mathbf{u} by \mathbf{S} and integrating leads to $\psi(\|\mathbf{t}\|^2) = \mathbb{E}_{\mathbf{S}}[\phi_{\mathbf{Y}}(\|\mathbf{t}\|\mathbf{S})] = \mathbb{E}_{\mathbf{S}}[\mathbb{E}_{\mathbf{Y}}[e^{i\|\mathbf{t}\|\mathbf{S}^\top \mathbf{Y}}]] \stackrel{\text{Fubini}}{=} \mathbb{E}_{\mathbf{Y}}[\mathbb{E}_{\mathbf{S}}[e^{i\|\mathbf{t}\|\mathbf{S}^\top \mathbf{Y}}]] = \mathbb{E}_{\mathbf{Y}}[\phi_{\mathbf{S}}(\|\mathbf{t}\|\mathbf{Y})] \stackrel{2)}{=} \mathbb{E}_{\mathbf{Y}}[\Omega_d(\|\mathbf{t}\|^2 \mathbf{Y}^\top \mathbf{Y})]$. We thus obtain that

$$\begin{aligned}\phi_{\mathbf{Y}}(\mathbf{t}) &\stackrel{2)}{=} \psi(\|\mathbf{t}\|^2) \underset{R:=\|\mathbf{Y}\|}{=} \mathbb{E}_R[\Omega_d(\|\mathbf{t}\|^2 R^2)] = \int_0^\infty \Omega_d(\|\mathbf{t}\|^2 r^2) dF_R(r) \\ &\stackrel{2)}{=} \int_0^\infty \phi_S(rt) dF_R(r) = \phi_{RS}(\mathbf{t}) \text{ for all } \mathbf{t} \in \mathbb{R}^d.\end{aligned}$$

“ \Leftarrow ” Let $\mathbf{Z} \sim N_d(\mathbf{0}, I_d)$. Since \mathbf{Z} is spherical and $\|\mathbf{Z}/\|\mathbf{Z}\|\| = \|\mathbf{Z}\|/\|\mathbf{Z}\| = 1$, $\mathbf{S} \stackrel{d}{=} \mathbf{Z}/\|\mathbf{Z}\|$. As such, \mathbf{S} itself is spherical, since $U\mathbf{S} \stackrel{d}{=} U\mathbf{Z}/\|\mathbf{Z}\| \stackrel{d}{=} \mathbf{Z}/\|\mathbf{Z}\| \stackrel{d}{=} \mathbf{S}$ for any orthogonal $U \in \mathbb{R}^{d \times d}$. Theorem 6.17 Part 2) implies that $\phi_S(\mathbf{t}) = \Omega_d(\|\mathbf{t}\|^2)$, so $\phi_{RS}(\mathbf{t}) = \mathbb{E}[\exp(i\mathbf{t}^\top RS)] = \mathbb{E}_R[\mathbb{E}[\exp(i\mathbf{t}^\top RS) | R]] = \mathbb{E}_R[\phi_S(R\mathbf{t})] = \mathbb{E}_R[\Omega_d(R^2\|\mathbf{t}\|^2)]$, which is a function in $\|\mathbf{t}\|^2$ and thus, by 2), RS is spherical. \square

Corollary 6.19

If $\mathbf{Y} \sim S_d(\psi)$ and $\mathbb{P}(\mathbf{Y} = \mathbf{0}) = 0$, then $(\|\mathbf{Y}\|, \frac{\mathbf{Y}}{\|\mathbf{Y}\|}) \stackrel{d}{=} (R, \mathbf{S})$ since $(\|\mathbf{Y}\|, \frac{\mathbf{Y}}{\|\mathbf{Y}\|}) \stackrel{d}{=} (\|RS\|, \frac{RS}{\|RS\|}) = (|R|\|\mathbf{S}\|, \frac{\mathbf{S}}{\|\mathbf{S}\|}) = (R, \mathbf{S})$.

In particular, $\|\mathbf{Y}\|$ and $\mathbf{Y}/\|\mathbf{Y}\|$ are independent (\Rightarrow goodness-of-fit).

- If $\mathbf{Y} \sim S_d(\psi)$ and admits a density $f_{\mathbf{Y}}$, then the *inversion formula* $f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-it^\top \mathbf{y}} \phi_{\mathbf{Y}}(\mathbf{t}) dt$ and Theorem 6.17 Part 2) show that for any orthogonal U ,

$$\begin{aligned}
 f_{\mathbf{Y}}(U\mathbf{y}) &\stackrel{\text{inv.}}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i(U^\top \mathbf{t})^\top \mathbf{y}} \phi_{\mathbf{Y}}(\mathbf{t}) dt \\
 &\stackrel{\text{subs.}}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-is^\top \mathbf{y}} \phi_{\mathbf{Y}}(Us) ds \\
 &\stackrel{2)}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-is^\top \mathbf{y}} \psi((Us)^\top Us) ds \\
 &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-is^\top \mathbf{y}} \psi(s^\top s) ds \stackrel{\text{backwards}}{=} \dots = f_{\mathbf{Y}}(\mathbf{y}).
 \end{aligned}$$

This implies that $f_{\mathbf{Y}}(\mathbf{y}) = g(\|\mathbf{y}\|^2)$ for a function $g : [0, \infty) \rightarrow [0, \infty)$ referred to as *density generator*. So $f_{\mathbf{Y}}(\mathbf{y})$ is constant on hyperspheres in \mathbb{R}^d .

- For $\mathbf{Y} \sim t_d(\nu, \mathbf{0}, I_d)$, $g(x) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)(\pi\nu)^{d/2}} (1 + \frac{x}{\nu})^{-(\nu+d)/2}$.

Example 6.20 (Standardized multivariate normal variance mixtures)

- $\mathbf{Y} \sim M_d(\mathbf{0}, \mathbf{I}_d, \hat{F}_W)$ is spherical (recall: $\mathbf{Y} \stackrel{d}{=} \mathbf{0} + \sqrt{W} I_d \mathbf{Z}$) since

$$\begin{aligned}\phi_{\mathbf{Y}}(\mathbf{t}) &= \mathbb{E}[\exp(i\mathbf{t}^\top \sqrt{W} \mathbf{Z})] = \mathbb{E}_W[\mathbb{E}[\exp(i(\mathbf{t}\sqrt{W})^\top \mathbf{Z}) | W]] \\ &= \mathbb{E}[\exp(-\frac{1}{2}W\mathbf{t}^\top \mathbf{t})] = \hat{F}_W(\frac{1}{2}\mathbf{t}^\top \mathbf{t}) = \hat{F}_W(\frac{1}{2}\|\mathbf{t}\|^2),\end{aligned}$$

so $\mathbf{Y} \sim S_d(\psi)$ by Theorem 6.17 Part 2). We see that the characteristic generator of \mathbf{Y} is $\psi(t) = \hat{F}_W(t/2)$.

- For $\mathbf{Y} \sim N_d(\mathbf{0}, I_d)$, $\psi(t) = \exp(-t/2)$. By Corollary 6.19, simulating $\mathbf{S} \sim U(\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\})$ can thus be done via $\mathbf{S} \stackrel{d}{=} \mathbf{Y}/\|\mathbf{Y}\|$. Fang et al. (1990, pp. 48) show that ψ generates $S_d(\psi)$ for all $d \in \mathbb{N}$ if and only if it is the characteristic generator of a normal mixture.
- Standardized normal variance mixtures \subseteq spherical distributions. They do not coincide, however, since $\mathbf{S} \sim U(\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\})$ is spherical but not a normal variance mixture (if it was, $\mathbf{S} = \sqrt{W} \mathbf{Z}$, so \sqrt{W} would have to scale Z_1, \dots, Z_d differently in order for $\|\mathbf{S}\| = 1$).

Example 6.21 (R , S , Cov, Cor)

- It follows from $\mathbf{Y} \sim N_d(\mathbf{0}, I_d)$ and $R^2 = \mathbf{Y}^\top \mathbf{Y} \sim \chi_d^2$ that

$$\mathbf{0} = \mathbb{E}\mathbf{Y} = \underset{\text{Th. 6.18}}{\mathbb{E}R\mathbb{E}\mathbf{S}} \Rightarrow \mathbb{E}\mathbf{S} = \mathbf{0},$$

$$I_d = \text{Cov } \mathbf{Y} = \underset{\text{Th. 6.18}}{\mathbb{E}[R^2]} \text{Cov } \mathbf{S} = d \text{Cov } \mathbf{S} \Rightarrow \text{Cov } \mathbf{S} = I_d/d. \quad (28)$$

- For $\mathbf{Y} \sim S_d(\psi)$ with $\mathbb{E}[R^2] < \infty$, it follows that $\text{Cov } \mathbf{Y} = \underset{\text{Th. 6.18}}{\mathbb{E}[R^2]} \text{Cov } \mathbf{S} = \frac{\mathbb{E}[R^2]}{d} I_d$ and thus $\text{Cor } \mathbf{Y} = I_d$.
- For $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Y}$ with $\mathbb{E}[R^2] < \infty$ and Cholesky factor A of a covariance matrix Σ , we have $\text{Cov } \mathbf{X} = \frac{\mathbb{E}[R^2]}{d} \Sigma$ and $\text{Cor } \mathbf{X} = P$ (the correlation matrix corresponding to Σ).
- Example:** For $\mathbf{Y} \sim t_d(\nu, \mathbf{0}, I_d)$, $R^2 = \mathbf{Y}^\top \mathbf{Y} = W\mathbf{Z}^\top \mathbf{Z}$ for $\mathbf{Z} \sim N_d(\mathbf{0}, I_d)$. Therefore, $\frac{R^2}{d} = \frac{\mathbf{Z}^\top \mathbf{Z}/d}{(\nu/W)/\nu} = \frac{\chi_d^2/\nu}{\chi_\nu^2/\nu} \sim F(d, \nu)$ and thus $\mathbb{E}[R^2/d] = \frac{\nu}{\nu-2}$. It follows that $\mathbf{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$ has $\text{Cov } \mathbf{X} = \frac{\nu}{\nu-2} \Sigma$ and $\text{Cor } \mathbf{X} = P$ which we already know from Section 6.2.1.

6.3.2 Elliptical distributions

Definition 6.22 (Elliptical distribution)

A random vector $\mathbf{X} = (X_1, \dots, X_d)$ has an *elliptical distribution* if

$$\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + A\mathbf{Y}, \quad (\text{multivariate affine transformation})$$

where $\mathbf{Y} \sim S_k(\psi)$, $A \in \mathbb{R}^{d \times k}$ (*scale matrix* $\Sigma = AA^\top$), and (*location vector*) $\boldsymbol{\mu} \in \mathbb{R}^d$.

- By Theorem 6.18, an elliptical random vector **admits the stochastic representation** $\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + RAS$, with R and S as given in (27).
- The **characteristic function** of an elliptical random vector \mathbf{X} is $\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[e^{i\mathbf{t}^\top \mathbf{X}}] = \mathbb{E}[e^{i\mathbf{t}^\top (\boldsymbol{\mu} + A\mathbf{Y})}] = e^{i\mathbf{t}^\top \boldsymbol{\mu}} \mathbb{E}[e^{i(A^\top \mathbf{t})^\top \mathbf{Y}}] = e^{i\mathbf{t}^\top \boldsymbol{\mu}} \psi(\mathbf{t}^\top \Sigma \mathbf{t})$.
Notation: $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ ($= E_d(\boldsymbol{\mu}, c\Sigma, \psi(\cdot/c))$, $c > 0$).
- If Σ is positive definite with Cholesky factor A , then $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ if and only if $\mathbf{Y} = A^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim S_d(\psi)$.

- Normal variance mixture distributions are (all) elliptical (most useful examples) since $\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + \sqrt{W} A \mathbf{Z} = \boldsymbol{\mu} + \sqrt{W} \|\mathbf{Z}\| A \mathbf{Z} / \|\mathbf{Z}\| = \boldsymbol{\mu} + R A S$ with $R = \sqrt{W} \|\mathbf{Z}\|$ and $S = \mathbf{Z} / \|\mathbf{Z}\|$. By Corollary 6.19, R and S are indeed independent.
- If $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ with $\mathbb{P}(\mathbf{X} = \boldsymbol{\mu}) = 0$, then $\mathbf{Y} = A^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim S_d(\psi)$. Corollary 6.19 implies that

$$\left(\sqrt{(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})}, \frac{A^{-1}(\mathbf{X} - \boldsymbol{\mu})}{\sqrt{(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})}} \right) \stackrel{d}{=} (R, S), \quad (29)$$

which can be used for testing elliptical symmetry. One can also use the following result for testing.

Proposition 6.23

Let $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ for positive definite Σ and $\mathbb{E}[R^2] < \infty$ (i.e., $\text{Cov}[\mathbf{X}]$ finite). For any $c \geq 0$ such that $\mathbb{P}((\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \geq c) > 0$,

$$\text{Cor}[\mathbf{X} | (\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \geq c] = \text{Cor}[\mathbf{X}].$$

Proof. $\mathbf{X} | ((\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \geq c) \stackrel{(29)}{\stackrel{\text{d}}{=}} \boldsymbol{\mu} + R\mathbf{A}\mathbf{S} | (R^2 \geq c) \stackrel{\text{ind.}}{=} \boldsymbol{\mu} + \tilde{R}\mathbf{A}\mathbf{S}$ where $\tilde{R} \stackrel{\text{d}}{=} (R | R^2 \geq c)$. Therefore, the conditional distribution remains elliptical with scale matrix Σ and thus the claim holds. \square

6.3.3 Properties of elliptical distributions

- **Density:** Let Σ be positive definite and $\mathbf{Y} \sim S_d(\psi)$ have density generator g . The [Density Transformation Theorem](#) implies that $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Y}$ has density

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{\det \Sigma}} g((\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})),$$

which depends on \mathbf{x} only through $(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$, i.e., is constant on ellipsoids (hence the name “elliptical”).

- **Linear combinations:** For $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$, $B \in \mathbb{R}^{k \times d}$ and $\mathbf{b} \in \mathbb{R}^k$,

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

If $\mathbf{a} \in \mathbb{R}^d$ (take $\mathbf{b} = \mathbf{0}$ and $B = \mathbf{a}^\top \in \mathbb{R}^{1 \times d}$),

$$\mathbf{a}^\top \mathbf{X} \sim E_1(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \Sigma \mathbf{a}, \psi) \quad (\text{as for } N(\boldsymbol{\mu}, \Sigma)). \quad (30)$$

From $\mathbf{a} = \mathbf{e}_j = (0, \dots, 0, 1, 0, \dots, 0)$ we see that **all marginal distributions are of the same type**.

Proof. Similarly as for multivariate normal variance mixtures,

$$\begin{aligned}\phi_{B\mathbf{X}+\mathbf{b}}(\mathbf{t}) &= \mathbb{E}[\exp(it^\top(B\mathbf{X} + \mathbf{b}))] = e^{it^\top \mathbf{b}} \phi_{\mathbf{X}}(B^\top \mathbf{t}) \\ &= e^{it^\top(\mathbf{b} + B\boldsymbol{\mu})} \psi(\mathbf{t}^\top B\Sigma B^\top \mathbf{t}).\end{aligned}$$

□

- **Marginal dfs:** As for $N_d(\boldsymbol{\mu}, \Sigma)$, it immediately follows that $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ satisfies $\mathbf{X}_1 \sim E_k(\boldsymbol{\mu}_1, \Sigma_{11}, \psi)$ and $\mathbf{X}_2 \sim E_{d-k}(\boldsymbol{\mu}_2, \Sigma_{22}, \psi)$.
- **Conditional distributions:** One can show that

$\mathbf{X}_2 | \mathbf{X}_1 = \mathbf{x}_1 \sim E_{d-k}(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}, \tilde{\psi}),$
where the characteristic generator $\tilde{\psi}$ is given in Embrechts et al. (2002).

For $N_d(\boldsymbol{\mu}, \Sigma)$ the characteristic generator remains the same.

- **Quadratic forms:** It follows from (29) that $(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \stackrel{d}{=} R^2$. If $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$, then $R^2 \sim \chi_d^2$; and if $\mathbf{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$, then $R^2/d \sim F(d, \nu)$.
- **Convolutions:** Let $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ and $\mathbf{Y} \sim E_d(\tilde{\boldsymbol{\mu}}, c\Sigma, \tilde{\psi})$ be independent. Then

$$a\mathbf{X} + b\mathbf{Y} \sim E_d(a\boldsymbol{\mu} + b\tilde{\boldsymbol{\mu}}, \Sigma, \psi^*)$$

for $a, b \in \mathbb{R}$, $c > 0$, and $\psi^*(t) = \psi(a^2 t) \tilde{\psi}(b^2 ct)$. Therefore, if $a = b = c = 1$, then $\mathbf{X} + \mathbf{Y} \sim E_d(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \Sigma, \psi(\cdot)\tilde{\psi}(\cdot))$.

Proof. $\phi_{a\mathbf{X}}(\mathbf{t}) = e^{it^\top a\boldsymbol{\mu}} \psi(a^2 \mathbf{t}^\top \Sigma \mathbf{t})$ and $\phi_{b\mathbf{Y}}(\mathbf{t}) = e^{it^\top b\tilde{\boldsymbol{\mu}}} \tilde{\psi}(b^2 c \mathbf{t}^\top \Sigma \mathbf{t})$. By independence of \mathbf{X} and \mathbf{Y} , $\phi_{a\mathbf{X}+b\mathbf{Y}}(\mathbf{t}) = \phi_{a\mathbf{X}}(\mathbf{t})\phi_{b\mathbf{Y}}(\mathbf{t}) = e^{it^\top (a\boldsymbol{\mu}+b\tilde{\boldsymbol{\mu}})} \psi^*(\mathbf{t}^\top \Sigma \mathbf{t})$, so $a\mathbf{X} + b\mathbf{Y} \sim E_d(a\boldsymbol{\mu} + b\tilde{\boldsymbol{\mu}}, \Sigma, \psi^*)$. \square

- We see that many nice properties of $N_d(\boldsymbol{\mu}, \Sigma)$ are preserved.

Proposition 6.24 (Subadditivity of VaR in elliptical models)

Let $L_i = \boldsymbol{\lambda}_i^\top \mathbf{X}$, $\boldsymbol{\lambda}_i \in \mathbb{R}^d$, $i \in \{1, \dots, n\}$, with $\mathbf{X} \sim \text{E}_d(\boldsymbol{\mu}, \Sigma, \psi)$. Then $\text{VaR}_\alpha(\sum_{i=1}^n L_i) \leq \sum_{i=1}^n \text{VaR}_\alpha(L_i)$ for all $\alpha \in [1/2, 1]$.

Proof. Consider a generic $L = \boldsymbol{\lambda}^\top \mathbf{X} \stackrel{d}{=} \boldsymbol{\lambda}^\top \boldsymbol{\mu} + \boldsymbol{\lambda}^\top A \mathbf{Y}$ for $\mathbf{Y} \sim S_k(\psi)$. By Theorem 6.17 Part 3), $\boldsymbol{\lambda}^\top A \mathbf{Y} \stackrel{d}{=} \|\boldsymbol{\lambda}^\top A\| Y_1$, so $L \stackrel{d}{=} \boldsymbol{\lambda}^\top \boldsymbol{\mu} + \|\boldsymbol{\lambda}^\top A\| Y_1$ (all of the same type). By Translation Invariance and Positive Homogeneity,

$$\text{VaR}_\alpha(L) = \boldsymbol{\lambda}^\top \boldsymbol{\mu} + \|\boldsymbol{\lambda}^\top A\| \text{VaR}_\alpha(Y_1). \quad (31)$$

Applying (31) to $L = \sum_{i=1}^n L_i$ and $L = L_i$, $i \in \{1, \dots, n\}$, and using that $\text{VaR}_\alpha(Y_1) \geq 0$ for $\alpha \in [1/2, 1]$, we obtain $\text{VaR}_\alpha(\sum_{i=1}^n L_i)$

$$\begin{aligned} &= \text{VaR}_\alpha((\sum_{i=1}^n \boldsymbol{\lambda}_i)^\top \mathbf{X}) \stackrel{(31)}{=} \sum_{i=1}^n \boldsymbol{\lambda}_i^\top \boldsymbol{\mu} + \|\sum_{i=1}^n \boldsymbol{\lambda}_i^\top A\| \text{VaR}_\alpha(Y_1) \\ &\leq \sum_{i=1}^n \boldsymbol{\lambda}_i^\top \boldsymbol{\mu} + (\sum_{i=1}^n \|\boldsymbol{\lambda}_i^\top A\|) \text{VaR}_\alpha(Y_1) = \sum_{i=1}^n (\boldsymbol{\lambda}_i^\top \boldsymbol{\mu} + \|\boldsymbol{\lambda}_i^\top A\| \text{VaR}_\alpha(Y_1)) \\ &\stackrel{(31)}{=} \sum_{i=1}^n \text{VaR}_\alpha(L_i). \end{aligned}$$

Note: For $\boldsymbol{\lambda}_i = \mathbf{e}_i$, $\text{VaR}_\alpha(\sum_{i=1}^d X_i) \leq \sum_{i=1}^d \text{VaR}_\alpha(X_i)$.

□

6.3.4 Estimating scale and correlation

- Suppose $\mathbf{X}_1, \dots, \mathbf{X}_n \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$. How can we estimate $\boldsymbol{\mu}, \Sigma$ and P ? (P is the correlation matrix corresponding to Σ ; this always exists)
- $\bar{\mathbf{X}}, S, R$ may not be the best options for heavy-tailed data (e.g., concerning robustness against contamination).

M-estimators for $\boldsymbol{\mu}, \Sigma$ (see Maronna (1976))

- **Goal:** Improve given estimators $\hat{\boldsymbol{\mu}}, \hat{\Sigma}$.
- **Idea:** Compute improved estimates by downweighting observations with large $D_i = \sqrt{(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^\top \hat{\Sigma}^{-1} (\mathbf{X}_i - \hat{\boldsymbol{\mu}})}$ (these are the ones which tend to distort $\hat{\boldsymbol{\mu}}, \hat{\Sigma}$ most).
- This can be turned into an iterative procedure that converges to so-called *M-estimates* of location and scale ($\hat{\Sigma}$ is in general biased).

Algorithm 6.25 (M-estimators of location and scale)

1) Set $k = 1$, $\hat{\mu}^{[1]} = \bar{X}$ and $\hat{\Sigma}^{[1]} = S$.

2) Repeat until convergence:

2.1) For $i \in \{1, \dots, n\}$ set $D_i = \sqrt{(\mathbf{X}_i - \hat{\mu}^{[k]})^\top \hat{\Sigma}^{[k]-1} (\mathbf{X}_i - \hat{\mu}^{[k]})}$.

2.2) Update:

$$\hat{\mu}^{[k+1]} = \frac{\sum_{i=1}^n w_1(D_i) \mathbf{X}_i}{\sum_{i=1}^n w_1(D_i)},$$

where w_1 is a weight function, e.g., $w_1(x) = (d + \nu)/(x^2 + \nu)$ (or $\mathbb{1}_{x \leq a} + (a/x)\mathbb{1}_{x > a}$ for some value a).

2.3) Update:

$$\hat{\Sigma}^{[k+1]} = \frac{1}{n} \sum_{i=1}^n w_2(D_i^2) (\mathbf{X}_i - \hat{\mu}^{[k]}) (\mathbf{X}_i - \hat{\mu}^{[k]})^\top,$$

where w_2 is a weight function, e.g., $w_2(x) = w_1(\sqrt{x})$ (or $(w_1(\sqrt{x}))^2$).

2.4) Set k to $k + 1$.

Estimating P via Kendall's tau

- One can show (see later) that if $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$, then

$$\tau(X_i, X_j) = \frac{2}{\pi} \arcsin(P_{ij}), \quad i \neq j, \quad (32)$$

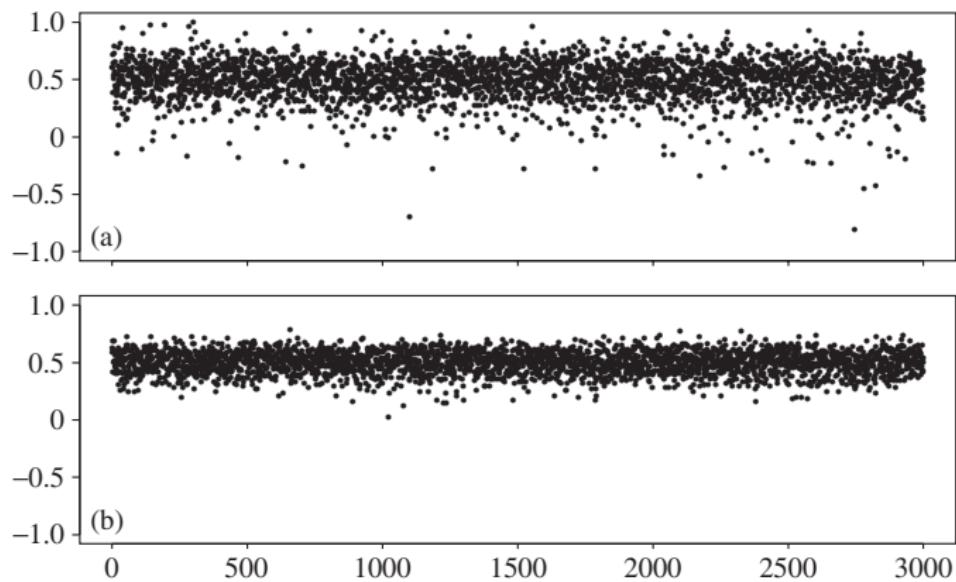
where P is the matrix of pairwise correlations corresponding to Σ (always existing, no matter whether the second moments of X_1, \dots, X_d do).

- Estimate $\tau(X_i, X_j)$ by $\hat{\tau}_{ij}$ (see later) and solve $\hat{\tau}_{ij}$ w.r.t. P_{ij} to obtain \hat{P}_{ij} (this does not require estimating variances/covariances).
- $(\hat{P}_{ij})_{ij}$ is not necessarily positive definite. There are various methods for finding a “near” matrix which is positive definite, see, e.g., Higham (2002) (or `Matrix:::nearPD()`).

Example 6.26 (Correlation estimation for heavy-tailed data)

Consider $n = 3000$ realizations of independent samples of size 90 from $t_2(3, \mathbf{0}, (\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}))$ (\Rightarrow linear correlation 0.5).

(a) Pearson's correlation; (b) Inversion of pairwise Kendall's tau estimator



The Kendall's tau transform method produces estimates that show less variation (and thus provides a more efficient way of estimating ρ).

6.4 Dimension reduction techniques

6.4.1 Factor models

Explain the variability of \mathbf{X} in terms of common factors.

Definition 6.27 (p -factor model)

\mathbf{X} follows a *p-factor model* if

$$\mathbf{X} = \mathbf{a} + B\mathbf{F} + \boldsymbol{\varepsilon}, \quad (33)$$

where

- 1) $B \in \mathbb{R}^{d \times p}$ is a *matrix of factor loadings* and $\mathbf{a} \in \mathbb{R}^d$;
- 2) $\mathbf{F} = (F_1, \dots, F_p)$ is the random vector of *(common) factors* with $p < d$ and existing $\Omega := \text{Cov}[\mathbf{F}]$, (*systematic risk*);
- 3) $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_d)$ is the random vector of *idiosyncratic error terms* with $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, $\Upsilon := \text{Cov}[\boldsymbol{\varepsilon}]$ diag., $\text{Cov}[\mathbf{F}, \boldsymbol{\varepsilon}] = (0)$ (*idiosync. risk*).

- **Goals:** Identify or estimate \mathbf{F}_t , $t \in \{1, \dots, n\}$, then model the distribution/dynamics of the (lower-dimensional) factors (instead of \mathbf{X}_t , $t \in \{1, \dots, n\}$).
- Factor models imply that $\Sigma := \text{Cov}[\mathbf{X}] = B\Omega B^\top + \Upsilon$.
- With $B^* = B\Omega^{1/2}$ and $\mathbf{F}^* = \Omega^{-1/2}(\mathbf{F} - \mathbb{E}[\mathbf{F}])$, we have

$$\mathbf{X} = \boldsymbol{\mu} + B^* \mathbf{F}^* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$. We have $\Sigma = B^*(B^*)^\top + \Upsilon$. Conversely, if $\text{Cov}[\mathbf{X}] = BB^\top + \Upsilon$ for some $B \in \mathbb{R}^{d \times p}$ with $\text{rank}(B) = p < d$ and diagonal matrix Υ , then \mathbf{X} has a factor-model representation for a p -dimensional \mathbf{F} and d -dimensional $\boldsymbol{\varepsilon}$.

Example 6.28 (One-factor/equicorrelation model)

Let $\mathbb{E}[\mathbf{X}] = \mathbf{0}$, $\Sigma = \text{Cov}[\mathbf{X}] = \rho J_d + (1 - \rho)I_d$ ($J_d = (1) \in \mathbb{R}^{d \times d}$).

- Then $\Sigma = BB^\top + \Upsilon$ for $B = \sqrt{\rho}\mathbf{1}$ and $\Upsilon = (1 - \rho)I_d$.
- Any Y with $\mathbb{E}Y = 0$, $\text{Var } Y = 1$ independent of \mathbf{X} leads to the *factor decomposition* of \mathbf{X}

$$F = \frac{\sqrt{\rho}}{1 + \rho(d-1)} \sum_{j=1}^d \textcolor{blue}{X}_j + \sqrt{\frac{1-\rho}{1+\rho(d-1)}} \mathbf{Y}, \quad \varepsilon_j = X_j - \sqrt{\rho}F.$$

We have $\mathbb{E}[F] = 0$, $\text{Var}[F] = 1$, so $\mathbf{X} = \mathbf{0} + BF + \boldsymbol{\varepsilon} = \sqrt{\rho}\mathbf{1}F + \boldsymbol{\varepsilon}$.

- The requirements of Definition 6.27 are fulfilled since $\text{Cov}[F, \varepsilon_j] = 0$, $\text{Cov}[\varepsilon_j, \varepsilon_k] = 0$ for all $j \neq k$.
- $\text{Var}[\bar{X}_n] = \text{Var}[\sqrt{\rho}F + \bar{\varepsilon}_d] = \rho + \frac{1-\rho}{d} \xrightarrow{(d \rightarrow \infty)} \rho$ (systematic factor matters!)
- If $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, take $Y \sim N(0, 1)$ (then F is also normal). One typically writes this (one-factor) equicorrelation model as $\mathbf{X} = \sqrt{\rho}F + \sqrt{1-\rho}\mathbf{Z}$, where $F, Z_1, \dots, Z_d \stackrel{\text{ind.}}{\sim} N(0, 1)$.

6.4.2 Statistical estimation strategies

Consider $\mathbf{X}_t = \mathbf{a} + B\mathbf{F}_t + \boldsymbol{\varepsilon}_t$, $t \in \{1, \dots, n\}$. Three types of factor model are commonly used:

- 1) **Macroeconomic factor models:** Here we assume that \mathbf{F}_t is observable, $t \in \{1, \dots, n\}$. Fitting B, \mathbf{a} is accomplished by time series regression (see later).
- 2) **Fundamental factor models:** Here we assume that the matrix of factor loadings B is known but the factors \mathbf{F}_t are unobserved (and have to be estimated from \mathbf{X}_t , $t \in \{1, \dots, n\}$, using cross-sectional regression at each t).
- 3) **Fundamental factor models:** Here we assume that neither the factors \mathbf{F}_t nor the factor loadings B are observed (both have to be estimated from \mathbf{X}_t , $t \in \{1, \dots, n\}$). The factors can be found with principal component analysis (see later).

6.4.3 Estimating macroeconomic factor models

There are two equivalent approaches.

Univariate regression

- Consider the (univariate) *time series regression* model

$$X_{t,j} = a_j + \mathbf{b}_j^\top \mathbf{F}_t + \varepsilon_{t,j}, \quad t \in \{1, \dots, n\}.$$

- To justify the use of the *ordinary least-squares* (OLS) method to derive statistical properties of the method it is usually assumed that, conditional on the factors, the errors $\varepsilon_{1,j}, \dots, \varepsilon_{n,j}$ form a white noise process (i.e., are identically distributed and serially uncorrelated).
- \hat{a}_j estimates a_j , $\hat{\mathbf{b}}_j$ estimates the j th row of B .

Multivariate regression

- Here, construct large matrices:

$$X = \underbrace{\begin{pmatrix} \mathbf{X}_1^\top \\ \vdots \\ \mathbf{X}_n^\top \end{pmatrix}}_{n \times d}, \quad F = \underbrace{\begin{pmatrix} 1 & \mathbf{F}_1^\top \\ \vdots & \vdots \\ 1 & \mathbf{F}_n^\top \end{pmatrix}}_{n \times (p+1)}, \quad \tilde{B} = \underbrace{\begin{pmatrix} \mathbf{a}^\top \\ B^\top \end{pmatrix}}_{(p+1) \times d}, \quad E = \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_1^\top \\ \vdots \\ \boldsymbol{\varepsilon}_n^\top \end{pmatrix}}_{n \times d}.$$

This model can be expressed by $X = F\tilde{B} + E$ (estimate \tilde{B}).

- Assume the unobserved $\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n$ form a white noise process. Then, conditional on $\mathbf{F}_1, \dots, \mathbf{F}_n$, we have a multivariate linear regression, see, e.g., Mardia et al. (1979), with estimator $\hat{B} = (F^\top F)^{-1} F^\top X$.
- Now examine the conditions of Definition 6.27: Do the errors vectors $\boldsymbol{\varepsilon}_t$ come from a distribution with diagonal covariance matrix, and are they uncorrelated with the factors?

- Consider the sample correlation matrix of $\hat{E} = X - F\hat{B}$ (model residual matrix; hopefully shows that there is little correlation in the errors) and take the diagonal elements as an estimator $\hat{\Upsilon}$ of Υ .

6.4.4 Estimating fundamental factor models

- Consider the cross-sectional regression model $X_t = BF_t + \varepsilon_t$ (B known; F_t to be estimated; $\text{Cov}[\varepsilon] = \Upsilon$); note that a can be absorbed into F_t . To obtain precision in estimating F_t , we need $d \gg p$.
- First estimate F_t via OLS by $\hat{F}_t^{\text{OLS}} = (B^\top B)^{-1} B^\top X_t$. This is the best linear unbiased estimator if the ε is homoskedastic. However, it is possible to obtain linear unbiased estimates with a smaller covariance matrix via generalized least squares (GLS).
- To this end, estimate Υ by $\hat{\Upsilon}$ via the diagonal of the sample covariance matrix of the residuals $\hat{\varepsilon}_t = X_t - B\hat{F}_t^{\text{OLS}}$, $t \in \{1, \dots, n\}$.
- Then estimate F_t via $\hat{F}_t = (B^\top \Upsilon^{-1} B)^{-1} B^\top \Upsilon^{-1} X_t$.

6.4.5 Principal component analysis

- **Goal:** Reduce the dimensionality of highly correlated data by finding a small number of uncorrelated linear combinations which account for most of the variance in the data; this can be used for finding factors.
- **Key:** Any symmetric A admits a *spectral decomposition*

where
$$A = \Gamma \Lambda \Gamma^\top,$$

- 1) $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$ is the diagonal matrix of eigenvalues of A which, w.l.o.g., are ordered so that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$; and
 - 2) Γ is an orthogonal matrix whose columns are eigenvectors of A standardized to have length 1.
- Let $\Sigma = \Gamma \Lambda \Gamma^\top$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$ (positive semidefiniteness \Rightarrow all eigenvalues ≥ 0) and $\mathbf{Y} = \Gamma^\top (\mathbf{X} - \boldsymbol{\mu})$ (the so-called *principal component transform*). The j th component $Y_j = \gamma_j^\top (\mathbf{X} - \boldsymbol{\mu})$ is the j th *principal component of \mathbf{X}* (where γ_j is the j th column of Γ).

- We have $\mathbb{E}\mathbf{Y} = \mathbf{0}$ and $\text{Cov}[\mathbf{Y}] = \Gamma^\top \Sigma \Gamma = \Gamma^\top \Gamma \Lambda \Gamma^\top \Gamma = \Lambda$, so the principal components are uncorrelated with $\text{Var}[Y_j] = \lambda_j$, $j \in \{1, \dots, d\}$. The principal components are thus ordered by variance (from largest to smallest).
- One can show:
 - ▶ The first principal component is that standardized linear combination of \mathbf{X} which has maximal variance among all such combinations, i.e., $\text{Var}(\gamma_1^\top \mathbf{X}) = \max\{\text{Var}(\mathbf{a}^\top \mathbf{X}) : \mathbf{a}^\top \mathbf{a} = 1\}$.
 - ▶ For $j \in \{2, \dots, d\}$, the j th principal component is that standardized linear combination of \mathbf{X} which has maximal variance among all such linear combinations which are orthogonal to (and hence uncorrelated with) the first $j - 1$ -many linear combinations.
- $\sum_{j=1}^d \text{Var}(Y_j) = \sum_{j=1}^d \lambda_j = \text{trace}(\Sigma) = \sum_{j=1}^d \text{Var}(X_j)$, so we can interpret $\sum_{j=1}^k \lambda_j / \sum_{j=1}^d \lambda_j$ as the fraction of total variance explained by the first k principal components.

Principal components as factors

- Inverting the principal component transform $\mathbf{Y} = \Gamma^\top(\mathbf{X} - \boldsymbol{\mu})$, we have

$$\mathbf{X} = \boldsymbol{\mu} + \Gamma\mathbf{Y} = \boldsymbol{\mu} + \Gamma_1\mathbf{Y}_1 + \Gamma_2\mathbf{Y}_2 =: \boldsymbol{\mu} + \Gamma_1\mathbf{Y}_1 + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}_1 \in \mathbb{R}^k$ contains the first k principal components. This is reminiscent of the basic factor model.

- Although $\varepsilon_1, \dots, \varepsilon_d$ will tend to have small variances, the assumptions of the factor model are generally violated (since they need not have a diagonal covariance matrix and need not be uncorrelated with \mathbf{Y}_1). Nevertheless, principal components are often interpreted as factors.

Sample principal components

- Assume $\mathbf{X}_1, \dots, \mathbf{X}_n$ with identical distribution, unknown mean vector $\boldsymbol{\mu}$ and covariance matrix Σ with the spectral decomposition $\Sigma = \Gamma\Lambda\Gamma^\top$ as before.
- Estimate $\boldsymbol{\mu}$ by $\bar{\mathbf{X}}$ and Σ by $S_x = \frac{1}{n} \sum_{t=1}^n (\mathbf{X}_t - \bar{\mathbf{X}})(\mathbf{X}_t - \bar{\mathbf{X}})^\top$.

- Apply the spectral decomposition to S_x to get $S_x = GLG^\top$, where G is the eigenvector matrix and $L = \text{diag}(l_1, \dots, l_d)$ is the diagonal matrix consisting of ordered eigenvalues.
- Define the “sample principle component transforms” $\mathbf{Y}_t = G^\top(\mathbf{X}_t - \bar{\mathbf{X}})$, $t \in \{1, \dots, n\}$. The j th component $Y_{t,j} = g_j^\top(\mathbf{X}_t - \bar{\mathbf{X}})$ is the *jth sample principal component at time t* (g_j is the j th column of G).
- The rotated vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ have sample covariance matrix L :

$$\begin{aligned} S_y &= \frac{1}{n} \sum_{t=1}^n (\mathbf{Y}_t - \bar{\mathbf{Y}})(\mathbf{Y}_t - \bar{\mathbf{Y}})^\top = \frac{1}{n} \sum_{t=1}^n \mathbf{Y}_t \mathbf{Y}_t^\top \\ &= \frac{1}{n} \sum_{t=1}^n G^\top(\mathbf{X}_t - \bar{\mathbf{X}})(\mathbf{X}_t - \bar{\mathbf{X}})^\top G = G^\top S_x G = L. \end{aligned}$$

Thus the rotated vectors show **no correlation between components** and the components are **ordered by their sample variances**, from largest to smallest.

- Now use G and Y_t to calibrate an approximate factor model. We assume our data are realisations from the model

$$\mathbf{X}_t = \bar{\mathbf{X}} + G_1 \mathbf{F}_t + \boldsymbol{\varepsilon}_t, \quad t \in \{1, \dots, n\},$$

where G_1 consists of the first k columns of G and $\mathbf{F}_t = (Y_{t,1}, \dots, Y_{t,k})$, $t \in \{1, \dots, n\}$.

- In practice, the errors $\boldsymbol{\varepsilon}_t$ do not have a diagonal covariance matrix and are not uncorrelated with \mathbf{F}_t . Nevertheless the method is a popular approach to constructing time series of statistically explanatory factors from multivariate time series of risk-factor changes.

7 Copulas and dependence

7.1 Copulas

7.2 Dependence concepts and measures

7.3 Normal mixture copulas

7.4 Archimedean copulas

7.5 Fitting copulas to data

7.1 Copulas

- We now look more closely at modeling the dependence among the components of a random vector $\mathbf{X} \sim H$ (risk-factor changes).
- In short: H “=” marginal dfs F_1, \dots, F_d “+” dependence structure C
- Advantages:
 - ▶ Most natural in a static distributional context (no time dependence; apply, e.g., to residuals of an ARMA-GARCH model)
 - ▶ Copulas allow us to understand and study dependence independently of the margins (first part of Sklar's Theorem; see later)
 - ▶ Copulas allow for a bottom-up approach to multivariate model building (second part of Sklar's Theorem; see later). This is often useful for constructing tailored H , e.g., when we have more information about the margins than C or for stress testing purposes.

7.1.1 Basic properties

Definition 7.1 (Copula)

A *copula* C is a df with $\text{U}[0, 1]$ margins.

Characterization

$C : [0, 1]^d \rightarrow [0, 1]$ is a copula if and only if

1) C is *grounded*, that is,

$$C(u_1, \dots, u_d) = 0 \text{ if } u_j = 0 \text{ for at least one } j \in \{1, \dots, d\}.$$

2) C has standard *uniform* univariate *margins*, that is,

$$C(1, \dots, 1, u_j, 1, \dots, 1) = u_j \text{ for all } u_j \in [0, 1] \text{ and } j \in \{1, \dots, d\}.$$

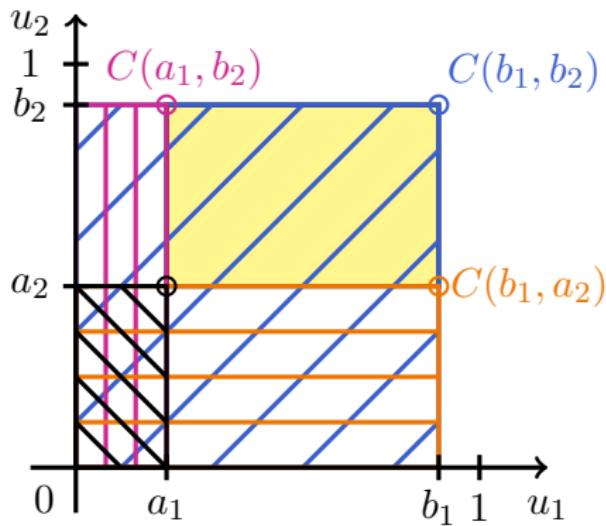
3) C is *d-increasing*, that is, for all $\mathbf{a}, \mathbf{b} \in [0, 1]^d$, $\mathbf{a} \leq \mathbf{b}$,

$$\Delta_{(\mathbf{a}, \mathbf{b})} C = \sum_{\mathbf{i} \in \{0, 1\}^d} (-1)^{\sum_{j=1}^d i_j} C(a_1^{i_1} b_1^{1-i_1}, \dots, a_d^{i_d} b_d^{1-i_d}) \geq 0.$$

Equivalently (if existent): density $c(\mathbf{u}) \geq 0$ for all $\mathbf{u} \in (0, 1)^d$.

2-increasingness explained in a picture:

$$\begin{aligned}\Delta_{(a,b]} C &= C(b_1, b_2) - C(b_1, a_2) - C(a_1, b_2) + C(a_1, a_2) \\ &= \mathbb{P}(U \in (a, b]) \stackrel{!}{\geq} 0\end{aligned}$$



$\Rightarrow \Delta_{(a,b]} C$ is the probability of a random vector $U \sim C$ to be in $(a, b]$.

Lemma 7.2

For all $\mathbf{a}, \mathbf{b} \in [0, 1]^d$,

$$|C(\mathbf{b}) - C(\mathbf{a})| \leq \sum_{j=1}^d |b_j - a_j| \quad (34)$$

In particular, copulas are uniformly equi-continuous.

Proof. Expanding $C(\mathbf{b}) - C(\mathbf{a})$ in a telescoping sum and using the triangle inequality leads to

$$\begin{aligned} |C(\mathbf{b}) - C(\mathbf{a})| &\leq \sum_{j=1}^d |C(b_1, \dots, b_{j-1}, b_j, a_{j+1}, \dots, a_d) \\ &\quad - C(b_1, \dots, b_{j-1}, a_j, a_{j+1}, \dots, a_d)| \end{aligned}$$

W.l.o.g. let $\mathbf{a} \leq \mathbf{b}$. By d -increasingness, $C \nearrow$ in each component, so omit $|\cdot|$. Since, again by d -increasingness, the j th summand is \nearrow in each component $\neq j$, let $b_1, \dots, b_{j-1}, a_{j+1}, \dots, a_d \nearrow 1$ to obtain the upper bound $\sum_{j=1}^d C(1, \dots, 1, b_j, 1, \dots, 1) - C(1, \dots, 1, a_j, 1, \dots, 1) = b_j - a_j$ for summand j . \square

A first “warm-up” example:

Let C_1, C_2 be copulas. Then $C(\mathbf{u}) = \lambda C_1(\mathbf{u}) + (1 - \lambda)C_2(\mathbf{u})$ is a copula for all $\lambda \in (0, 1)$, i.e., convex combinations of copulas are copulas.

Proof (analytic).

- 1) Let $\mathbf{u}_j = (u_1, \dots, u_{j-1}, 0, u_{j+1}, \dots, u_d)$. Then

$$C(\mathbf{u}_j) = \lambda C_1(\mathbf{u}_j) + (1 - \lambda)C_2(\mathbf{u}_j) = \lambda \cdot 0 + (1 - \lambda) \cdot 0 = 0$$

since C_1, C_2 are grounded. Hence, C is grounded.

- 2) Let $\mathbf{u}_j = (1, \dots, 1, u_j, 1, \dots, 1)$. Then

$$C(\mathbf{u}_j) = \lambda C_1(\mathbf{u}_j) + (1 - \lambda)C_2(\mathbf{u}_j) = \lambda u_j + (1 - \lambda)u_j = u_j$$

since C_1, C_2 have $U[0, 1]$ margins. Hence, C has $U[0, 1]$ margins.

- 3) $\Delta_{(a,b]} C = \lambda \Delta_{(a,b]} C_1 + (1 - \lambda) \Delta_{(a,b]} C_2 \geq 0$, so C is d -increasing. \square

Proof (probabilistic). Let $\mathbf{U}_k \sim C_k$, $k \in \{1, 2\}$ and let $X \sim \text{B}(1, \lambda)$, independent of each other. Furthermore, let

$$\mathbf{U} = \begin{cases} \mathbf{U}_1, & \text{if } X = 1, \\ \mathbf{U}_2, & \text{if } X = 0. \end{cases}$$

The Law of Total Probability implies

$$\begin{aligned}\mathbb{P}(\mathbf{U} \leq \mathbf{u}) &= \mathbb{P}(\mathbf{U} \leq \mathbf{u}, X = 1) + \mathbb{P}(\mathbf{U} \leq \mathbf{u}, X = 0) \\ &= \mathbb{P}(\mathbf{U}_1 \leq \mathbf{u}, X = 1) + \mathbb{P}(\mathbf{U}_2 \leq \mathbf{u}, X = 0) \\ &= \mathbb{P}(\mathbf{U}_1 \leq \mathbf{u}) \mathbb{P}(X = 1) + \mathbb{P}(\mathbf{U}_2 \leq \mathbf{u}) \mathbb{P}(X = 0) \\ &= C_1(\mathbf{u}) \lambda + C_2(\mathbf{u}) (1 - \lambda) = C(\mathbf{u}).\end{aligned}$$

So $\mathbf{U} \sim C$ and hence C is a df. From the same calculation it follows that \mathbf{U} has uniform margins, hence C is a copula. □

Preliminaries

$T \nearrow$ means that T is *increasing*; $T \uparrow$ means that T is *strictly increasing*;
 $\text{ran } T = \{T(x) : x \in \mathbb{R}\}$ denotes the *range* of T .

Proposition 7.3 (Working with generalized inverses)

Let $T : \mathbb{R} \rightarrow \mathbb{R} \nearrow$ with $T(-\infty) = \lim_{x \downarrow -\infty} T(x)$ and $T(\infty) = \lim_{x \uparrow \infty} T(x)$ and let $x, y \in \mathbb{R}$. Then,

- (GI1) $T^-(y) = -\infty$ if and only if $T(x) \geq y$ for all $x \in \mathbb{R}$. Similarly,
 $T^-(y) = \infty$ if and only if $T(x) < y$ for all $x \in \mathbb{R}$.
- (GI2) $T^- \nearrow$. If $T^-(y) \in (-\infty, \infty)$, T^- is left-continuous at y and admits a limit from the right at y .
- (GI3) $T^-(T(x)) \leq x$. If $T \uparrow$, then $T^-(T(x)) = x$.
- (GI4) Let T be right-continuous. $T^-(y) < \infty$ implies $T(T^-(y)) \geq y$. Furthermore, $y \in \text{ran } T \cup \{\inf T, \sup T\}$ implies $T(T^-(y)) = y$.

Moreover, if $y < \inf T$ then $T(T^-(y)) > y$ and if $y > \sup T$ then $T(T^-(y)) < y$.

- (GI5) $T(x) \geq y$ implies $x \geq T^-(y)$. The other implication holds if T is right-continuous. Furthermore, $T(x) < y$ implies $x \leq T^-(y)$.
- (GI6) $(T^-(y-), T^-(y+)) \subseteq \{x \in \mathbb{R} : T(x) = y\} \subseteq [T^-(y-), T^-(y+)]$, where $T^-(y-) = \lim_{z \uparrow y} T^-(z)$ and $T^-(y+) = \lim_{z \downarrow y} T^-(z)$.
- (GI7) T is continuous if and only if $T^- \uparrow$ on $[\inf T, \sup T]$.
 $T \uparrow$ if and only if T^- is continuous on $\text{ran } T$.
- (GI8) If T_1 and T_2 are right-continuous transformations with properties as T , then $(T_1 \circ T_2)^- = T_2^- \circ T_1^-$.

Proof. See Embrechts and Hofert (2013a). □

It is not important to know them, but rather to be aware of these “rules” when manipulating probabilities.

Lemma 7.4 (Probability transformation)

Let $X \sim F$, F continuous. Then $F(X) \sim U[0, 1]$.

Proof. Note that the *range* of a rv X is defined by

$$\text{ran } X = \{x \in \mathbb{R} : \mathbb{P}(X \in (x - h, x]) > 0 \text{ for all } h > 0\}.$$

Since F is continuous on \mathbb{R} , (GI7) implies that $F^- \uparrow$ on $[\inf F, \sup F] = [0, 1]$. Thus

$$\begin{aligned}\mathbb{P}(F(X) \leq u) &\stackrel{\text{(GI7)}}{=} \mathbb{P}(F^-(F(X)) \leq F^-(u)) \stackrel{\text{(GI3)}}{=} \mathbb{P}(X \leq F^-(u)) \\ &= F(F^-(u)) \stackrel{\text{(GI4)}}{=} u, \quad u \in [0, 1],\end{aligned}$$

where (GI3) applies since $F \uparrow$ on $\text{ran } X$. □

- Note that we need F to be **continuous** (otherwise $F(X)$ would not reach all intervals $\subseteq [0, 1]$).

Lemma 7.5 (Quantile transformation)

Let $U \sim U[0, 1]$ and F be any df. Then $X = F^-(U) \sim F$.

Proof. $\mathbb{P}(F^-(U) \leq x) \stackrel{(GI5)}{=} \mathbb{P}(U \leq F(x)) = F(x), \quad x \in \mathbb{R}$. □

- Probability and quantile transformations are the key to all applications involving copulas, e.g., goodness-of-fit testing (probability transformation) or simulation (quantile transformation). They will allow us to go from \mathbb{R}^d to $[0, 1]^d$ and back, respectively.
- Both transformations have multivariate equivalents (but then involving the underlying dependence structure as well!).

Sklar's Theorem

Theorem 7.6 (Sklar's Theorem)

- 1) For any df H with margins F_1, \dots, F_d , there exists a copula C such that

$$H(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)), \quad \mathbf{x} \in \mathbb{R}^d. \quad (35)$$

C is uniquely defined on $\prod_{j=1}^d \text{ran } F_j$ and given by

$$C(u_1, \dots, u_d) = H(F_1^-(u_1), \dots, F_d^-(u_d)), \quad \mathbf{u} \in \prod_{j=1}^d \text{ran } F_j.$$

- 2) Conversely, given any copula C and univariate dfs F_1, \dots, F_d , H defined by (35) is a df with margins F_1, \dots, F_d .

Proof.

- 1) **Proof for continuous F_1, \dots, F_d only.** Let $\mathbf{X} \sim H$ and define $U_j = F_j(X_j)$, $j \in \{1, \dots, d\}$. By the probability transformation, $U_j \sim U[0, 1]$ (continuity!), $j \in \{1, \dots, d\}$, so the df C of \mathbf{U} is a copula. Since $F_j \uparrow$ on $\text{ran } X_j$, (GI3) implies that $X_j = F_j^-(F_j(X_j)) \stackrel{\text{a.s.}}{=} F_j^-(U_j)$, $j \in \{1, \dots, d\}$. Therefore,

$$\begin{aligned} H(\mathbf{x}) &= \mathbb{P}(X_j \leq x_j \ \forall j) = \mathbb{P}(F_j^-(U_j) \leq x_j \ \forall j) \stackrel{(\text{GI5})}{=} \mathbb{P}(U_j \leq F_j(x_j) \ \forall j) \\ &= C(F_1(x_1), \dots, F_d(x_d)), \quad \mathbf{x} \in \mathbb{R}^d. \end{aligned}$$

Hence C is a copula and satisfies (35).

(GI4) implies that $F_j(F_j^-(u_j)) = u_j$ for all $u_j \in \text{ran } F_j$, so

$$\begin{aligned} C(u_1, \dots, u_d) &= C(F_1(F_1^-(u_1)), \dots, F_d(F_d^-(u_d))) \\ &\stackrel{(35)}{=} H(F_1^-(u_1), \dots, F_d^-(u_d)), \quad \mathbf{u} \in \prod_{j=1}^d \text{ran } F_j. \end{aligned}$$

2) For $\mathbf{U} \sim C$, define $\mathbf{X} = (F_1^-(U_1), \dots, F_d^-(U_d))$. Then

$$\begin{aligned}\mathbb{P}(\mathbf{X} \leq \mathbf{x}) &= \mathbb{P}(F_j^-(U_j) \leq x_j \ \forall j) \stackrel{\text{(GI5)}}{=} \mathbb{P}(U_j \leq F_j(x_j) \ \forall j) \\ &= C(F_1(x_1), \dots, F_d(x_d)), \quad \mathbf{x} \in \mathbb{R}^d.\end{aligned}$$

Therefore, H defined by (35) is a df (that of \mathbf{X}), with (by the quantile transformation) margins F_1, \dots, F_d . \square

Remark 7.7

- The general proof of Part 1) of Sklar's Theorem can be found, e.g., in Rüschedorf (2009). It works similarly but utilizes the *generalized distributional transform* $\mathbf{U} = (U_1, \dots, U_d) = (F_1(X_1, V_1), \dots, F_d(X_d, V_d))$, where $V_1, \dots, V_d \sim \text{U}[0, 1]$ are independent of \mathbf{X} and

$$F_j(x, v) = \mathbb{P}(X < x) + v\mathbb{P}(X = x) = F(x-) + v(F(x) - F(x-)).$$

One can show: $\mathbf{U} \sim \text{U}[0, 1]^d$; $(F_1^-(U_1), \dots, F_d^-(U_d)) \stackrel{\text{a.s.}}{=} \mathbf{X}$.

- Non-uniqueness follows from different choices of \mathbf{V} , e.g., independent components or $\mathbf{V} = (V, \dots, V)$. See also the following example.

Example 7.8 (Bivariate Bernoulli distribution)

Let (X_1, X_2) follow a bivariate Bernoulli distribution with $\mathbb{P}(X_1 = k, X_2 = l) = 1/4$, $k, l \in \{0, 1\}$. $\Rightarrow \mathbb{P}(X_j = k) = 1/2$, $k \in \{0, 1\}$, $\text{ran } F_j = \{0, 1/2, 1\}$, $j \in \{1, 2\}$. Any copula with $C(1/2, 1/2) = 1/4$ satisfies (35) (e.g., $C(u_1, u_2) = \Pi(u_1, u_2)$ or the diagonal copula $C(u_1, u_2) = \min\{u_1, u_2, (\delta(u_1) + \delta(u_2))/2\}$ with $\delta(u) = u^2$). See Genest and Nešlehová (2007) for more details on copulas for discrete data.

Interpretation of Sklar's Theorem

- 1) Allows one to decompose any df H into its margins and a copula.
This allows one to study multivariate distributions independently of the margins (interesting for statistical applications, e.g., parameter estimation or goodness-of-fit).
- 2) Allows one to compose new multivariate distributions (interesting for constructing flexible models, sampling, stress testing).

- A copula model for \mathbf{X} means $H(\mathbf{x}) = C(F_1(x_1), \dots, F_d(x_d))$ for some (parametric) copula C and (parametric) marginals F_1, \dots, F_d .
- \mathbf{X} (or H) with margins F_1, \dots, F_d has copula C if (35) holds.

Invariance principle

Lemma 7.9 (Core of the invariance principle)

Let $X_j \sim F_j$, F_j continuous, $j \in \{1, \dots, d\}$. Then

$$\mathbf{X} \sim H \text{ has copula } C \iff (F_1(X_1), \dots, F_d(X_d)) \sim C.$$

Proof.

$$\begin{aligned} \Rightarrow & \mathbb{P}(F_j(X_j) \leq u_j \forall j) \underset{\text{cont.}}{=} \mathbb{P}(F_j(X_j) < u_j \forall j) \underset{(GI5)}{=} \mathbb{P}(X_j < F_j^-(u_j) \forall j) \\ & \underset{\text{cont.}}{=} \mathbb{P}(X_j \leq F_j^-(u_j) \forall j) = H(F_1^-(u_1), \dots, F_d^-(u_d)) \underset{\text{Sklar}}{=} C(\mathbf{u}). \end{aligned}$$

“ \Leftarrow ” Since $F_j \uparrow$ on $\text{ran } X_j$, $j \in \{1, \dots, d\}$,

$$\begin{aligned} H(\mathbf{x}) &\stackrel{\text{(GI3)}}{=} \mathbb{P}(F_j^-(F_j(X_j)) \leq x_j \forall j) \stackrel{\text{(GI5)}}{=} \mathbb{P}(F_j(X_j) \leq F_j(x_j) \forall j) \\ &= C(F_1(x_1), \dots, F_d(x_d)) \quad \underset{\text{ass.}}{\Rightarrow} \quad \mathbf{X} \text{ has copula } C \end{aligned}$$

□

Theorem 7.10 (Invariance principle)

Let $\mathbf{X} \sim H$ with continuous margins F_1, \dots, F_d and copula C . If $T_j \uparrow$ on $\text{ran } X_j$ for all j , then $(T_1(X_1), \dots, T_d(X_d))$ (also) has copula C .

Proof. W.l.o.g. assume T_j to be right-continuous at its at most countably many discontinuities (since X_j is continuously distributed, we only change $T_j(X_j)$ on a null set). Since $T_j \uparrow$ on $\text{ran } X_j$ and X_j is continuously distributed, $T_j(X_j)$ is continuously distributed and we have

$$\begin{aligned} F_{T_j(X_j)}(\mathbf{x}) &= \mathbb{P}(T_j(X_j) \leq x) = \mathbb{P}(T_j(X_j) < x) \stackrel{\text{(GI5)}}{=} \mathbb{P}(X_j < T_j^-(x)) \\ &= \mathbb{P}(X_j \leq T_j^-(x)) = F_j(T_j^-(x)), \quad x \in \mathbb{R}. \end{aligned}$$

This implies that

$$\begin{aligned}\mathbb{P}(F_{T_j(X_j)}(T_j(X_j)) \leq u_j \forall j) &= \mathbb{P}(F_j(T_j^-(T_j(X_j))) \leq u_j \forall j) \\ &\stackrel{(GI3)}{=} \mathbb{P}(F_j(X_j) \leq u_j \forall j) \stackrel{\text{L.7.9}}{\underset{\text{"only if"}}{=}} C(\mathbf{u}).\end{aligned}$$

The claim follows by the if part (" \Leftarrow ") of Lemma 7.9. \square

- The invariance principle allows us to study dependence in terms of the margin-free $\mathbf{U} = (F_1(X_1), \dots, F_d(X_d))$ instead of $\mathbf{X} = (X_1, \dots, X_d)$, \mathbf{X}, \mathbf{U} have the same copula!

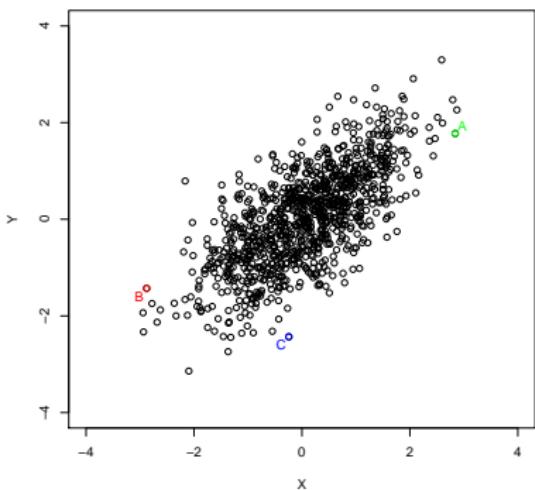
How does $(F_1(\cdot), \dots, F_d(\cdot))$ act on \mathbf{X} ? A picture is worth a thousand words...

Visualizing the first part of Sklar's Theorem

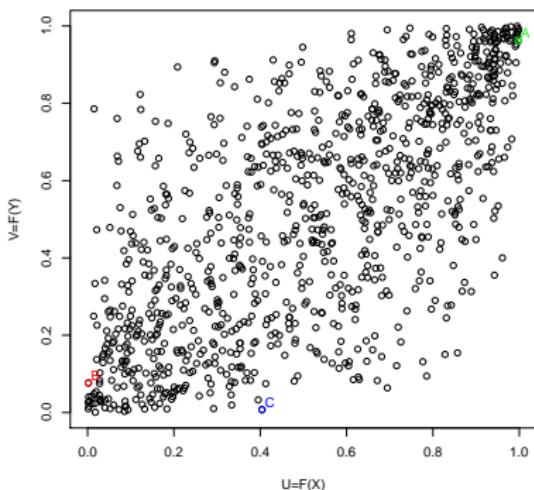
Left: Scatter plot of $n = 1000$ samples from $(X_1, X_2) \sim N_2(\mathbf{0}, P)$, where $P = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$. We mark three points A, B, C.

Right: After applying the F_j 's (the df Φ of $N(0, 1)$), scatter plot of the corresponding Gauss copula. Note how the points A, B, C change.

1000 realizations of (X,Y) for a joint normal distribution with rho = 0.7



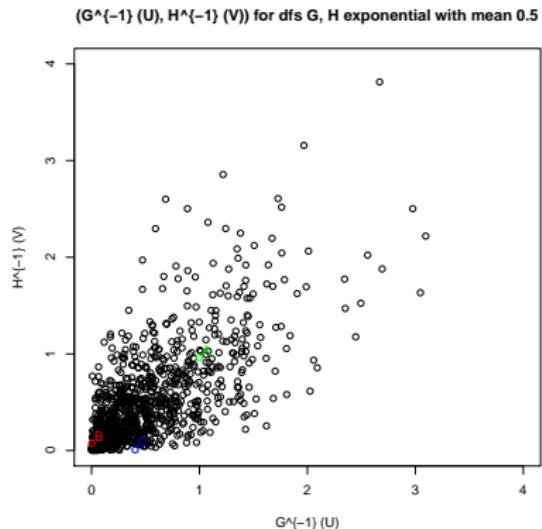
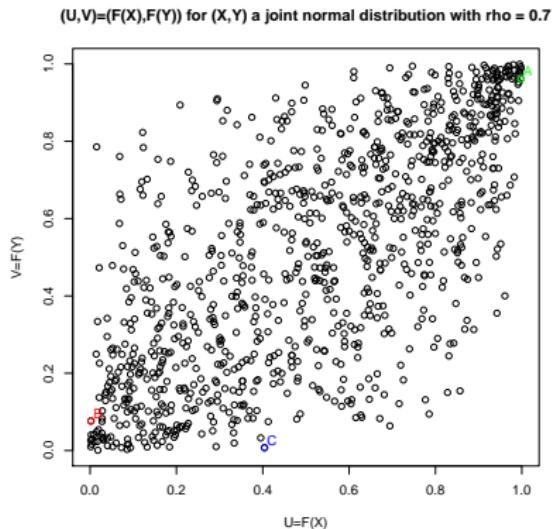
(U,V)=(F(X),F(Y)) for (X,Y) a joint normal distribution with rho = 0.7



Visualizing the second part of Sklar's Theorem

Left: Same Gauss copula scatter plot as before. Apply marginal $\text{Exp}(2)$ -quantile functions ($F_j^{-1}(u) = -\log(1-u)/2$, $j \in \{1, 2\}$).

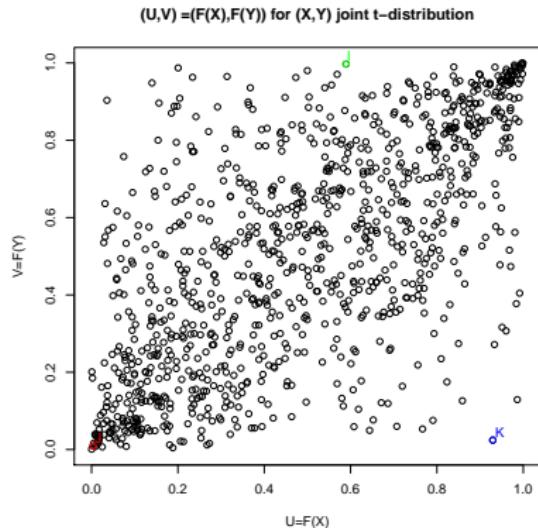
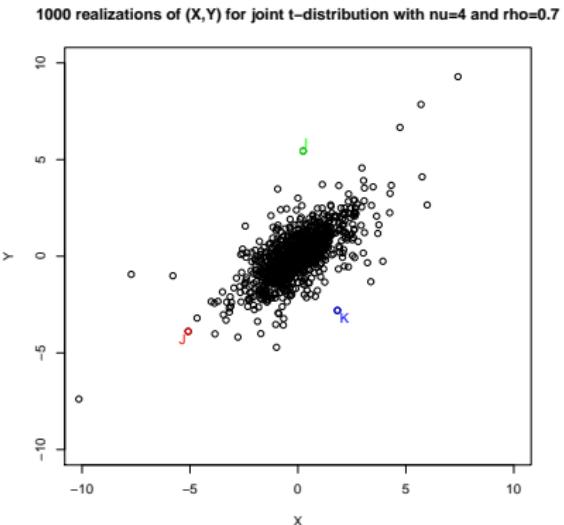
Right: The corresponding transformed random variates. Again, note the three points A, B, C.



Visualizing the first part of Sklar's Theorem

Left: Scatter plot of $n = 1000$ samples from $(X_1, X_2) \sim t_2(4, \mathbf{0}, P)$, where $P = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$. We mark three points I, J, K.

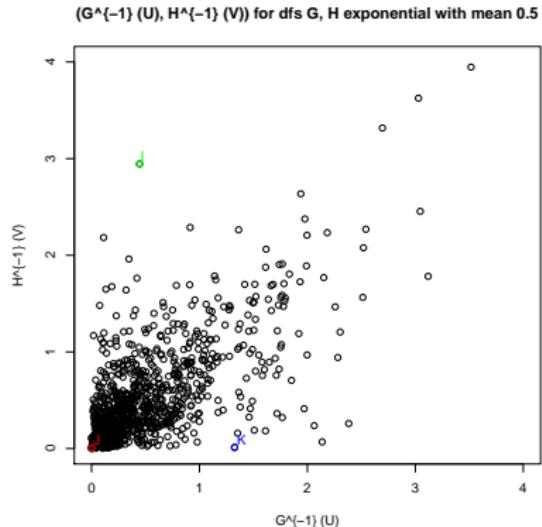
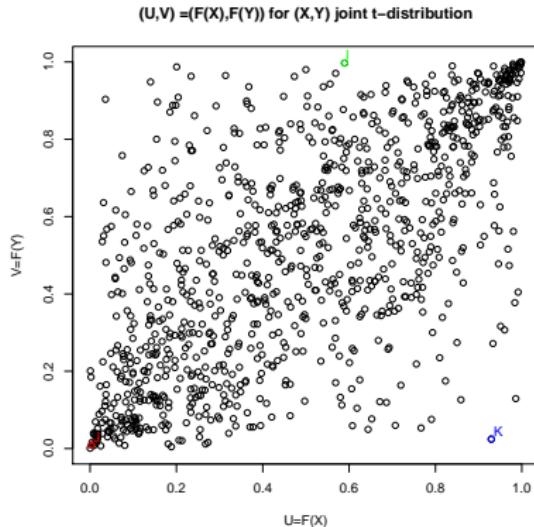
Right: After applying the F_j 's (the df of t_4), scatter plot of the corresponding t_4 copula. Note how the points I, J, K change.



Visualizing the second part of Sklar's Theorem

Left: Same t_4 copula scatter plot as before. Apply marginal $\text{Exp}(2)$ -quantile functions ($F_j^{-1}(u) = -\log(1-u)/2$, $j \in \{1, 2\}$).

Right: The corresponding transformed random variates. Again, note the three points I, J, K.



Fréchet–Hoeffding bounds

Theorem 7.11 (Fréchet–Hoeffding bounds)

Let $W(\mathbf{u}) = \max\{\sum_{j=1}^d u_j - d + 1, 0\}$ and $M(\mathbf{u}) = \min_{1 \leq j \leq d}\{u_j\}$.

1) For any d -dimensional copula C ,

$$W(\mathbf{u}) \leq C(\mathbf{u}) \leq M(\mathbf{u}), \quad \mathbf{u} \in [0, 1]^d.$$

2) W is a copula if and only if $d = 2$.

3) M is a copula for all $d \geq 2$.

Proof.

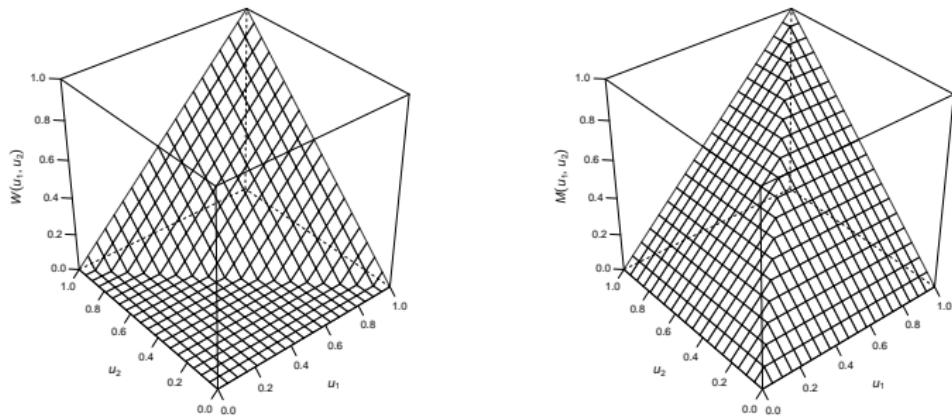
- 1) ■ By (34), $1 - C(\mathbf{u}) = C(\mathbf{1}) - C(\mathbf{u}) \leq \sum_{j=1}^d (1 - u_j) = d - \sum_{j=1}^d u_j$, so $C(\mathbf{u}) \geq \sum_{j=1}^d u_j - d + 1$. Also, $C(\mathbf{u}) \geq 0$. So $C(\mathbf{u}) \geq W(\mathbf{u})$.
- Since copulas are componentwise increasing, $C(\mathbf{u}) \leq C(1, \dots, 1, u_j, 1, \dots, 1) = u_j$ for all j . Hence, $C(\mathbf{u}) \leq \min_{1 \leq j \leq d}\{u_j\} = M(\mathbf{u})$.

2) W is a copula for $d = 2$ since $(U, 1 - U) \sim W$ for $U \sim \text{U}[0, 1]$. W is not a copula for $d \geq 3$ since

$$\begin{aligned}
& \Delta_{(\frac{1}{2}, 1]} W \\
&= \sum_{\boldsymbol{i} \in \{0,1\}^d} (-1)^{\sum_{j=1}^d i_j} W\left(\frac{1}{2}^{i_1}, \dots, \frac{1}{2}^{i_d}\right) \\
&= \max\{1 + 1 + 1 + \dots + 1 - d + 1, 0\} \quad (i_j = 0 \ \forall j) \\
&\quad - d \max\{\frac{1}{2} + 1 + 1 + \dots + 1 - d + 1, 0\} \quad (\exists! j : i_j = 1) \\
&\quad + \binom{d}{2} \max\{\frac{1}{2} + \frac{1}{2} + 1 + \dots + 1 - d + 1, 0\} \quad (\exists! \text{ two } j : i_j = 1) \\
&\quad - \dots + (-1)^d \max\{\frac{1}{2} + \dots + \frac{1}{2} - d + 1, 0\} \quad (i_j = 1 \ \forall j) \\
&= 1 - \frac{d}{2} < 0 \quad \text{for } d \geq 3.
\end{aligned}$$

3) M is a copula for all $d \geq 2$ since $(U, \dots, U) \sim M$ for $U \sim \text{U}[0, 1]$. \square

- Plot of W, M for $d = 2$ (compare with $(U, 1 - U) \sim W, (U, U) \sim M$)



- The Fréchet–Hoeffding bounds correspond to perfect dependence (negative for W ; positive for M); see Proposition 7.18 later.
- The Fréchet–Hoeffding bounds lead to bounds for any df H , via

$$\max\left\{ \sum_{j=1}^d F_j(x_j) - d + 1, 0 \right\} \leq H(\mathbf{x}) \leq \min_{1 \leq j \leq d} \{F_j(x_j)\}.$$

We will use them later to derive bounds for the correlation coefficient.

7.1.2 Examples of copulas

- *Fundamental copulas*: important special copulas;
- *Implicit copulas*: extracted from known H via Sklar's Theorem;
- *Explicit copulas*: have simple closed-form expressions and follow construction principles of copulas.

Fundamental copulas

- $\Pi(\mathbf{u}) = \prod_{j=1}^d u_j$ is the *independence copula* since $C(F_1(x_1), \dots, F_d(x_d)) = H(\mathbf{x}) = \prod_{j=1}^d F_j(x_j)$ if and only if $C(\mathbf{u}) = \Pi(\mathbf{u})$ (now replace x_j by $F_j^{-1}(u_j)$ and apply (GI4)). Therefore, X_1, \dots, X_d are independent if and only if their copula is Π .
- The Fréchet–Hoeffding bound W is the *countermonotonicity copula*. It is the df of $(U, 1 - U)$. If X_1, X_2 are perfectly negatively dependent (X_2 is a.s. a strictly decreasing function in X_1), their copula is W .

- The Fréchet–Hoeffding bound M is the *comonotonicity copula*. It is the df of (U, \dots, U) . If X_1, \dots, X_d are perfectly positively dependent (X_2, \dots, X_{d-1} are a.s. strictly increasing functions in X_1), their copula is M .

Implicit copulas

Elliptical copulas are implicit copulas arising from elliptical distributions via Sklar's Theorem. The two most prominent parametric families in this class are the *Gauss copula* and the *t copula*.

Gauss copulas

- Consider (w.l.o.g.) $\mathbf{X} \sim N_d(\mathbf{0}, P)$. The *Gauss copula* (family) is given by

$$\begin{aligned} C_P^{\text{Ga}}(\mathbf{u}) &= \mathbb{P}(\Phi(X_1) \leq u_1, \dots, \Phi(X_d) \leq u_d) \\ &= \Phi_P(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)) \end{aligned}$$

where Φ_P is the df of $N_d(\mathbf{0}, P)$ and Φ the df of $N(0, 1)$.

- $P = I_d \Rightarrow \mathcal{C} = \Pi$; and $P = J_d = \mathbf{1}\mathbf{1}^\top \Rightarrow \mathcal{C} = M$;
- $d = 2$ and $\rho = P_{12} = -1 \Rightarrow \mathcal{C} = W$.
- For $d > 3$, C_P^{Ga} is evaluated by (randomized quasi-)Monte Carlo.
- Sklar's Theorem \Rightarrow The density of $C(\mathbf{u}) = H(F_1^-(u_1), \dots, F_d^-(u_d))$ is

$$c(\mathbf{u}) = \frac{h(F_1^-(u_1), \dots, F_d^-(u_d))}{\prod_{j=1}^d f_j(F_j^-(u_j))}, \quad \mathbf{u} \in (0, 1)^d.$$

In particular, the density of C_P^{Ga} is

$$c_P^{\text{Ga}}(\mathbf{u}) = \frac{1}{\sqrt{\det P}} \exp\left(-\frac{1}{2} \mathbf{x}^\top (P^{-1} - I_d) \mathbf{x}\right), \quad (36)$$

where $\mathbf{x} = (\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))$.

t copulas (“lonely island copula”)

- Consider (w.l.o.g.) $\mathbf{X} \sim t_d(\nu, \mathbf{0}, P)$. The *t copula* (family) is given by

$$\begin{aligned} \mathcal{C}_{\nu, P}^t(\mathbf{u}) &= \mathbb{P}(t_\nu(X_1) \leq u_1, \dots, t_\nu(X_d) \leq u_d) \\ &= t_{\nu, P}(t_\nu^{-1}(u_1), \dots, t_\nu^{-1}(u_d)) \end{aligned}$$

where $t_{\nu,P}$ is the df of $t_d(\nu, \mathbf{0}, P)$ and t_ν the df of the univariate t distribution with ν degrees of freedom.

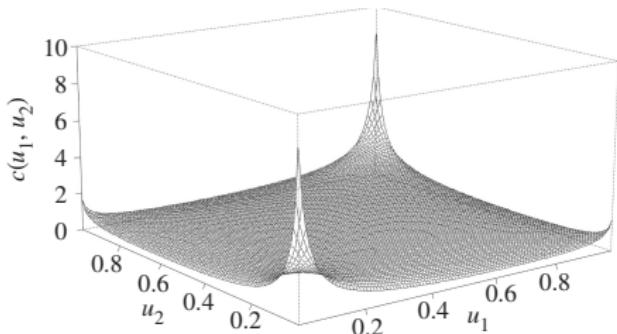
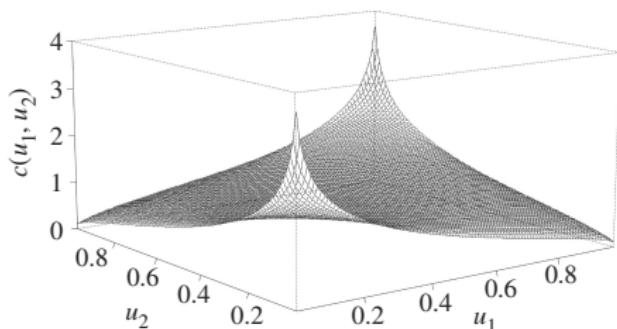
- $P = J_d = \mathbf{1}\mathbf{1}^\top \Rightarrow C = M$; and $d = 2$ and $\rho = P_{12} = -1 \Rightarrow C = W$. However, $P = I_d \Rightarrow C \neq \Pi$ (unless $\nu = \infty$ in which case $C_{\nu,P}^t = C_P^{\text{Ga}}$).
- For $d > 3$, $C_{\nu,P}^t$ is evaluated by (randomized quasi-)Monte Carlo.
- Sklar's Theorem \Rightarrow The density of $C_{\nu,P}^t$ is

$$c_{\nu,P}^t(\mathbf{u}) = \frac{\Gamma((\nu + d)/2)}{\Gamma(\nu/2)\sqrt{\det P}} \left(\frac{\Gamma(\nu/2)}{\Gamma((\nu + 1)/2)} \right)^d \frac{(1 + \mathbf{x}^\top P^{-1} \mathbf{x}/\nu)^{-(\nu+d)/2}}{\prod_{j=1}^d (1 + x_j^2/\nu)^{-(\nu+1)/2}},$$

for $\mathbf{x} = (t_\nu^{-1}(u_1), \dots, t_\nu^{-1}(u_d))$.

- For more details, see Demarta and McNeil (2005).
- For scatter plots, see the visualization of Sklar's Theorem above. Note the difference in the tails: The smaller ν , the more mass is concentrated in the joint tails.

Perspective plots of the densities of $C_{\rho=0.3}^{\text{Ga}}$ (left) and $C_{4,\rho=0.3}^t(\mathbf{u})$ (right).



Advantages and drawbacks of elliptical copulas (see later, too):

Advantages:

- Modeling pairwise dependencies (comparably flexible)
- Density available
- Sampling (typically) simple

Drawbacks:

- Typically, C is not explicit
- Radially symmetric, so same lower/upper tail behavior ($\lambda_L = \lambda_U$)

Explicit copulas

Archimedean copulas are copulas of the form

$$C(\mathbf{u}) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)), \quad \mathbf{u} \in [0, 1]^d,$$

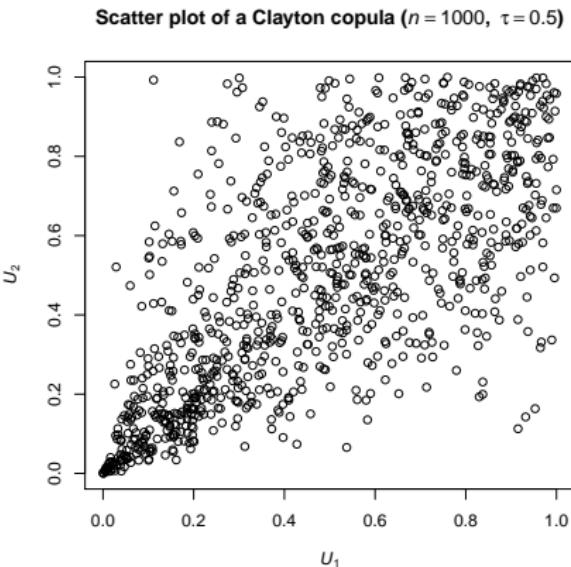
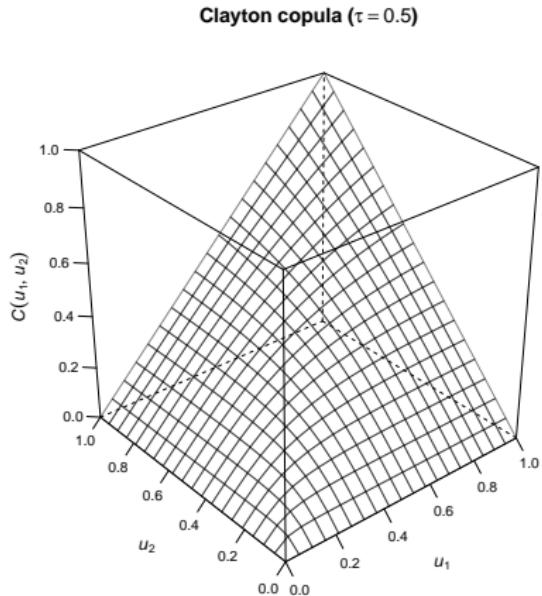
where the (*Archimedean*) generator $\psi : [0, \infty) \rightarrow [0, 1]$ is \downarrow on $[0, \inf\{t : \psi(t) = 0\}]$ and satisfies $\psi(0) = 1$, $\psi(\infty) = \lim_{t \rightarrow \infty} \psi(t) = 0$; we set $\psi^{-1}(0) = \inf\{t : \psi(t) = 0\}$. The set of all generators is denoted by Ψ . If $\psi(t) > 0$, $t \in [0, \infty)$, we call ψ *strict*.

Examples

- **Clayton copula:** Obtained for $\psi(t) = (1+t)^{-1/\theta}$, $t \in [0, \infty)$, $\theta \in (0, \infty)$
 $\Rightarrow C_\theta^c(\mathbf{u}) = (u_1^{-\theta} + \cdots + u_d^{-\theta} - d + 1)^{-1/\theta}$. For $\theta \downarrow 0$, $C \rightarrow \Pi$; and for $\theta \uparrow \infty$, $C \rightarrow M$.
- **Gumbel copula:** Obtained for $\psi(t) = \exp(-t^{1/\theta})$, $t \in [0, \infty)$, $\theta \in [1, \infty)$
 $\Rightarrow C_\theta^G(\mathbf{u}) = \exp(-((- \log u_1)^\theta + \cdots + (- \log u_d)^\theta)^{1/\theta})$. For $\theta = 1$, $C = \Pi$; and for $\theta \rightarrow \infty$, $C \rightarrow M$.

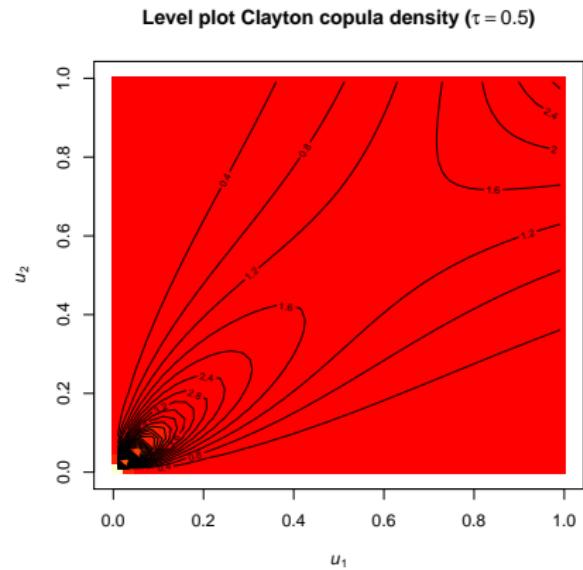
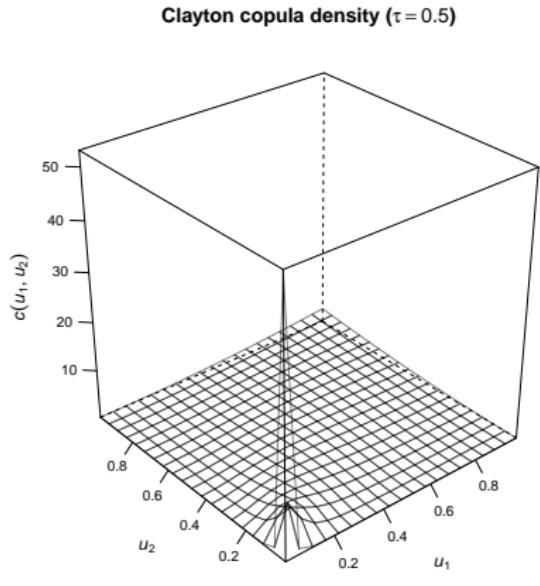
Left: Plot of a bivariate Clayton copula (Kendall's tau 0.5; see later).

Right: Corresponding scatter plot (sample size $n = 1000$)



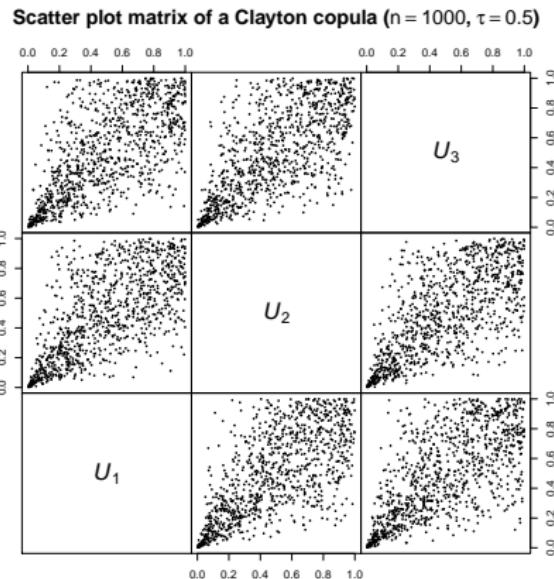
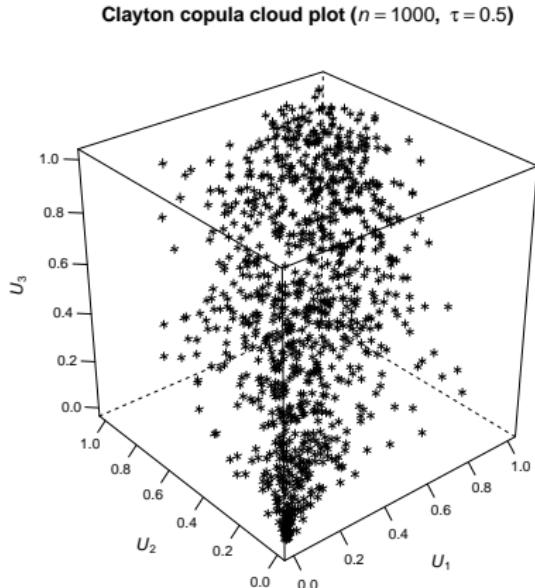
Left: Plot of the corresponding density.

Right: Level plot of the density (with heat colors).



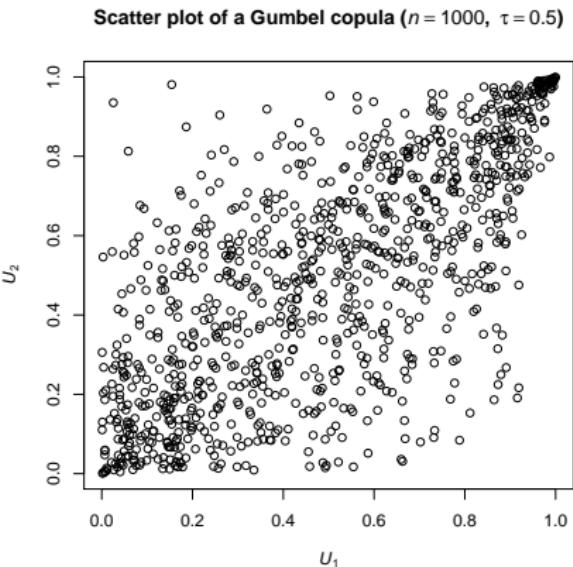
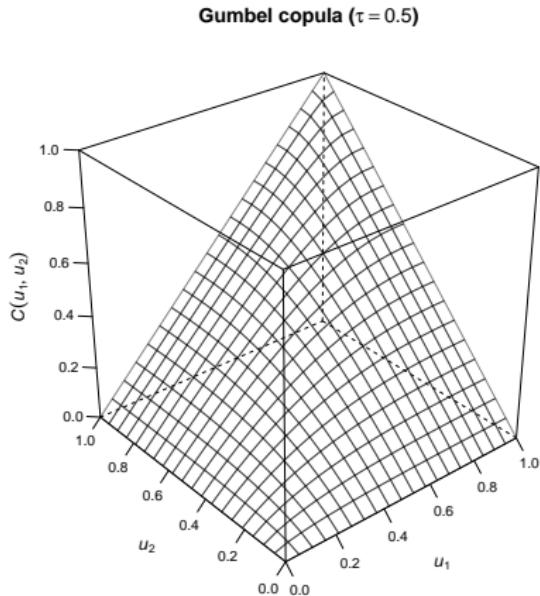
Left: Cloud plot of a trivariate Clayton copula (sample size $n = 1000$; Kendall's tau 0.5).

Right: Corresponding scatter plot matrix.



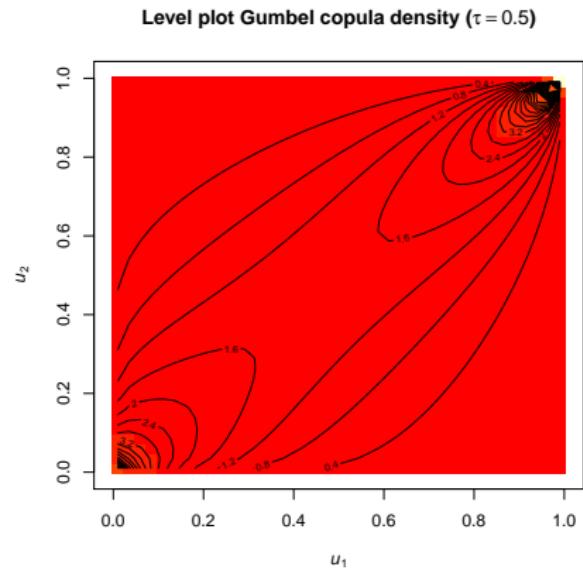
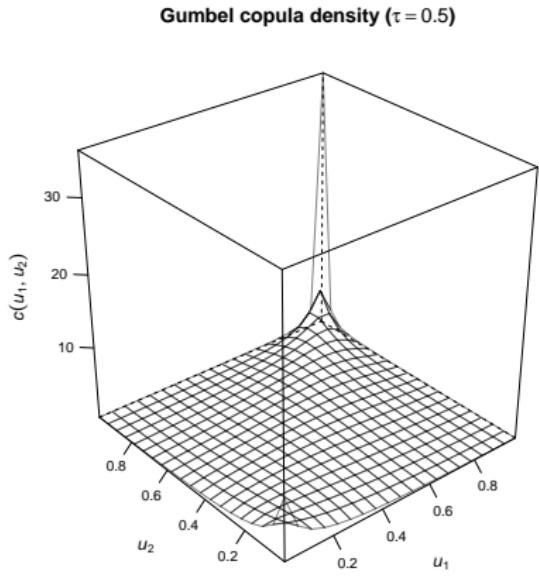
Left: Plot of a bivariate Gumbel copula (Kendall's tau 0.5).

Right: Corresponding scatter plot (sample size $n = 1000$)



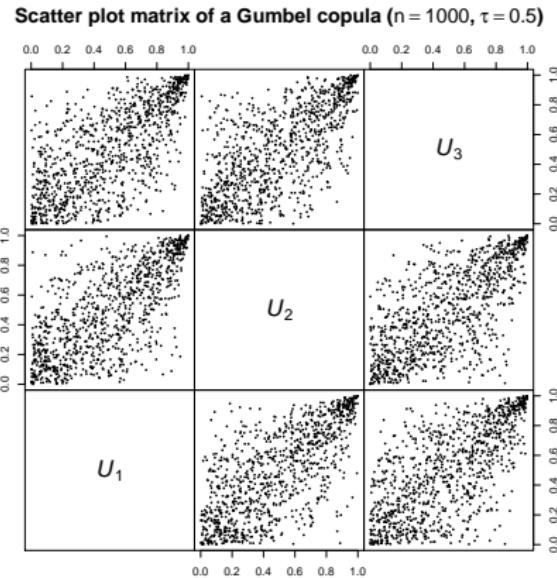
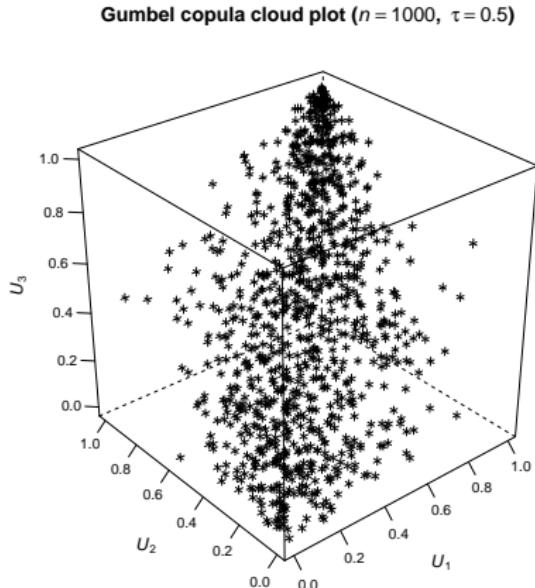
Left: Plot of the corresponding density.

Right: Level plot of the density (with heat colors).



Left: Cloud plot of a trivariate Gumbel copula (sample size $n = 1000$; Kendall's tau 0.5).

Right: Corresponding scatter plot matrix.



Advantages and drawbacks of Archimedean copulas (see later, too):

Advantages:

- Typically explicit
(if ψ^{-1} is available)
- Useful in calculations:
Properties can typically be expressed in terms of ψ
- Densities of various examples available
- Sampling often simple
- Not restricted to radial symmetry
($\lambda_L \neq \lambda_U$ possible)

Drawbacks:

- Exchangeable, i.e., (functionally) symmetric (all pairs have the same dependence)
- Often used only with a small number of parameters (some extensions available, but still less than $d(d - 1)/2$)

Excursion: Other animals in the zoo (of copulas)

Extreme value copulas

- *Extreme value copulas* are the copulas C of limiting distributions of properly location-scale transformed componentwise maxima of a sequence of random vectors.
- They are given by

$$C(\mathbf{u}) = \left(\prod_{j=1}^d u_j \right)^{A\left(\frac{\log u_1}{\log \Pi(\mathbf{u})}, \dots, \frac{\log u_d}{\log \Pi(\mathbf{u})}\right)}$$

for a *Pickands dependence function* A ; see Ressel (2013) for a characterization of A .

- Examples: Gumbel copula, Marshall-Olkin copulas.
- For more details, see Jaworski et al. (2010, Chapter 6).

Marshall–Olkin copulas

- Here we focus on $d = 2$ only.
- **Motivation:** Consider a system of two components affected by three types of fatal shocks: One hitting the first, one the second, and one both components simultaneously. Assuming the shocks to follow independent homogeneous Poisson processes, the times of occurrence are given by $T_j \sim \text{Exp}(\lambda_j)$, $j \in \{1, 2\}$, $T_{12} \sim \text{Exp}(\lambda_{12})$ (all independent). The lifetimes of the components are thus

$$X_j = \min\{T_j, T_{12}\}, \quad j \in \{1, 2\}.$$

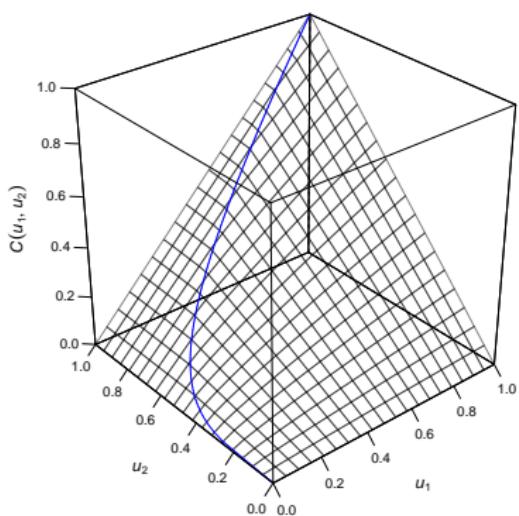
- The survival copula \hat{C} (see later) of \mathbf{X} is known as *Marshall–Olkin copula*. Derivation:

$$\begin{aligned}\bar{H}(x_1, x_2) &= \mathbb{P}(X_1 > x_1, X_2 > x_2) \\ &= \mathbb{P}(T_1 > x_1, T_2 > x_2, T_{12} > \max\{x_1, x_2\}) \\ &= \exp(-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_{12} \max\{x_1, x_2\}),\end{aligned}$$

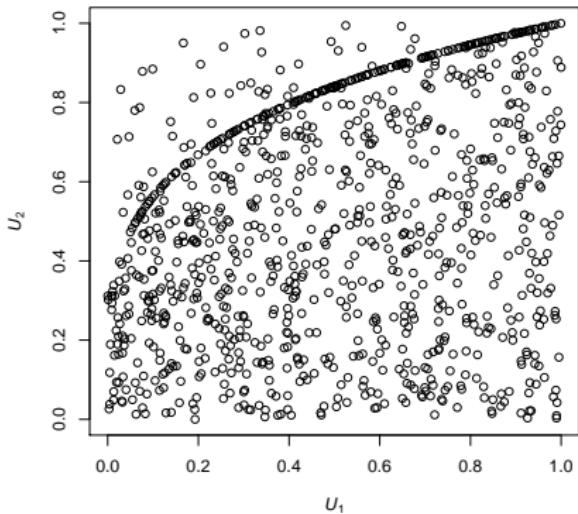
from which we obtain that $\bar{F}_j(x) = \exp(-(\lambda_j + \lambda_{12})x)$, $j \in \{1, 2\}$. With $\alpha_j = \frac{\lambda_{12}}{\lambda_j + \lambda_{12}}$, $j \in \{1, 2\}$, it follows that

$$\begin{aligned}\hat{C}(u_1, u_2) &\stackrel{\text{Sklar}}{=} \bar{H}(\bar{F}_1^-(u_1), \bar{F}_2^-(u_2)) = u_1^{1-\alpha_1} u_2^{1-\alpha_2} \min\{u_1^{\alpha_1}, u_2^{\alpha_2}\} \\ &= \min\{u_1 u_2^{1-\alpha_2}, u_1^{1-\alpha_1} u_2\}, \quad \alpha_1, \alpha_2 \in [0, 1].\end{aligned}$$

MO copula with singular component ($\alpha_1 = 0.2$, $\alpha_2 = 0.8$, $\tau = 0.19$)



Scatter plot MO copula ($n = 1000$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, $\tau = 0.19$)



- Any copula C can be decomposed into

$$C(\mathbf{u}) = A_C(\mathbf{u}) + S_C(\mathbf{u}) \quad (\text{Lebesgue Decomposition}),$$

where

$$A_C(\mathbf{u}) = \int_{(\mathbf{0}, \mathbf{u}]} D_{d, \dots, 1} C(\mathbf{v}) d\mathbf{v}$$

is the *absolutely continuous component* of C and

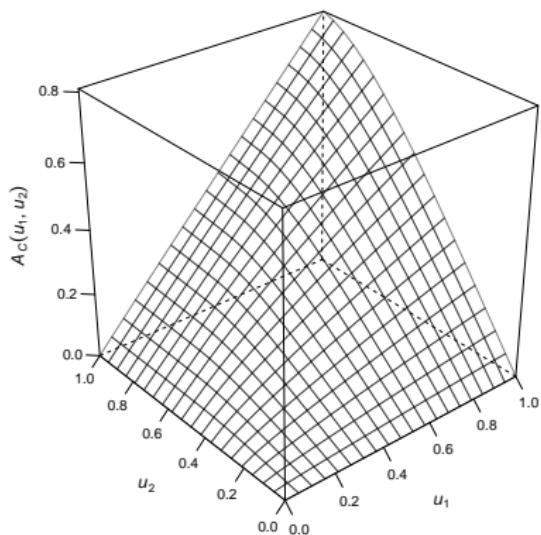
$$S_C(\mathbf{u}) = C(\mathbf{u}) - A_C(\mathbf{u})$$

is the *singular component* of C .

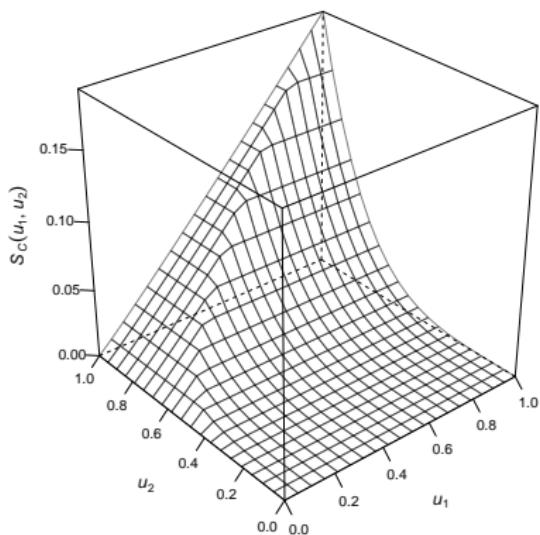
- For Marshall–Olkin copulas, integrating $D_{21}C(u_1, u_2)$ (exists for $u_1^{\alpha_1} \neq u_2^{\alpha_2}$) yields A_C and thus S_C . The mass on the singular component is $\mathbb{P}(U_1^{\alpha_1} = U_2^{\alpha_2}) = S_C(1, 1) = \frac{\alpha_1 \alpha_2}{\alpha_1 + \alpha_2 - \alpha_1 \alpha_2}$.

Absolutely continuous and singular components of a Marshall–Olkin copula:

Abs. cont. comp. A_C of a MO copula ($\alpha_1 = 0.2, \alpha_2 = 0.8, \tau = 0.19$)

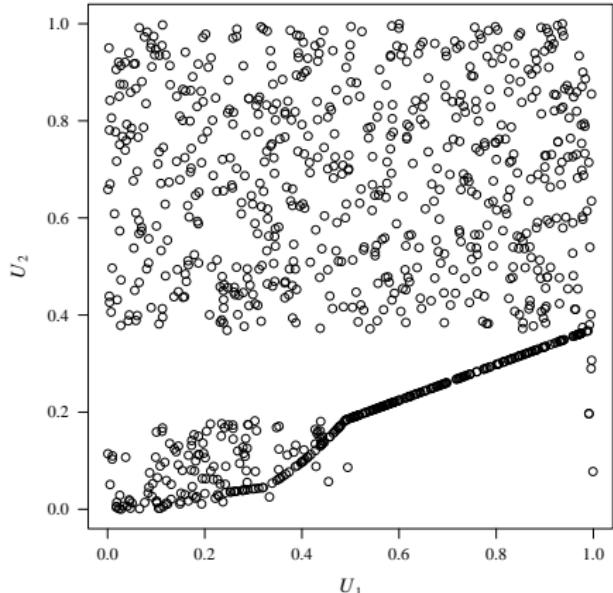
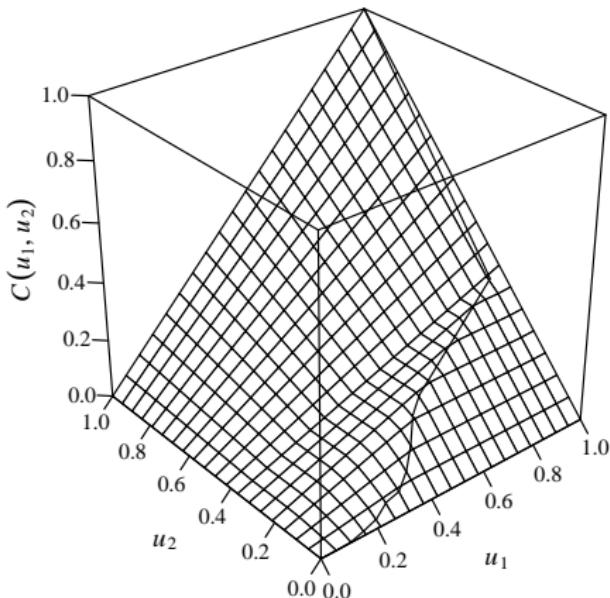


Singular comp. S_C of a MO copula ($\alpha_1 = 0.2, \alpha_2 = 0.8, \tau = 0.19$)



A rather exotic example (with singular component)

A Sibuya copula; see Hofert and Vrins (2013).



7.1.3 Meta distributions

- *Fréchet class*: Class of all dfs H with given marginal dfs F_1, \dots, F_d ;
- *Meta- C models*: All dfs H with the same given copula C .
- **Example:** A meta-Gauss model is a multivariate df H with Gauss copula C and some margins F_1, \dots, F_d . Such a model, with exponential margins, is used in Li's model in credit risk.

7.1.4 Simulation of copulas and meta distributions

Sampling implicit copulas

Due to their construction via Sklar's Theorem, implicit copulas can be sampled via Lemma 7.9.

Algorithm 7.12 (Simulation of implicit copulas)

- 1) Sample $\mathbf{X} \sim H$, where H is a df with continuous margins F_1, \dots, F_d .
- 2) Return $\mathbf{U} = (F_1(X_1), \dots, F_d(X_d))$ (**probability transformation**).

Example 7.13

- Sampling **Gauss copulas** C_P^{Ga} :

- 1) Sample $\mathbf{X} \sim \text{N}_d(\mathbf{0}, P)$ ($\mathbf{X} \stackrel{d}{=} A\mathbf{Z}$ for $AA^\top = P$, $\mathbf{Z} \sim \text{N}_d(\mathbf{0}, I_d)$).
- 2) Return $\mathbf{U} = (\Phi(X_1), \dots, \Phi(X_d))$.

- Sampling **t_ν copulas** $C_{\nu, P}^t$:

- 1) Sample $\mathbf{X} \sim t_d(\nu, \mathbf{0}, P)$ ($\mathbf{X} \stackrel{d}{=} \sqrt{W}A\mathbf{Z}$ for $W = \frac{1}{V}$, $V \sim \Gamma(\frac{\nu}{2}, \frac{\nu}{2})$).
- 2) Return $\mathbf{U} = (t_\nu(X_1), \dots, t_\nu(X_d))$.

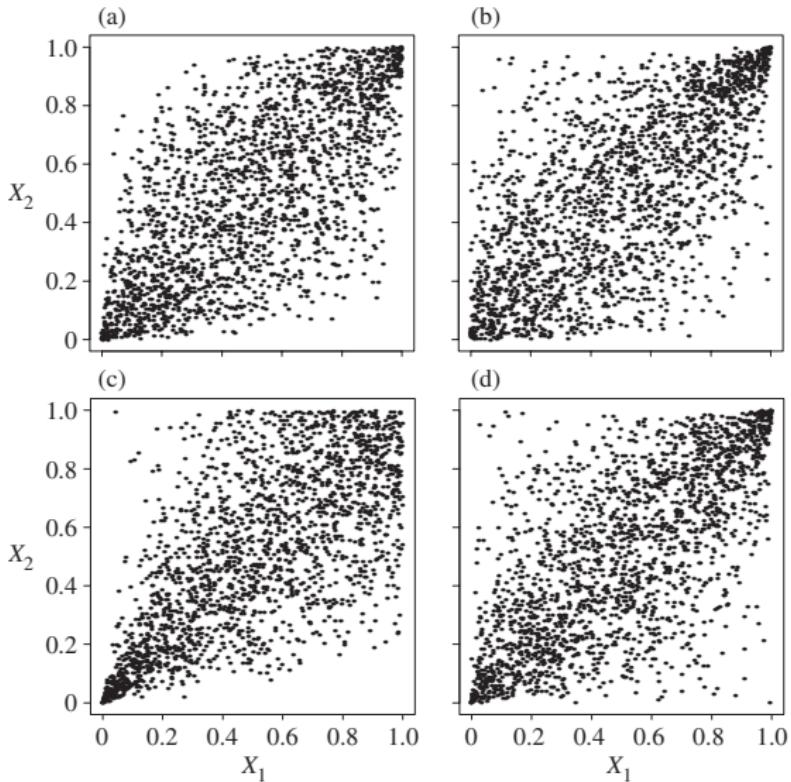
Sampling meta distributions

Meta- C distributions can be sampled via Sklar's Theorem, Part 2).

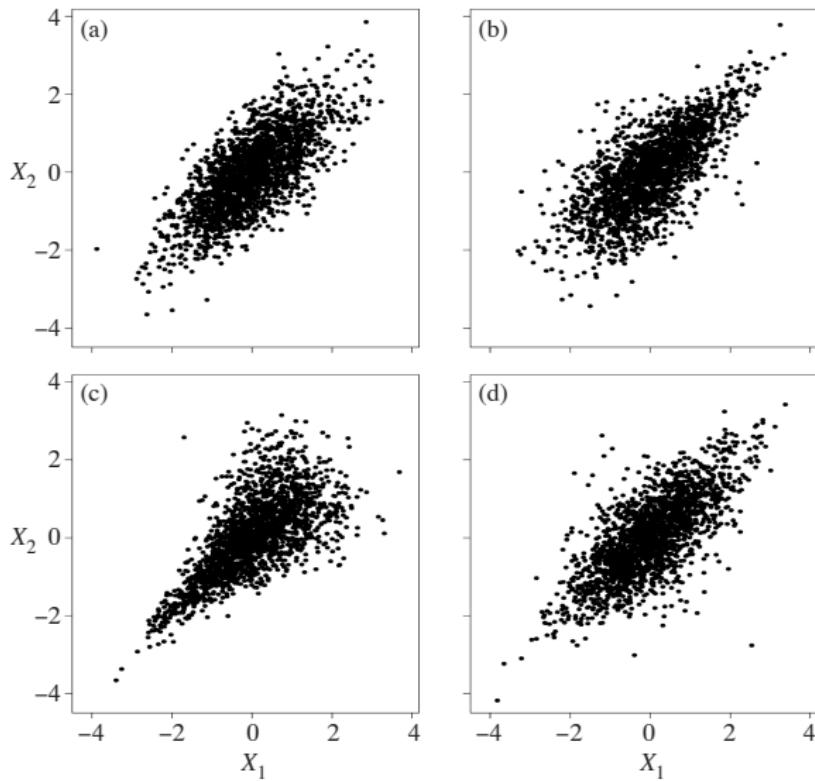
Algorithm 7.14 (Sampling)

- 1) Sample $\mathbf{U} \sim C$.
- 2) Return $\mathbf{X} = (F_1^-(U_1), \dots, F_d^-(U_d))$ (**quantile transformation**).

2000 samples from (a): $C_{\rho=0.7}^{\text{Ga}}$; (b): $C_{\theta=2}^{\text{G}}$; (c): $C_{\theta=2.2}^{\text{C}}$; (d): $C_{\nu=4, \rho=0.71}^t$



... transformed to $N(0, 1)$ margins; all have linear correlation ≈ 0.7 !



A general sampling algorithm

For a general copula C (without further information), the only known sampling algorithm is the *conditional distribution method*; see Embrechts et al. (2003) and Hofert (2010, p. 41).

Theorem 7.15 (Conditional distribution method)

If C is a d -dimensional copula and $\mathbf{U}' \sim \text{U}[0, 1]^d$, let

$$U_1 = U'_1,$$

$$U_2 = C^-(U'_2 | U_1),$$

$$\vdots$$

$$U_d = C^-(U'_d | U_1, \dots, U_{d-1}).$$

Then $\mathbf{U} \sim C$.

This typically involves numerical root-finding and the following result.

Theorem 7.16 (Schmitz (2003))

Let C be a d -dimensional copula which admits, for $d \geq 3$, continuous partial derivatives w.r.t. the first $d - 1$ arguments. Then

$$C(u_j | u_1, \dots, u_{j-1}) = \frac{D_{j-1, \dots, 1} C^{(1, \dots, j)}(u_1, \dots, u_j)}{D_{j-1, \dots, 1} C^{(1, \dots, j-1)}(u_1, \dots, u_{j-1})}$$

for a.e. $u_1, \dots, u_{j-1} \in [0, 1]$, where the superscripts denote the corresponding marginal copulas and $D_{j-1, \dots, 1}$ the differential operator w.r.t. the first $j - 1$ components.

Example 7.17 (Conditional distribution method for Clayton copula)

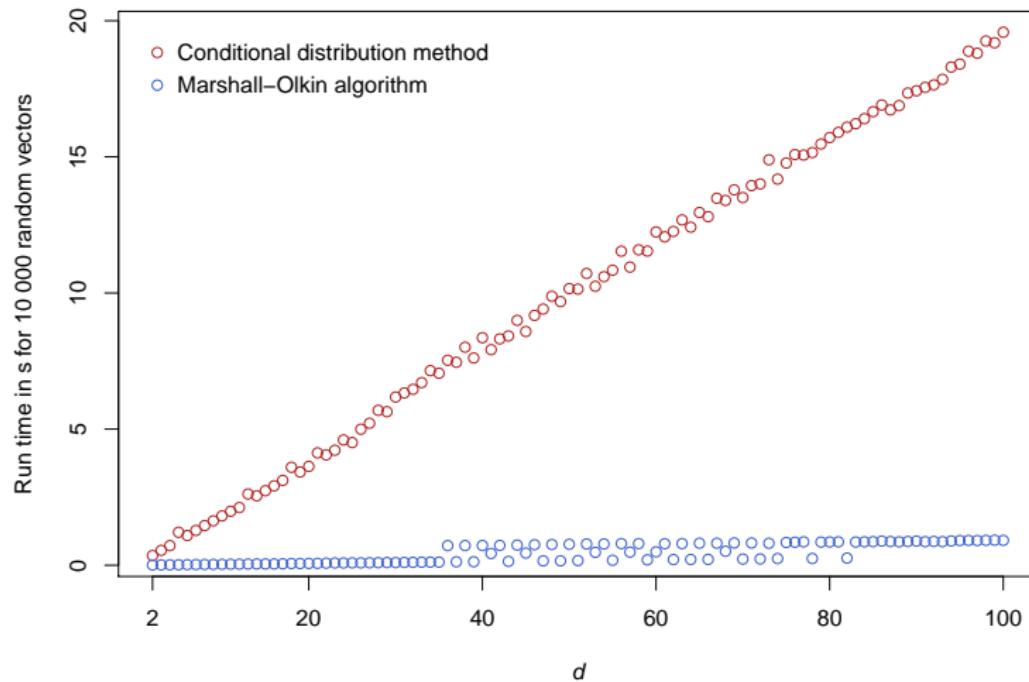
- For a Clayton copula, $\psi^{(k)}(t) = (-1)^k (1 + t)^{-k+1/\theta} \prod_{l=0}^{k-1} (l + 1/\theta)$.
- By Theorem 7.16, $C(u_j | u_1, \dots, u_{j-1}) = \frac{\psi^{(j-1)} \left(\sum_{k=1}^j \psi^{-1}(u_k) \right)}{\psi^{(j-1)} \left(\sum_{k=1}^{j-1} \psi^{-1}(u_k) \right)}$
 $= \left(\frac{1-(j-1)+\sum_{k=1}^{j-1} u_k^{-\theta}}{1-j+\sum_{k=1}^j u_k^{-\theta}} \right)^{j-1+1/\theta}.$

- Thus $C^-(u_j | u_1, \dots, u_{j-1})$ equals

$$\left(1 + \left(1 - (j-1) + \sum_{k=1}^{j-1} u_k^{-\theta} \right) (u_j^{-1/(j-1+1/\theta)} - 1) \right)^{-1/\theta}$$

- The Clayton copula is one of the rare cases where both $C(u_j | u_1, \dots, u_{j-1})$ and $C^-(u_j | u_1, \dots, u_{j-1})$ can be computed explicitly. These quantities are often difficult to compute for $d > 2$.
- For all well-known copula families, the conditional distribution method is neither simple to apply nor fast \Rightarrow Efficient sampling algorithms are typically family-specific.

A comparison with the conditional distribution method (here: Clayton)



This is the **best-case scenario for applying the conditional distribution method!** But even here there are faster algorithms (see below).

7.1.5 Further properties of copulas

Survival copulas

- If $\mathbf{U} \sim C$, then $\mathbf{1} - \mathbf{U} \sim \hat{C}$, the *survival copula* of C .
- Survival copulas transform a copula into another copula (the one of $\mathbf{1} - \mathbf{U}$).
- \hat{C} can be expressed as

$$\hat{C}(\mathbf{u}) = \sum_{J \subseteq \{1, \dots, d\}} (-1)^{|J|} C((1 - u_1)^{\mathbb{1}_J(1)}, \dots, (1 - u_d)^{\mathbb{1}_J(d)})$$

in terms of its corresponding copula (essentially an application of the Poincaré-Sylvester sieve formula). For $d = 2$,

$$\begin{aligned}\hat{C}(u_1, u_2) &= 1 - (1 - u_1) - (1 - u_2) + C(1 - u_1, 1 - u_2) \\ &= -1 + u_1 + u_2 + C(1 - u_1, 1 - u_2).\end{aligned}$$

- If C admits a density, $\hat{c}(\mathbf{u}) = c(\mathbf{1} - \mathbf{u})$.

- If $\hat{C} = C$, C is called *radially symmetric*. Check that W , Π , and M are radially symmetric.
- One can show: If X_j is symmetrically distributed about a_j , $j \in \{1, \dots, d\}$, then \mathbf{X} is radially symmetric about \mathbf{a} if and only if $C = \hat{C}$.
- Sklar's Theorem can also be formulated for survival functions. In this case, the main part reads

$$\bar{H}(\mathbf{x}) = \hat{C}(\bar{F}_1(x_1), \dots, \bar{F}_d(x_d)),$$

where $H(\mathbf{x}) = \mathbb{P}(\mathbf{X} > \mathbf{x})$ with corresponding marginal survival functions $\bar{F}_1, \dots, \bar{F}_d$ (with $\bar{F}_j(x) = \mathbb{P}(X_j > x)$).

- ⇒ Survival copulas combine marginal survival functions to joint survival functions. Note that \hat{C} is a df, whereas \bar{H} and $\bar{F}_1, \dots, \bar{F}_d$ are not!
- From this we derived Marshall–Olkin copulas.

Copula densities

- By Sklar's Theorem, if F_j has density f_j , $j \in \{1, \dots, d\}$, and C has density c , then the density h of H satisfies

$$h(\mathbf{x}) = c(F_1(x_1), \dots, F_d(x_d)) \prod_{j=1}^d f_j(x_j) \quad (37)$$

As seen before, we can recover c via

$$c(\mathbf{u}) = \frac{h(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \cdots f_d(F_d^{-1}(u_d))}.$$

- It follows from (37) that the log-density splits into

$$\log h(\mathbf{x}) = \log c(F_1(x_1), \dots, F_d(x_d)) + \sum_{j=1}^d \log f_j(x_j).$$

which allows for a *two-stage estimation* (marginal/copula parameters separately); see Section 7.5.

Exchangeability

- X is exchangeable if

$$(X_1, \dots, X_d) \stackrel{d}{=} (X_{\pi(1)}, \dots, X_{\pi(d)})$$

for any permutation $(\pi(1), \dots, \pi(d))$ of $(1, \dots, d)$.

- A copula C is exchangeable if it is the df of an exchangeable \mathbf{U} with $U[0, 1]$ margins. This holds if only if $C(u_1, \dots, u_d) = C(u_{\pi(1)}, \dots, u_{\pi(d)})$ for all possible permutations of arguments, i.e., if C is symmetric.
- Exchangeable/symmetric copulas are useful for modeling homogeneous portfolios or as (possibly crude) approximations of the underlying (possibly non-homogeneous) dependence structure.
- Examples:

- ▶ Archimedean copulas
- ▶ Elliptical copulas (such as Gauss/t) for equicorrelated P (i.e., $P = \rho J_d + (1 - \rho)I_d$ for $\rho \geq -1/(d - 1)$); in particular, $d = 2$

7.2 Dependence concepts and measures

Measures of association/dependence are scalar measures which **summarize** the dependence in terms of a **single number**. There are better and worse examples of such measures, which we will study in this section.

7.2.1 Perfect dependence

X_1, X_2 are *countermonotone* if (X_1, X_2) has copula W .

X_1, \dots, X_d are *comonotone* if (X_1, \dots, X_d) has copula M .

Proposition 7.18 (Perfect dependence)

- 1) $X_2 = T(X_1)$ almost surely with $T(x) = F_2^-(1 - F_1(x))$ (*counter-monotone*) if and only if $C(u_1, u_2) = W(u_1, u_2)$, $u_1, u_2 \in [0, 1]$.
- 2) $X_j = T_j(X_1)$ almost surely with $T_j(x) = F_j^-(F_1(x))$, $j \in \{2, \dots, d\}$ (*comonotone*), if and only if $C(\mathbf{u}) = M(\mathbf{u})$, $\mathbf{u} \in [0, 1]^d$.

Proof. We only consider Part 1) as Part 2) works similarly.

“ \Rightarrow ” By assumption, $\mathbb{P}(X_2 \leq x)$ equals $\mathbb{P}(F_2^-(1 - F_1(X_1)) \leq x)$ $\stackrel{\text{(GI5)}}{=}$ $\mathbb{P}(1 - F_1(X_1) \leq F_2(x)) = F_2(x)$. If (X_1, X_2) has copula C , then

$$\begin{aligned} C(\mathbf{u}) &\stackrel{\text{L.7.9}}{\underset{\text{"only if"}}{=}} \mathbb{P}(F_1(X_1) \leq u_1, F_2(F_2^-(1 - F_1(X_1))) \leq u_2) \\ &\stackrel{\text{(GI4)}}{=} \mathbb{P}(F_1(X_1) \leq u_1, 1 - F_1(X_1) \leq u_2) \\ &= \mathbb{P}(1 - u_2 < U \leq u_1) = W(u_1, u_2) \quad \text{for } U \sim \text{U}[0, 1]. \end{aligned}$$

“ \Leftarrow ” Let $\mathbf{u} = (u_1, u_2)$ and note that $W(\mathbf{u}) = 0$ for all $\mathbf{u} \in [0, 1]^2$ such that $u_1 + u_2 - 1 < 0$, so W puts no mass below the secondary diagonal. Similarly (exercise!) one shows that W puts no mass above the diagonal. This implies that W puts mass 1 on the secondary diagonal. Since $F_2 \uparrow \text{ran } X_2$, we thus obtain $\mathbb{P}(X_2 = F_2^-(1 - F_1(X_1))) = \mathbb{P}(F_2(X_2) = F_2(F_2^-(1 - F_1(X_1)))) \stackrel{\text{(GI4)}}{=} \mathbb{P}(F_2(X_2) = 1 - F_1(X_1)) = \mathbb{P}(U_2 = 1 - U_1) = 1$. \square

Comonotone additivity of quantiles

Proposition 7.19 (Comonotone additivity)

Let $\alpha \in (0, 1)$ and $X_j \sim F_j$, $j \in \{1, \dots, d\}$, be comontone. Then $F_{X_1 + \dots + X_d}^-(\alpha) = F_1^-(\alpha) + \dots + F_d^-(\alpha)$.

Proof. Consider $T(u) = F_1^-(u) + \dots + F_d^-(u)$, left-continuous and let $U \sim U[0, 1]$. We first show that $F_{T(U)}^-(u) = T(u)$, for all $u \in [0, 1]$.

$$1) \quad T \text{ left-continuous} \Rightarrow T(\textcolor{brown}{u}) \leq x \Leftrightarrow \textcolor{brown}{u} \leq \textcolor{blue}{u}_x := \sup\{u : T(u) \leq x\}$$

$$2) \quad 1) \Rightarrow \{T(\textcolor{brown}{U}) \leq x\} = \{\textcolor{brown}{U} \leq \textcolor{blue}{u}_x\} \Rightarrow F_{T(U)}(x) = F_U(\textcolor{blue}{u}_x) = \textcolor{blue}{u}_x.$$

$$\Rightarrow F_{T(U)}^-(u) \leq x \stackrel{(\text{GI5})}{\Leftrightarrow} F_{T(U)}(x) \geq u \stackrel{2)}{\Leftrightarrow} \textcolor{blue}{u}_x \geq u \stackrel{1)}{\Leftrightarrow} T(u) \leq x$$

\Rightarrow Choosing $x = T(u)$ and $x = F_{T(U)}^-(u)$, we see that $F_{T(U)}^-(u) = T(u)$.

Now Proposition 7.18 1) implies that $(X_1, \dots, X_d) \stackrel{d}{=} (F_1^-(U), \dots, F_d^-(U))$,

so that $F_{\sum_{j=1}^d X_j}^-(\alpha) = \textcolor{blue}{F}_{T(U)}^-(\alpha) = T(\alpha) = \sum_{j=1}^d F_j^-(\alpha)$. □

7.2.2 Linear correlation

For two random variables X_1 and X_2 with $\mathbb{E}[X_j^2] < \infty$, $j \in \{1, 2\}$, the (*linear* or *Pearson's*) *correlation coefficient* ρ is defined by

$$\rho(X_1, X_2) = \frac{\text{Cov}[X_1, X_2]}{\sqrt{\text{Var } X_1} \sqrt{\text{Var } X_2}} = \frac{\mathbb{E}[(X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)]}{\sqrt{\mathbb{E}[(X_1 - \mathbb{E}X_1)^2]} \sqrt{\mathbb{E}[(X_2 - \mathbb{E}X_2)^2]}}.$$

Proposition 7.20 (Hoeffding's identity)

Let $X_j \sim F_j$, $j \in \{1, 2\}$, be two random variables with $\mathbb{E}[X_j^2] < \infty$, $j \in \{1, 2\}$, and joint distribution function H . Then

$$\text{Cov}[X_1, X_2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (H(x_1, x_2) - F_1(x_1)F_2(x_2)) dx_1 dx_2.$$

Proof. Let (X'_1, X'_2) be an iid-copy of (X_1, X_2) . Consider

$$2 \text{Cov}[X_1, X_2]$$

$$= \mathbb{E}[(X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)] + \mathbb{E}[(X'_1 - \mathbb{E}X'_1)(X'_2 - \mathbb{E}X'_2)]$$

$$\begin{aligned}
 &= \mathbb{E}^{\text{check}}[((X_1 - \mathbb{E}X_1) - (X'_1 - \mathbb{E}X'_1)) \cdot ((X_2 - \mathbb{E}X_2) - (X'_2 - \mathbb{E}X'_2))] \\
 &= \mathbb{E}[(X_1 - X'_1)(X_2 - X'_2)].
 \end{aligned}$$

With $b - a = \int_{-\infty}^{\infty} (\mathbb{1}_{\{a \leq x\}} - \mathbb{1}_{\{b \leq x\}}) dx$ for all $a, b \in \mathbb{R}$, we obtain that

$$\begin{aligned}
 &2 \operatorname{Cov}[X_1, X_2] \\
 &= \mathbb{E}\left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathbb{1}_{\{X'_1 \leq x_1\}} - \mathbb{1}_{\{X_1 \leq x_1\}})(\mathbb{1}_{\{X'_2 \leq x_2\}} - \mathbb{1}_{\{X_2 \leq x_2\}}) dx_1 dx_2\right] \\
 &\stackrel{\text{Fubini}}{=} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{E}[\dots] dx_1 dx_2 \stackrel{\substack{\text{multiply} \\ \text{ind.}}}{=} 2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (H(x_1, x_2) - F_1(x_1)F_2(x_2)) dx_1 dx_2.
 \end{aligned}$$

□

We now collect well-known and not so well-known properties of ρ (which also reveal ρ 's deficiencies to quantify dependence).

Proposition 7.21 (Properties of linear correlation)

Let X_1 and X_2 be two random variables with $\mathbb{E}[X_j^2] < \infty$, $j \in \{1, 2\}$.

- 1) $|\rho| \leq 1$. Furthermore, $|\rho| = 1$ if and only if there are constants $a \in \mathbb{R} \setminus \{0\}, b \in \mathbb{R}$ with $X_2 = aX_1 + b$ a.s. with $a \geq 0$ if and only if $\rho = \pm 1$.
- 2) If X_1 and X_2 are independent, then $\rho = 0$ (the converse is not true in general, see Example 7.23 below).
- 3) ρ is invariant under strictly increasing linear transformations on $\text{ran } X_1 \times \text{ran } X_2$ (but not necessarily invariant under strictly increasing functions in general, see Example 7.22 below).
- 4) If (X_{i1}, X_{i2}) has copula C_i and $\mathbb{E}[X_{ij}^2] < \infty$, $i, j \in \{1, 2\}$, then $C_1(u_1, u_2) \leq C_2(u_1, u_2) \forall u_1, u_2$ implies $\rho(X_{11}, X_{12}) \leq \rho(X_{21}, X_{22})$.

Proof.

- 1) Cauchy–Schwarz inequality $|\langle X, Y \rangle| \leq \|X\| \|Y\|$ ($L^2 = \{rv X : \mathbb{E}[X^2] < \infty\}$ is a Hilbert space with inner product $\langle X, Y \rangle = \mathbb{E}[XY]$).
- 2) $\text{Cov}[X_1, X_2] = \mathbb{E}[(X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)] \stackrel{\text{ind.}}{=} \mathbb{E}[X_1 - \mathbb{E}X_1]\mathbb{E}[X_2 - \mathbb{E}X_2] = 0 \Rightarrow \rho = 0$.
- 3) $\text{Cov}[a_1X_1 + b_1, a_2X_2 + b_2] = a_1a_2 \text{Cov}[X_1, X_2]$ and $\sqrt{\text{Var}[a_jX_j + b_j]} = \sqrt{a_j^2 \text{Var } X_j} = |a_j| \sqrt{\text{Var } X_j}$, $j \in \{1, 2\}$, imply that for $a_1, a_2 > 0$,
$$\rho(a_1X_1 + b_1, a_2X_2 + b_2) = \frac{a_1a_2}{|a_1||a_2|} \rho(X_1, X_2) = \rho(X_1, X_2).$$
- 4) $\text{Cov}[X_{11}, X_{12}] \stackrel{\substack{\text{Hoeff.} \\ \text{Sklar}}}{=} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathcal{C}_1(F_1(x_1), F_2(x_2)) - F_1(x_1)F_2(x_2)) dx_1 dx_2$
$$\leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathcal{C}_2(F_1(x_1), F_2(x_2)) - F_1(x_1)F_2(x_2)) dx_1 dx_2 = \text{Cov}[X_{21}, X_{22}].$$

□

Drawbacks of linear correlation

- 1) ρ depends on the marginal distributions. In particular, for ρ to exist, $\text{Var } X_1, \text{Var } X_2$ must exist!

Example 7.22 (ρ not invariant under general increasing functions)

Consider $X_1, X_2 \stackrel{\text{ind.}}{\sim} \text{Par}(3)$ (df $F(x) = 1 - x^{-3}$, $x \geq 1$; $\text{Var } X_1, \text{Var } X_2$ exist) $\Rightarrow \rho(X_1, X_2) = 0$ but $\rho(X_1^2, X_2)$ does not even exist since $X_1^2 \sim \text{Par}(3/2)$ and thus $\text{Var}[X_1^2] = \infty$.

- 2) Linear correlation only gives a scalar summary of linear dependence. It is not invariant with respect to strictly increasing transformations in general, i.e.,

$$\rho(T_1(X_1), T_2(X_2)) \neq \rho(X_1, X_2).$$

(assuming the left-hand side to exist; see Example 7.22).

- 3) X_1, X_2 independent $\Rightarrow \rho(X_1, X_2) = 0$. But $\rho(X_1, X_2) = 0 \not\Rightarrow X_1, X_2$ independent

Example 7.23 (Uncorrelated $\not\Rightarrow$ independent)

Consider the two risks

$$X_1 = Z \quad (\text{Profit \& Loss Country A}),$$

$$X_2 = VZ \quad (\text{Profit \& Loss Country B}),$$

where V, Z are independent with $Z \sim N(0, 1)$ and $\mathbb{P}(V = -1) = \mathbb{P}(V = 1) = 1/2$ (Rademacher). Then $X_2 \sim N(0, 1)$ and $\rho(X_1, X_2) = \text{Cov}[X_1, X_2] = \mathbb{E}[X_1 X_2] = \underset{\text{ind.}}{\mathbb{E}[V]\mathbb{E}[Z^2]} = 0$, but X_1 and X_2 are not independent (in fact, V switches between counter- and comonotonicity).

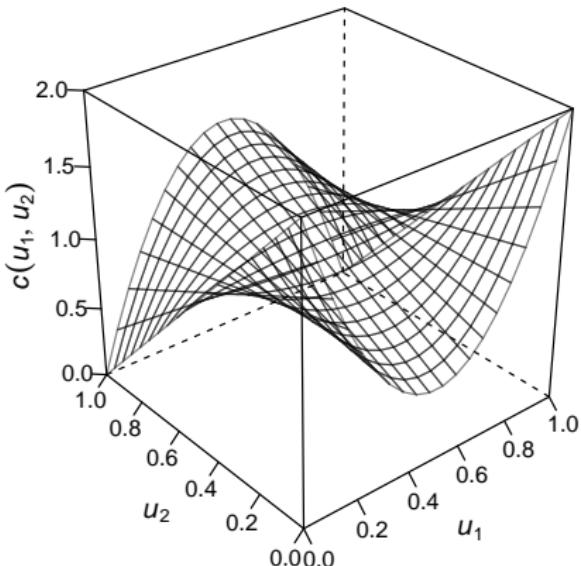
- Only known exception: $N_2(\mu, \Sigma)$ (uncorrelatedness $\xrightarrow[\text{density}]{} \Rightarrow$ independence)
- Compare Example 7.23 with $\mathbf{X}' = (X'_1, X'_2) \sim N_2(\mathbf{0}, I_2)$. Both have $N(0, 1)$ margins and $\rho = 0$, but the copula of \mathbf{X}' is the Π whereas the copula of $\mathbf{X} = (X_1, X_2)$ is $C(\mathbf{u}) = 0.5W(\mathbf{u}) + 0.5M(\mathbf{u}) \neq \Pi$. Thus F_1, F_2 and ρ do not uniquely determine H . Another example of this type is given by Fallacy 1 below.

Correlation fallacies

Fallacy 1: F_1, F_2 , and ρ uniquely determine H

This is true for bivariate elliptical distributions, but wrong in general.

Counter-example: $C(u_1, u_2) = u_1 u_2 (1 - 2\theta(u_1 - \frac{1}{2})(u_1 - 1)(u_2 - 1))$



Properties

- 1) Take $F_1 = F_2 = U[0, 1]$
(extends to $\mathbb{E}[X_j^2] < \infty$ and
 F_1 symmetric about 0).
- 2) $\rho = 0$ for all $\theta \in [-1, 1]$ (see
below)
- 3) There are ∞ -many such models
 H (∞ -many θ ; plot: $\theta = 1$)

In particular, $\rho = 0 \not\Rightarrow C = \Pi!$

Reasoning for $\rho = 0$ for all $\theta \in [-1, 1]$:

By Hoeffding's identity and since $F_1 = F_2 = U[0, 1]$ (for other margins, use Sklar's Theorem),

$$\rho = \frac{\text{Cov}[X_1, X_2]}{\sqrt{\text{Var } X_1} \sqrt{\text{Var } X_2}} \stackrel{U[0,1]}{\underset{\text{Hoeff.}}{=}} 12 \int_0^1 \int_0^1 (C(u_1, u_2) - u_1 u_2) du_1 du_2.$$

Now consider

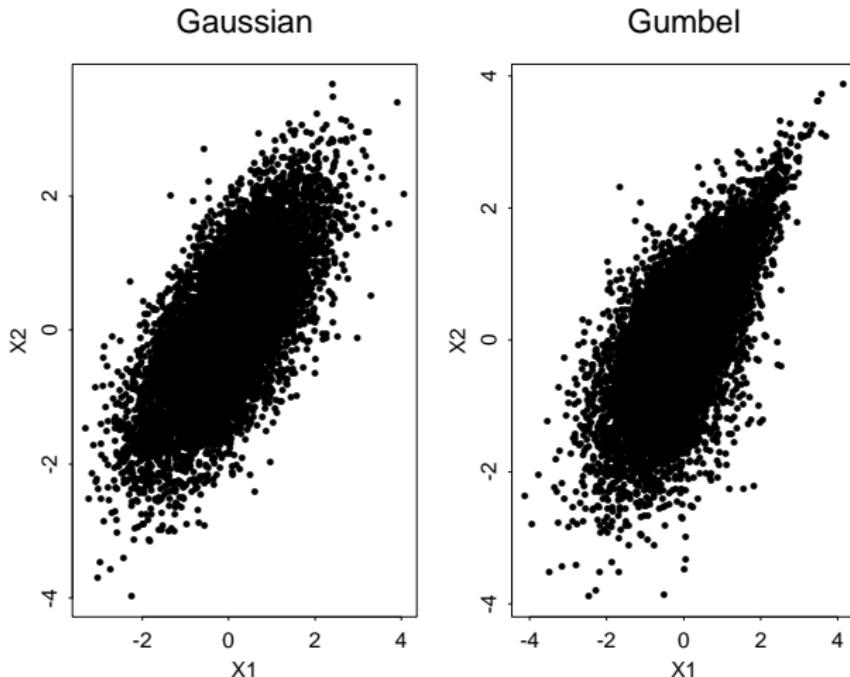
$$C(u_1, u_2) = u_1 u_2 + g_1(u_1)g_2(u_2)$$

with $g_j(0) = g_j(1) = 0$ and $g'_1(u_1)g'_2(u_2) \geq -1$ (check that C is a copula by computing its density). Then

$$\rho = 12 \int_0^1 g_1(u_1) du_1 \int_0^1 g_2(u_2) du_2$$

⇒ If g_1 is point symmetric about $1/2$, then $\rho = 0$.

Another counter-example: Gauss and Gumbel copulas compared



Margins are $N(0, 1)$, $\rho = 0.7$; yet different multivariate models.

Fallacy 2: Given F_1, F_2 , any $\rho \in [-1, 1]$ is attainable

- This is true if (X_1, X_2) is elliptically distributed (we have seen that if $\mathbb{E}[R^2] < \infty$, $\text{Cov } \mathbf{X} = \frac{\mathbb{E}[R^2]}{d} \Sigma$ and $\text{Cor } \mathbf{X} = \mathbf{P}$, the correlation matrix corresponding to the scale matrix Σ ; see Example 6.21) but wrong in general; see below.
- Reasoning 1: If F_1 and F_2 are not of the same type, $\rho(X_1, X_2) < 1$; see Proposition 7.21 1).
- Reasoning 2: Hoeffding's identity

$$\text{Cov}[X_1, X_2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (C(F_1(x_1), F_2(x_2)) - F_1(x_1)F_2(x_2)) dx_1 dx_2.$$

implies bounds on attainable ρ via

$$\rho_{\min} \leq \rho \leq \rho_{\max} \quad (\text{attained for } C = W \text{ and } C = M, \text{ respectively}).$$

Example 7.24 (Bounds for a model with $\text{LN}(0, \sigma_j^2)$ margins)

Let $X_j \sim \text{LN}(0, \sigma_j^2)$, $j \in \{1, 2\}$.

$C = W : (X_1, X_2) = (\exp(\sigma_1 Z), \exp(-\sigma_2 Z))$ for $Z \sim N(0, 1)$. Then

$$\begin{aligned}\text{Cov}[X_1, X_2] &= \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] \\ &= e^{(\sigma_1 - \sigma_2)^2/2} - e^{(\sigma_1^2 + \sigma_2^2)/2}.\end{aligned}$$

$\mathbb{E}[\exp(tZ)] = e^{t^2/2}$

With $\text{Var}[\text{LN}(\mu, \sigma^2)] = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$, it follows that

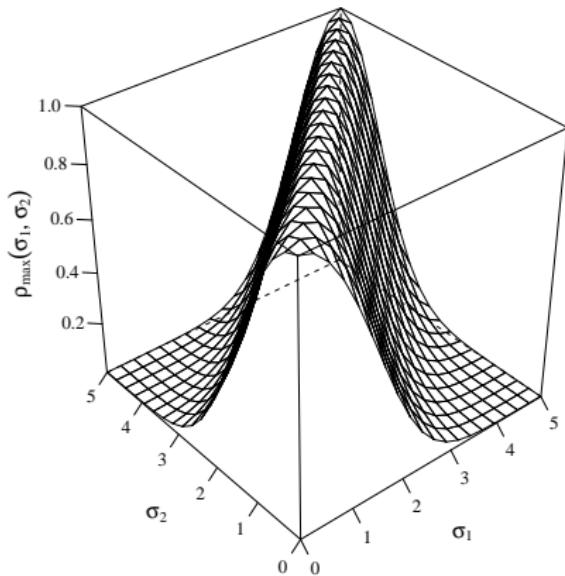
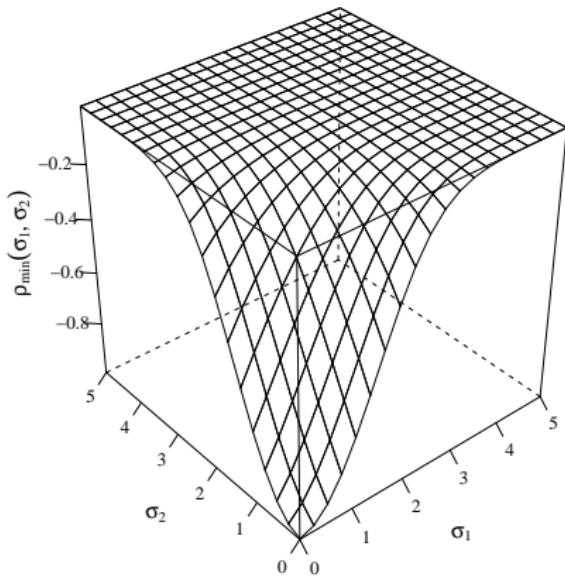
$$\rho_{\min} = \tilde{\rho}(\sigma_1, -\sigma_2),$$

where

$$\tilde{\rho}(\sigma_1, \sigma_2) = \frac{e^{(\sigma_1 + \sigma_2)^2/2} - e^{(\sigma_1^2 + \sigma_2^2)/2}}{\sqrt{(e^{\sigma_1^2} - 1)e^{\sigma_1^2}} \sqrt{(e^{\sigma_2^2} - 1)e^{\sigma_2^2}}}$$

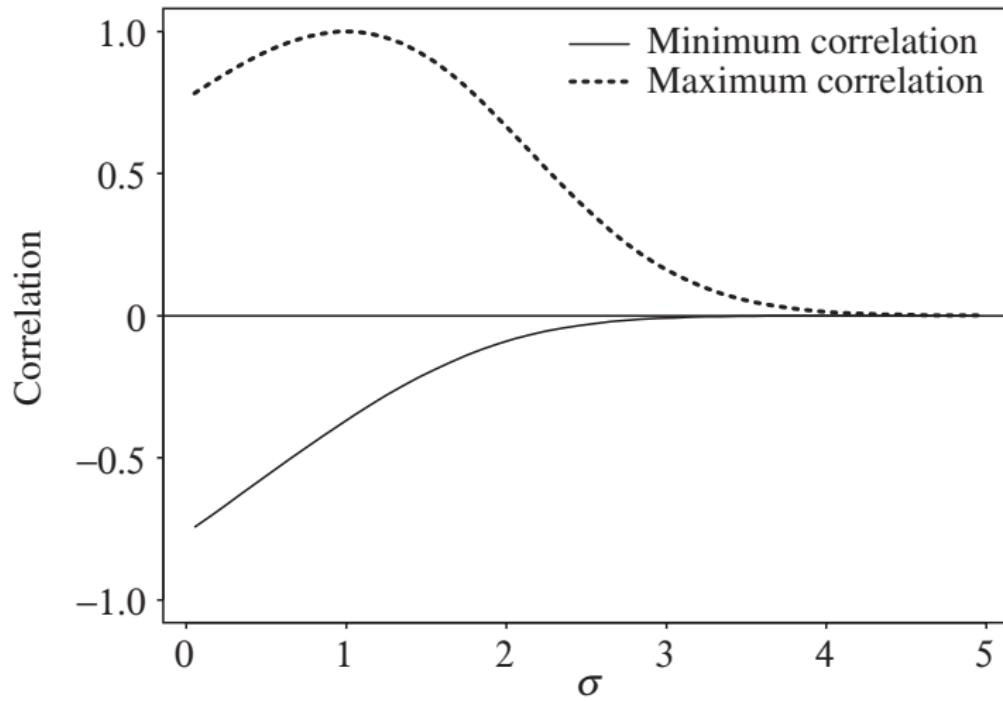
$C = M$: Similarly, $\rho_{\max} = \tilde{\rho}(\sigma_1, \sigma_2)$.

A picture is worth a thousand words...

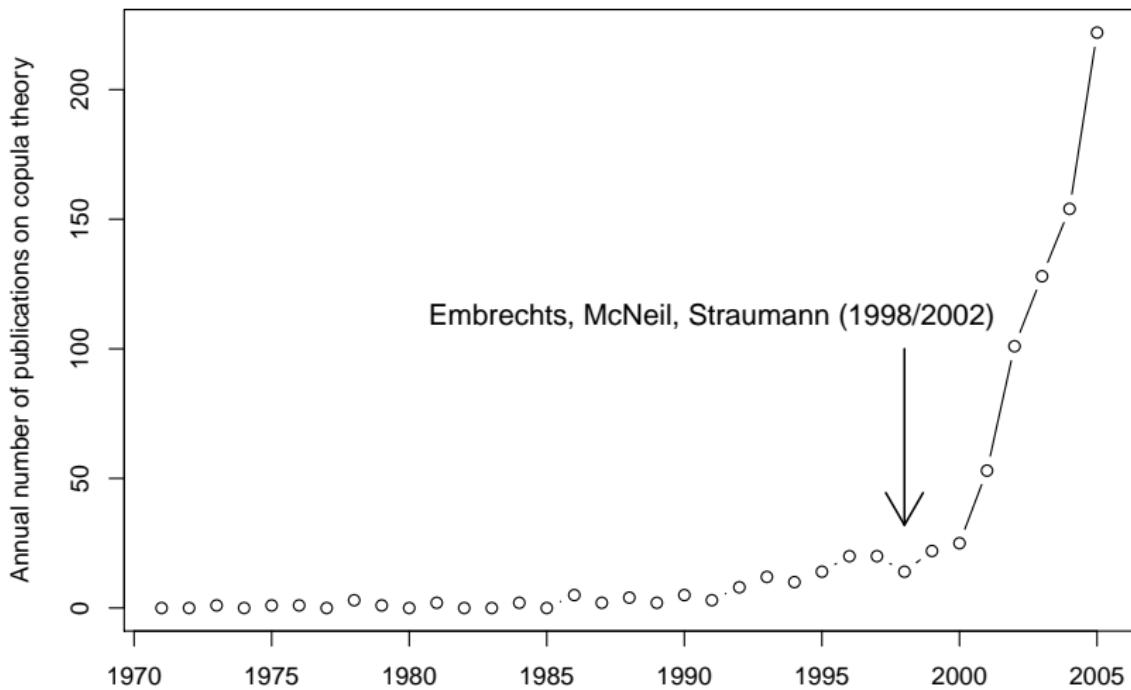


Example: For $\sigma_1^2 = 1$, $\sigma_2^2 = 16$ one has $\rho \in [-0.0003, 0.0137]!$

Specifically, let $X_1 \sim \text{LN}(0, 1)$ and $X_2 \sim \text{LN}(0, \sigma^2)$. Now let σ vary and plot ρ_{\min} and ρ_{\max} against σ :



Note: This industry-inspired example has been influential...



The dataset stems from Genest et al. (2009).

Fallacy 3: ρ maximal (i.e., $C = M$) $\Rightarrow \text{VaR}_\alpha(X_1 + X_2)$ maximal

- This is true if (X_1, X_2) is elliptically distributed (since the maximal $\rho = 1$ implies that X_1, X_2 are comonotone, see Proposition 7.21 1), VaR_α is subadditive (so additivity provides the largest possible bound), and VaR_α is comonotone additive; see Propositions 6.24 and 7.19).
- Any superadditivity example $\text{VaR}_\alpha(X_1 + X_2) > \text{VaR}_\alpha(X_1) + \text{VaR}_\alpha(X_2)$ ($= \text{VaR}_\alpha(X_1 + X_2)$ under maximal correlation, i.e., $C = M$) serves as a counterexample; see Section 2.3.5.

7.2.3 Rank correlation

To overcome (some) of the deficiencies of ρ , Scarsini (1984) introduced:

Definition 7.25 (Rank correlation coefficient)

A measure of association $\kappa = \kappa(X_1, X_2) = \kappa(C)$ between two continuously distributed random variables X_1 and X_2 with copula C is a *rank correlation coefficient* if

- 1) κ exists for every pair (X_1, X_2) of cont. distributed random variables;
- 2) $-1 \leq \kappa \leq 1$, $\kappa(W) = -1$, and $\kappa(M) = 1$;
- 3) $\kappa(X_1, X_2) = \kappa(X_2, X_1)$;
- 4) X_1 and X_2 being independent implies $\kappa(X_1, X_2) = \kappa(\Pi) = 0$;
- 5) $\kappa(-X_1, X_2) = -\kappa(X_1, X_2)$;
- 6) $C_1(\mathbf{u}) \leq C_2(\mathbf{u})$ for all $\mathbf{u} \in [0, 1]^2$ implies $\kappa(C_1) \leq \kappa(C_2)$;
- 7) $C_n \rightarrow C$ ($n \rightarrow \infty$) pointwise implies $\lim_{n \rightarrow \infty} \kappa(C_n) = \kappa(C)$.

Proposition 7.26 (Basic properties of κ)

Let κ be a rank correlation coefficient for two continuously distributed random variables $X_1 \sim F_1$ and $X_2 \sim F_2$. Then

- 1) $\kappa(X_1, X_2) = \kappa(C)$ (κ only depends on C).
- 2) if T_j is a strictly increasing function on $\text{ran } X_j$, $j \in \{1, 2\}$, then $\kappa(T_1(X_1), T_2(X_2)) = \kappa(X_1, X_2)$.

Proof.

- 1) Set $(U_1, U_2) = (F_1(X_1), F_2(X_2))$. By the invariance principle, (X_1, X_2) and (U_1, U_2) have the same copula C . Thus, by 6), $\kappa(U_1, U_2) \leq \kappa(X_1, X_2)$, but also $\kappa(X_1, X_2) \leq \kappa(U_1, U_2)$, so $\kappa(X_1, X_2) = \kappa(U_1, U_2)$ (\Rightarrow only depends on C).
- 2) Invariance principle \Rightarrow The copula C of (X_1, X_2) equals the copula of $(T_1(X_1), T_2(X_2))$. Hence $\kappa(T_1(X_1), T_2(X_2)) = \kappa(C) = \kappa(X_1, X_2)$.

□

Rank correlation coefficients are...

- ... always defined;
- ... invariant under strictly increasing transformations of the random variables (hence only depend on the underlying copula).

Examples: Kendall's tau and Spearman's rho

Definition 7.27 (Kendall's tau)

Let $X_j \sim F_j$ with F_j continuous, $j \in \{1, 2\}$. Let (X'_1, X'_2) be an independent copy of (X_1, X_2) . *Kendall's tau* is defined by

$$\begin{aligned}\tau &= \mathbb{E}[\text{sign}((X_1 - X'_1)(X_2 - X'_2))] \\ &= \mathbb{P}((X_1 - X'_1)(X_2 - X'_2) > 0) - \mathbb{P}((X_1 - X'_1)(X_2 - X'_2) < 0),\end{aligned}$$

where $\text{sign}(x) = \mathbb{1}_{(0,\infty)}(x) - \mathbb{1}_{(-\infty,0)}(x)$.

Proposition 7.28 (Formula for Kendall's tau)

Let $X_j \sim F_j$ with F_j continuous, $j \in \{1, 2\}$, and copula C . Then

$$\tau = 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1.$$

Proof. Let (X'_1, X'_2) be an independent copy of (X_1, X_2) . Then

$$\begin{aligned}\tau &= \mathbb{P}((X_1 - X'_1)(X_2 - X'_2) > 0) - \mathbb{P}((X_1 - X'_1)(X_2 - X'_2) < 0) \\ &= 2 \underbrace{\mathbb{P}((X_1 - X'_1)(X_2 - X'_2) > 0)}_{=2\mathbb{P}(X_1 < X'_1, X_2 < X'_2)} - 1 = 4\mathbb{P}(U_1 \leq U'_1, U_2 \leq U'_2) - 1 \\ &= 4 \int_0^1 \int_0^1 \mathbb{P}(U_1 \leq u_1, U_2 \leq u_2) dC(u_1, u_2) - 1\end{aligned}$$

□

- For computing τ , $\int_{[0,1]^2} C(\mathbf{u}) d\tilde{C}(\mathbf{u}) = \frac{1}{2} - \int_{[0,1]^2} D_1 C(\mathbf{u}) D_2 \tilde{C}(\mathbf{u}) d\mathbf{u}$ is often helpful; see Li et al. (2002).

- An estimator of τ is provided by the sample version of Kendall's tau

$$\hat{\tau}_n = \frac{1}{\binom{n}{2}} \sum_{1 \leq i_1 < i_2 \leq n} \text{sign}((X_{i_1 1} - X_{i_2 1})(X_{i_1 2} - X_{i_2 2})). \quad (38)$$

- This is often used to compute estimators for P (the correlation matrix corresponding to the scale matrix Σ) for elliptical copulas:

- 1) Compute $\hat{\tau}_{n,j_1 j_2}$ for all $1 \leq j_1 < j_2 \leq d$ (column indices of all pairs).
- 2) Invert (32) to get $\hat{P}'_{j_1 j_2} = \sin(\frac{\pi}{2}\tau)$, $j_1 \neq j_2$ (see later why).
- 3) Use, e.g., `nearPD(, corr=TRUE)` in the R package `Matrix` to get a positive-definite matrix \hat{P} close to \hat{P}' .

Definition 7.29 (Spearman's rho)

Let $X_j \sim F_j$ with F_j continuous, $j \in \{1, 2\}$. *Spearman's rho* is defined by $\rho_S = \rho(F_1(X_1), F_2(X_2))$.

Proposition 7.30 (Formula for Spearman's rho)

Let $X_j \sim F_j$ with F_j continuous, $j \in \{1, 2\}$, and copula C . Then

$$\rho_S = 12 \int_0^1 \int_0^1 C(u_1, u_2) du_1 du_2 - 3.$$

Proof. By Hoeffding's identity, we have $\rho_S(X_1, X_2) = \rho(F_1(X_1), F_2(X_2)) = 12 \int_0^1 \int_0^1 (C(u_1, u_2) - u_1 u_2) du_1 du_2 = 12 \int_0^1 \int_0^1 C(u_1, u_2) du_1 du_2 - 3$. \square

- An estimator $\hat{\rho}_{S,n}$ is given by the sample correlation computed from pseudo-observations (marginal empirical dfs applied componentwise).
- For κ being either Spearman's rho or Kendall's tau, Embrechts et al. (2002) show that $\kappa = \pm 1$ if and only if X_1, X_2 are co-/countermonotonic.
- Fallacy 2 (For F_1, F_2 , any $\rho \in [-1, 1]$ is attainable) is solved. Take

$$H(x_1, x_2) = \lambda \textcolor{brown}{W}(F_1(x_1), F_2(x_2)) + (1 - \lambda) \textcolor{brown}{M}(F_1(x_1), F_2(x_2)).$$

This is a model with $\rho_S = \tau = 1 - 2\lambda$ (choose λ as desired).

- Fallacy 1 (F_1, F_2, ρ uniquely determine H) is not solved by replacing ρ by rank correlation coefficients κ (it is easy to construct several copulas with the same Kendall's tau, e.g., via Archimedean copulas).
- Fallacy 3 ($C = M$ implies $\text{VaR}_\alpha(X_1 + X_2)$ maximal) is also not solved by rank correlation coefficients $\kappa = 1$: Although $\kappa = 1$ corresponds to $C = M$, this copula does not necessarily provide the largest $\text{VaR}_\alpha(X_1 + X_2)$; see our superadditivity examples.
- Note that for $S \sim U(\{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| = 1\})$, $S_1 \stackrel{d}{=} -S_1$, so $\kappa(S_1, S_2) = \kappa(-S_1, S_2) = -\kappa(S_1, S_2)$ and hence $\kappa(S_1, S_2) = 0$ although (S_1, S_2) are clearly not independent.
- Nevertheless, rank correlations are useful to overall summarize dependence or parameterize copula families to make dependence comparable (they can obviously not replace the underlying copula, though).

7.2.4 Coefficients of tail dependence

Goal: Measure **extremal dependence**, i.e., dependence in the **joint tails**.

Definition 7.31 (Tail dependence)

Let $X_j \sim F_j$, $j \in \{1, 2\}$, be continuously distributed random variables. Provided that the limits exist, the *lower tail-dependence coefficient* λ_L and *upper tail-dependence coefficient* λ_U of X_1 and X_2 are defined by

$$\lambda_L = \lim_{u \downarrow 0} \mathbb{P}(X_2 \leq F_2^-(u) \mid X_1 \leq F_1^-(u)),$$

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

If $\lambda_L \in (0, 1]$ ($\lambda_U \in (0, 1]$), then (X_1, X_2) is *lower (upper) tail dependent*. If $\lambda_L = 0$ ($\lambda_U = 0$), then (X_1, X_2) is *lower (upper) tail independent*.

- As (conditional) probabilities, we clearly have $\lambda_L, \lambda_U \in [0, 1]$.
- Tail dependence is a copula property, since (note that $F_j \uparrow$ on $\text{ran } X_j$)

$$\begin{aligned} \mathbb{P}(X_2 \leq F_2^-(u) \mid X_1 \leq F_1^-(u)) &= \frac{\mathbb{P}(X_2 \leq F_2^-(u), X_1 \leq F_1^-(u))}{\mathbb{P}(X_1 \leq F_1^-(u))} \\ &= \frac{\mathbb{P}(F_2(X_2) \leq F_2(F_2^-(u)), F_1(X_1) \leq F_1(F_1^-(u)))}{\mathbb{P}(F_1(X_1) \leq F_1(F_1^-(u)))} \\ &\stackrel{(GI4)}{=} \frac{\mathbb{P}(F_1(X_1) \leq u, F_2(X_2) \leq u)}{\mathbb{P}(F_1(X_1) \leq u)} = \frac{C(u, u)}{u}, \quad u \in (0, 1), \end{aligned}$$

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

- If $u \mapsto C(u, u)$ is differentiable in a neighborhood of 0 and the limit exists, then $\lambda_L = \lim_{u \downarrow 0} \frac{d}{du} C(u, u)$ (l'Hôpital's Rule).
- If C is totally differentiable in a neighborhood of 0 and the limit exists, then $\lambda_L = \lim_{u \downarrow 0} (\mathbf{D}_1 C(u, u) + \mathbf{D}_2 C(u, u))$ (Chain Rule). If C is symmetric, then $\lambda_L = 2 \lim_{u \downarrow 0} \mathbf{D}_1 C(u, u)$.

- Similarly as above, for the upper tail-dependence coefficient,

$$\lambda_U = \lim_{u \uparrow 1} \frac{1 - 2u + C(u, u)}{1 - u} = \lim_{u \downarrow 0} \frac{\hat{C}(u, u)}{u}$$

Also, $\lambda_U = 2 - \lim_{u \uparrow 1} \frac{d}{du} C(u, u)$ and $\lambda_U = \lim_{u \uparrow 1} (D_1 C(u, u) + D_2 C(u, u))$ if the corresponding limits exist.

- λ_L, λ_U not necessarily exist, but they do in all practical cases.
- For elliptical copulas, $\lambda_L = \lambda_U =: \lambda$ (true for all radially symmetric copulas); for C_P^{Ga} , $\lambda = 0$, and for $C_{\nu, P}^t$, $\lambda > 0$ (see later).
- For Archimedean copulas with strict ψ , a substitution and l'Hôpital's Rule show:

$$\lambda_L = \lim_{u \downarrow 0} \frac{\psi(2\psi^{-1}(u))}{u} = \lim_{t \rightarrow \infty} \frac{\psi(2t)}{\psi(t)} = 2 \lim_{t \rightarrow \infty} \frac{\psi'(2t)}{\psi'(t)},$$

$$\lambda_U = 2 - \lim_{u \uparrow 1} \frac{1 - \psi(2\psi^{-1}(u))}{1 - u} = 2 - \lim_{t \downarrow 0} \frac{1 - \psi(2t)}{1 - \psi(t)} = 2 - 2 \lim_{t \downarrow 0} \frac{\psi'(2t)}{\psi'(t)}$$

Clayton: $\lambda_L = 2^{-1/\theta}$, $\lambda_U = 0$; Gumbel: $\lambda_L = 0$, $\lambda_U = 2 - 2^{1/\theta}$

7.3 Normal mixture copulas

... are the copulas of multivariate normal (mean-)variance mixtures $\mathbf{X} \stackrel{\text{d}}{=} \mu + \sqrt{W}\mathbf{A}\mathbf{Z}$ ($\mathbf{X} \stackrel{\text{d}}{=} \mathbf{m}(W) + \sqrt{W}\mathbf{A}\mathbf{Z}$); e.g., Gauss, t copulas.

7.3.1 Tail dependence

Coefficients of tail dependence

Let $(U_1, U_2) \sim C$ for a normal variance mixture copula C . Then

$$\lambda \stackrel{\substack{\text{radial} \\ \text{symm.}}}{=} \lambda_L \stackrel{\text{symm.}}{=} 2 \lim_{q \downarrow 0} D_1 C(q, q) \stackrel{\text{Th. 7.16}}{=} 2 \lim_{q \downarrow 0} \mathbb{P}(U_2 \leq q \mid U_1 = q).$$

Example 7.32 (λ for the Gauss and t copula)

- This result goes back to Sibuya (1959). Let $\mathbf{U} = (\Phi(X_1), \Phi(X_2))$ with $\mathbf{X} \sim N_2(\mathbf{0}, P)$, where $P = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, i.e., $\mathbf{U} \sim C_P^{Ga}$. Recall that $X_2 \mid X_1 = x \sim N(\rho x, 1 - \rho^2)$. This implies that $\lambda = 2 \lim_{x \downarrow -\infty} \mathbb{P}(X_2 \leq x \mid X_1 = x) = 2 \lim_{x \downarrow -\infty} \Phi\left(\frac{x(1-\rho)}{\sqrt{1-\rho^2}}\right) = \mathbb{1}_{\{\rho=1\}}$ (essentially no tail dependence).

- For $C_{\nu, P}^t$, one can show that (via $f_{X_2|X_1}(x_2 | x_1) = \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_1}(x_1)}$)

$$X_2 | X_1 = x \sim t_{\nu+1}\left(\rho x, \frac{(1-\rho^2)(\nu+x^2)}{\nu+1}\right)$$

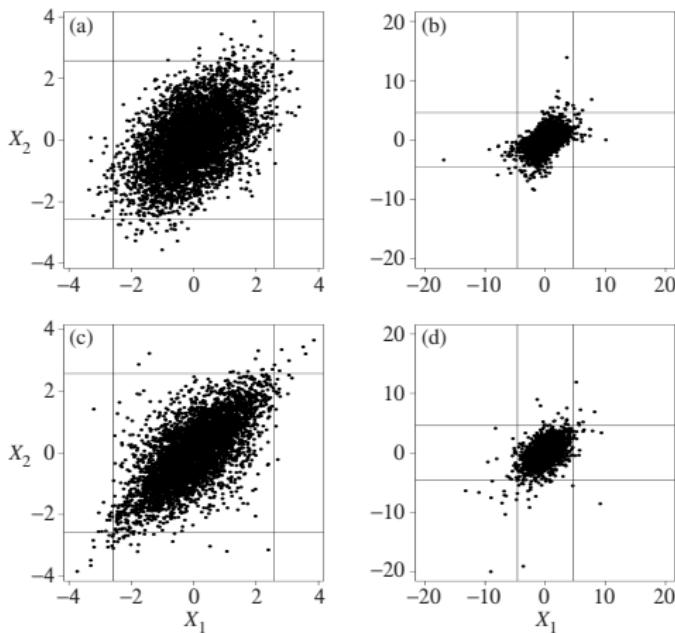
and thus $\mathbb{P}(X_2 \leq x | X_1 = x) = t_{\nu+1}\left(\frac{x-\rho x}{\sqrt{\frac{(1-\rho^2)(\nu+x^2)}{\nu+1}}}\right)$. Hence

$$\lambda = 2t_{\nu+1}\left(-\sqrt{\frac{(\nu+1)(1-\rho)}{1+\rho}}\right) \quad (\text{tail dependence}).$$

ν	$\rho = -0.5$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 1$
∞	0	0	0	0	1
10	0.00	0.01	0.08	0.46	1
4	0.01	0.08	0.25	0.63	1
2	0.06	0.18	0.39	0.72	1

- What drives tail dependence of normal variance mixtures is W . If W has a power tail, we get tail dependence, otherwise not.

Joint quantile exceedance probabilities



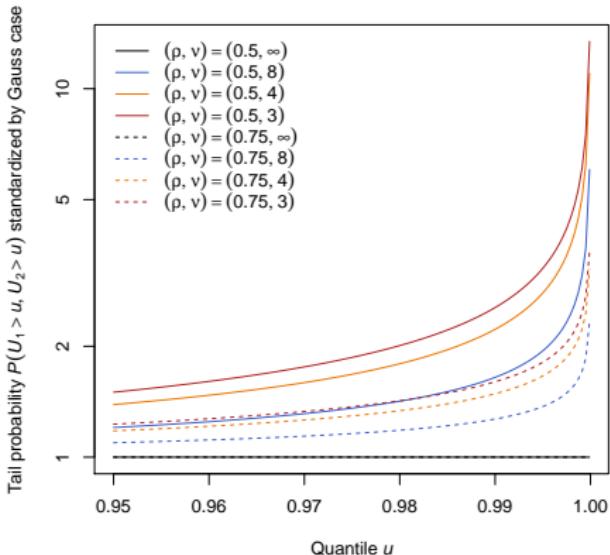
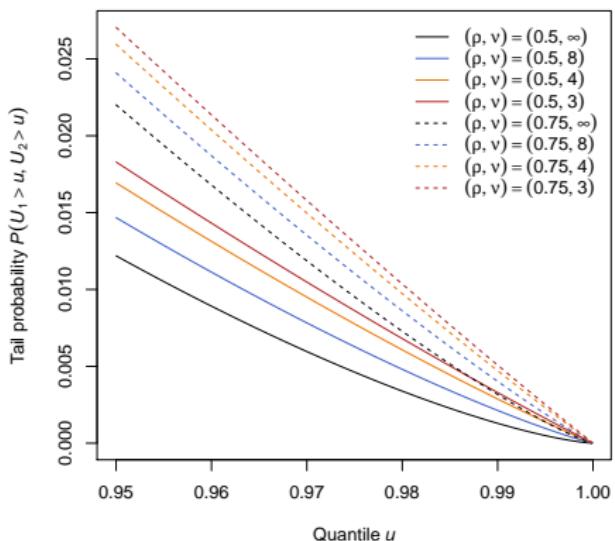
5000 samples from

- (a) $N_2(\mathbf{0}, P = (\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}))$, $\rho = 0.5$;
- (b) C_{ρ}^{Ga} with t_4 margins (same dependence as in (a));
- (c) $C_{4,\rho}^t$ with $N(0, 1)$ margins;
- (d) $t_2(4, \mathbf{0}, P)$ (same dependence as in (c)).

Lines denote 0.005- and 0.995-quantiles.

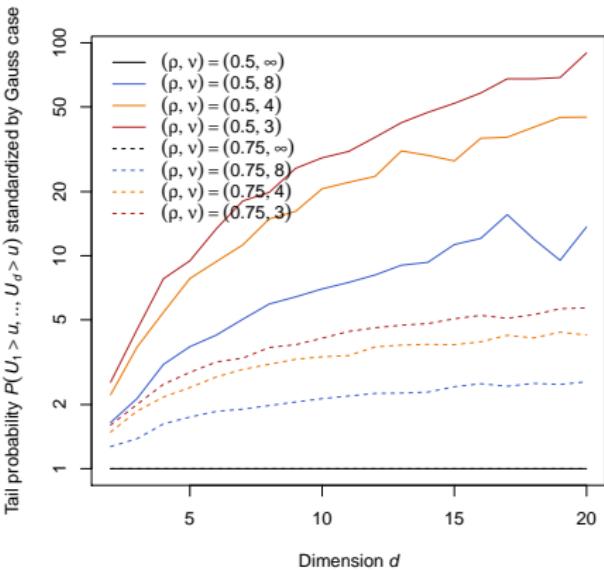
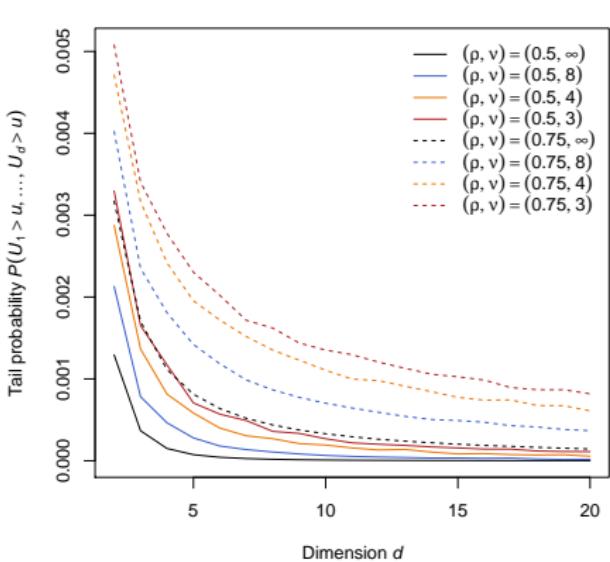
Note the different number of points in the bivariate tails (all models have the same Kendall's tau!)

Joint tail probabilities $\mathbb{P}(U_1 > u, U_2 > u)$ for $d = 2$



- **Left:** The higher ρ or the smaller ν , the larger $\mathbb{P}(U_1 > u, U_2 > u)$.
- **Right:** $u \mapsto \frac{\mathbb{P}(U_1 > u, U_2 > u)}{\mathbb{P}(V_1 > u, V_2 > u)}$ $\stackrel{\text{radial}}{=} \frac{C_{\nu, \rho}^t(u, u)}{C_{\rho}^{\text{Ga}}(u, u)}$ $\stackrel{\text{symm.}}{=}$

Joint tail probabilities $\mathbb{P}(U_1 > u, \dots, U_d > u)$ for $u = 0.99$



- Homogeneous P (off-diagonal entry ρ). Note the MC randomness.
- **Left:** Clear, less mass in corners in higher dimensions.
- **Right:** $d \mapsto \frac{\mathbb{P}(U_1 > u, \dots, U_d > u)}{\mathbb{P}(V_1 > u, \dots, V_d > u)}$ $\stackrel{\text{radial}}{=} \frac{C_{\nu, \rho}^t(u, \dots, u)}{C_{\rho}^{\text{Ga}}(u, \dots, u)}$ for $u = 0.99$.

Example 7.33 (Joint tail probabilities: an interpretation)

- Consider 5 daily returns $\mathbf{X} = (X_1, \dots, X_5)$ with pairwise correlations (all) $\rho = 0.5$. However, we are unsure about the best joint model.
- If the copula of \mathbf{X} is $C_{\rho=0.5}^{\text{Ga}}$, the probability that on any day all 5 returns fall below their $u = 0.01$ quantiles is

$$\begin{aligned}\mathbb{P}(X_1 \leq F_1^-(u), \dots, X_5 \leq F_5^-(u)) &= \mathbb{P}(U_1 \leq u, \dots, U_5 \leq u) \\ &\stackrel{\text{MC error}}{\approx} 7.48 \times 10^{-5}.\end{aligned}$$

In the long run such an event will happen once every $1/7.48 \times 10^{-5} \approx 13369$ trading days on average (\approx once every 51.4 years; assuming 260 trading days in a year).

- If the copula of \mathbf{X} is $C_{\nu=4, \rho=0.5}^t$, however, such an event will happen approximately 7.68 times more often, i.e., \approx once every 6.7 years. This gets worse the larger d !

Felix Salmon: “Recipe for Disaster: The Formula That Killed Wall Street”

$$\Pr[T_A < 1, T_B < 1] = \Phi_2(\Phi^{-1}(F_A(1)), \Phi^{-1}(F_B(1)), \gamma)$$

Here's what killed your 401(k) David X. Li's Gaussian copula function as first published in 2000. Investors exploited it as a quick—and fatally flawed—way to assess risk. A shorter version appears on this month's cover of Wired.

Probability

Specifically, this is a joint default probability—the likelihood that any two members of the pool (A and B) will both default. It's what investors are looking for, and the rest of the formula provides the answer.

Copula

This couples (hence the Latinate term copula) the individual probabilities associated with A and B to come up with a single number. Errors here massively increase the risk of the whole equation blowing up.

Survival times

The amount of time between now and when A and B can be expected to default. Li took the idea from a concept in actuarial science that charts what happens to someone's life expectancy when their spouse dies.

Distribution functions

The probabilities of how long A and B are likely to survive. Since these are not certainties, they can be dangerous: Small miscalculations may leave you facing much more risk than the formula indicates.

Equality

A dangerously precise concept, since it leaves no room for error. Clean equations help both quants and their managers forget that the real world contains a surprising amount of uncertainty, fuzziness, and precariousness.

Gamma

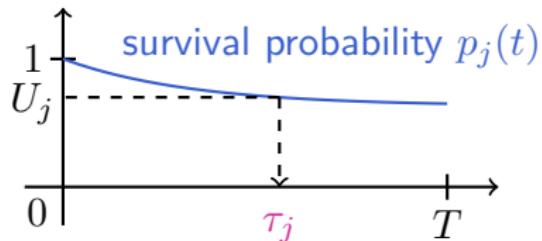
The all-powerful correlation parameter, which reduces correlation to a single constant—something that should be highly improbable, if not impossible. This is the magic number that made Li's copula function irresistible.

Application to credit risk

Intensity-based default model:

$$p_j(t) = \exp\left(-\int_0^t \lambda_j(s) ds\right)$$

$$\tau_j = \inf\{t \geq 0 : p_j(t) \leq U_j\}$$



Note: $\lambda_U = 0$ (as for the Gauss copula!)

⇒ (Almost) no joint defaults! (p_j typically very flat)

Copulas for the triggers U :

- 1) Li (2000): Gauss (Sibuya (1959)): $\lambda_U = 0$)
- 2) Schönbucher and Schubert (2001): Archimedean ($\lambda_U > 0$)
- 3) Hofert and Scherer (2011): nested Archimedean ($\lambda_U > 0$, hierarchies)

Typical application: CDO pricing models based on iTraxx data.

7.3.2 Rank correlations

Lemma 7.34

Let $\mathbf{X} \sim E_2(\mathbf{0}, \Sigma, \psi)$ with $\mathbb{P}(\mathbf{X} = \mathbf{0}) = 0$ and $\rho = P_{12} = \text{Cor}[\Sigma]_{12}$. Then

$$\mathbb{P}(X_1 > 0, X_2 > 0) = \frac{1}{4} + \frac{\arcsin \rho}{2\pi}.$$

Proof.

- Note that $\mathbf{Y} = \begin{pmatrix} 1/\sqrt{\sigma_{11}} & 0 \\ 0 & 1/\sqrt{\sigma_{22}} \end{pmatrix} \mathbf{X} \sim E_2(\mathbf{0}, P, \psi)$ with $P = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.
- Let $A = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix}$ so that $AA^\top = P$. Then $\mathbf{Y} \stackrel{d}{=} RA\mathbf{U} \stackrel{d}{=} RA \begin{pmatrix} \cos \Theta \\ \sin \Theta \end{pmatrix}$, $\Theta \sim U[-\pi, \pi]$ independent of R .
- With $\varphi = \arcsin \rho$, we have $\mathbf{Y} \stackrel{d}{=} R \begin{pmatrix} \cos \Theta \\ \sin \varphi \cos \Theta + \cos \varphi \sin \Theta \end{pmatrix} = \begin{pmatrix} \cos \Theta \\ \sin(\varphi + \Theta) \end{pmatrix}$.
- Thus $\mathbb{P}(X_1 > 0, X_2 > 0) = \mathbb{P}(Y_1 > 0, Y_2 > 0) = \mathbb{P}(\cos \Theta > 0, \sin(\varphi + \Theta) > 0) = \mathbb{P}(\Theta \in (-\frac{\pi}{2}, \frac{\pi}{2}), \varphi + \Theta \in (0, \pi)) = \mathbb{P}(\Theta \in (-\frac{\pi}{2}, \frac{\pi}{2}), \Theta \in (-\varphi, \pi - \varphi)) = \mathbb{P}(\Theta \in (-\varphi, \frac{\pi}{2})) = (\frac{\pi}{2} - (-\varphi))/(2\pi)$. □

Lemma 7.35 (Representation of Spearman's rho)

Let $(U_1, U_2) \sim C$ and $\tilde{U}_1, \bar{U}_2 \stackrel{\text{ind.}}{\sim} U[0, 1]$ be independent. Then $\rho_S = \rho_S(U_1, U_2) = 12\mathbb{P}(U_1 \leq \tilde{U}_1, U_2 \leq \bar{U}_2) - 3$.

Proof. $12\mathbb{P}(U_1 \leq \tilde{U}_1, U_2 \leq \bar{U}_2) - 3 = 12\mathbb{E}[\mathbb{P}(\tilde{U}_1 > U_1, \bar{U}_2 > U_2 | U_1, U_2)] - 3 = 12\mathbb{E}[(1 - U_1)(1 - U_2)] - 3 = 12\mathbb{E}[U_1 U_2] - 3 = \rho_S(U_1, U_2)$. \square

Theorem 7.36 (Spearman's rho for Gauss copulas)

Let $(U_1, U_2) \sim C_\rho^{\text{Ga}}$. Then $\rho_S = \rho_S(U_1, U_2) = \frac{6}{\pi} \arcsin \frac{\rho}{2}$.

Proof. Let $(X_1, X_2) \sim N_2(\mathbf{0}, P)$ with $P_{12} = \rho$, indep. of $\tilde{X}_1, \bar{X}_2 \stackrel{\text{ind.}}{\sim} N(0, 1)$.

With $\mathbf{Y} = (\tilde{X}_1 - X_1, \bar{X}_2 - X_2) \sim N_2(\mathbf{0}, I_2 + P)$ ($\text{Cor}[Y_1, Y_2] = \rho/2$)

$$\rho_S \stackrel{\text{L. 7.35}}{\underset{\Phi^{-1}}{=}} 12\mathbb{P}(X_1 \leq \tilde{X}_1, X_2 \leq \bar{X}_2) - 3 = 12\mathbb{P}(Y_1 \geq 0, Y_2 \geq 0) - 3$$

$$\stackrel{\text{cont.}}{\underset{\text{L. 7.34}}{=}} 12\left(\frac{1}{4} + \frac{\arcsin(\rho/2)}{2\pi}\right) - 3 = \frac{6}{\pi} \arcsin \frac{\rho}{2}. \quad \square$$

Proposition 7.37 (Spearman's rho for normal variance mixtures)

Let $\mathbf{X} \sim M_2(\mathbf{0}, P, \hat{H})$ with $\mathbb{P}(\mathbf{X} = \mathbf{0}) = 0$, $\rho = P_{12}$. Then

$$\rho_S = \frac{6}{\pi} \mathbb{E} \left[\arcsin \frac{W\rho}{\sqrt{(W + \tilde{W})(W + \bar{W})}} \right],$$

for $W, \tilde{W}, \bar{W} \stackrel{\text{ind.}}{\sim} H$ with Laplace–Stieltjes transform \hat{H} .

Proof. $\mathbf{X} \stackrel{d}{=} \sqrt{W}\mathbf{Z}$ for $\mathbf{Z} \sim N_2(\mathbf{0}, P)$. Let $\tilde{Z}, \bar{Z} \sim N(0, 1)$ and assume $\mathbf{Z}, \tilde{Z}, \bar{Z}, W, \tilde{W}$ and \bar{W} are all independent. Let

$$\tilde{X} = \sqrt{\tilde{W}}\tilde{Z}, \quad \bar{X} = \sqrt{\bar{W}}\bar{Z},$$

$$Y_1 = X_1 - \tilde{X} = \sqrt{W}Z_1 - \sqrt{\tilde{W}}\tilde{Z},$$

$$Y_2 = X_2 - \bar{X} = \sqrt{W}Z_2 - \sqrt{\bar{W}}\bar{Z}.$$

$$\begin{aligned}
\rho_S(X_1, X_2) &\stackrel{\text{L. 7.35}}{\underset{\Phi^{-1}}{=}} 12\mathbb{P}(X_1 \leq \tilde{X}_1, X_2 \leq \bar{X}_2) - 3 \\
&= 6\mathbb{P}((X_1 - \tilde{X}_1)(X_2 - \bar{X}_2) > 0) - 3 \\
&= 3(2\mathbb{E}[\mathbb{P}(Y_1 Y_2 > 0 \mid W, \tilde{W}, \bar{W})] - 1) \\
&= 3(4\mathbb{E}[\mathbb{P}(Y_1 > 0, Y_2 > 0 \mid W, \tilde{W}, \bar{W})] - 1).
\end{aligned}$$

Now note that $\mathbf{Y} \mid W, \tilde{W}, \bar{W} \sim N_2(\mathbf{0}, (\begin{smallmatrix} W+\tilde{W} & W\rho \\ W\rho & W+\bar{W} \end{smallmatrix}))$ with $\rho(Y_1, Y_2) = \frac{W\rho}{\sqrt{(W+\tilde{W})(W+\bar{W})}}$. Apply Lemma 7.34 to see that this equals

$$\rho_S(X_1, X_2) = 3\left(4\mathbb{E}\left[\frac{1}{4} + \frac{\arcsin \rho}{2\pi}\right] - 1\right) = \frac{12}{2\pi}\mathbb{E}[\arcsin \rho(Y_1, Y_2)]. \quad \square$$

Proposition 7.38 (Kendall's tau for elliptical distributions)

Let $\mathbf{X} \sim E_2(\mathbf{0}, P, \psi)$ with $\mathbb{P}(\mathbf{X} = \mathbf{0}) = 0$, $\rho = P_{12}$. Then $\tau = \frac{2}{\pi} \arcsin \rho$.

Proof.

- Let (X'_1, X'_2) be an independent copy of (X_1, X_2) . We have already seen that $\tau = 2\mathbb{P}((X_1 - X'_1)(X_2 - X'_2) > 0) - 1$.
- With $\mathbf{X} \stackrel{d}{=} R\mathbf{AU}$ and $\mathbf{X}' \stackrel{d}{=} R'\mathbf{AU}' (\stackrel{d}{=} -\mathbf{X}')$ we have $\mathbf{Y} = \mathbf{X} - \mathbf{X}' \stackrel{d}{=} \mathbf{0} + A(R\mathbf{U} - R'\mathbf{U}')$. Note that the characteristic function of $-\mathbf{X}'$ is $\phi_{-\mathbf{X}'}(\mathbf{t}) = \phi_{\mathbf{X}'}(\mathbf{t}) = \phi_{\mathbf{X}}(\mathbf{t})$ so that $\phi_{\mathbf{Y}}(\mathbf{t}) \underset{\text{ind.}}{=} \phi_{\mathbf{X}}(\mathbf{t})\phi_{-\mathbf{X}'}(\mathbf{t}) = \phi_{\mathbf{X}}(\mathbf{t})^2$, hence $\mathbf{Y} \sim E_2(\mathbf{0}, P, \psi^2)$.
- We thus obtain that

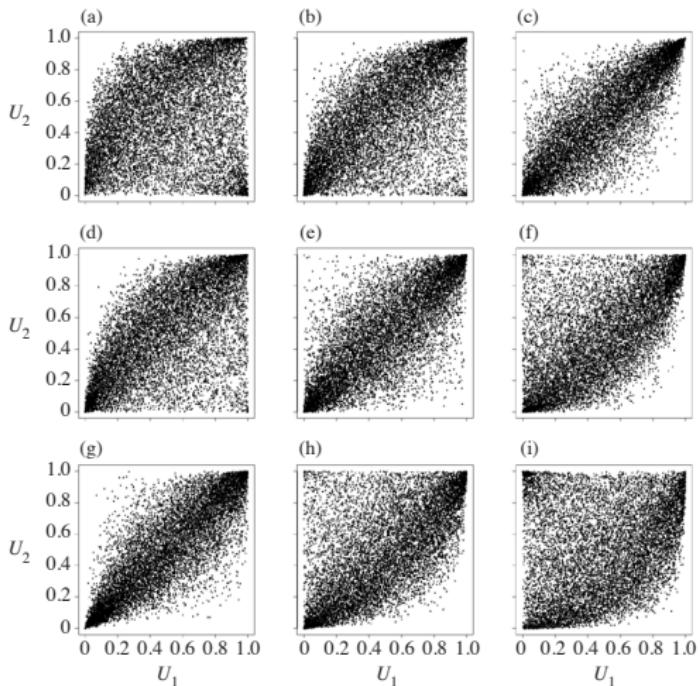
$$\begin{aligned}\tau &= 2\mathbb{P}(Y_1 Y_2 > 0) - 1 = 2(\mathbb{P}(Y_1 > 0, Y_2 > 0) + \mathbb{P}(Y_1 < 0, Y_2 < 0)) - 1 \\ &= 4\mathbb{P}(\mathbf{Y} > \mathbf{0}) - 1 \stackrel{\substack{\text{cont.} \\ \text{L.7.34}}}{=} \frac{2}{\pi} \arcsin \rho.\end{aligned}$$

For a generalization to componentwise n.d. \mathbf{X} , see Lindskog et al. (2002).

7.3.3 Skewed normal mixture copulas

- . . . are the copulas of normal mixture distributions which are not elliptical, e.g., the *skewed t copula* $C_{\nu, P, \gamma}^t$ is the copula of $\mathbf{X} \sim \text{GH}_d(-\frac{\nu}{2}, \nu, 0, \boldsymbol{\mu}, \Sigma, \gamma)$.
- The main advantage of $C_{\nu, P, \gamma}^t$ over $C_{\nu, P}^t$ is its radial asymmetry (e.g., for modeling $\lambda_L \neq \lambda_U$)
- $C_{\nu, P, \gamma}^t$ can be sampled as other implicit copulas; see Algorithm 7.12 (the evaluation of the margins requires numerical integration of a skewed *t* density; note that $X_j \sim \text{GH}_1(-\frac{\nu}{2}, \nu, 0, \mu_j, \Sigma_{jj}, \gamma_j)$, $j \in \{1, \dots, d\}$).

10 000 samples from $C^t_{\nu=5, \rho=0.8, \gamma=0.8(\mathbb{1}_{\{i<2\}}-\mathbb{1}_{\{i>2\}}, \mathbb{1}_{\{j>2\}}-\mathbb{1}_{\{j<2\}})}$



- (a) $\gamma = (-0.8, -0.8)$
(b) $\gamma = (-0.8, 0)$
(c) $\gamma = (-0.8, 0.8)$
(d) $\gamma = (0, -0.8)$
(e) $\gamma = (0, 0)$
(f) $\gamma = (0, 0.8)$
(g) $\gamma = (-0.8, -0.8)$
(h) $\gamma = (-0.8, 0)$
(i) $\gamma = (-0.8, 0.8)$

7.3.4 Grouped normal mixture copulas

- ... are copulas which attach together a set of normal mixture copulas, e.g., a *grouped t copula* is the copula of

$$\mathbf{X} = (\sqrt{W_1}Y_1, \dots, \sqrt{W_1}Y_{s_1}, \dots, \sqrt{W_S}Y_{s_1+\dots+s_{S-1}+1}, \dots, \sqrt{W_S}Y_d)$$

for $(W_1, \dots, W_S) \sim M(\text{IG}(\frac{\nu_1}{2}, \frac{\nu_1}{2}), \dots, \text{IG}(\frac{\nu_S}{2}, \frac{\nu_S}{2}))$ and $\mathbf{Y} \sim N_d(\mathbf{0}, P)$ (so $\mathbf{Y} \stackrel{d}{=} A\mathbf{Z}$ as before); see Demarta and McNeil (2005) for more details.

- Clearly, the marginals are *t* distributed, hence

$$\mathbf{U} = (t_{\nu_1}(X_1), \dots, t_{\nu_1}(X_{s_1}), \dots, t_{\nu_S}(X_{s_1+\dots+s_{S-1}+1}), \dots, t_{\nu_S}(X_d))$$

follows a *grouped t copula*. This is straightforward to simulate.

- It can be fitted with pairwise inversion of Kendall's tau.
- If $S = d$, grouped *t* copulas are also known as *generalized t copulas*; see Luo and Shevchenko (2010).

7.4 Archimedean copulas

- *Archimedean copulas* are explicit copulas, arising from a construction principle of copulas directly (*conditional independence approach*).
- **Recall:** An (*Archimedean*) generator ψ is a function $\psi : [0, \infty) \rightarrow [0, 1]$ which is \downarrow on $[0, \inf\{t : \psi(t) = 0\}]$ and satisfies $\psi(0) = 1$, $\psi(\infty) = \lim_{t \rightarrow \infty} \psi(t) = 0$; the set of all generators is denoted by Ψ .

7.4.1 Bivariate Archimedean copulas

Theorem 7.39 (Bivariate Archimedean copulas)

For $\psi \in \Psi$, $C(u_1, u_2) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_2))$ is a copula if and only if ψ is convex.

Proof. See Nelsen (1999, pp. 91). □

- **Recall:** If ψ is strict, $\lambda_L = 2 \lim_{t \rightarrow \infty} \frac{\psi'(2t)}{\psi'(t)}$ and $\lambda_U = 2 - 2 \lim_{t \downarrow 0} \frac{\psi'(2t)}{\psi'(t)}$.

Proposition 7.40 (Kendall's tau for Archimedean copulas)

For a strict and twice-continuously differentiable ψ ,

$$\tau = 1 - 4 \int_0^\infty t(\psi'(t))^2 dt = 1 + 4 \int_0^1 \frac{\psi^{-1}(t)}{(\psi^{-1}(t))'} dt.$$

Proof. See Nelsen (1999, pp. 130). □

Example 7.41 (Outer power Clayton copula)

An *outer power Clayton copula* is generated by $\tilde{\psi}(t) = \psi(t^{1/\beta})$, $\beta \in [1, \infty)$, where $\psi(t) = (1+t)^{-1/\theta}$, $\theta \in (0, \infty)$, denotes the Clayton generator (with $\tau = \theta/(\theta+2)$). The quantities corresponding to $\tilde{\psi}$ are

$$\tilde{\tau} = 1 - \frac{1 - \tau}{\beta} = 1 - \frac{2}{\beta(\theta+2)}, \quad \tilde{\lambda}_L = 2^{-1/(\theta\beta)}, \quad \tilde{\lambda}_U = 2 - 2^{1/\beta}.$$

The most widely used one-parameter Archimedean copulas are (see the R package `copula`):

Family	θ	$\psi(t)$	$V \sim F = \mathcal{L}\mathcal{S}^{-1}[\psi]$
A	$[0, 1)$	$(1 - \theta)/(\exp(t) - \theta)$	$\text{Geo}(1 - \theta)$
C	$(0, \infty)$	$(1 + t)^{-1/\theta}$	$\Gamma(1/\theta, 1)$
F	$(0, \infty)$	$-\log(1 - (1 - e^{-\theta}) \exp(-t))/\theta$	$\text{Log}(1 - e^{-\theta})$
G	$[1, \infty)$	$\exp(-t^{1/\theta})$	$S(1/\theta, 1, \cos^\theta(\pi/(2\theta)), \mathbb{1}_{\{\theta=1\}}; 1)$
J	$[1, \infty)$	$1 - (1 - \exp(-t))^{1/\theta}$	$\text{Sibuya}(1/\theta)$

Family	τ	λ_L	λ_U
A	$1 - 2(\theta + (1 - \theta)^2 \log(1 - \theta))/(3\theta^2)$	0	0
C	$\theta/(\theta + 2)$	$2^{-1/\theta}$	0
F	$1 + 4(D_1(\theta) - 1)/\theta$	0	0
G	$(\theta - 1)/\theta$	0	$2 - 2^{1/\theta}$
J	$1 - 4 \sum_{k=1}^{\infty} 1/(k(\theta k + 2)(\theta(k - 1) + 2))$	0	$2 - 2^{1/\theta}$

7.4.2 Multivariate Archimedean copulas

ψ is *completely monotone (c.m.)* if $(-1)^k \psi^{(k)}(t) \geq 0$ for all $t \in (0, \infty)$ and all $k \in \mathbb{N}_0$. The set of all c.m. generators is denoted by Ψ_∞ .

Theorem 7.42 (Kimberling (1974))

If $\psi \in \Psi$, $C(\mathbf{u}) = \psi\left(\sum_{j=1}^d \psi^{-1}(u_j)\right)$ is a copula $\forall d$ if and only if $\psi \in \Psi_\infty$.

Proof. See Kimberling (1974) or Hofert (2010, p. 54). □

Bernstein's Theorem characterizes all $\psi \in \Psi_\infty$.

Theorem 7.43 (Bernstein (1928))

$\psi(0) = 1$, ψ c.m. if and only if $\psi(t) = \mathbb{E}[\exp(-tV)]$ for $V \sim F$, $V \geq 0$.

Proof. See Feller (1971, pp. 439). □

We thus use the notation $\psi(t) = \mathcal{LS}[F](t)$ or $F(x) = \mathcal{LS}^{-1}[\psi](x)$.

Proposition 7.44 (Stochastic representation, related properties)

Let $\psi \in \Psi_\infty$ with $V \sim F = \mathcal{LS}^{-1}[\psi]$ and $E_1, \dots, E_d \stackrel{\text{ind.}}{\sim} \text{Exp}(1)$ independent of V . Then

- 1) The survival copula of $\mathbf{X} = (E_1/V, \dots, E_d/V)$ is Archimedean with generator ψ .
- 2) $\mathbf{U} = (\psi(X_1), \dots, \psi(X_d)) \sim C$
- 3) The components of \mathbf{U} are conditionally independent given V with conditional df $\mathbb{P}(U_j \leq u | V = v) = \exp(-v\psi^{-1}(u))$.

Proof.

- 1) The joint survival function of \mathbf{X} is given by

$$\begin{aligned}\bar{H}(\mathbf{x}) &= \mathbb{P}(X_j > x_j \ \forall j) = \int_0^\infty \mathbb{P}(E_j/V > x_j \ \forall j | V = v) dF(v) \\ &= \int_0^\infty \mathbb{P}(E_j > vx_j \ \forall j) dF(v) = \int_0^\infty \prod_{j=1}^d \exp(-vx_j) dF(v)\end{aligned}$$

$$= \int_0^\infty \exp\left(-v \sum_{j=1}^d x_j\right) dF(v) = \psi\left(\sum_{j=1}^d x_j\right).$$

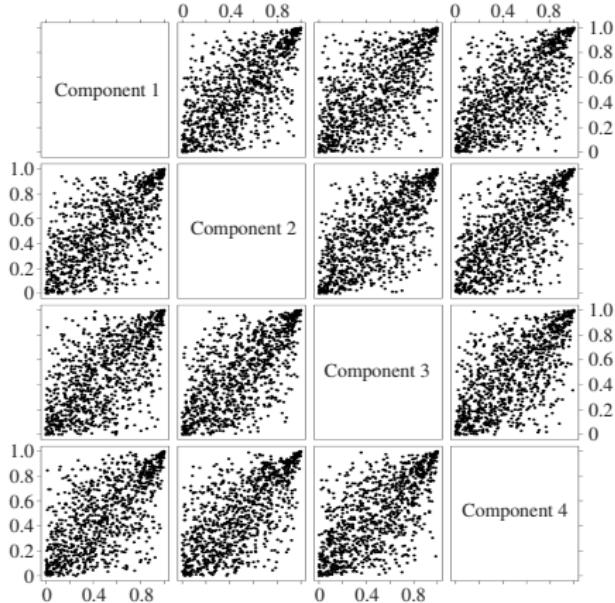
The j th marginal survival function is thus (set $x_k = 0 \forall k \neq j$)
 $\bar{F}_j(x_j) = \mathbb{P}(X_j > x_j) = \psi(x_j)$ (\downarrow and continuous) and therefore
 $\hat{C}(\mathbf{u}) = \bar{H}(\bar{F}_1^-(u_1), \dots, \bar{F}_d^-(u_d)) = \psi(\sum_{j=1}^d \psi^{-1}(u_j)).$

- 2) $\mathbb{P}(\mathbf{U} \leq \mathbf{u}) = \mathbb{P}(X_j > \psi^{-1}(u_j) \forall j) \stackrel{\text{1)}}{=} \psi(\sum_{j=1}^d \psi^{-1}(u_j)).$
 - 3) Cond. indep. is clear. Furthermore, $\mathbb{P}(U_j \leq u | V = v) = \mathbb{P}(X_j > \psi^{-1}(u) | V = v) = \mathbb{P}(E_j > v\psi^{-1}(u)) = \exp(-v\psi^{-1}(u))$. \square
- We call all Archimedean copulas with $\psi \in \Psi_\infty$ *LT-Archimedean copulas*.

Algorithm 7.45 (Marshall and Olkin (1988))

- 1) Sample $V \sim F = \mathcal{LS}^{-1}[\psi]$ (df corresponding to ψ ; see tables above).
- 2) Sample $E_1, \dots, E_d \stackrel{\text{ind.}}{\sim} \text{Exp}(1)$ independently of V .
- 3) Return $\mathbf{U} = (\psi(E_1/V), \dots, \psi(E_d/V))$ (conditional independence).

1000 samples of a 4-dim. Gumbel copula ($\tau = 0.5$; $\lambda_U \approx 0.5858$)



- For fixed d , c.m. can be relaxed to d -monotonicity; see McNeil and Nešlehová (2009).
- Various non-exchangeable extensions to Archimedean copulas exist.

7.5 Fitting copulas to data

- Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. copies of $\mathbf{X} \sim H$ with continuous margins F_1, \dots, F_d and corresponding copula C .
- Assume
 - ▶ $F_j = F_j(\cdot; \boldsymbol{\theta}_{0,j})$ for some $\boldsymbol{\theta}_{0,j} \in \Theta_j$, $j \in \{1, \dots, d\}$;
 $(F_j(\cdot; \boldsymbol{\theta}_j) \text{ continuous } \forall \boldsymbol{\theta}_j \in \Theta_j, j \in \{1, \dots, d\})$
 - ▶ $C = C(\cdot; \boldsymbol{\theta}_{0,C})$ for some $\boldsymbol{\theta}_{0,C} \in \Theta_C$.

Thus H has the true but unknown parameter vector $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,C}^\top, \boldsymbol{\theta}_{0,1}^\top, \dots, \boldsymbol{\theta}_{0,d}^\top)^\top$ to be estimated.

- Here, we focus particularly on $\boldsymbol{\theta}_{0,C}$. Whenever necessary, we assume that the margins F_1, \dots, F_d and the copula C are absolutely continuous with corresponding densities f_1, \dots, f_d and c , respectively.
- We assume the chosen copula to be appropriate (w.r.t. symmetry, tail dependence etc.).

7.5.1 Method-of-moments using rank correlation

- We focus on one-parameter copulas here, i.e., $\theta_{0,C} = \theta_{0,C}$.
- For $d = 2$, Genest and Rivest (1993) suggested to estimate $\theta_{0,C}$ by solving $\tau(\theta_C) = \hat{\tau}_n$ w.r.t. θ_C , i.e.,

$$\hat{\theta}_{n,C}^{\text{IKTE}} = \tau^{-1}(\hat{\tau}_n), \quad (\text{inversion of Kendall's tau estimator (IKTE)})$$

where $\tau(\cdot)$ denotes Kendall's tau as a function in θ and $\hat{\tau}_n$ is the sample version of Kendall's tau (computed via (38) from $\mathbf{X}_1, \dots, \mathbf{X}_n$ or pseudo-observations $\mathbf{U}_1, \dots, \mathbf{U}_n$; see later).

- The standardized dispersion matrix P for elliptical copulas can be estimated via *pairwise inversion of Kendall's tau*; see McNeil et al. (2015, Example 7.56). If $\hat{\tau}_{n,j_1j_2}$ denotes the sample version of Kendall's tau for data pair (j_1, j_2) , then $\hat{P}_{n,j_1j_2}^{\text{IKTE}} = \sin(\frac{\pi}{2}\hat{\tau}_{n,j_1j_2})$; see Proposition 7.38. For obtaining a proper correlation matrix P (positive semi-definite), see the **eigenvalue method** (McNeil et al. (2015, Algorithm 7.57)) or **Higham (2002) (Matrix::nearPD())**.

- Extensions to $d > 2$ are used for one-parameter exchangeable copula models; see Berg (2009) or Kojadinovic and Yan (2010). The corresponding *pairwise inversion of Kendall's tau estimator* is

$$\hat{\theta}_{n,C}^{\text{IKTE}} = \tau^{-1} \left(\binom{d}{2}^{-1} \sum_{1 \leq j_1 < j_2 \leq d} \hat{\tau}_{n,j_1 j_2} \right),$$

where $\hat{\tau}_{n,j_1 j_2}$ is the inversion of Kendall's tau estimator for pair (j_1, j_2) .

Remark 7.46

- Clayton copula:** $\tau^{-1}(x) = \frac{2x}{1-x}$; **Gumbel copula:** $\tau^{-1}(x) = \frac{1}{1-x}$.
- Gauss copula:** Here we could also use Spearman's rho based on

$$\rho_S \stackrel{\text{Th.7.36}}{=} \frac{6}{\pi} \arcsin \frac{\rho}{2} \approx \rho.$$

The latter approximation error is comparably small, so that the matrix of pairwise sample versions of Spearman's rho is an estimator for P .

- t copula:** Use \hat{P}_n^{IKTE} to estimate P and maximize the likelihood in ν .

7.5.2 Forming a pseudo-sample from the copula

- $\mathbf{X}_1, \dots, \mathbf{X}_n$ (as good as) never has $U[0, 1]$ margins. For applying the “copula approach” we thus need *pseudo-observations* from C .
- In general, we take $\hat{\mathbf{U}}_1, \dots, \hat{\mathbf{U}}_n$ with

$$\hat{\mathbf{U}}_i = (\hat{U}_{i1}, \dots, \hat{U}_{id}) = (\hat{F}_1(X_{i1}), \dots, \hat{F}_d(X_{id})),$$

where \hat{F}_j denotes an estimator of F_j ; see Lemma 7.9. Note that $\hat{\mathbf{U}}_1, \dots, \hat{\mathbf{U}}_n$ are typically neither independent (even if $\mathbf{X}_1, \dots, \mathbf{X}_n$ are) nor perfectly $U[0, 1]$.

- Possible choices for \hat{F}_j :
 - 1) Non-parametric estimators with scaled empirical dfs (to avoid density evaluation on the boundary of $[0, 1]^d$), so

$$\hat{U}_{ij} = \frac{n}{n+1} \hat{F}_{n,j}(X_{ij}) = \frac{R_{ij}}{n+1}, \quad (39)$$

where R_{ij} denotes the *rank* of X_{ij} among all X_{1j}, \dots, X_{nj} .

- 2) Parametric estimators (such as Student t , Pareto, etc.; typically if n is small). In this case, one often still uses (39) for estimating $\theta_{0,C}$ (to keep the error due to misspecification of the margins small).
- 3) EVT-based. Bodies are modeled empirically; tails semiparametrically via GPD.

7.5.3 Maximum likelihood estimation

The (classical) maximum likelihood estimator

- By Sklar's Theorem, the density of H is given by

$$h(\mathbf{x}; \boldsymbol{\theta}_0) = c(F_1(x_1; \boldsymbol{\theta}_{0,1}), \dots, F_d(x_d; \boldsymbol{\theta}_{0,d}); \boldsymbol{\theta}_{0,C}) \prod_{j=1}^d f_j(x_j; \boldsymbol{\theta}_{0,j}).$$

- The log-likelihood based on $\mathbf{X}_1, \dots, \mathbf{X}_n$ is thus

$$\ell(\boldsymbol{\theta}; \mathbf{X}_1, \dots, \mathbf{X}_n) = \sum_{i=1}^n \ell(\boldsymbol{\theta}; \mathbf{X}_i)$$

$$= \sum_{i=1}^n \ell_C(\boldsymbol{\theta}_C; F_1(X_{i1}; \boldsymbol{\theta}_1), \dots, F_d(X_{id}; \boldsymbol{\theta}_d)) + \sum_{i=1}^n \sum_{j=1}^d \ell_j(\boldsymbol{\theta}_j; X_{ij}), \quad (40)$$

where

$$\begin{aligned}\ell_C(\boldsymbol{\theta}_C; u_1, \dots, u_d) &= \log c(u_1, \dots, u_d; \boldsymbol{\theta}_C) \\ \ell_j(\boldsymbol{\theta}_j; x) &= \log f_j(x; \boldsymbol{\theta}_j), \quad j \in \{1, \dots, d\}.\end{aligned}$$

- The *maximum likelihood estimator (MLE)* of $\boldsymbol{\theta}_0$ is

$$\hat{\boldsymbol{\theta}}_n^{\text{MLE}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \ell(\boldsymbol{\theta}; \mathbf{X}_1, \dots, \mathbf{X}_n). \quad (41)$$

The optimization in (41) is typically done by numerical means. Note that this can be quite demanding, especially in high dimensions.

The inference functions for margins estimator

- Due to the decomposition (40), Joe and Xu (1996) suggested the two-step estimation approach:

Step 1: For $j \in \{1, \dots, d\}$, estimate $\theta_{0,j}$ by its MLE $\hat{\theta}_{n,j}^{\text{MLE}}$.

Step 2: Estimate $\theta_{0,C}$ by

$$\hat{\theta}_{n,C}^{\text{IFME}} = \underset{\theta_C \in \Theta_C}{\operatorname{argsup}} \ell(\theta_C, \hat{\theta}_{n,1}^{\text{MLE}}, \dots, \hat{\theta}_{n,d}^{\text{MLE}}; \mathbf{X}_1, \dots, \mathbf{X}_n).$$

The corresponding *inference functions for margins estimator (IFME)* of θ_0 is thus

$$\hat{\theta}_n^{\text{IFME}} = (\hat{\theta}_{n,C}^{\text{IFME}}, \hat{\theta}_{n,1}^{\text{MLE}}, \dots, \hat{\theta}_{n,d}^{\text{MLE}})$$

- This is typically much easier to compute than $\hat{\theta}_n^{\text{MLE}}$ while providing good results; see Joe and Xu (1996) or Kim et al. (2007). If H is $N_d(\mu, \Sigma)$, then $\hat{\theta}_n^{\text{IFME}} = \hat{\theta}_n^{\text{MLE}}$.
- $\hat{\theta}_n^{\text{IFME}}$ can also be used as initial value for computing $\hat{\theta}_n^{\text{MLE}}$.
- In terms of likelihood equations, $\hat{\theta}_n^{\text{IFME}}$ compares to $\hat{\theta}_n^{\text{MLE}}$ as follows:

$$\hat{\theta}_n^{\text{MLE}} \text{ solves } \left(\frac{\partial}{\partial \theta_C} \ell, \frac{\partial}{\partial \theta_1} \ell, \dots, \frac{\partial}{\partial \theta_d} \ell \right) = \mathbf{0},$$

$$\hat{\boldsymbol{\theta}}_n^{\text{IFME}} \text{ solves } \left(\frac{\partial}{\partial \boldsymbol{\theta}_C} \ell, \frac{\partial}{\partial \boldsymbol{\theta}_1} \ell_1, \dots, \frac{\partial}{\partial \boldsymbol{\theta}_d} \ell_d \right) = \mathbf{0},$$

where $\ell = \ell(\boldsymbol{\theta}; \mathbf{X}_1, \dots, \mathbf{X}_n)$, $\ell_j = \ell_j(\boldsymbol{\theta}_j; X_{1j}, \dots, X_{nj}) = \sum_{i=1}^n \ell_j(\boldsymbol{\theta}_j; X_{ij})$.

Example 7.47 (A computationally convincing example)

Suppose $X_j \sim N(\mu_j, \sigma_j^2)$, $j \in \{1, \dots, d\}$, for $d = 100$, and C has (just) one parameter.

- MLE requires to solve a 201-dimensional optimization problem.
- IFME only requires 100 optimizations in two dimensions and 1 one-dimensional optimization.

If the marginals are estimated parametrically one often still uses the pseudo-observations built from the marginal empirical dfs to estimate $\boldsymbol{\theta}_{0,C}$ (see MPLE below) in order to avoid misspecification of the margins (if n is sufficiently large).

The maximum pseudo-likelihood estimator

- The *maximum pseudo-likelihood estimator (MPLE)*, introduced by Genest et al. (1995), works similarly as $\hat{\theta}_n^{\text{IFME}}$, but estimates the margins non-parametrically:

Step 1: Compute rank-based pseudo-observations $\hat{U}_1, \dots, \hat{U}_n$.

Step 2: Estimate $\theta_{0,C}$ by

$$\hat{\theta}_{n,C}^{\text{MPLE}} = \underset{\theta_C \in \Theta_C}{\operatorname{argsup}} \sum_{i=1}^n \ell_C(\theta_C; \hat{U}_{i1}, \dots, \hat{U}_{id}) = \underset{\theta_C \in \Theta_C}{\operatorname{argsup}} \sum_{i=1}^n \log c(\hat{U}_i; \theta_C).$$

- Genest and Werker (2002) show that $\hat{\theta}_{n,C}^{\text{MPLE}}$ is not asymptotically efficient in general.
- Kim et al. (2007) compare $\hat{\theta}_n^{\text{MLE}}$, $\hat{\theta}_n^{\text{IFME}}$, and $\hat{\theta}_{n,C}^{\text{MPLE}}$ in a simulation study ($d = 2$ only!) and argue in favor of $\hat{\theta}_{n,C}^{\text{MPLE}}$ overall, especially w.r.t. robustness against misspecification of the margins; but see Embrechts and Hofert (2013b) for $d \gg 2$.

Example 7.48 (Fitting the Gauss copula)

- The (copula-related) log-likelihood ℓ_C is

$$\ell_C(P; \hat{U}_1, \dots, \hat{U}_n) = \sum_{i=1}^n \ell_C(P; \hat{U}_i) \stackrel{\text{Eq. (36)}}{=} \sum_{i=1}^n \log c_P^{\text{Ga}}(\hat{U}_i).$$

For maximization over all correlation matrices P , we can use the Cholesky factor A as reparameterization and maximize over all lower triangular matrices A with 1s on the diagonal; still this is $\mathcal{O}(d^2)$.

- An approximate solution can be found via $c_P^{\text{Ga}}(\mathbf{u}) \underset{\text{Sklar}}{=} \frac{h_P(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))}{\prod_{j=1}^d \varphi(\Phi^{-1}(u_j))}$, where h_P denotes the density of a $N(\mathbf{0}, P)$ distribution. Thus

$$\operatorname{argsup}_P \ell_C(P; \hat{U}_1, \dots, \hat{U}_n) = \operatorname{argsup}_P h_P(\hat{Y}_1, \dots, \hat{Y}_n), \quad \hat{Y}_i = \Phi^{-1}(\hat{U}_i).$$

Now optimize over all covariance matrices Σ (assuming \hat{Y}_i to be independent!) to obtain the analytic solution $\hat{\Sigma} = \frac{1}{n(-1)} \sum_{i=1}^n \hat{Y}_i \hat{Y}_i^\top$ which is typically close to being a correlation matrix. Thus take $\hat{P} = \text{Cor}[\hat{\Sigma}]$.

- Alternatively, use pairwise inversion of Spearman's rho or Kendall's tau.

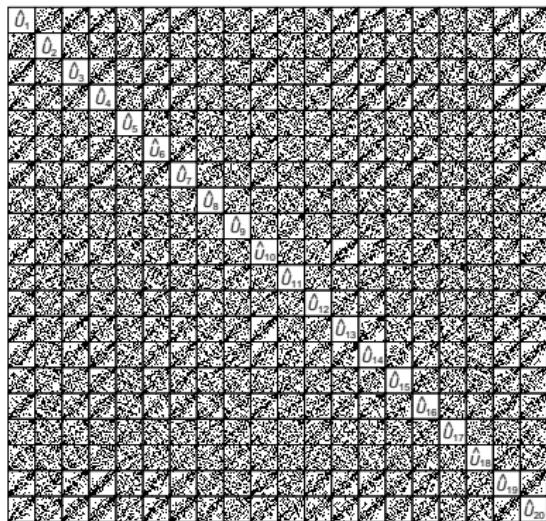
Example 7.49 (Fitting the t copula)

- For small d , maximize the likelihood (as for the Gauss copula case) over all correlation matrices (via reparameterization in terms of the Cholesky factor A) and the d.o.f. ν .
- For moderate/larger d , do:
 - 1) Estimate P via pairwise inversion of Kendall's tau (see above).
 - 2) Plug \hat{P} into the likelihood and maximize it w.r.t. ν to obtain $\hat{\nu}$.

Estimation is only one side of the coin. The other is *goodness-of-fit* (i.e., to find out whether our estimated model indeed represents the given data well) and *model selection* (i.e., to decide which model is best among all adequate fitted models). Goodness-of-fit can be (computationally) challenging, particularly for large d .

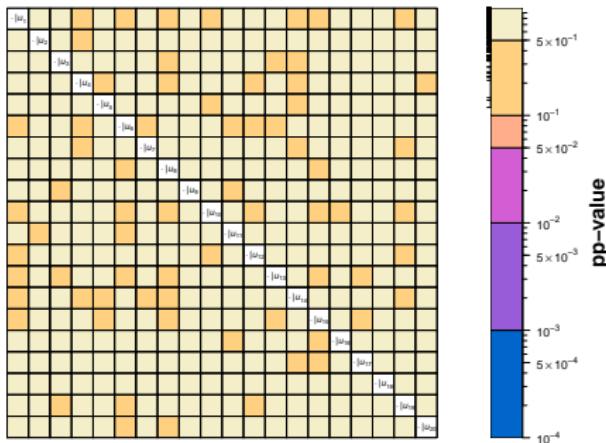
A graphical goodness-of-fit approach by Hofert and Mächler (2014); see `demo(gof_graph)` (daily log-returns of SMI, 2011-09-09–2012-03-28).

Pseudo-observations of the log-returns of the SMI



Pairwise Rosenblatt transformed pseudo-observations

to test $H_0: C$ is $t_{11.96}$



References

- Acharya, B. V., Cooley, V. V., Richardson, M., and Walter, I. (2009), Manufacturing tail risk: A perspective on the financial crisis of 2007–2009, *Foubdations and Trends in Finance*, 4(4), 247–325.
- Artzner, P., Delbaen, F., Eber, J. M., and Heath, D. (1999), Coherent measures of risk, *Mathematical Finance*, 9, 203–228.
- Balkema, A. A. and de Haan, L. (1974), Residual life time at great age, *The Annals of Probability*, 2, 792–804.
- Basel Committee on Banking Supervision (2004), Basel II: International convergence of capital measurement and capital standards: A revised framework, Bank of International Settlements.
- Basel Committee on Banking Supervision (2013), Fundamental review of the trading book: A revised market risk framework, Bank of International Settlements.

- Berg, D. (2009), Copula goodness-of-fit testing: an overview and power comparison, *The European Journal of Finance*, <http://www.informaworld.com/10.1080/13518470802697428> (2009-03-25).
- Bernstein, S. N. (1928), Sur les fonctions absolument monotones, *Acta Mathematica*, 52, 1–66.
- BIS (2012), Fundamental review of the trading book, Consultative document May 2012, <http://www.bis.org/publ/bcbs219.pdf> (2012-02-03).
- Black, F. and Scholes, M. (1973), The Pricing of Options and Corporate Liabilities, *Journal of Political Economy*, 81(3), 637–654.
- Brockwell, P. J. and Davis, R. A. (1991), Time Series: Theory and Methods, Springer.
- Brockwell, P. J. and Davis, R. A. (2002), Introduction to Time Series and Forecasting, Springer.

Chavez-Demoulin, V., Embrechts, P., and Hofert, M. (2013), An extreme value approach for modeling operational risk losses depending on covariates.

CME SPAN: Standard Portfolio Analysis of Risk (2010), www.cmegroup.com/c
Chicago Mercantile Exchange.

D'Agostino, R. B. and Stephens, M. A. (1986), Goodness-of-fit techniques, Dekker.

Davison, A. C. (2003), Statistical Models, Cambridge Series in Statistical and Probabilistic Mathematics.

Delbaen, F. (2000), Coherent Risk Measures, Cattedra Galiliana, Scuola Normale Superiore, Pisa.

Delbaen, F. (2002), Coherent risk measures on general probability spaces, *Advances in Finance and Stochastics*, ed. by K. Sandmann and P. Schönbucher, Berlin: Springer, 1–37.

Demarta, S. and McNeil, A. J. (2005), The t Copula and Related Copulas, *International Statistical Review*, 73(1), 111–129.

- Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997), Modelling Extreme Events for Insurance and Finance, Springer.
- Embrechts, P., McNeil, A. J., and Straumann, D. (2002), Correlation and Dependency in Risk Management: Properties and Pitfalls, *Risk Management: Value at Risk and Beyond*, ed. by M. Dempster, Cambridge University Press, 176–223.
- Embrechts, P., Lindskog, F., and McNeil, A. J. (2003), Modelling Dependence with Copulas and Applications to Risk Management, *Handbook of Heavy Tailed Distributions in Finance*, ed. by S. Rachev, Elsevier, 329–384.
- Embrechts, P. and Hofert, M. (2013a), A note on generalized inverses, *Mathematical Methods of Operations Research*, 77(3), 423–432, doi: <http://dx.doi.org/10.1007/s00186-013-0436-7>.
- Embrechts, P. and Hofert, M. (2013b), Statistical inference for copulas in high dimensions: A simulation study, *ASTIN Bulletin*, 43(2), 81–95, doi:<http://dx.doi.org/10.1017/asb.2013.6>.

- Fang, K.-T., Kotz, S., and Ng, K.-W. (1990), Symmetric Multivariate and Related Distributions, Chapman & Hall/CRC.
- Feller, W. (1971), An Introduction to Probability Theory and Its Applications, 2nd ed., vol. 2, Wiley.
- Fischer, T. (2003), Risk capital allocation by coherent risk measures based on one-sided moments, *Insurance: Mathematics and Economics*, 32, 135–146.
- Fisher, R. A. and Tippett, L. H. C. (1928), Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society*, 24, 180–190.
- Föllmer, H. and Schied, A. (2002), Convex measures of risk and trading constraints, *Finance and Stochastics*, 6, 429–447.
- Frey, R., McNeil, A. J., and Nyfeler, M. (2001), Copulas and Credit Models, *Risk*, 14(10), 111–114.
- Genest, C. and Nešlehová, J. (2007), A primer on copulas for count data, *The Astin Bulletin*, 37, 475–515.

- Genest, C. and Rivest, L.-P. (1993), Statistical Inference Procedures for Bivariate Archimedean Copulas, *Journal of the American Statistical Association*, 88(423), 1034–1043.
- Genest, C. and Werker, B. J. M. (2002), Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models, *Distributions with Given Marginals and Statistical Modelling*, ed. by C. M. Cuadras, J. Fortiana, and J. A. Rodríguez-Lallena, Kluwer, Dordrecht, 103–112.
- Genest, C., Ghoudi, K., and Rivest, L.-P. (1995), A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, *Biometrika*, 82(3), 543–552.
- Genest, C., Gendron, M., and Bourdeau-Brien, M. (2009), The Advent of Copulas in Finance, *The European Journal of Finance*, 15, 609–618.
- Gnanadesikan, R. and Kettenring, J. R. (1972), Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrics*, 28, 81–124.

- Gnedenko, B. V. (1943), Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of Mathematics*, 44, 423–453.
- Gneiting, T. (2011), Making and Evaluating Point Forecasts, *Journal of the American Statistical Association*, 106(494), 746–762.
- Harrison, J. M. and Kreps, D. M. (1979), Martingales and Arbitrage in Multiperiod Securities Markets, *Journal of Economic Theory*, 20, 381–408.
- Harrison, J. M. and Pliska, S. R. (1981), Martingales and Stochastic Integrals in the Theory of Continuous Trading, *Stochastic Processes and their Applications*, 11, 215–260.
- Higham, N. (2002), Computing the nearest correlation matrix – A problem from finance, *IMA Journal of Numerical Analysis*, 22, 329–343.
- Hofert, M. (2010), Sampling Nested Archimedean Copulas with Applications to CDO Pricing, PhD thesis, Südwestdeutscher Verlag für Hochschulschriften AG & Co. KG, ISBN 978-3-8381-1656-3.

- Hofert, M. and Mächler, M. (2014), A graphical goodness-of-fit test for dependence models in higher dimensions, *Journal of Computational and Graphical Statistics*, 23(3), 700–716, doi:<http://dx.doi.org/10.1080/10618600.2013.812518>.
- Hofert, M. and McNeil, A. J. (2014), On superadditivity of Value-at-Risk in portfolios of defaultable bonds.
- Hofert, M. and Scherer, M. (2011), CDO pricing with nested Archimedean copulas, *Quantitative Finance*, 11(5), 775–787, doi:<http://dx.doi.org/10.1080/14697680903508479>.
- Hofert, M. and Vrins, F. (2013), Sibuya copulas, *Journal of Multivariate Analysis*, 114, 318–337, doi:<http://dx.doi.org/10.1016/j.jmva.2012.08.007>.
- Jaworski, P., Durante, F., Härdle, W. K., and Rychlik, T., eds. (2010), Copula Theory and Its Applications, vol. 198, Lecture Notes in Statistics – Proceedings, Springer.

- Joe, H. and Xu, J. J. (1996), The Estimation Method of Inference Functions for Margins for Multivariate Models, *Technical Report 166, Department of Statistics, University of British Columbia*.
- Joenssen, D. W. and Vogel, J. (2014), A power study of goodness-of-fit tests for multivariate normality implemented in R, *Journal of Statistical Computation and Simulation*, 84, 1055–1078.
- Jorion, P. (2007), Value at Risk: The New Benchmark for Managing Financial Risk, 3rd ed., New York: McGraw-Hill.
- Kim, G., Silvapulle, M. J., and Silvapulle, P. (2007), Comparison of semiparametric and parametric methods for estimating copulas, *Computational Statistics & Data Analysis*, 51, 2836–2850.
- Kimberling, C. H. (1974), A probabilistic interpretation of complete monotonicity, *Aequationes Mathematicae*, 10, 152–164.
- Kloman, H. F. (1990), Risk management agonists, *Risk Analysis*, 10, 201–205.

- Kojadinovic, I. and Yan, J. (2010), Comparison of three semiparametric methods for estimating dependence parameters in copula models, *Insurance: Mathematics and Economics*, 47, 52–63.
- Kolmogorov, A. N. (1933), Grundbegriffe der Wahrscheinlichkeitsrechnung, Berlin: Ergebnisse der Mathematik.
- Kou, S. and Peng, X. (2014), On the Measurement of Economic Tail Risk, <http://arxiv.org/abs/1401.4787> (2014-06-09).
- Leadbetter, M. R. (1991), On a basis for “Peaks over Threshold” modeling, *Statistics & Probability Letters*, 12, 357–362.
- Li, D. X. (2000), On Default Correlation: A Copula Function Approach, *The Journal of Fixed Income*, 9(4), 43–54.
- Li, X., Mikusiński, P., and Taylor, M. D. (2002), Some integration-by-parts formulas involving 2-copulas, *Distributions with Given Marginals and Statistical Modelling*, ed. by C. M. Cuadras, J. Fortiana, and J. A. Rodríguez-Lallena, Kluwer Academic Publishers, 153–159.

- Lindskog, F., McNeil, A. J., and Schmock, U. (2002), Kendall's tau for elliptical distributions, *Credit Risk: Measurement, Evaluation and Management*, ed. by G. Bol, G. Nakhaeizadeh, S. T. Rachev, T. Ridder, and K.-H. Vollmer, Springer, 149–156.
- Lord Turner (2009), The Turner Review: A regulatory response to the global banking crisis, Financial Services Authority, London.
- Luo, X. and Shevchenko, P. V. (2010), The t copula with multiple parameters of degrees of freedom: bivariate characteristics and application to risk management, *Quantitative Finance*, 10(9), 1039–1054.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), Multivariate Analysis, London: Academic Press.
- Markowitz, H. M. (1952), Portfolio Selection, *The Journal of Finance*, 7, 77–91.
- Maronna, R. A. (1976), Robust M-Estimators of multivariate location and scatter, *The Annals of Statistics*, 4, 51–67.

- Marshall, A. W. and Olkin, I. (1988), Families of Multivariate Distributions, *Journal of the American Statistical Association*, 83(403), 834–841.
- McNeil, A. J. and Nešlehová, J. (2009), Multivariate Archimedean copulas, d -monotone functions and l_1 -norm symmetric distributions, *The Annals of Statistics*, 37(5b), 3059–3097.
- McNeil, A. J., Frey, R., and Embrechts, P. (2005), Quantitative Risk Management: Concepts, Techniques, Tools, Princeton University Press.
- McNeil, A. J., Frey, R., and Embrechts, P. (2015), Quantitative Risk Management: Concepts, Techniques, Tools, 2nd ed., Princeton University Press.
- Nelsen, R. B. (1999), An Introduction to Copulas, Springer Verlag.
- Pickands, J. (1975), Statistical inference using extreme order statistics, *The Annals of Statistics*, 3, 119–131.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992), Numerical Recipes in C, Cambridge: Cambridge University Press.

- Ressel, P. (2013), Homogeneous distributions – And a spectral representation of classical mean values and stable tail dependence functions, *Journal of Multivariate Analysis*, 117, 246–256.
- RiskMetrics (1996), RiskMetrics Technical Document, 3rd, J.P. Morgan, New York.
- Rüschedorf, L. (2009), On the distributional transform, Sklar's Theorem, and the empirical copula process, *Journal of Statistical Planning and Inference*, 139(11), 3921–3927.
- Scarsini, M. (1984), On measures of concordance, *Stochastica*, 8(3), 201–218.
- Schmitz, V. (2003), Copulas and Stochastic Processes, PhD thesis, Rheinisch-Westfälische Technische Hochschule Aachen.
- Schönbucher, P. J. and Schubert, D. (2001), Copula-Dependent Default Risk in Intensity Models, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=301968 (2009-12-30).

- Shreve, S. E. (2008), Don't blame the quants, Available at www.forbes.com/2008shreve.html.
- Sibuya, M. (1959), Bivariate extreme statistics, I, *Annals of the Institute of Statistical Mathematics*, 11(2), 195–210.
- Smith, R. L. (1985), Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, 72, 67–92.
- Smith, R. L. (1987), Estimating Tails of Probability Distributions, *The Annals of Statistics*, 15, 1174–1207.
- Tsay, R. S. and Tiao, G. C. (1984), Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models, *Journal of the American Statistical Association*, 79, 84–96.
- Van der Vaart, A. W. (2000), Asymptotic Statistics, Cambridge University Press.
- Williams, D. (1991), Probability with Martingales, Cambridge University Press.