

Quantitative Risk Management

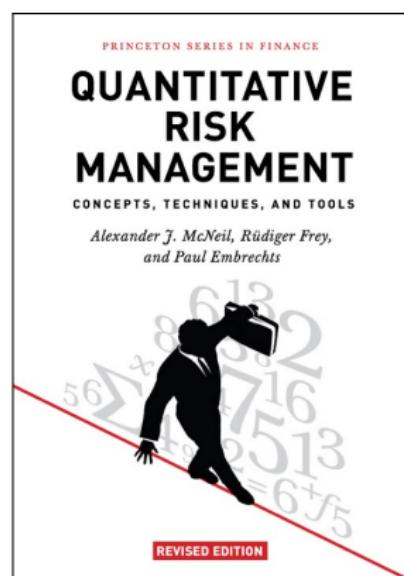
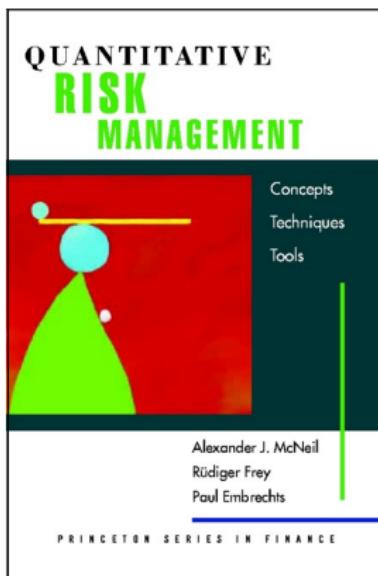
<http://www.qrmtutorial.org>

Last update: 2016-04-27

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**
- A Appendix**

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.
For us: *risk = chance of loss* \Rightarrow uncertainty \Rightarrow randomness

1.1.1 Risk and randomness

- Kolmogorov (1933) introduced the notion of a *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$:
 - ▶ Ω is the *sample space* which contains realizations $\omega \in \Omega$ (“state of nature”) of an experiment;
 - ▶ the *σ -algebra* \mathcal{F} contains all sets (“events”) to which we can assign probabilities; 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 modelled by a random vector $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$.
- Most of this modelling concerns the distribution functions

$$F_X(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}, \quad \text{and} \quad F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d,$$

of X and \mathbf{X} , respectively.

- If time matters, one can consider sequences of random variables $(X_t)_{t \geq 0}$, so-called stochastic processes.
- Our modelling tools will mainly come from probability and statistics.

1.1.2 Financial Risk

There are various types of risks. We focus on

Market risk Risk of loss in a financial position due to changes in the underlying components (e.g. stock/bond/commodity prices)

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

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

There are many other types of risks:

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/minimize a loss. **Funding liquidity risk** refers to the ease with which institutions can raise funding. The two often interact.

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 Risk of using a **misspecified** (inappropriate) model for measuring risk. This is **always present** to some degree!

Good risk management (RM) has to follow a **holistic approach**, i.e. all 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 $F^\leftarrow(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$).

- To this end, 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.
- Good risk measurement is essential (for good RM). For any product sold, the underlying risks need to be properly quantified and clearly communicated to stakeholders. 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 a bank or insurance company.
- What does managing risks involve?

- ▶ Determine the capital to hold to absorb losses, both for *regulatory purposes* (to satisfy regulators) and *economic capital* purposes (to survive as a company).
- ▶ Ensuring portfolios are well diversified.
- ▶ Optimizing portfolios according to risk-return considerations (for example, 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

Academic innovation in the 20th century

- Markowitz (1952): Theory of portfolio selection; Desirability of an investment was decided upon a risk-return diagram (x-axis: 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 (arbitrage-free/completeness conditions)
- By 1995: Nominal values outstanding in derivatives: 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 ruin: OpRisk losses + *straddle position* on the Nikkei (short in a call and put; allows for a gain if Nikkei does not move too far down or up) + Kobe earthquake = loss of \$1.3 billion
- 1998 Long-Term Capital Management (LTCM): hedge fund; losses due to derivatives trading, required a \$3.5 billion bail-out to prevent collapse; M. Scholes and R. Merton (Nobel Prize winners 1997) 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). By 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 fell by 50%*.
- 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)*. Credits were given to *borrowers with low credit ratings*. CDO issuance volume by 2008 was around *\$3 trillion*, for *credit default swaps (CDS)* around *\$30 trillion*.
- *CDSs were used by investors to speculate on (changing) credit risk.*
- The *consensus* was that all this activity was a *good thing*:
 - ▶ 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.”

- ▶ 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; was sold to JP Morgan Chase
 - ▶ September 2008: Lehman Brothers filed for bankruptcy (⇒ worldwide panic, markets tumbled, liquidity vanished, many banks 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 ⇒ needed an 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 (e.g. US TARP = Troubled Asset Relief Program).

- Mathematicians/financial engineers were also blamed due to the failure of 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.
- Mathematicians had also warned 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.
- Trades are executed by computer (algorithms) in fractions of a second (no testing), computer centers are build near stock markets for faster trading. One casualty of algorithmic trading: Knight Capital Group (financial services firm) lost \$460 million due to trading errors on 2012-08-01.
- Ongoing concern: Systemic 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 (G10) in 1974. The Basel Committee does not have legal force but it formulates standards/best practices/guidelines, 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 measurement of risk
 - ▶ Claims were divided into 3 categories only, counterparties being governments, regulated banks and others;
 - ▶ Risk weighting identical for all corporate borrowers, independent of their credit rating;
 - ▶ Unsatisfactory treatment of derivatives.

The birth of VaR

- 1993: G30 (international body of leading financiers and academics) 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 JPMorgan methodology (which became known as *RiskMetrics*), set an industry-wide standard.
- Banks pushed to be allowed to use netting (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 *models* for more sophisticated banks
- Coarseness problem for credit risk remained (not enough incentives to diversify credit portfolios; regulatory capital rules too risk insensitive).

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 modelling process; 3) disclosure requirements.
- Important themes were:
 - ▶ Under Pillar 1, banks are 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, which delayed the implementation of Basel II.

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** (estimate of default/migration risk of unsecuritized credit products in the trading book). 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 (ex-changes);

- 3) Introduction of a leverage ratio to prevent excessive leverage (technique to multiply gains/losses; often by buying more of an asset with borrowed capital);
- 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.
- Basel III works alongside Basel II and 2.5, not replacing it. Its targeted end date of implementation is 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.

- 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”)

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* ⇒ Single,

robust system, **easy to understand**, **inexpensive to monitor**. However, it is **mainly volume based** and **not explicitly risk based**.

- Solvency II was **initiated** in 2001 (publication of the Sharma report); adopted by the Council of the European Union and the European Parliament in November 2009; **application** of the framework from 2016-01-01.
- 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 (Basel Committee)

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).

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 perform *stress tests* of a bank's capital adequacy.

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*)) and the *trading book* (assets held that are regularly traded; marked-to-market every day) 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 (e.g. regulatory arbitrage).

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)
- 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.
- To calculate risk weights, banks use either the *standardized approach* (risk weights prescribed by regulator) or one of the more advanced *internal-ratings-based (IRB)* approaches.
- 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 (RWAs) 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 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*.
- 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 to ES.

1.3.2 The Solvency II Framework

Main features

- Solvency II also 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 firm 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 both 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 market risk are interest-rate/equity/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. annual 0.5% probability).

- 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 correlations.

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 determine the SCR.

ORSA (Own risk and solvency assessment)

- ORSA = 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.
- ORSA (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 approach 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: 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.
- ▶ 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.”
- ▶ 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*).
- 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.

- 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

- 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 incorporate RM 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*.
- ▶ RM can reduce the impact of *costly external financing*.

1.5 Quantitative Risk Management

1.5.1 The Q in QRM

- We 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 a suitable language and with appropriate concepts for describing financial risks.
- We also point out assumptions and limitations of the methodology used.
- We should also be aware that 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 RM 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. 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.
- ▶ The problem of scale. A portfolio may represent the entire position in risky assets of a financial institution, but 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 modelled, e.g. correlation in credit risk models).

- ▶ 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.**

2 Basics concepts in risk management

2.1 Risk management for a financial firm

2.2 Modelling 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

A stylized balance sheet for a bank is:

Assets		Liabilities	
Investments of the firm		Obligations from fundraising	
Cash	£10M	Customer deposits	£80M
(and central bank balance)			
Securities	£50M	Bonds issued	
- bonds, stocks, derivatives		- senior bond issues	£25M
Loans and mortgages	£100M	- subordinated bond issues	£15M
- corporates		Short-term borrowing	£30M
- retail and smaller clients		Reserves (for losses on loans)	£20M
- government			
Other assets	£20M	Debt (sum of above)	£170M
- property			
- investments in companies		Equity	£30M
Short-term lending	£20M		
Total	£200M	Total	£200M

A stylized balance sheet for an **insurer** is:

Assets	Liabilities	
Investments		Reserves for policies written (technical provisions) £80M
- bonds	£50M	Bonds issued £10M
- stocks	£5M	
- property	£5M	
Investments for unit-linked contracts	£30M	Debt (sum of above) £90M
Other assets	£10M	Equity £10M
- property		
Total	£100M	Total £100M

- Balance sheet equation: $\text{Assets} = \text{Liabilities} = \text{Debt} + \text{Equity}$.
If equity > 0, the company is *solvent*, otherwise *insolvent*.
- Valuation of the items on the balance sheet is a non-trivial task.
 - ▶ *Amortized cost accounting* values a position a *book value* at its inception and this is carried forward/progressively reduced over time.

- ▶ *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 industry to move towards fair-value accounting. Market consistent valuation in Solvency II follows similar principles.

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 or a bank run).
- The prime risk for 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 firms: MCR + SCR;
- A regulatory framework also specifies the **capital quality**. 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).

Economic capital

- Capital required to control the probability of **becoming insolvent** (typically over one year);
- **Internal assessment** of 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 obligations to outside creditors; they **can** thus **serve as buffer against losses**.

2.2 Modelling value and value change

2.2.1 Mapping of risks

We now set up a general mathematical model for (changes in) value caused by financial risks. For this we 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).

- 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 and *assume*:
 - 1) the *portfolio composition* remains *fixed* over Δt ;
 - 2) there are *no intermediate payments* during Δt

⇒ Fine for small Δt but *unlikely to hold for large Δ .*

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

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

and we define the (random) *loss* by 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 appropriate, but we will mostly neglect this issue.

- V_t is typically modelled 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).

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

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

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

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

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*

$$L_{t+1}^\Delta = - \left(\underbrace{f_t(t, \mathbf{Z}_t)}_{=: c_t} + \sum_{j=1}^d \underbrace{f_{z_j}(t, \mathbf{Z}_t) X_{t+1,j}}_{=: b_{t,j}} \right) = -(c_t + \mathbf{b}'_t \mathbf{X}_{t+1}),$$

a linear function of $X_{t+1,1}, \dots, X_{t+1,d}$ (indices denote partial derivatives). The approximation is best if the risk-factor changes 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

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

- The one-period ahead loss is then given by

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

which is non-linear in $X_{t+1,j}$.

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

$$\begin{aligned} L_{t+1}^{\Delta} &= -\left(f_t(t, \mathbf{Z}_t) + \sum_{j=1}^d f_{z_j}(t, \mathbf{Z}_t) X_{t+1,j}\right) = -\left(0 + \sum_{j=1}^d w_{t,j} X_{t+1,j}\right) \\ &= -\mathbf{w}'_t \mathbf{X}_{t+1}. \end{aligned}$$

- Note that

$$L_{t+1}^{\Delta} = -(c_t + \mathbf{b}'_t \mathbf{X}_{t+1})$$

for $c_t = 0$ and $\mathbf{b}_t = \mathbf{w}_t$.

- 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

$$\mathbb{E} L_{t+1}^{\Delta} = -\sum_{j=1}^d w_{t,j} \mathbb{E}(X_{t+1,j}) = -\mathbf{w}'_t \mu,$$

$$\text{var } L_{t+1}^{\Delta} = \text{var}(\mathbf{w}'_t \mathbf{X}_{t+1}) = \mathbf{w}'_t \text{cov}(\mathbf{X}_{t+1}) \mathbf{w}_t = \mathbf{w}'_t \Sigma \mathbf{w}_t.$$

Example 2.3 (European call option)

Consider a portfolio consisting of a European call option on a non-dividend-paying stock S_t with maturity T and strike (exercise price) K . The Black–Scholes formula says that today's value is

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

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 of S_t (standard deviation).

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

$$Z_t = (\log S_t, r_t, \sigma_t) \Rightarrow X_{t+1} = (\log(S_{t+1}/S_t), r_{t+1} - r_t, \sigma_{t+1} - \sigma_t).$$

This implies that the mapping f (in terms of the risk factors) is given by

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

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

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

If our risk management horizon is 1 d (as opposed to 1 y), we need to introduce $\Delta t = 1/250$ here. Note that 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} ⇒ higher-order (Taylor) approximations such as the *delta-gamma-approximation* (second-order) can be used.

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/paid* on an *active market*. One distinguishes:

Level 1 *Mark-to-market*. The *fair value* of an investment is determined from *quoted prices* for the *same instrument*; see Example 2.2.

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; see Example 2.3.

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 spreads 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. the *risk-neutral pricing measure Q* (also called *equivalent martingale measure (EMM)*); it turns discounted prices into martingales, so fair bets) as opposed to the real world/physical measure \mathbb{P} .
- 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 (fair bet).
- 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 .

- \mathbb{P} is estimated from historical data; Q is calibrated to market prices.

Example 2.4 (European call option continued)

- Suppose that options with strike K or maturity T are not traded, but other options on the same stock are.
- Under \mathbb{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 $\mu \in \mathbb{R}$ (drift) and $\sigma > 0$ (volatility), and a standard Brownian motion (W_t).
- Under the EMM Q , $(e^{-rt} S_t)$ is a martingale and S_t follows a GBM with drift r and volatility σ .
- The European call option payoff is $H = (S_T - K)^+ = \max\{S_T - K, 0\}$ and the risk-neutral valuation formula may be shown to be

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

- One typically uses quoted prices $C^{\text{BS}}(t, S_t; r, \sigma, K^*, T^*)$ (for different K^*, T^*) to infer the unknown σ . Then plug this so-called *implied volatility* into (3).

2.2.3 Loss distributions

Having determined the mapping f (may involve *valuation models*, e.g. Black–Scholes, or numerical approximation), we can identify the following key statistical tasks of QRM:

- 1) Find a statistical model for \mathbf{X}_{t+1} (typically a model for forecasting \mathbf{X}_{t+1} , estimated based on historical data);
- 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 these challenges.

1) Analytical method

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

Prime example: *Variance-covariance method*, see RiskMetrics (1996):

Assumption 1 $\mathbf{X}_{t+1} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (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 $F_{L_{t+1}}$.

$$L_{t+1}^\Delta = -(c_t + \mathbf{b}'_t \mathbf{X}_{t+1}) \stackrel{\text{Ass. 1}}{\Rightarrow} L_{t+1}^\Delta \sim N(-c_t - \mathbf{b}'_t \boldsymbol{\mu}, \mathbf{b}'_t \boldsymbol{\Sigma} \mathbf{b}_t).$$

- Advantages:
- $F_{L_{t+1}^\Delta}$ explicit (\Rightarrow typically explicit risk measures)
 - Easy to implement

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

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 (edf)*

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

based on $\tilde{L}_k = -(f(t+1, \mathbf{Z}_t + \mathbf{X}_k) - f(t, \mathbf{Z}_t))$. $\tilde{L}_{t-n+1}, \dots, \tilde{L}_t$ show what would happen to the current portfolio if the past n risk-factor changes were to recur.

Advantages:

- Easy to implement
- No estimation of the distribution of \mathbf{X}_{t+1} required

Drawbacks:

- Sufficient 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 model for \mathbf{X}_{t+1} , simulate from it, compute the corresponding simulated losses and estimate $F_{L_{t+1}}$ (typically via edf).

Advantages: ■ Quite general (applicable to any model of \mathbf{X}_{t+1} which is easy to sample)

Drawbacks: ■ Unclear how to find an appropriate model for \mathbf{X}_{t+1} (any result is only as good as the chosen $F_{\mathbf{X}_{t+1}}$)
■ 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)

So-called *economic scenario generators* (i.e. economically motivated dynamic models for the evolution and interaction of risk factors) used in insurance also fall under the heading of Monte Carlo methods.

2.3 Risk measurement

- A *risk measure* for a financial position with (random) loss L is a *real number* which measures the “riskiness of L ”. In the Basel or Solvency context, it is often interpreted as the amount of *capital required to make a position with loss L acceptable* to an (internal/external) regulator.
- 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).
 - ▶ 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* = summed notational values of the securities times their riskiness factor
- 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** (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

- 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 under all scenarios.

- If $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ denote the risk-factor changes (*scenarios*) with corresponding *weights* $\mathbf{w} = (w_1, \dots, w_n)$, the risk is

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

where $L(\mathbf{x})$ denotes the loss the portfolio would suffer if the hypothetical scenario \mathbf{x} were to occur. Many risk measures are of the form (4); see *CME SPAN: Standard Portfolio Analysis of Risk* (2010).

- Mathematical interpretation of (4):
 - ▶ Assume $L(\mathbf{0}) = 0$ (okay if Δt small) and $w_i \in [0, 1] \forall i$.
 - ▶ $w_i L(\mathbf{x}_i) = w_i L(\mathbf{x}_i) + (1 - w_i)L(\mathbf{0}) = \mathbb{E}_{\mathbb{P}_i}(L(\mathbf{X}_i))$ 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 a *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.

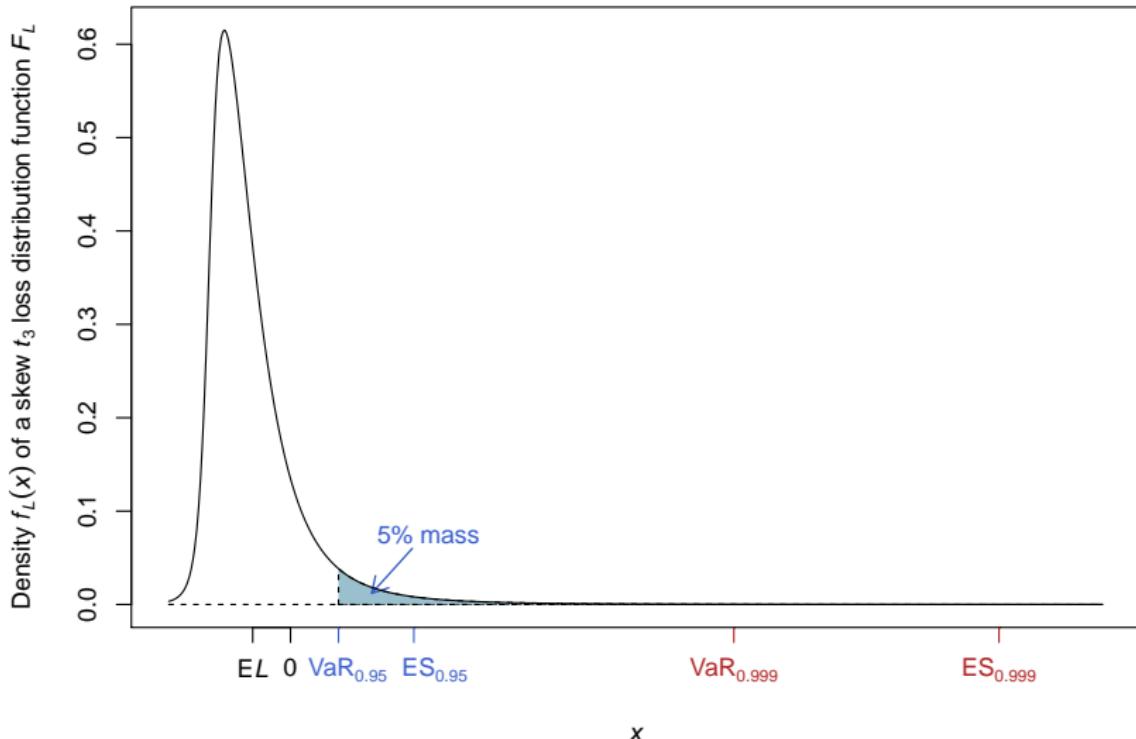
2.3.2 Value-at-risk

Definition 2.5 (Value-at-risk)

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

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

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



Example 2.6 (VaR $_{\alpha}$ for $N(\mu, \sigma^2)$ and $t_{\nu}(\mu, \sigma^2)$)

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

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

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

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

If $\nu > 1$, $\mathbb{E}X = 0$; if $\nu > 2$, $\text{var } X = \frac{\nu}{\nu-2}$.

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: $\alpha = 0.999$, $\Delta t = 1 \text{ y}$
 - ▶ 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 strictly interpreting $\text{VaR}_\alpha(L)$ (and other risk measure) estimates, 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 $x < y$. $T \uparrow$ means that T is *strictly increasing*, i.e. $T(x) < T(y)$ for all $x < y$.

Definition 2.7 (Generalized inverse)

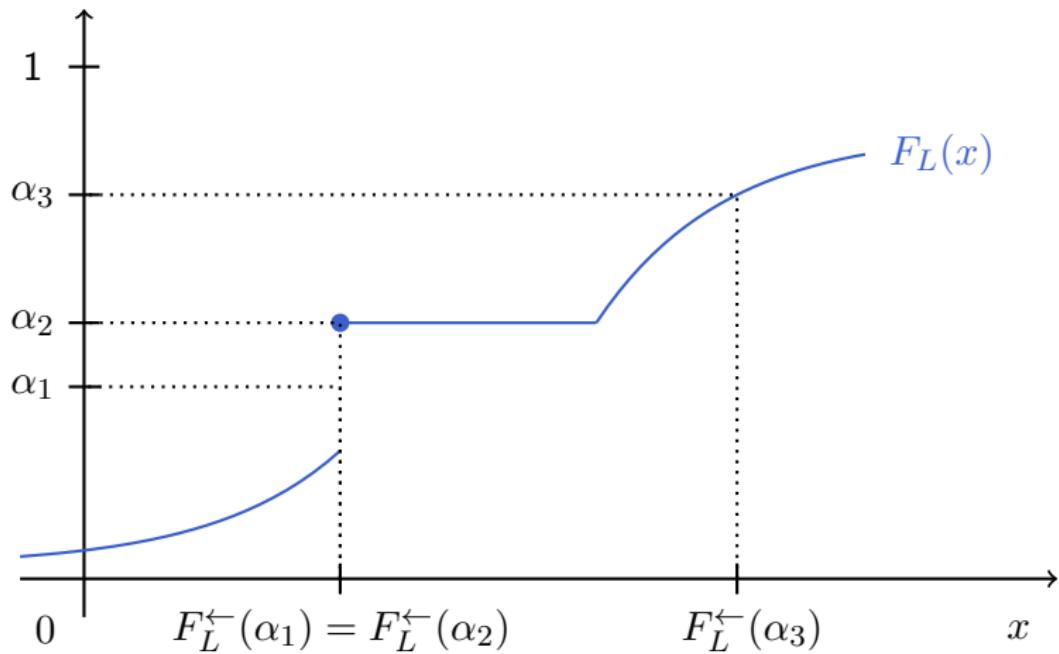
For any increasing function $T : \mathbb{R} \rightarrow \mathbb{R}$, with $T(-\infty) = \lim_{x \downarrow -\infty} T(x)$ and $T(\infty) = \lim_{x \uparrow \infty} T(x)$, the *generalized inverse* $T^\leftarrow : \mathbb{R} \rightarrow \bar{\mathbb{R}} = [-\infty, \infty]$ of T is defined by

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

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

- If T is continuous and \uparrow , then $T^\leftarrow \equiv T^{-1}$ (ordinary inverse).
- There are *rules for working with T^\leftarrow* (similar to T^{-1}); see Proposition A.15.

F_L^\leftarrow visualized 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 very 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 $V_t = A_t - B_t$ 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 $(\geq)\alpha$. 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 $(\geq)\alpha$). If $x_0 < 0$, the company is already well capitalized.

2.3.4 Other risk measures based on loss distributions

1) Variance (or standard deviation)

- $\text{var}_\alpha(L)$ (or standard deviation) has a long history as a 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 if F_L is roughly symmetric around $\mathbb{E}L$, but F_L is typically skewed in credit and operational risk)

2) Expected shortfall

Definition 2.8 (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

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

- 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) $\Rightarrow \text{ES}_\alpha \geq \text{VaR}_\alpha$
- Besides VaR, ES is the **most important risk measure** in practice.
- 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).
- ES_α is more difficult to estimate and backtest than VaR_α (larger sample size required; the variance of estimators is typically larger).
- $\text{ES}_\alpha(L) < \infty$ requires $\mathbb{E}|L| < \infty$.
- If F_L is continuous, $\text{ES}_\alpha(L) = \mathbb{E}(L | L > \text{VaR}_\alpha(L))$; see the appendix.

- Subadditivity and *elicability* (see the appendix). One can show:
 - ▶ 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 exists if $\mathbb{E}|L| = \infty$ and is elicitable (see also the appendix).

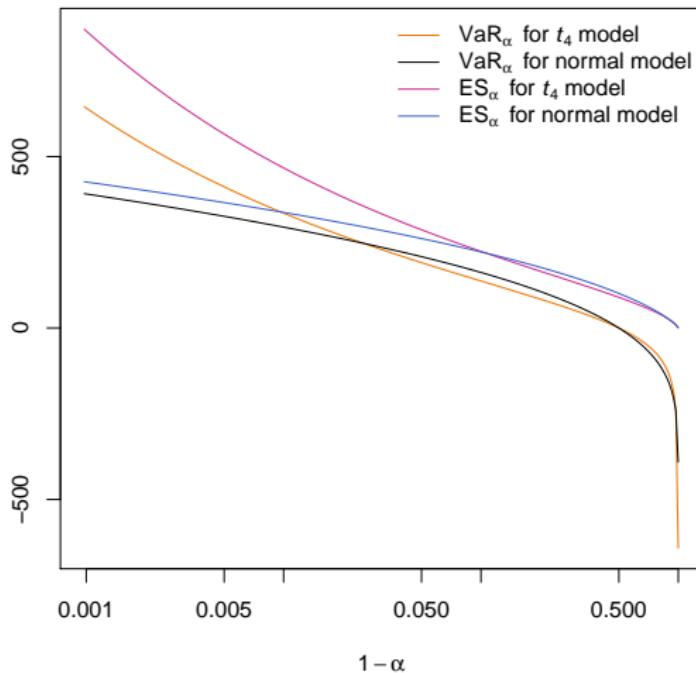
Example 2.9 (VaR and ES for stock returns)

- Consider Example 2.2 with a portfolio consisting of a single stock $V_t = S_t = 10\,000$. In this case, $L_{t+1}^\Delta = -S_t X_{t+1}$, where $X_{t+1} = \log(S_{t+1}/S_t)$.
- Let $\sigma = 0.2/\sqrt{250}$ (annualized volatility of 20%) and assume
 - 1) $X_{t+1} \sim N(0, \sigma^2) \Rightarrow L_{t+1}^\Delta \sim N(0, S_t^2 \sigma^2)$;
 - 2) $X_{t+1} \sim t_\nu(0, \sigma^2 \frac{\nu-2}{\nu})$ (so that $\text{var } X_{t+1}$ is also σ^2). Then

$$X_{t+1} = \sqrt{\sigma^2 \frac{\nu-2}{\nu}} Y \quad \text{for } Y \sim t_\nu.$$

$$\Rightarrow L_{t+1}^\Delta = -S_t \sqrt{\sigma^2 \frac{\nu-2}{\nu}} Y \sim t_\nu(0, S_t^2 \sigma^2 \frac{\nu-2}{\nu}) \quad (\text{var}(L_{t+1}^\Delta) = S_t^2 \sigma^2).$$

Consider $\nu = 4$ and note that $\text{VaR}_{\alpha}^{t_4} \geq \text{VaR}_{\alpha}^{\text{normal}}$ and $\text{ES}_{\alpha}^{t_4} \geq \text{ES}_{\alpha}^{\text{normal}}$ only hold for sufficiently large α .



⇒ The t_4 model is not always “riskier” than the normal model.

Example 2.10 (Example 2.6 continued; ES_α for $N(\mu, \sigma^2)$ and $t_\nu(\mu, \sigma^2)$)

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

$$\text{ES}_\alpha(\tilde{L}) = \frac{1}{1 - \alpha} \int_{\alpha}^1 \Phi^{-1}(u) du \underset{x = \Phi^{-1}(u)}{=} \frac{1}{1 - \alpha} \int_{\Phi^{-1}(\alpha)}^{\infty} x \varphi(x) dx,$$

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

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

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

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

2) 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 2.6.

By l'Hôpital's Rule (case “0/0”), one can show that

$$1 \leq \lim_{\alpha \uparrow 1} \frac{\text{ES}_\alpha(L)}{\text{VaR}_\alpha(L)} = \frac{\nu}{\nu - 1}.$$

- 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).
- For $\nu \uparrow \infty$, $\lim_{\alpha \uparrow 1} \frac{\text{ES}_\alpha(L)}{\text{VaR}_\alpha(L)} \downarrow 1$; for $\nu \downarrow 1$, $\lim_{\alpha \uparrow 1} \frac{\text{ES}_\alpha(L)}{\text{VaR}_\alpha(L)} \uparrow \infty$.

Conclusion:

For losses with *heavy tails* (power-like), 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 of a good risk measure**.
- Assume that risk measures ϱ are defined on a **linear space of random variables \mathcal{M}** (including constants; we can thus add rvs, multiply them with constants etc.), so $\varrho : \mathcal{M} \rightarrow \mathbb{R}$.
- There are **two possible interpretations** of elements of \mathcal{M} :
 - 1) Elements of \mathcal{M} are random variables V_{t+1} ; $\tilde{\varrho}(V_{t+1})$ denotes the amount of **additional capital** that needs to be added to a position with future value V_{t+1} to make it acceptable to a regulator.
 - 2) Elements of \mathcal{M} are losses $L_{t+1} = -(V_{t+1} - V_t)$; $\varrho(L_{t+1})$ denotes the **total amount of capital** necessary to back a position with loss L .
- 1) and 2) are related via $\varrho(L_{t+1}) = V_t + \tilde{\varrho}(V_{t+1})$ (total capital = available capital + additional capital). We focus on 2) and drop “ $t + 1$ ”.

Axioms of coherence

Axiom 1 (**monotonicity**) $L_1, L_2 \in \mathcal{M}$, $L_1 \leq L_2$ (a.s., i.e. almost surely)
 $\Rightarrow \varrho(L_1) \leq \varrho(L_2)$

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

Criticism: none

Axiom 2 (**translation invar.**) $\varrho(L + l) = \varrho(L) + l$ 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 $\varrho(L) > 0$, and $l = -\varrho(L)$, then $\varrho(L - \varrho(L)) = \varrho(L + l) = \varrho(L) + l = 0$ so that adding $\varrho(L)$ to a position with loss L makes it acceptable.

Criticism: Most people believe this to be reasonable.

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

Interpr.:

- Reflects the idea of **diversification**
- Using a non-subadditive ϱ encourages institutions to legally break up into subsidiaries to reduce regulatory capital requirements.
- Subadditivity makes decentralization possible: Assume $L = L_1 + L_2$ and that we want to bound $\varrho(L)$ by M . Choose M_j such that $\varrho(L_j) \leq M_j$, $j \in \{1, 2\}$, and $M_1 + M_2 \leq M$. Then $\varrho(L) \leq_{\text{subadd.}} \varrho(L_1) + \varrho(L_2) \leq M_1 + M_2 \leq M$.

Criticism: VaR is ruled out under certain scenarios (see later).
VaR is monotone, translation invariant, and positive homogeneous, but in general not subadditive.

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

Interpr.: (or motivation): $\lambda = n \in \mathbb{N}$ and subadditivity imply $\varrho(nL) \leq n\varrho(L)$, 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 $\varrho(\lambda L) > \lambda \varrho(L)$ (also to penalize concentration or risk), but this contradicts subadditivity. This has led to convex risk measures (see later), i.e. risk measures ϱ satisfying $\varrho(\lambda L_1 + (1 - \lambda)L_2) \leq \lambda \varrho(L_1) + (1 - \lambda)\varrho(L_2)$ for all $L_1, L_2 \in \mathcal{M}$, $0 \leq \lambda \leq 1$.

Definition 2.11 (Coherent risk measure)

A risk measure ϱ which satisfies Axioms 1–4 is called *coherent*.

Example 2.12 (Generalized scenario risk measures)

Let $L(\mathbf{x})$ denote the hypothetical loss under scenario \mathbf{x} (risk-factor change).

The generalized scenario risk measure

$$\psi_{\mathcal{X}, \mathbf{w}}(L) = \max\{\mathbb{E}_{\mathbb{P}}(L(\mathbf{X})) : \mathbf{X} \sim \mathbb{P} \in \{\mathbb{P}_1, \dots, \mathbb{P}_n\}\}$$

is coherent. Monotonicity, translation invariance, positive homogeneity are clear (by monotonicity and linearity of $\mathbb{E}(\cdot)$); for subadditivity, note that

$$\begin{aligned}\psi_{\mathcal{X}, \mathbf{w}}(L_1 + L_2) &= \max\{\underbrace{\mathbb{E}_{\mathbb{P}}(L_1(\mathbf{X}) + L_2(\mathbf{X}))}_{=\mathbb{E}_{\mathbb{P}}(L_1(\mathbf{X})) + \mathbb{E}_{\mathbb{P}}(L_2(\mathbf{X}))} : \mathbf{X} \sim \mathbb{P} \in \{\mathbb{P}_1, \dots, \mathbb{P}_n\}\} \\ &\leq \psi_{\mathcal{X}, \mathbf{w}}(L_1) + \psi_{\mathcal{X}, \mathbf{w}}(L_2).\end{aligned}$$

One can show that all coherent risk measures can be represented as generalized scenarios via

$$\varrho(L) = \sup\{\mathbb{E}_{\mathbb{P}}(L) : \mathbb{P} \in \mathcal{P}\}$$

for a suitable set \mathcal{P} of probability measures.

Definition 2.13 (Convex risk measure)

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

- Justification: Again diversification but they don't have to be positive homogeneous.
- Any coherent risk measure is also a convex risk measure. The *converse is not true in general*, but for positive homogeneous risk measures, convexity and subadditivity are equivalent.

Theorem 2.14 (Coherence of ES)

ES is a coherent risk measure.

Proof. Monotonicity, translation invariance and positive homogeneity follow from VaR. Subadditivity is more involved but can be shown in various ways; see Embrechts and Wang (2015). □

Superadditivity scenarios for VaR

Under the following scenarios, VaR_α is typically superadditive:

- 1) L_1, L_2 have skewed distributions;
- 2) Independent, light-tailed L_1, L_2 and small α ;
- 3) L_1, L_2 have special dependence;
- 4) L_1, L_2 have heavy tailed distributions.

Exercise 2.15 (Skewed loss distributions)

Consider a portfolio of two independently defaultable zero-coupon bonds (maturity $T = 1$ year, nominal/face value 100, paid interest of 5%, default probability $p = 0.009$, no recovery). The loss of bond j (from the lender's/investor's perspective) is thus

$$L_j = \begin{cases} -5, & \text{with prob. } 1 - p = 0.991, \\ 100, & \text{with prob. } p = 0.009, \end{cases} \quad j \in \{1, 2\}.$$

Set $\alpha = 0.99$. Then $\text{VaR}_\alpha(L_j) = -5$, $j \in \{1, 2\}$.

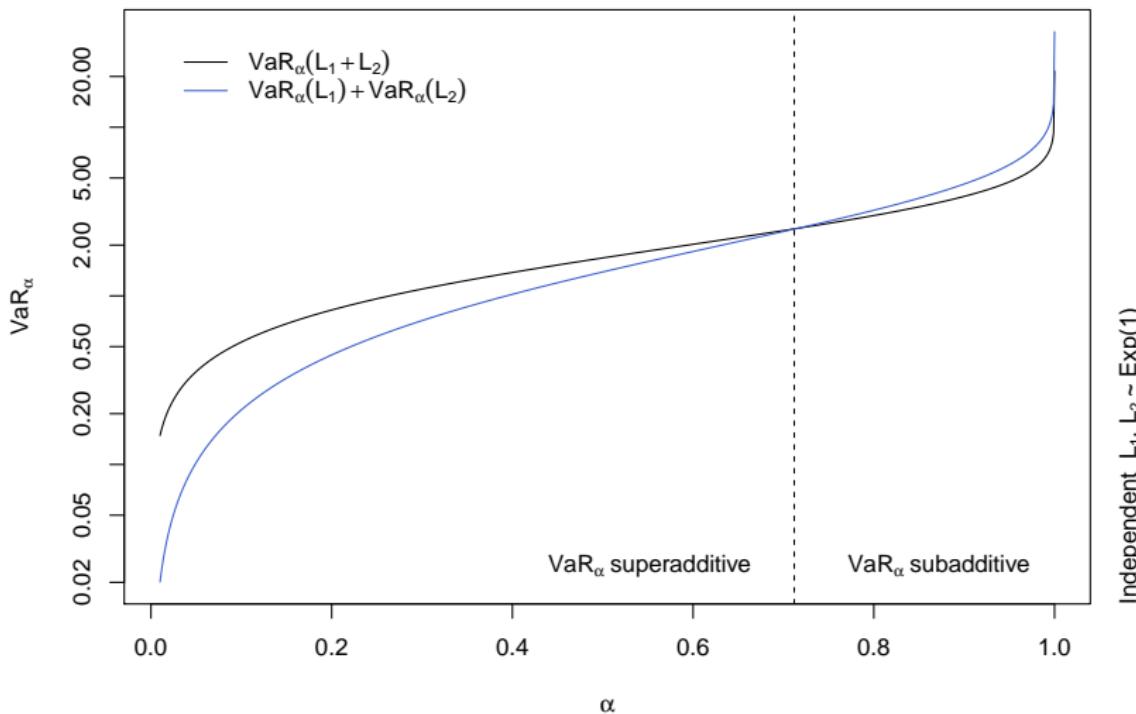
The loss $L_1 + L_2$ is given by

$$L_1 + L_2 = \begin{cases} -10, & \text{with prob. } (1 - p)^2 = 0.982081, \\ 95, & \text{with prob. } 2p(1 - p) = 0.017838, \\ 200, & \text{with prob. } p^2 = 0.000081. \end{cases}$$

Therefore, $\text{VaR}_\alpha(L_1 + L_2) = 95 > -10 = \text{VaR}_\alpha(L_1) + \text{VaR}_\alpha(L_2)$. Hence VaR_α is superadditive.

Exercise 2.16 (Independent, light-tailed 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 2.17 (Special dependence)

Let $\alpha \in (0, 1)$, $L_1 \sim U(0, 1)$ and define $L_2 \stackrel{\text{a.s.}}{=} \begin{cases} L_1, & \text{if } L_1 < \alpha, \\ 1 + \alpha - L_1, & \text{if } L_1 \geq \alpha. \end{cases}$

One can show that $L_2 \sim U(0, 1)$. Also, $L_1 + L_2 = \begin{cases} 2L_1, & \text{if } L_1 < \alpha, \\ 1 + \alpha, & \text{if } L_1 \geq \alpha, \end{cases}$ from which one can show that

$$F_{L_1+L_2}(x) = \begin{cases} 0, & \text{if } x < 0, \\ x/2, & \text{if } x \in [0, 2\alpha), \\ \alpha, & \text{if } x \in [2\alpha, 1 + \alpha), \\ 1, & \text{if } x \geq 1 + \alpha. \end{cases}$$

For all $\varepsilon \in (0, \frac{1-\alpha}{2})$, we thus obtain that

$$\text{VaR}_{\alpha+\varepsilon}(L_1 + L_2) = \underset{\varepsilon \in (0, \frac{1-\alpha}{2})}{1 + \alpha} > 2(\alpha + \varepsilon) = \text{VaR}_{\alpha+\varepsilon}(L_1) + \text{VaR}_{\alpha+\varepsilon}(L_2).$$

Exercise 2.18 (Heavy tailed loss distributions)

Let $L_1, L_2 \stackrel{\text{ind.}}{\sim} F(x) = 1 - x^{-1/2}$, $x \in [1, \infty)$. By deriving the distribution function

$$F_{L_1+L_2}(x) = 1 - 2\sqrt{x-1}/x, \quad x \geq 2,$$

of $L_1 + L_2$ (via the density convolution formula; tedious), one can show (via solving a quadratic equation) that VaR_α is superadditive for all $\alpha \in (0, 1)$.

Remark 2.19 (Special case of comonotone risks; elliptical risks)

- In comparison to Exercise 2.17, $L_1 \stackrel{\text{a.s.}}{=} L_2$ does not lead to the largest $\text{VaR}_\alpha(L_1 + L_2)$ since

$$\text{VaR}_\alpha(L_1 + L_2) \stackrel{\text{pos. hom.}}{=} 2 \text{VaR}_\alpha(L_1) = \text{VaR}_\alpha(L_1) + \text{VaR}_\alpha(L_2),$$

so “only” equality.

- VaR_α is subadditive and thus coherent for a certain class of multivariate distributions (strictly including the multivariate normal and t); see later.

3 Empirical properties of financial data

3.1 Stylized facts of financial return series

3.2 Multivariate stylized facts

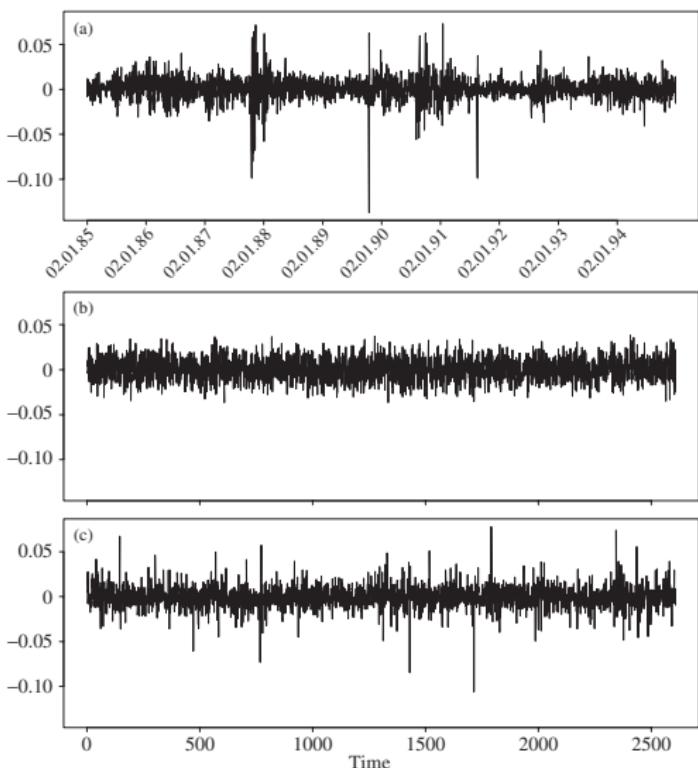
3.1 Stylized facts of financial return series

- 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).
- 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).
- Consider discrete-time risk-factor changes $X_t = Z_t - Z_{t-1}$, e.g. $Z_t = \log S_t$, in which case

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

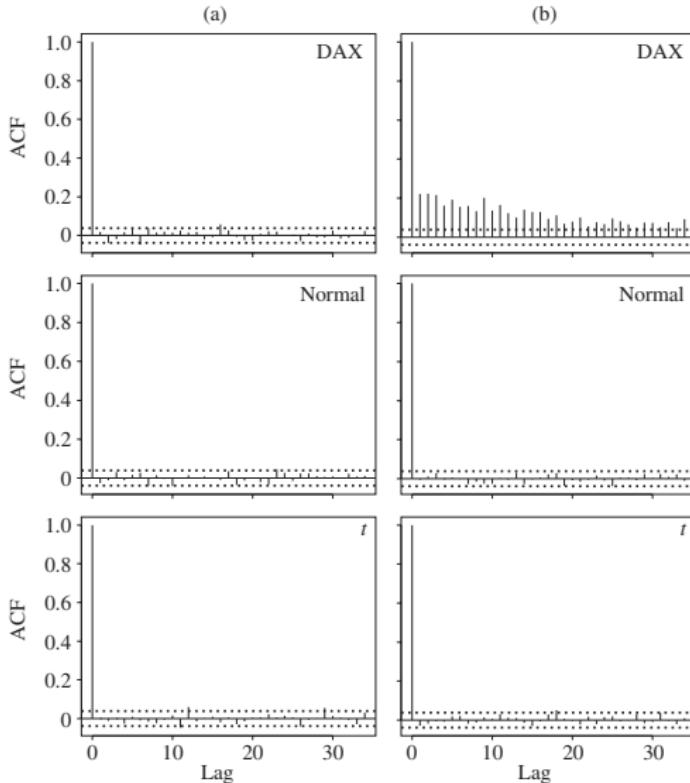
the former is often called *(log-)return*, the latter *(classical) return*.

3.1.1 Volatility Clustering



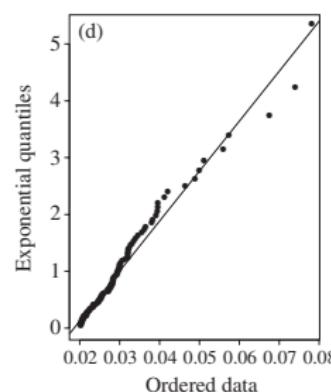
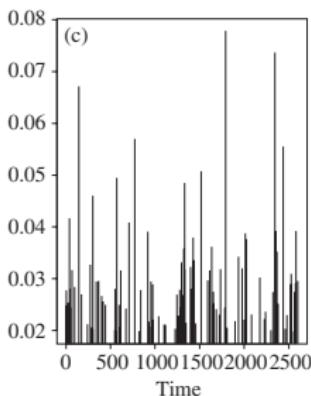
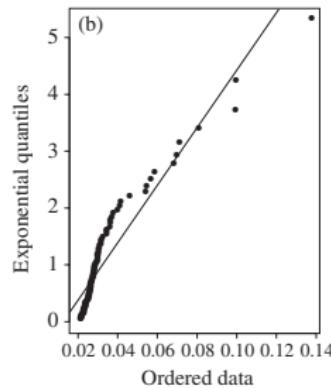
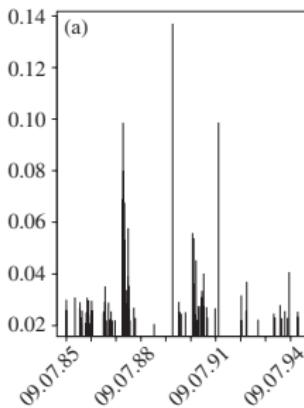
- (a) Log-returns for the DAX index from 1985-01-02 to 1994-12-30 ($n = 2608$)
Simulated iid data from a fitted normal ($\hat{\mu} = \bar{X}_n$, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$)
⇒ Shows too few extremes
- (b) Simulated iid data from a fitted $t_{3.8}$ (num. max. of log-likelihood;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

Autocorrelation function (ACF) $\rho(h) = \text{corr}(X_0, X_h)$ for $h \in \mathbb{Z}$



- (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 \Rightarrow Predicted return ≈ 0
- iid data $(X_t)_{t \in \mathbb{Z}}$ implies $\rho_X(h) = \rho_{|X|}(h) = I_{\{h=0\}}$; not the case here (confirm with a Ljung–Box test $H_0 : \rho(k) = 0, k = 1, \dots, h$)
- $(X_t)_{t \in \mathbb{Z}}$ not a random walk (e.g. no geometric BM)

Concerning clustering of extremes, consider the **100 largest losses** of the...



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

3.1.2 Non-normality and heavy tails

Formal statistical tests of normality

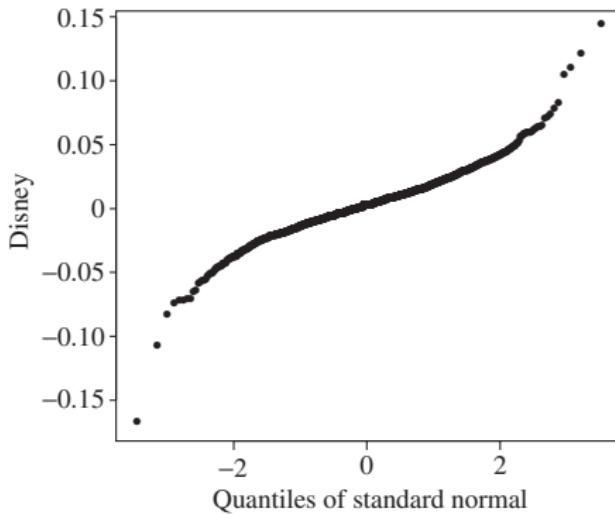
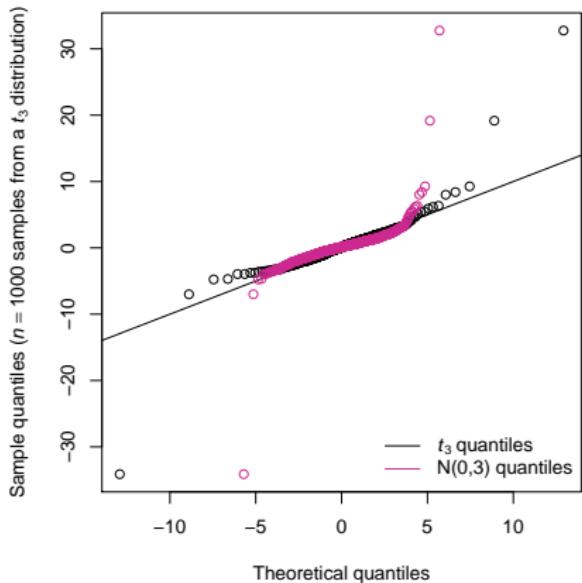
- For general univariate df F :
 - ▶ Kolmogorov–Smirnov (test statistic $T_n = \sup_x |\hat{F}_n(x) - F(x)|$)
 - ▶ Cramér–von Mises ($T_n = n \int_{-\infty}^{\infty} (\hat{F}_n(x) - F(x))^2 dF(x)$)
 - ▶ Anderson–Darling ($T_n = n \int_{-\infty}^{\infty} \frac{(\hat{F}_n(x) - F(x))^2}{F(x)(1-F(x))} dF(x)$; recommended by D'Agostino and Stephens (1986))
- For $F = N(\mu, \sigma^2)$:
 - ▶ Shapiro–Wilk (idea: quantify Q-Q plot in one number; see later)
 - ▶ D'Agostino (based on skewness and kurtosis as Jarque–Bera)
 - ▶ **Jarque–Bera test:** Compares skewness $\beta = \frac{\mathbb{E}((X-\mu)^3)}{\sigma^3}$ and kurtosis $\kappa = \frac{\mathbb{E}((X-\mu)^4)}{\sigma^4}$ with sample versions. The test statistic is

$$T_n = \frac{n}{6} \left(\hat{\beta}^2 + \frac{1}{4} (\hat{\kappa} - 3)^2 \right) \underset{n \text{ large}}{\overset{H_0}{\sim}} \chi_2^2.$$

Graphical tests

- Suppose we want to graphically test whether $X_1, \dots, X_n \sim F$ for some df F based on realizations x_1, \dots, x_n of iid X_1, \dots, X_n .
- Let $x_{(1)} \leq \dots \leq x_{(n)}$ denote the corresponding order statistics and note that $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I_{\{x_i \leq x\}} = \frac{1}{n} \sum_{i=1}^n I_{\{x_{(i)} \leq x\}}$, $x \in \mathbb{R}$, i.e. the order statistics contain all relevant information about x_1, \dots, x_n .
- Possible graphical tests (see also the appendix):
 - ▶ P-P plot: Plot $\{(p_i, F(x_{(i)})) : i = 1, \dots, n\}$, where $p_i \approx \frac{i-1/2}{n} \approx \frac{i}{n}$ $\approx \hat{F}_n(x_{(i)})$. If $F \approx \hat{F}_n$, the points lie roughly on a line with slope 1; this also applies to Q-Q plots.
 - ▶ Q-Q plot: Plot $\{(F^{-1}(p_i), x_{(i)}) : i = 1, \dots, n\}$ (tail differences better visible).

Interpreting Q-Q plots (**S-shape** hints at **heavier tails** than $N(\mu, \sigma^2)$):



Daily returns typically have kurtosis $\kappa > 3$ (*leptokurtic*; narrower center, heavier tails than $N(\mu, \sigma^2)$ for which $\kappa = 3$). They have typically **power-like tails** rather than exponential.

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 iid, less heavy-tailed).
- The (non-overlapping) h -period log-return at $t \in \{h, 2h, \dots, \lfloor \frac{n}{h} \rfloor h\}$ is

$$X_t^{(h)} = \log\left(\frac{S_t}{S_{t-h}}\right) = \log\left(\frac{S_t}{S_{t-1}} \frac{S_{t-1}}{S_{t-2}} \dots \frac{S_{t-h+1}}{S_{t-h}}\right) = \sum_{k=0}^{h-1} X_{t-k}$$

A Central Limit Theorem (CLT) effect takes place (less heavy-tailed, less evidence of serial correlation).

- Problem: The larger h , the less data is available.
- Possible remedy: Consider overlapping returns

$$\left\{ X_t^{(h)} : t \in \left\{ h, h+k, \dots, h + \left\lfloor \frac{n-h}{k} \right\rfloor k \right\} \right\}, \quad 1 \leq k < h.$$

⇒ More data but serially dependent now.

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

- (U1) Return series are not iid 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

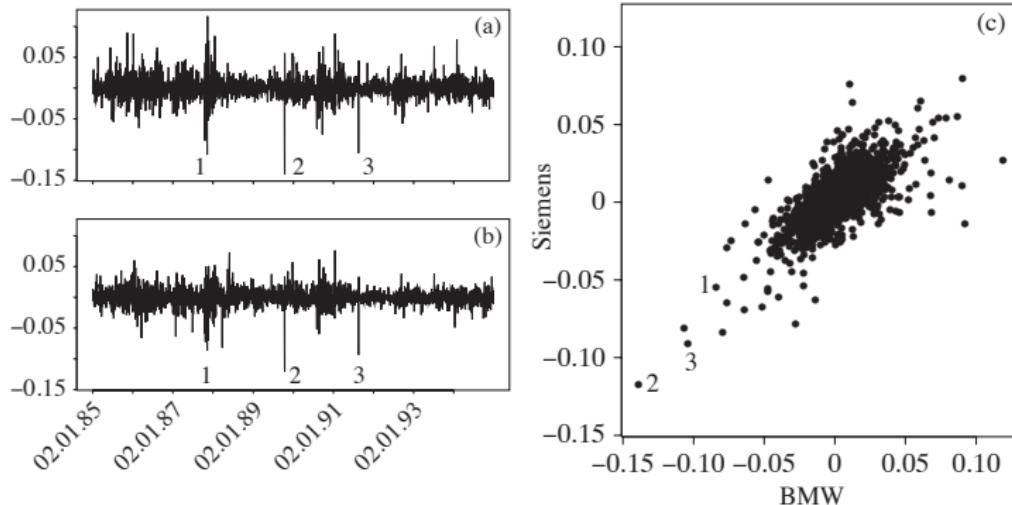
Consider multivariate log-return data $\mathbf{X}_1, \dots, \mathbf{X}_n$.

3.2.1 Correlation between series

- By (U1), the returns of stock A at t and $t + h$ show little correlation. The same applies to 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 (*contemporaneous dependence*).
- 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).
- 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 $t + h$.

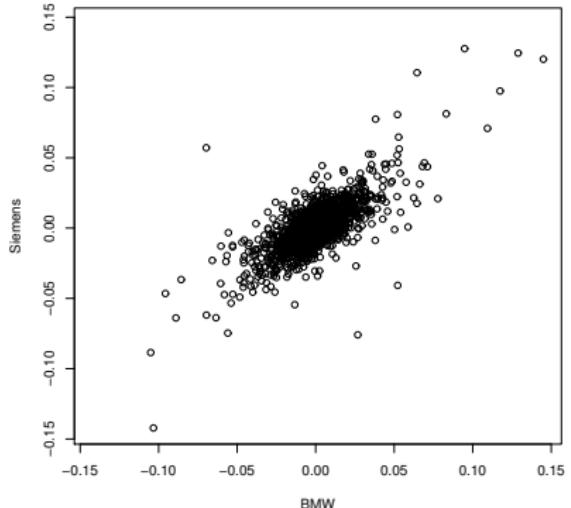
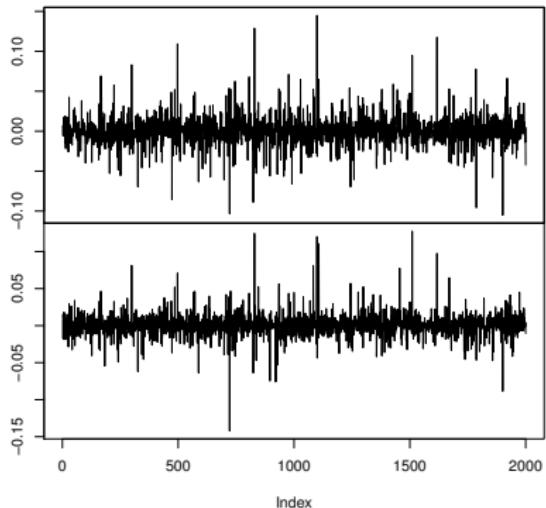
3.2.2 Tail dependence

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



In volatile/extreme periods, dependence is stronger (1: 1987-10-19 Black Monday (DJ drop by 22%); 2: 1989-10-16 Monday demonstrations in Leipzig (Wende); 3: 1991-08-19 coup against soviet president M. 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 in a symmetric way.
- The multivariate normal distribution cannot replicate such behaviour, known as tail dependence; see Chapter 7.

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 (i.e. at the same t);
- (M2) Multivariate series of absolute returns show profound cross-correlation;
- (M3) Correlations between contemporaneous returns vary over time;
- (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

4.1.1 Basic definitions

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.

Definition 4.1 (Mean function, autocovariance function)

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

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

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

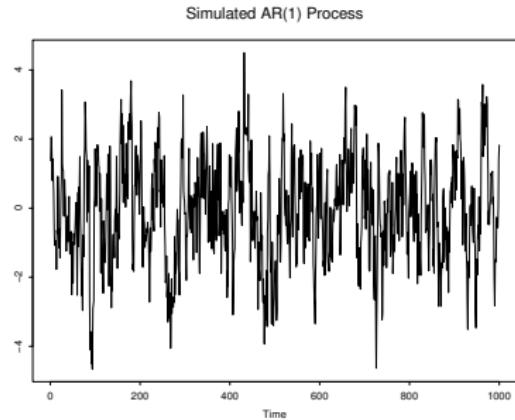
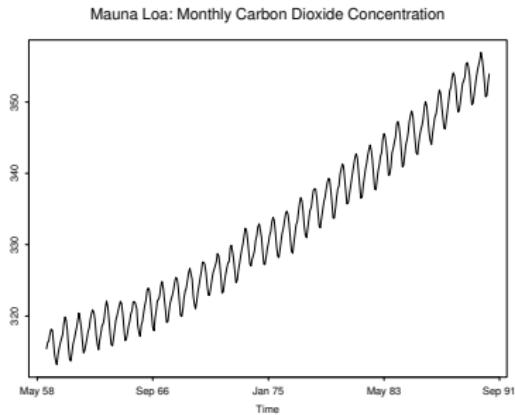
Definition 4.2 ((Weak/strict) stationarity)

- 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 4.3

- 1) Both types of stationarity formalize that $(X_t)_{t \in \mathbb{Z}}$ behaves similarly in any epoch.
- 2)
 - Strict stationarity $\not\Rightarrow$ stationarity if $\mathbb{E}(X_t^2)$ doesn't exist (e.g. GARCH processes). If it does, " \Rightarrow " holds.
 - Stationarity $\not\Rightarrow$ strict stationarity because $\mathbb{E}(|X_t|^p)$, $p > 2$, could change.
- 3) $\gamma(0, t - s) = \gamma(s, t) = \gamma(t, s) = \gamma(0, s - t)$, so $\gamma(t, s)$ only depends on the lag $h = |t - s|$. We can thus write $\gamma(h) := \gamma(0, |h|)$, $h \in \mathbb{Z}$.

Stationary?



(Partial) autocorrelation in stationary time series

Definition 4.4 (ACF)

The *autocorrelation function (ACF)* (or *serial correlation*) of a stationary time series $(X_t)_{t \in \mathbb{Z}}$ is defined by

$$\rho(h) := \text{corr}(X_0, X_h) = \gamma(h)/\gamma(0), \quad h \in \mathbb{Z}.$$

The study of autocorrelation is known as *analysis in the time domain*.

Another important quantity is the *partial autocorrelation function (PACF)* ϕ , defined by

$$\phi(h) := \text{corr}(X_0 - P_{\mathcal{H}_{h-1}}X_0, X_h - P_{\mathcal{H}_{h-1}}X_h),$$

where $P_{\mathcal{H}_{h-1}}X_t$ denotes the best approximation/prediction of X_t from an element of $\mathcal{H}_{h-1} = \{\sum_{k=1}^{h-1} \alpha_k X_{h-k} : \alpha_1, \dots, \alpha_{h-1} \in \mathbb{R}\}$. Note that $\phi(1) = \phi_{1,1} = \gamma(1)/\gamma(0) = \rho(1)$.

- The PACF is the corr between X_0 and X_h with the linear dependence of X_1, \dots, X_{h-1} removed.
- It can be used for model identification of $\text{AR}(p)$ processes similarly to how the ACF is used for $\text{MA}(q)$ processes (see later).
- It can be computed with the Durbin-Levinson algorithm; see the appendix.

White noise processes

Definition 4.5 ((Strict) white noise)

- 1) $(X_t)_{t \in \mathbb{Z}}$ is a *white noise* process if $(X_t)_{t \in \mathbb{Z}}$ is stationary with $\rho(h) = I_{\{h=0\}}$ (*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 $(X_t)_{t \in \mathbb{Z}}$ is a sequence of iid rvs with $\gamma(0) = \sigma^2 < \infty$. If $\mu(t) = 0$, we write $\text{SWN}(0, \sigma^2)$.

For GARCH processes (see later), we need another notion of noise.

Let $(X_t)_{t \in \mathbb{Z}}$ be a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$. A sequence $(\mathcal{F}_t)_{t \in \mathbb{Z}}$ of σ -algebras is called *filtration* if $\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}$, $t \in \mathbb{Z}$. 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 $X_t \in \mathcal{F}_t$, $t \in \mathbb{Z}$ (X_t is \mathcal{F}_t -measurable).

Definition 4.6 (MGDS)

$(X_t)_{t \in \mathbb{Z}}$ 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 $(\varepsilon_t)_{t \in \mathbb{Z}} \sim \text{WN}(0, \sigma^2)$;
 - ▶ $\mathbb{E}(\varepsilon_{t+1+k} | \mathcal{F}_t) = \mathbb{E}(\mathbb{E}(\varepsilon_{t+1+k} | \mathcal{F}_{t+k}) | \mathcal{F}_t) = 0$, $k \in \mathbb{N}$.

4.1.2 ARMA processes

Definition 4.7 (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 $t \in \mathbb{Z}$,

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}. \quad (6)$$

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

Remark 4.8

- 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 (6) can be written as $\phi(B)X_t = \theta(B)\varepsilon_t$, $t \in \mathbb{Z}$, where B denotes the *backshift operator* (such that $B^k X_t = X_{t-k}$) and $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$.

Causal processes

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

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

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

Proposition 4.9 (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

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

Theorem 4.10 (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 (see the appendix for an idea of the proof)

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

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

- If $\theta(z) \neq 0$, $|z| \leq 1$ (known as *invertibility condition*), we can recover ε_t from $(X_s)_{s \leq t}$ via $\varepsilon_t = \phi(B)X_t/\theta(B)$, so ε_t is \mathcal{F}_t -measurable for $\mathcal{F}_t = \sigma(\{X_s : s \leq t\})$ if $(X_t)_{t \in \mathbb{Z}}$ is invertible.
- An ARMA(p, q) process with mean μ can be written as $X_t = \mu_t + \varepsilon_t$ for $\mu_t = \mu + \sum_{k=1}^p \phi_k (X_{t-k} - \mu) + \sum_{k=1}^q \theta_k \varepsilon_{t-k}$. If $(X_t)_{t \in \mathbb{Z}}$ is invertible, $\mu_t \in \mathcal{F}_{t-1}$. If $(\varepsilon_t)_{t \in \mathbb{Z}}$ is a MGDS w.r.t. $(\mathcal{F}_t)_{t \in \mathbb{Z}}$, then $\mu_t = \mathbb{E}(X_t | \mathcal{F}_{t-1})$. Therefore, ARMA processes put structure on the conditional mean μ_t given the past. We will see that GARCH processes put structure on $\sigma_t^2 = \text{var}(X_t | \mathcal{F}_{t-1})$ (helpful for modeling volatility clustering).

Example 4.11

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

- **ACF**: Proposition 4.9 $\Rightarrow \rho(h) = \frac{\sum_{k=0}^{q-|h|} \theta_k \theta_{k+|h|}}{\sum_{k=0}^q \theta_k^2}, |h| \in \{1, \dots, q\}$,
and $\rho(h) = 0$ for all $|h| > q \Rightarrow$ ACF cuts off after lag q .
 - **PACF**: 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) $\text{AR}(p) = \text{ARMA}(p, 0)$: $X_t - \sum_{k=1}^p \phi_k X_{t-k} = \varepsilon_t$. **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 4.9.

Example: By Theorem 4.10, an $\text{AR}(1)$ has a stationary and causal solution if and only if $1 - \phi_1 z \neq 0$ for all $z \in \mathbb{C} : |z| \leq 1$, so $|\phi_1| < 1$.
In this case, $X_t = \phi_1 X_{t-1} + \varepsilon_t = \phi_1(\phi_1 X_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \dots$

$= \phi_1^n X_{t-n} + \sum_{k=0}^{n-1} \phi_1^k \varepsilon_{t-k} \rightarrow \sum_{k=0}^{\infty} \phi_1^k \varepsilon_{t-k}$, so $\psi_k = \phi_1^k$, $k \in \mathbb{N}_0$. By Proposition 4.9,

$$\rho(h) = \frac{\sum_{k=0}^{\infty} \phi_1^{2k+|h|}}{\sum_{k=0}^{\infty} \phi_1^{2k}} = \phi_1^{|h|}, \quad h \in \mathbb{Z},$$

which decreases exponentially.

For AR(p), one can show this from a general form of ψ_k (see Brockwell and Davis (1991, p. 92)), possibly with damped sine waves. Furthermore, one can show that the PACF of an AR(p) cuts off after lag p ; it can be computed with the Durbin–Levinson algorithm; see the appendix.

- 3) ARMA(1, 1): $X_t - \phi_1 X_{t-1} = \varepsilon_t + \theta_1 \varepsilon_{t-1}$ for $|\phi_1| < 1$ has a stationary and causal solution (by Theorem 4.10). For determining the ACF, we first write $X_t = \psi(B)\varepsilon_t$, where

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

$$= \sum_{k=0}^{\infty} \phi_1^k z^k + \sum_{k=1}^{\infty} \theta_1 \phi_1^{k-1} z^k = 1 + \sum_{k=1}^{\infty} \phi_1^{k-1} (\phi_1 + \theta_1) z^k,$$

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 4.9 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.

Remark 4.12

$(X_t)_{t \in \mathbb{Z}}$ is an ARIMA(p, d, q) (Integrated) process if

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

We see that this is also an ARMA($d+p, q$) process. Extensions to SARIMA (Seasonal) models are available; see the appendix.

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 4.13

Let $X_t - \mu = \sum_{k=0}^{\infty} \psi_k Z_{t-k}$ and $(Z_t) \sim \text{SWN}(0, \sigma^2)$. Under suitable conditions,

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

for some covariance matrix W depending on ρ ; see McNeil et al. (2015, Theorem 4.13).

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 (iid) hypothesis of SWN \Rightarrow serial correlation.

Portmanteau tests

- As a formal test of the SWN hypothesis, 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, so is $(X_t^2)_{t \in \mathbb{Z}}$. It is a good idea to also apply the correlogram and Ljung–Box tests to $(|X_t|)_{t \in \mathbb{Z}}$ or $(X_t^2)_{t \in \mathbb{Z}}$.

4.1.4 Statistical analysis of time series

The Box–Jenkins approach

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

- Preliminary analysis

- i) Plot the time series \Rightarrow Does it look stationary?

- ii) 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 estimated via **smoothing with local averages**:

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

or **exponentially weighted moving averages**.

- A **seasonal component** s_t can be estimated by considering

$(\tilde{X}_s)_{s=1}^S$ (e.g. for monthly data, $S = 12$) with

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

Overall, removing μ_t, s_t can be done non-parametrically, via regression, or by taking differences.

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 $(X_t^2)_{t \in \mathbb{Z}}$ (hints at an GARCH). 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) If possible, identify the order and fit the corresponding model; or
- 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 the model assumptions via plots, ACF, Ljung–Box, etc.

4.1.5 Prediction

Let X_{t-n+1}, \dots, X_t denote the available data 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). Two approaches are possible.

Conditional expectation ($\mathbb{E}(X_{t+h} | \mathcal{F}_t)$ is best L^2 approx. to X_{t+h})

Let the ARMA $(X_t)_{t \in \mathbb{Z}}$ be invertible and $(\varepsilon_t)_{t \in \mathbb{Z}}$ be a MGDS w.r.t. $(\mathcal{F}_t)_{t \in \mathbb{Z}}$. Since $\mathbb{E}(X_{t+h} | \mathcal{F}_t)$ minimizes $\mathbb{E}((X_{t+h} - \cdot)^2)$, $P_t X_{t+h} = \mathbb{E}(X_{t+h} | \mathcal{F}_t)$ \Rightarrow 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 4.14 (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

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

$$\mathbb{E}(X_{t+2} | \mathcal{F}_t) = \mu + \phi_1 \mathbb{E}(X_{t+1} | \mathcal{F}_t) - \phi_1 \mu \stackrel{\text{MGDS}}{=} 0$$

$$+ \theta_1 \underbrace{\mathbb{E}(\varepsilon_{t+1} | \mathcal{F}_t)}_{=0} + \underbrace{\mathbb{E}(\varepsilon_{t+2} | \mathcal{F}_t)}_{=0}$$

$$= \mu + \phi_1(\mathbb{E}(X_{t+1} | \mathcal{F}_t) - \mu) = \mu + \phi_1^2(X_t - \mu) + \phi_1 \theta_1 \varepsilon_t;$$

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

Exponentially weighted moving averages

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

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

Increasing $\alpha \in (0, 1)$ puts more weight on the last observation.

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 4.15 (ARCH(p))

Let $(Z_t)_{t \in \mathbb{Z}} \sim \text{SWN}(0, 1)$. $(X_t)_{t \in \mathbb{Z}}$ 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 $\alpha_0 > 0$, $\alpha_k \geq 0$, $k \in \{1, \dots, p\}$.

Typical examples: $Z_t \stackrel{\text{ind.}}{\sim} N(0, 1)$ or $Z_t \stackrel{\text{ind.}}{\sim} t_\nu(0, (\nu - 2)/\nu)$.

Remark 4.16

- 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 σ_t** (conditional standard deviation) **is changing in time**, depending on past values of the process. 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). This is where “autoregressive conditionally heteroscedastic” comes from.

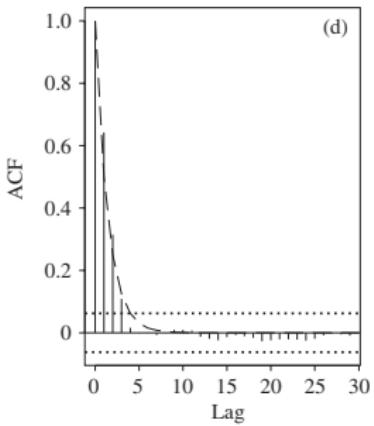
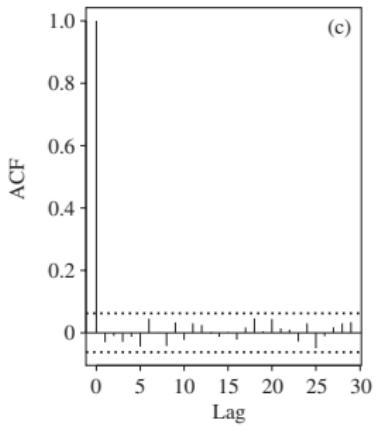
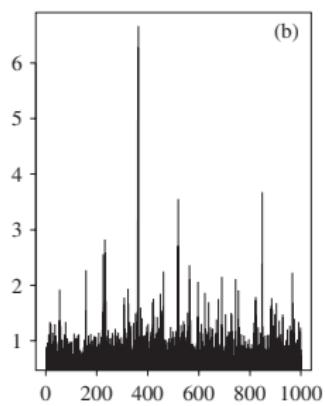
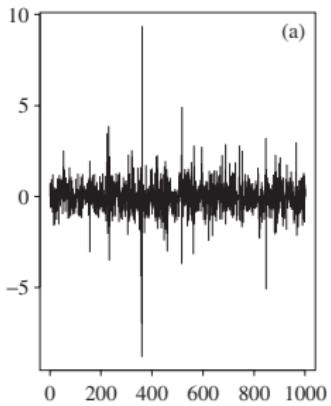
Example 4.17 (ARCH(1))

- One can show 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, $X_t^2 = \alpha_0 \sum_{k=0}^{\infty} \alpha_1^k \prod_{j=0}^k Z_{t-j}^2$.
- $(X_t)_{t \in \mathbb{Z}}$ is stationary $\Leftrightarrow \alpha_1 < 1$. In this case, $\text{var}(X_t) = \alpha_0 / (1 - \alpha_1)$.

Proof of necessity. $X_t^2 = \sigma_t^2 Z_t^2 = (\alpha_0 + \alpha_1 X_{t-1}^2) Z_t^2 \Rightarrow \sigma_X^2 = \mathbb{E}(X_t^2) = \alpha_0 + \alpha_1 \mathbb{E}(X_{t-1}^2 Z_t^2) = \alpha_0 + \alpha_1 \sigma_X^2 \Rightarrow \sigma_X^2 = \frac{\alpha_0}{1 - \alpha_1}, \alpha_1 < 1$. \square

For sufficiency, see McNeil et al. (2015, Proposition 4.18).

- Provided that $\mathbb{E}(Z_t^4) < \infty$ and $\alpha_1 < (\mathbb{E}(Z_t^4))^{-1/2}$, one can show that $\kappa(X_t) = \frac{\mathbb{E}(X_t^4)}{\mathbb{E}(X_t^2)^2} = \frac{\kappa(Z_t)(1 - \alpha_1^2)}{(1 - \alpha_1^2 \kappa(Z_t))}$. If $\kappa(Z_t) > 1$, $\kappa(X_t) > \kappa(Z_t)$. For Gaussian or t innovations, $\kappa(X_t) > 3$ (leptokurtic).
- Parallels with the AR(1) process: If $\mathbb{E}(X_t^4) < \infty$, $\alpha_1 < 1$ and $\varepsilon_t = \sigma_t^2(Z_t^2 - 1)$, one can show that $(X_t^2)_{t \in \mathbb{Z}}$ is an AR(1) of the form $X_t^2 - \frac{\alpha_0}{1 - \alpha_1} = \alpha_1(X_{t-1}^2 - \frac{\alpha_0}{1 - \alpha_1}) + \varepsilon_t$.



- a) $n = 1000$ realizations 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 4.16 1);
- d) Correlogram of $(X_t^2)_{t \in \mathbb{Z}}$ (AR(1)); dashed line = true ACF

4.2.2 GARCH processes

Definition 4.18 (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

$$X_t = \sigma_t Z_t,$$

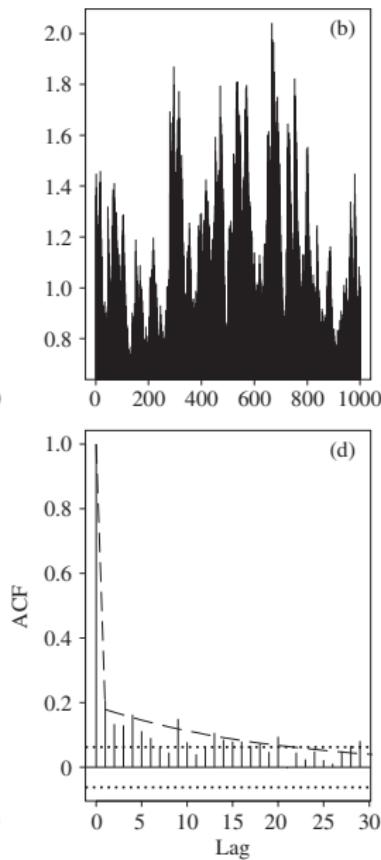
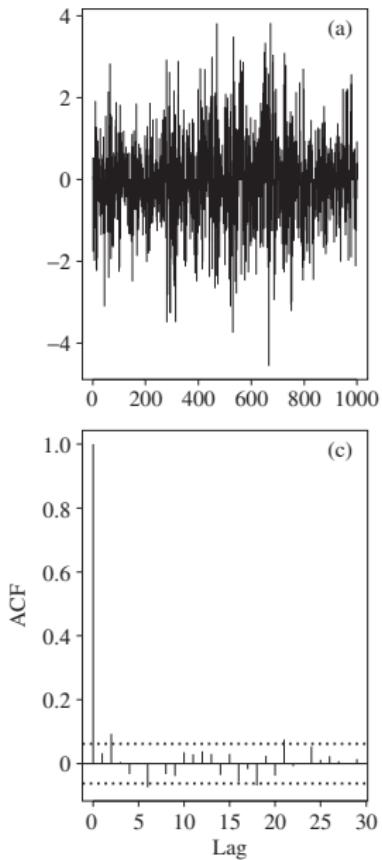
$$\sigma_t^2 = \alpha_0 + \sum_{k=1}^p \alpha_k X_{t-k}^2 + \sum_{k=1}^q \beta_k \sigma_{t-k}^2,$$

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 4.19 (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, $\text{var}(X_t) = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}$.
- Provided that $\mathbb{E}((\alpha_1 Z_t^2 + \beta_1)^2) < 1$ (or $(\alpha_1 + \beta_1)^2 < 1 - (\kappa(Z_t) - 1)\alpha_1^2$), one can show that $\kappa(X_t) = \frac{\kappa(Z_t)(1 - (\alpha_1 + \beta_1)^2)}{1 - (\alpha_1 + \beta_1)^2 - (\kappa(Z_t) - 1)\alpha_1^2}$. If $\kappa(Z_t) > 1$ (Gaussian, scaled t innovations), $\kappa(X_t) > \kappa(Z_t)$.
- Parallels with the ARMA(1,1) process: If $\mathbb{E}(X_t^4) < \infty$, $\alpha_1 + \beta_1 < 1$ and $\varepsilon_t = \sigma_t^2(Z_t^2 - 1)$, one can show that $(X_t^2)_{t \in \mathbb{Z}}$ is an ARMA(1,1) of the form $X_t^2 - \frac{\alpha_0}{1 - \alpha_1 - \beta_1} = (\alpha_1 + \beta_1)(X_{t-1}^2 - \frac{\alpha_0}{1 - \alpha_1 - \beta_1}) + \varepsilon_t - \beta_1 \varepsilon_{t-1}$.



- a) $n = 1000$ realization 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 4.16 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

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

The GARCH(p,q) model

- Higher-order GARCH models have the same general behaviour 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).

4.2.3 Simple extensions of the GARCH model

Consider stationary GARCH processes as white noise for ARMA processes.

Definition 4.20 (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 $\alpha_0 > 0$, $\alpha_k \geq 0$, $k \in \{1, \dots, p_2\}$, $\beta_k \geq 0$, $k \in \{1, \dots, q_2\}$, $\sum_{k=1}^{p_2} \alpha_k + \sum_{k=1}^{q_2} \beta_k < 1$.

- 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:
 - ▶ *GJR-GARCH*. 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 GJR-GARCH) 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 or skewed t distribution.

4.2.4 Fitting GARCH 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

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

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 on 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.

- 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 $\hat{\boldsymbol{\theta}}_n$.

Model checking

- After model fitting, **check its residuals**. We consider an **ARMA** model with **GARCH** errors $X_t = \mu_t + \varepsilon_t = \mu_t + \sigma_t Z_t$; see Definition 4.20.
- We distinguish **two kinds of residuals**:
 - 1) **Unstandardized residuals**. These are the residuals $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ and 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**

$$\hat{Z}_t = \hat{\varepsilon}_t / \hat{\sigma}_t, \quad \hat{\sigma}_t^2 = \hat{\alpha}_0 + \sum_{k=1}^{p_2} \hat{\alpha}_k \hat{\varepsilon}_{t-k}^2 + \sum_{k=1}^{q_2} \hat{\beta}_k \hat{\sigma}_{t-k}^2; \quad (7)$$

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 modelling and modelling of shocks that drive the process;
- ▶ Practical in higher dimensions.

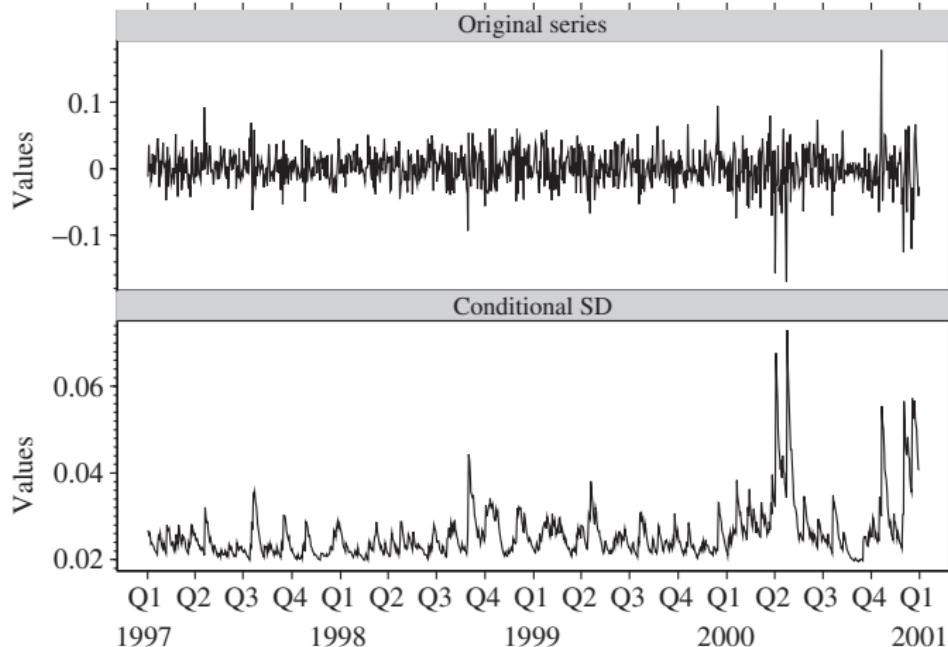
Drawbacks:

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

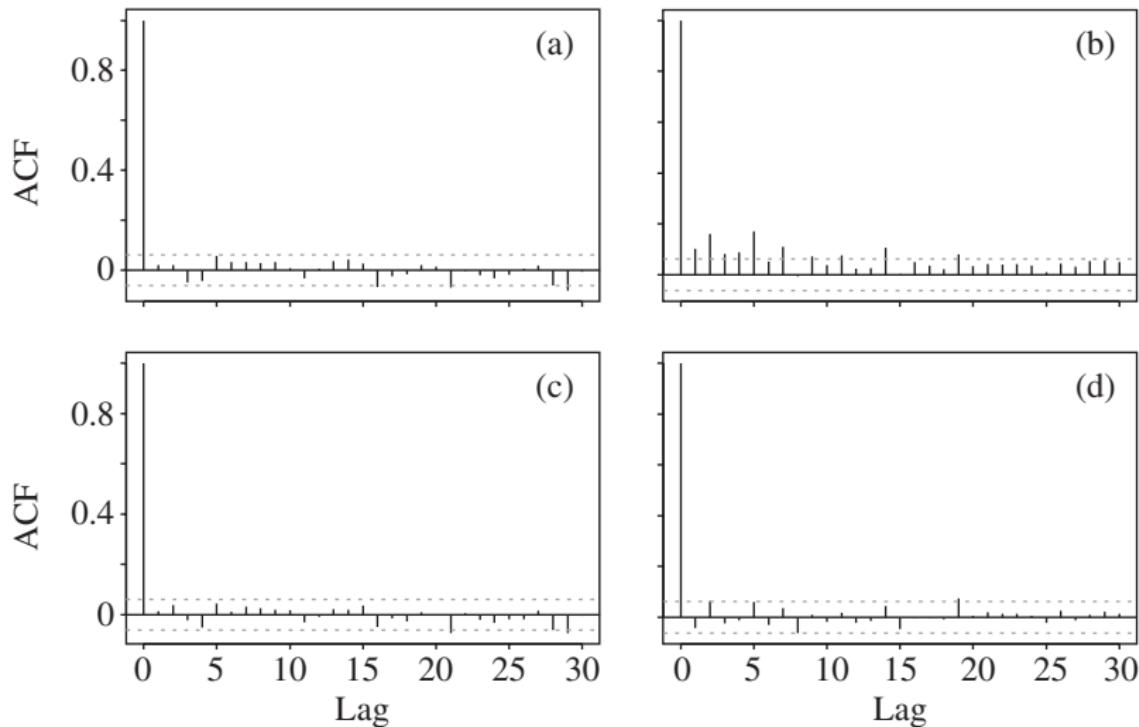
Example 4.21 (GARCH model for Microsoft log-returns)

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

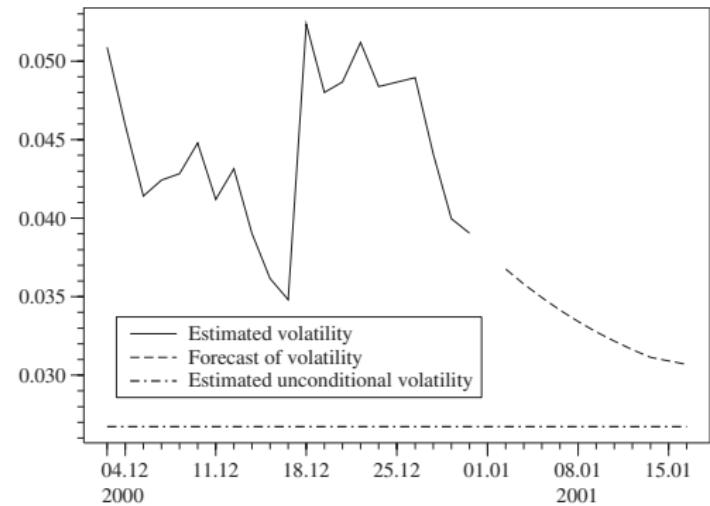
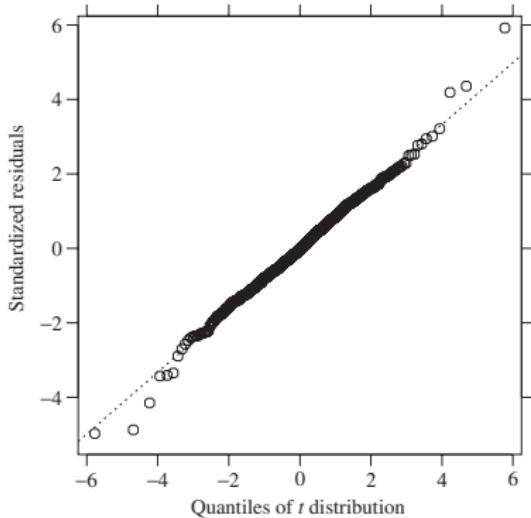
Microsoft log-returns 1997–2000: Data (top) and estimated volatility (bottom) from a GJR-GARCH(1, 1).



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 for a GARCH(1, 1) model.



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 GARCH 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:** Via conditional expectations and via exponentially weighted moving averages.

Conditional expectation

The general procedure becomes clear from the following two examples.

Example 4.22 (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

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

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 4.23 (Prediction in the ARMA(1, 1)–GARCH(1, 1) model)

Let $X_t = \mu_t + \sigma_t Z_t = \mu_t + \varepsilon_t$ as before. It follows from Examples 4.14 and 4.22 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 (7).

Exponentially weighted moving averages

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

$$P_t X_{t+1} = \alpha X_t + (1 - \alpha) P_{t-1} X_t. \quad (8)$$

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

$$P_t (X_{t+1} - \mu_{t+1})^2 = \alpha (X_t - \mu_t)^2 + (1 - \alpha) P_{t-1} (X_t - \mu_t)^2. \quad (9)$$

- Since $\sigma_{t+1}^2 = \mathbb{E}((X_{t+1} - \mu_{t+1})^2 | \mathcal{F}_t)$, we can use (9) 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 Chapter 3). Alternatively, apply exponential smoothing to μ_t via $P_{t-1} X_t$ in (8).

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^t , ES_α^t based on $F_{X_{t+1}|\mathcal{F}_t}$. 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

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

- Then $\text{VaR}_\alpha^t = \mu_{t+1} + \sigma_{t+1} F_Z^\leftarrow(\alpha)$ and $\text{ES}_\alpha^t = \mu_{t+1} + \sigma_{t+1} \text{ES}_\alpha(Z)$.
- If we can estimate μ_{t+1} , σ_{t+1} (parametrically/non-parametrically/semi-parametrically), we only have left to estimate $F_Z^\leftarrow(\alpha)$ and $\text{ES}_\alpha(Z)$.
 - For GARCH-type models it is easy to calculate $F_Z^\leftarrow(\alpha)$ 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

$$\hat{Z}_s = (X_s - \hat{\mu}_s)/\hat{\sigma}_s, \quad s \in \{t-n+1, \dots, n\},$$

to estimate $F_Z^\leftarrow(\alpha)$ and $\text{ES}_\alpha(Z)$.

5 Extreme value theory

5.1 Maxima

5.2 Threshold exceedances

5.1 Maxima

Consider a series of financial losses $(X_k)_{k \in \mathbb{N}}$.

5.1.1 Generalized extreme value distribution

Convergence of sums

Let $(X_k)_{k \in \mathbb{N}}$ be iid with $\mathbb{E}(X_1^2) < \infty$ (mean μ , variance σ^2) and $S_n = \sum_{k=1}^n X_k$. As $n \rightarrow \infty$, $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$ by the Strong Law of Large Numbers (SLLN), so $(\bar{X}_n - \mu)/\sigma \xrightarrow{\text{a.s.}} 0$. By the CLT,

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow[n \uparrow \infty]{\text{d}} N(0, 1) \text{ or } \lim_{n \rightarrow \infty} \mathbb{P}\left(\frac{S_n - d_n}{c_n} \leq x\right) = \Phi(x),$$

where the sequences $c_n = \sqrt{n}\sigma$ and $d_n = n\mu$ give normalization and where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-z^2/2} dz$. 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 (orst-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 maximum* is given by

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

One can show that, for $n \rightarrow \infty$, $M_n \xrightarrow{\text{a.s.}} x_F$ (similar as in the SLLN) where $x_F := \sup\{x \in \mathbb{R} : F(x) < 1\} = F^\leftarrow(1) \leq \infty$ denotes the *right endpoint of F* (similar to the SLLN).

Question: 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* df H (not a unit jump). Then F is in the *maximum domain of attraction* of H ($F \in \text{MDA}(H)$).

One can show that H is determined up to location/scale, i.e. H specifies a unique *type* of distribution. This is guaranteed by the *convergence to types theorem*; see the appendix.

Question: What does H look like?

Definition 5.2 (Generalized extreme value (GEV) distribution)

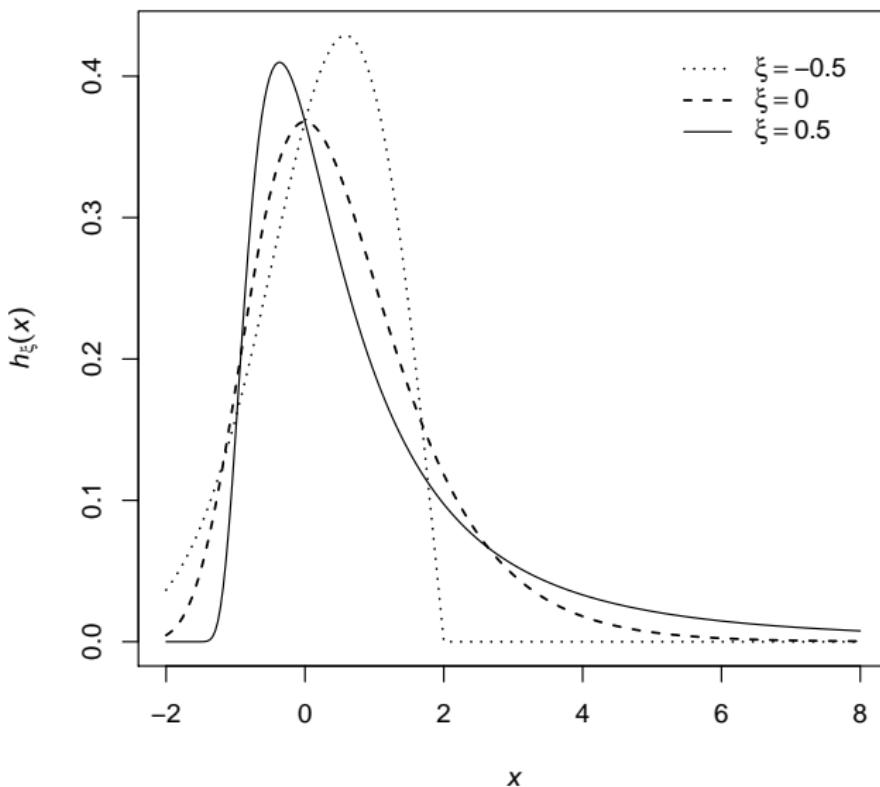
The (standard) *generalized extreme value (GEV) distribution* is given by

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

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 modelling).
- 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

Density h_ξ for $\xi \in \{-0.5, 0, 0.5\}$ (dotted, dashed, solid)



Theorem 5.3 (Fisher–Tippett–Gnedenko)

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

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 non-degenerate limit, the limiting 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.

Example 5.4 (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

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

Example 5.5 (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 $c_n = \kappa n^{1/\theta}/\theta$, $d_n = \kappa(n^{1/\theta} - 1)$, $F^n(c_n x + d_n)$ equals

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

Therefore, $F \in \text{MDA}(H_{1/\theta})$.

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 5.6 (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 \mathcal{R}_0 ; e.g. $c, \log \in \mathcal{R}_0$.
- 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 \mathcal{R}_α ; e.g. $x^\alpha L(x) \in \mathcal{R}_\alpha$.

If $\bar{F} \in \mathcal{R}_{-\alpha}$, $\alpha > 0$, the tail of F decays like a power function (Pareto like).

The Fréchet case

Theorem 5.7 (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 $L \in \mathcal{R}_0$. If $F \in \text{MDA}(H_\xi)$, $\xi > 0$, the normalizing sequences can be chosen as $c_n = F^\leftarrow(1 - 1/n)$ and $d_n = 0$, $n \in \mathbb{N}$.

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

- **Interpretation:** Distributions in $\text{MDA}(H_\xi)$, $\xi > 0$, are those whose tails decay like power functions; $\alpha = 1/\xi$ is known as *tail index*.
- If $X \sim F \in \text{MDA}(H_\xi)$, $\xi > 0$, $X \geq 0$, 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).
- **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 5.8 (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 $L(x) = (\kappa^{-1} + x^{-1})^{-\theta} \in \mathcal{R}_0$. We (again) see that $F \in \text{MDA}(H_\xi)$, $\xi > 0$.

The Gumbel case

- The characterization of this class is more complicated; see the appendix and Embrechts et al. (1997, p. 142).
- 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 (except Student t).

The Weibull case

Theorem 5.9 (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 $L \in \mathcal{R}_0$; the normalizing sequences can be chosen as $c_n = x_F - F^\leftarrow(1 - 1/n)$ and $d_n = x_F$, $n \in \mathbb{N}$.

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

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 $X_k \sim F$, $k \in \mathbb{Z}$.

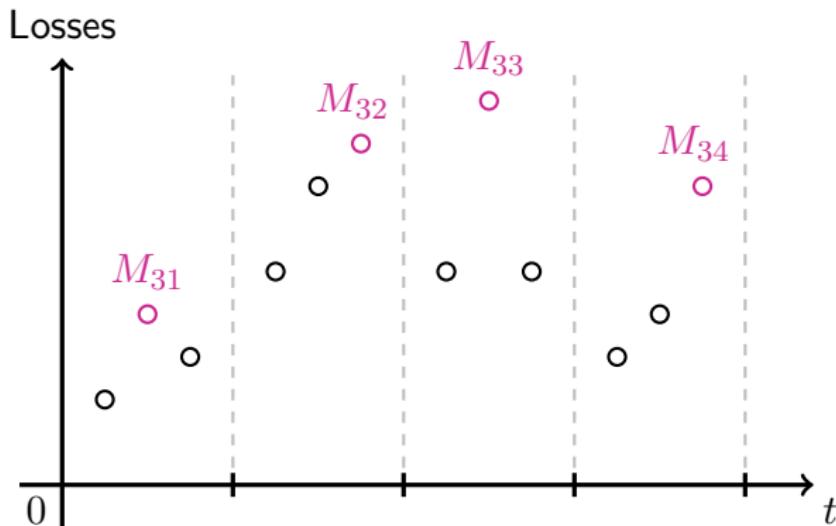
- Let $\tilde{X}_k \stackrel{\text{ind.}}{\sim} F$, $k \in \mathbb{Z}$, and $\tilde{M}_n = \max\{\tilde{X}_1, \dots, \tilde{X}_n\}$. For many processes one can show that there exists a real number $\theta \in (0, 1]$ such that $\lim_{n \uparrow \infty} \mathbb{P}((M_n - d_n)/c_n \leq x) = H^\theta(x)$ if and only if $\lim_{n \uparrow \infty} \mathbb{P}((\tilde{M}_n - d_n)/c_n \leq x) = H(x)$ (non-degenerate); θ is known as the *extremal index*.
- If $F \in \text{MDA}(H_\xi)$ for some $\xi \Rightarrow M_n$ 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 iid 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 iid 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 iid 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 $\text{MDA}(H_\xi)$, $\xi > 0$);
 - ▶ GARCH 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 $\theta = (\xi, \mu, \sigma)$** , divide the realizations into m blocks of size n denoted by M_{n1}, \dots, M_{nm} (e.g. daily log-returns \Rightarrow monthly 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 **density h_ξ** of H_ξ is

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

The log-likelihood is thus

$$\ell(\boldsymbol{\theta}; M_{n1}, \dots, M_{nm}) = \sum_{i=1}^m \log\left(\frac{1}{\sigma} h_\xi\left(\frac{M_{ni} - \mu}{\sigma}\right) I_{\{1+\xi(M_{ni}-\mu)/\sigma > 0\}}\right).$$

Maximize w.r.t. $\boldsymbol{\theta} = (\xi, \mu, \sigma)$ to get $\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\mu}, \hat{\sigma})$.

Remark 5.10

- 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) MLE regularity conditions for consistency and asymptotic efficiency were shown by Smith (1985) for $\xi > -1/2$ (fine for practice).

Return levels and stress losses (exceedances)

Let $M_n \sim H$ (exact or estimated). H can be used to estimate the...

1) ... size of an event with prescribed frequency (*return-level problem*)

- The level $r_{n,k}$ which is expected to be exceeded in one out of every k blocks of size n satisfies $\mathbb{P}(M_n > r_{n,k}) = 1/k$ (e.g. 10-year return level $r_{260,10}$ = level exceeded in one out of every 10y; 260d \approx 1y).
- $r_{n,k} = H^\leftarrow(1 - 1/k)$ is known as *k n-block return level* with parametric estimator $\hat{r}_{n,k} = H_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}^\leftarrow(1 - 1/k) = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}((- \log(1 - 1/k))^{-\hat{\xi}} - 1)$.

2) ... frequency of an event with prescribed size (*return-period problem*)

- The number $k_{n,u}$ of n -blocks for which we expect to see a single n -block exceeding u satisfies $r_{n,k_{n,u}} = u$.
- $k_{n,u} = 1/\bar{H}(u)$ is known as *return period of* the event $\{M_n > u\}$ with parametric estimator $\hat{k}_{n,u} = 1/\bar{H}_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}(u)$.

Example 5.11 (Block maxima analysis of S&P500)

Suppose it is Friday 1987-10-16; the Friday before Black Monday (1987-10-19). The S&P 500 index fell by 9.12% this week. On that Friday alone the index is down 5.16%. We fit a GEV distribution to (bi)annual maxima of daily negative log-returns $X_t = \log(S_t/S_{t-1})$ since 1960-01-01.

Analysis 1: Annual maxima ($m = 28$; including the latest from the incomplete year 1987): $\hat{\theta} = (0.30, 0.02, 0.007) \Rightarrow$ heavy-tailed Fréchet distribution (infinite fourth moment). The corresponding standard errors are $(0.21, 0.002, 0.001) \Rightarrow$ High uncertainty (m small) for estimating ξ .

Analysis 2: Biannual maxima ($m = 56$): $\hat{\theta} = (0.34, 0.02, 0.006)$ with standard errors $(0.14, 0.0009, 0.0008) \Rightarrow$ Even heavier tails.
In what follows we work with the annual maxima.

- What is the probability that next year's maximal risk-factor change exceeds all previous ones? $1 - H_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}(\text{"all previous maxima"})$

- Was a risk-factor change of the size as of Black Monday foreseeable?
 - ▶ Based on data up to and including Friday 1987-10-16, the 10-year return level $r_{260,10}$ is estimated as $\hat{r}_{260,10} = 4.42\%$.
 - ▶ Index drop Black Monday: $20.47\% \Rightarrow X_{t+1} = 22.9\% \gg \hat{r}_{260,10}$.
 - ▶ One can show that 22.9% is in the 95% confidence interval of $r_{260,50}$ (estimated as $\hat{r}_{260,50} = 7.49\%$), but the 28 maxima are too few to get a reliable estimate of a once-in-50-years event.
 - Based on the available data, what is the (estimated) return period of a risk-factor change at least as large as on Black Monday?
 - ▶ The estimated return period $k_{260,0.229}$ is $\hat{k}_{260,0.229} = 1877$ years.
 - ▶ One can show that the 95% confidence interval encompasses everything from 45y to essentially never! \Rightarrow Very high uncertainty!
- \Rightarrow On 1987-10-16 we did not have enough data to say anything meaningful about such an event. Quantifying such events is difficult.

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 5.12 (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 $\xi < 0$.

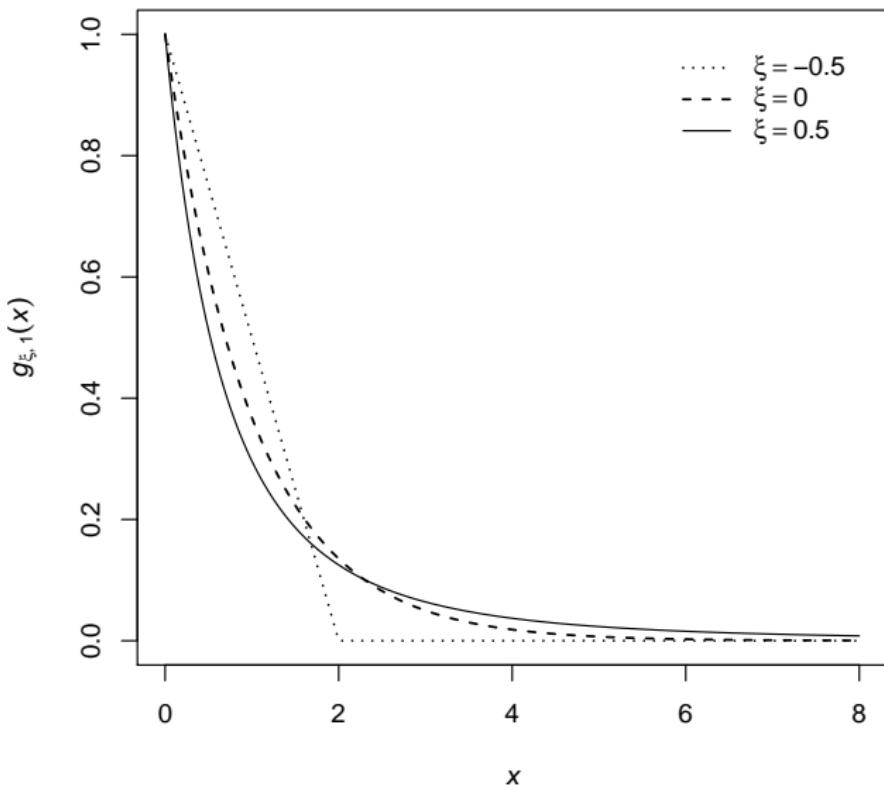
- The parameterization is continuous in ξ .
- The larger ξ , the heavier tailed $G_{\xi,\beta}$ (if $\xi > 0$, $\mathbb{E}(X^k) = \infty$ iff $k \geq \frac{1}{\xi}$; if $\xi < 1$, then $\mathbb{E}X = \beta/(1 - \xi)$).
- ξ is known as *shape*; β as *scale*. Special cases:
 - 1) $\xi > 0$: $\text{Par}(1/\xi, \beta/\xi)$
 - 2) $\xi = 0$: $\text{Exp}(1/\beta)$
 - 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!).

- $G_{\xi,\beta} \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$.

Density $g_{\xi,1}$ for $\xi \in \{-0.5, 0, 0.5\}$ (dotted, dashed, solid)



Definition 5.13 (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

$$e(u) = \mathbb{E}(X - u \mid X > u) \quad (\text{i.e. the mean w.r.t. } F_u)$$

Interpretation

F_u describes the distribution of the excess loss over u , given that u is exceeded. $e(u)$ is the mean of F_u as a function in u .

- One can show the useful formula $e(u) = \frac{1}{F(u)} \int_u^{x_F} \bar{F}(x) dx$.
- For continuous $X \sim F$ with $\mathbb{E}|X| < \infty$, the following formula holds:

$$\text{ES}_\alpha(X) = e(\text{VaR}_\alpha(X)) + \text{VaR}_\alpha(X), \quad \alpha \in (0, 1) \quad (10)$$

Example 5.14 (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). 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 5.15 (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 $F \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$.

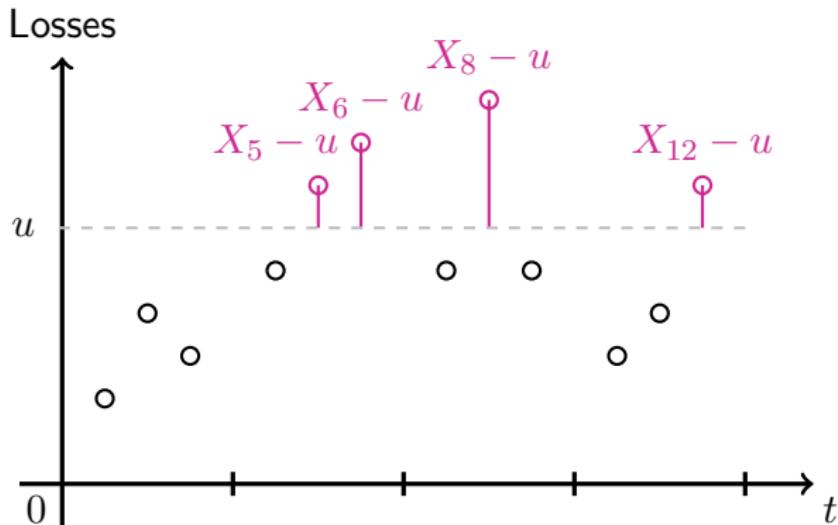
Proof. Non-trivial; see, e.g. Pickands (1975) and Balkema and de Haan (1974). \square

Interpretation

- GPD = Canonical df for modelling 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 Modelling 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 *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 iid 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 $k \in \{1, \dots, N_u\}$.

Excesses over higher thresholds

Once a model is fitted to F_u , we can infer a model for F_v , $v \geq u$.

Lemma 5.16

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

$$\begin{aligned}\bar{F}_v(x) &= \frac{\bar{F}(v+x)}{\bar{F}(v)} = \frac{\bar{F}(u + (v+x-u))}{\bar{F}(u)} \frac{\bar{F}(u)}{\bar{F}(u + (v-u))} \\ &= \frac{\bar{F}_u(v+x-u)}{\bar{F}_u(v-u)} = \frac{\bar{G}_{\xi,\beta}(x+v-u)}{\bar{G}_{\xi,\beta}(v-u)} \stackrel{\text{check}}{=} \bar{G}_{\xi,\beta+\xi(v-u)}(x) \quad \square\end{aligned}$$

\Rightarrow 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 $\xi < 1$ (so if it exists), the mean excess function is given by

$$e(v) = \frac{\xi}{1-\xi} v + \frac{\beta - \xi u}{1-\xi}, \quad v \in [u, \infty) \text{ if } \xi \in [0, 1), \quad (11)$$

and $v \in [u, u - \beta/\xi]$ if $\xi < 0$. This forms the basis for a graphical method for choosing u .

Sample mean excess plot and choice of the threshold

Definition 5.17 (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

$$e_n(v) = \frac{\sum_{i=1}^n (X_i - v) I_{\{X_i > v\}}}{\sum_{i=1}^n I_{\{X_i > v\}}}, \quad v < X_{(n)}.$$

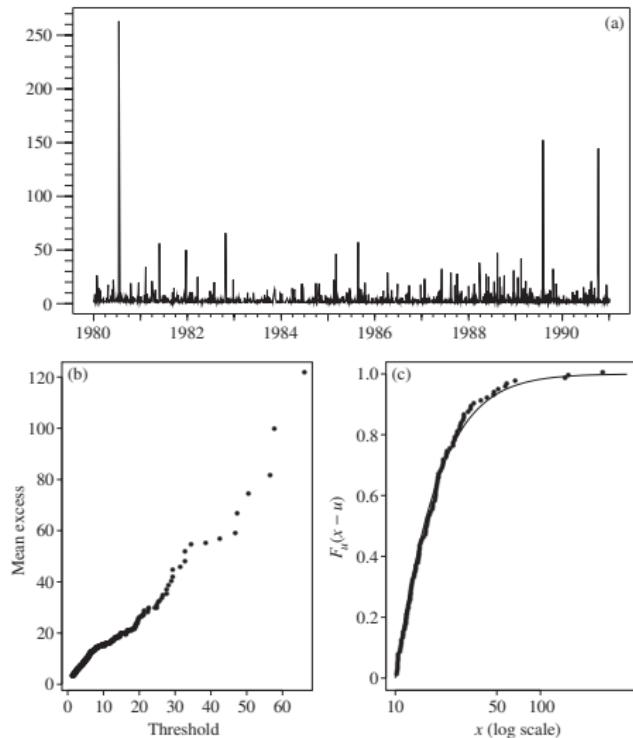
The *mean excess plot* is the plot of $\{(X_{(i)}, e_n(X_{(i)})) : 1 \leq i \leq n-1\}$, where $X_{(i)}$ denotes the i th order statistic.

- If the data supports the GPD model over u , $e_n(v)$ should become increasingly “linear” for higher values of $v \geq u$. An upward/zero/downward trend indicates whether $\xi > 0/\xi = 0/\xi < 0$.
- 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.
- 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 $L \in \mathcal{R}_0$).
- One should always analyze the data for several u and check the sensitivity of the choice of u .

Example 5.18 (Danish fire loss data)

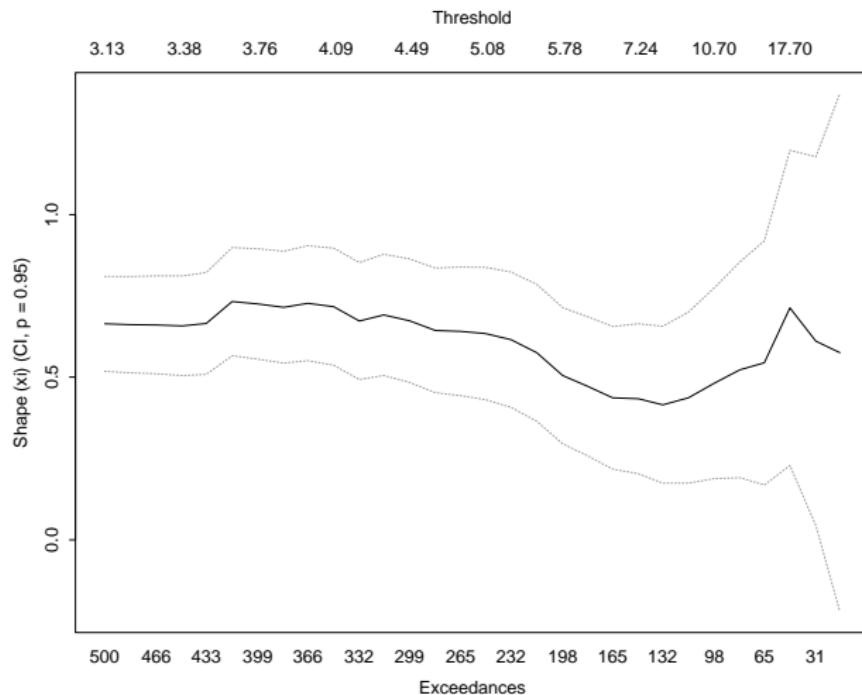
- 2156 fire insurance losses over 1M Danish kroner from 1980-01-03 to 1990-12-31; combined loss for a building and its contents, in some cases also a loss of business earnings. The losses are inflation adjusted to reflect values as of 1985.
- The mean excess function shows a “kink” below 10; “straightening 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 (11) 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 :



⇒ The higher u , the wider the confidence intervals (also support $u = 10$).

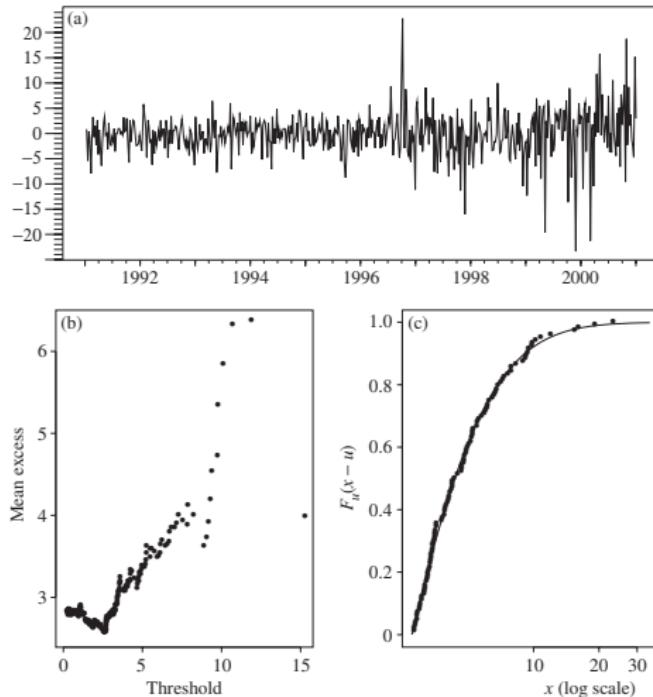
Example 5.19 (AT&T weekly loss data)

- 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 data over 1993–2000 is not consistent with the iid 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 Modelling 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:

$$\begin{aligned}\bar{F}(x) &= \mathbb{P}(X > u)\mathbb{P}(X > x | X > u) \\ &= \bar{F}(u)\mathbb{P}(X - u > x - u | X > u) = \bar{F}(u)\bar{F}_u(x - u) \\ &= \bar{F}(u)\left(1 + \xi \frac{x - u}{\beta}\right)^{-1/\xi}, \quad x \geq u.\end{aligned}$$

- Assuming we know $\bar{F}(u)$, inverting this formula for $\alpha \geq F(u)$ leads to

$$\text{VaR}_\alpha = F^\leftarrow(\alpha) = u + \frac{\beta}{\xi} \left(\left(\frac{1 - \alpha}{\bar{F}(u)} \right)^{-\xi} - 1 \right), \quad (12)$$

$$\text{ES}_\alpha = \frac{\text{VaR}_\alpha}{1 - \xi} + \frac{\beta - \xi u}{1 - \xi}, \quad \xi < 1. \quad (13)$$

The formula for ES_α can also be obtained from $e(\cdot)$ via (10) and (11).

- $\bar{F}(x)$, 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.
- In this spirit, Smith (1987) proposed the *tail estimator*

$$\hat{\bar{F}}(x) = \frac{N_u}{n} \left(1 + \hat{\xi} \frac{x - u}{\hat{\beta}} \right)^{-1/\hat{\xi}}, \quad x \geq u;$$

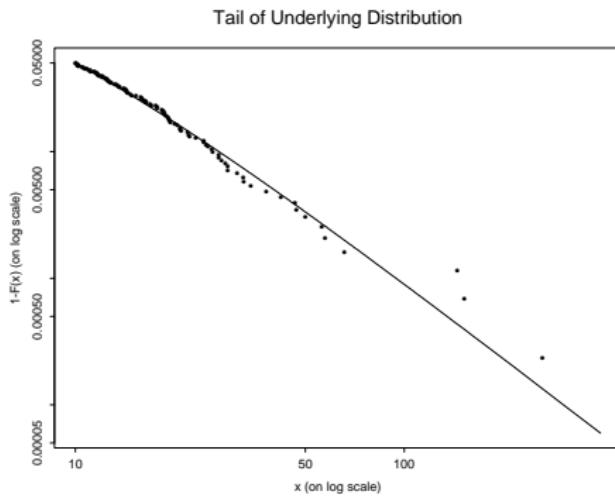
also known as the *Smith estimator* (note that it is only valid for $x \geq u$). It faces a **bias-variance tradeoff**: If u is increased, the bias of parametrically estimating $\bar{F}_u(x - u)$ decreases, but the variance of it and the nonparametrically estimated $\bar{F}(u)$ increases.

- Similarly, GPD-based $\widehat{\text{VaR}}_\alpha$, $\widehat{\text{ES}}_\alpha$ for $\alpha \geq 1 - N_u/n$ can be obtained from (12), (13).

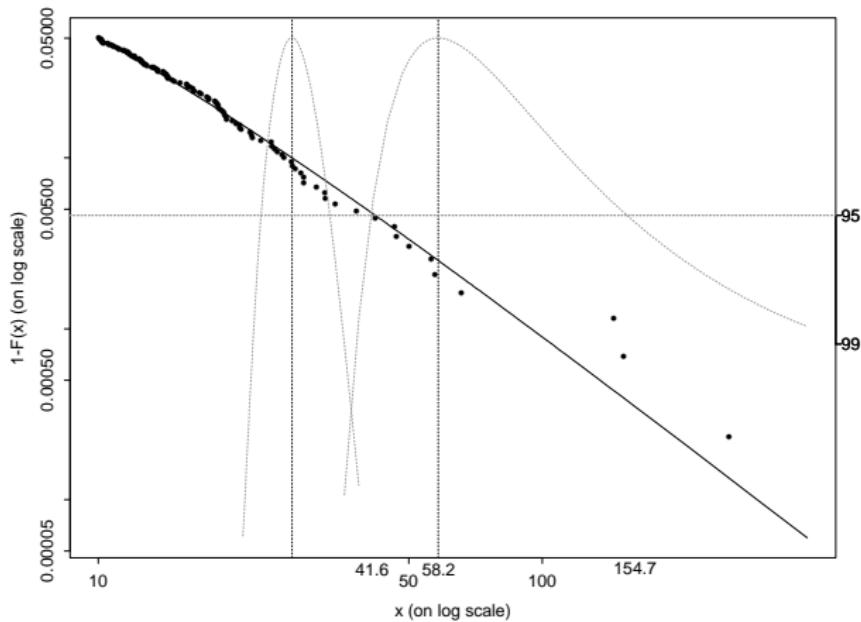
- Confidence intervals for $\bar{F}(x)$, $x \geq u$, VaR_α , ES_α can be obtained likelihood-based (neglecting the uncertainty in N_u/n): Reparametrize the GPD model in terms of $\phi = g(\xi, \beta, N_u/n)$ and construct a confidence interval for ϕ based on the likelihood ratio test.

Example 5.20 (Danish fire loss data (continued))

The semi-parametric Smith/tail estimator $\hat{\bar{F}}(x)$, $x \geq u$ is given by:

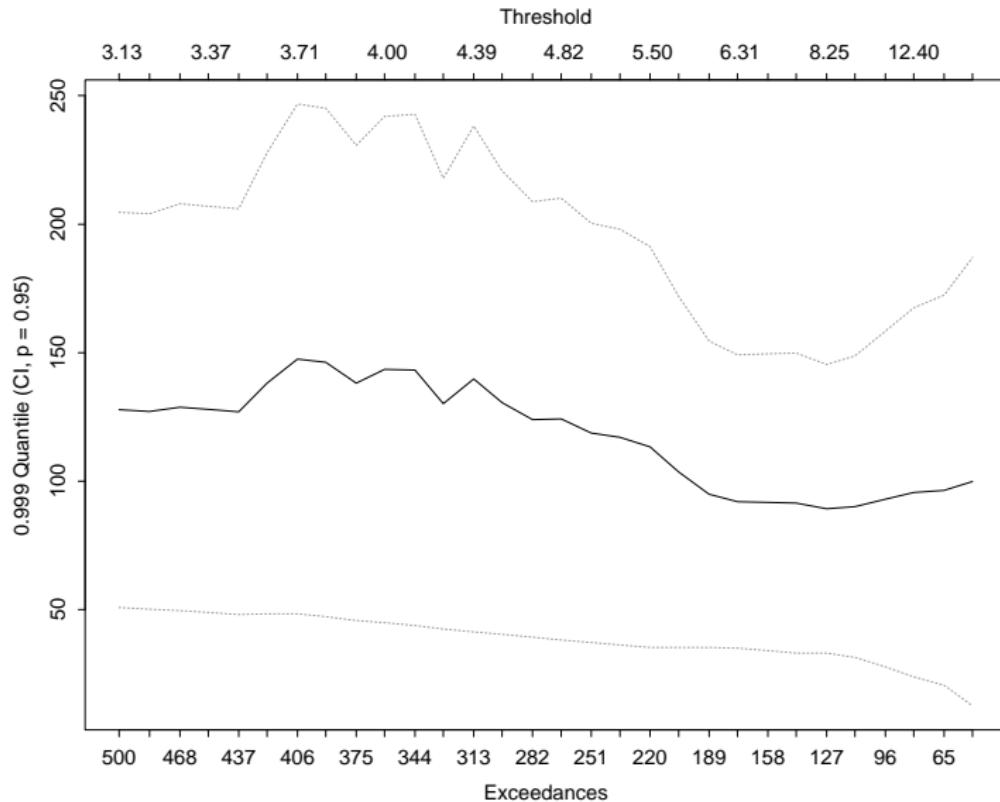


Here are $\hat{F}(x)$, $x \geq u$, $\widehat{\text{VaR}}_{0.99}$, $\widehat{\text{ES}}_{0.99}$ including confidence intervals.

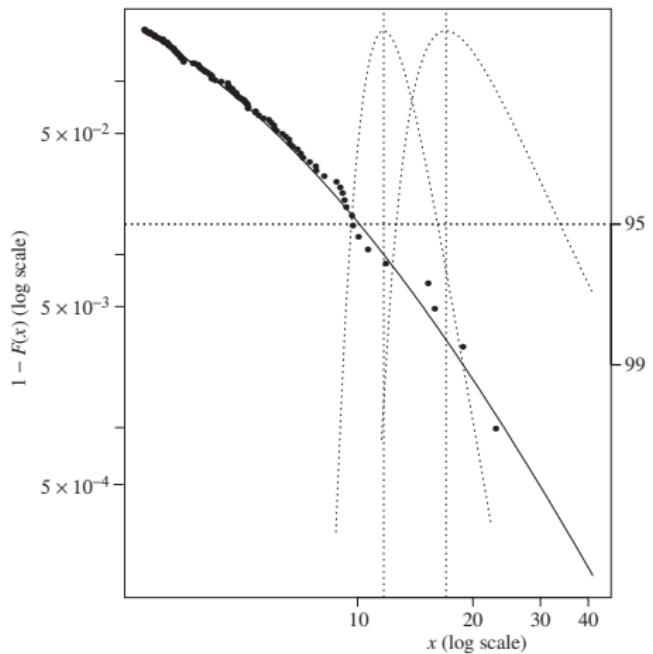


Log-log scale often helpful: If $\bar{F}(x) = x^{-\alpha} L(x)$, $\log \bar{F}(x) = -\alpha \log(x) + \log L(x)$ which is approximately linear in $\log x$.

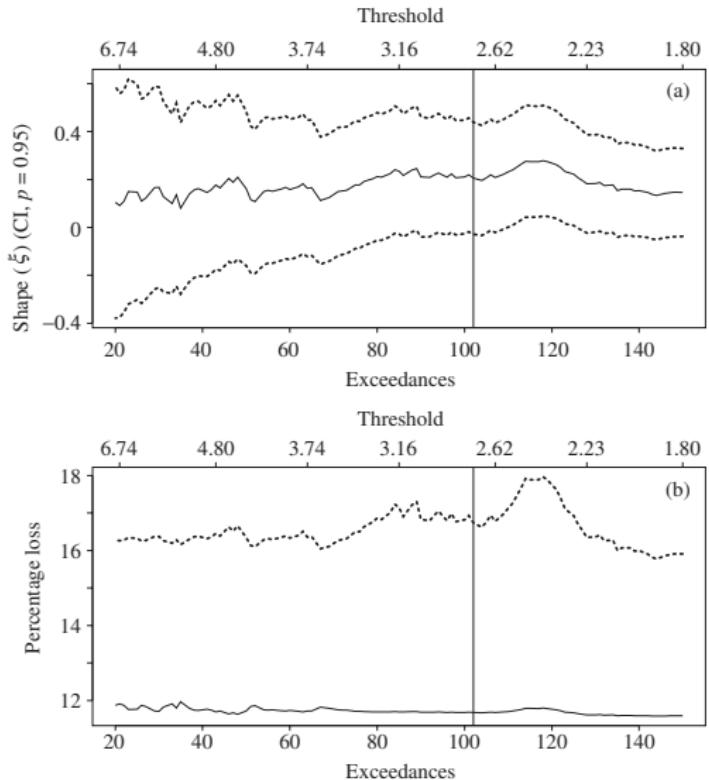
It is important to check the sensitivity of \hat{F} (or $\widehat{\text{VaR}}_\alpha$, $\widehat{\text{ES}}_\alpha$) w.r.t. u .



Example 5.21 (AT&T weekly loss data (continued))



- Fitted GPD model as in Example 5.19.
- Plot of $\hat{F}(x)$.
- Vertical lines: $\widehat{\text{VaR}}_{0.99}$, $\widehat{\text{ES}}_{0.99}$



- 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)

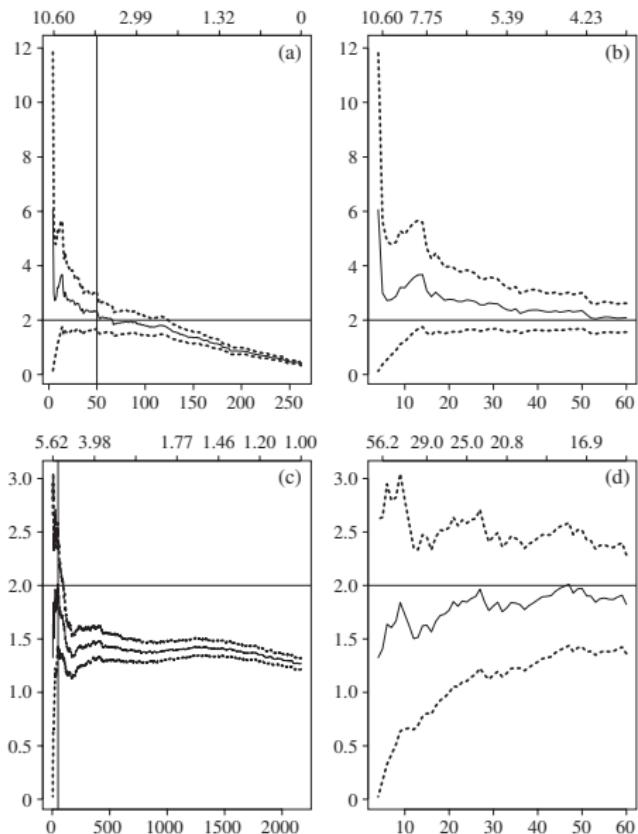
5.2.4 The Hill estimator

- Assume $F \in \text{MDA}(H_\xi)$, $\xi > 0$, so that $\bar{F}(x) = x^{-\alpha}L(x)$, $\alpha > 0$.
- The standard form of the *Hill estimator* of the tail index α is

$$\hat{\alpha}_{k,n}^{(H)} = \left(\frac{1}{k} \sum_{i=1}^k \log X_{i,n} - \log X_{k,n} \right)^{-1}, \quad 2 \leq k \leq n, \quad k \text{ sufficiently small.}$$

Idea: This can be derived by noting that the mean excess function $e(\log u)$ of $\log X$ at $\log u$ is roughly $1/\alpha$ for large u (by Karamata's Theorem), then using $e_n(\log X_{k,n})$ as an estimator for $e(\log u)$ and solving for α ; see the appendix. Note: $X_{1,n} \geq \dots \geq X_{n,n}$.

- Choosing k : Find a small k where the *Hill plot* $\{(k, \hat{\alpha}_{k,n}^{(H)}) : 2 \leq k \leq n\}$ 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 = zoomed-in version of the lhs).
- (a),(b) suggest estimates of $\alpha \in [2, 4]$ ($\xi \in [1/4, 1/2]$; larger than the estimated $\hat{\xi} = 0.22$, see Example 5.19); (c),(d) suggest estimates of $\alpha \in [1.5, 2]$ ($\xi \in [1/2, 2/3]$ (infinite variance!); close to the estimated $\hat{\xi} = 0.50$, see Example 5.18)

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 $X_{k,n}$ (for k sufficiently small).
- Note that $c = u^\alpha \bar{F}(u)$ so $\hat{c} = X_{k,n}^{\hat{\alpha}_{k,n}^{(H)}} \hat{F}_n(X_{k,n}) \approx X_{k,n}^{\hat{\alpha}_{k,n}^{(H)}} \frac{k}{n}$. We thus obtain the semi-parametric *Hill tail estimator*

$$\hat{\bar{F}}(x) = \frac{k}{n} \left(\frac{x}{X_{k,n}} \right)^{-\hat{\alpha}_{k,n}^{(H)}}, \quad x \geq X_{k,n}.$$

- From this result we obtain the semi-parametric *Hill VaR estimator*

$$\widehat{\text{VaR}}_\alpha(X) = \left(\frac{n}{k} (1 - \alpha) \right)^{-\frac{1}{\hat{\alpha}_{k,n}^{(H)}}} X_{k,n}, \quad \alpha \geq F(u) \approx 1 - \frac{k}{n},$$

and, for $\hat{\alpha}_{k,n}^{(H)} > 1$, $\alpha \geq F(u) \approx 1 - \frac{k}{n}$, the semi-param. *Hill ES estimator*

$$\widehat{\text{ES}}_\alpha(X) = \frac{\left(\frac{n}{k} \right)^{-\frac{1}{\hat{\alpha}_{k,n}^{(H)}}} X_{k,n}}{1 - \alpha} \int_\alpha^1 (1 - z)^{-\frac{1}{\hat{\alpha}_{k,n}^{(H)}}} dz = \frac{\hat{\alpha}_{k,n}^{(H)}}{\hat{\alpha}_{k,n}^{(H)} - 1} \widehat{\text{VaR}}_\alpha(X).$$

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

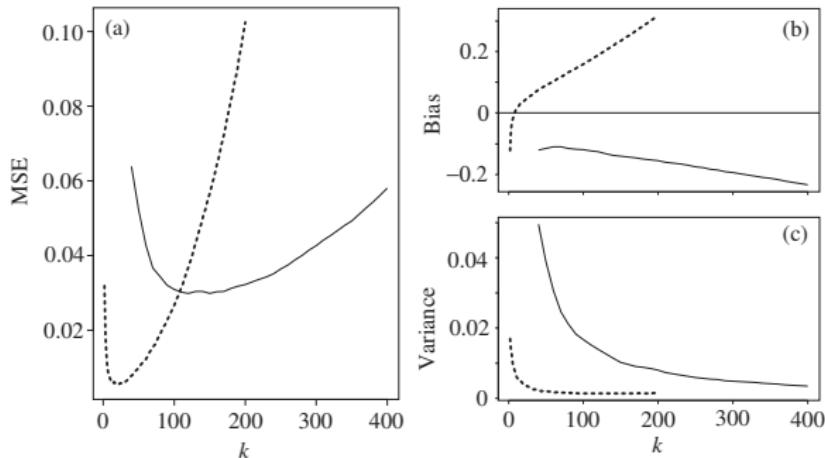
$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \theta)^2) = \mathbb{E}((\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}(\hat{\theta}) - \theta)^2) \\ &= \mathbb{E}((\hat{\theta} - \mathbb{E}[\hat{\theta}])^2) + \mathbb{E}(2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}(\hat{\theta}) - \theta)) + \mathbb{E}((\mathbb{E}[\hat{\theta}] - \theta)^2) \\ &= (\mathbb{E}(\hat{\theta}) - \theta)^2 + \text{var}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{var}(\hat{\theta})\end{aligned}$$

with a Monte Carlo study (based on 1000 samples from a t_4 distribution with corresponding true $\xi = 1/4$) since 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)

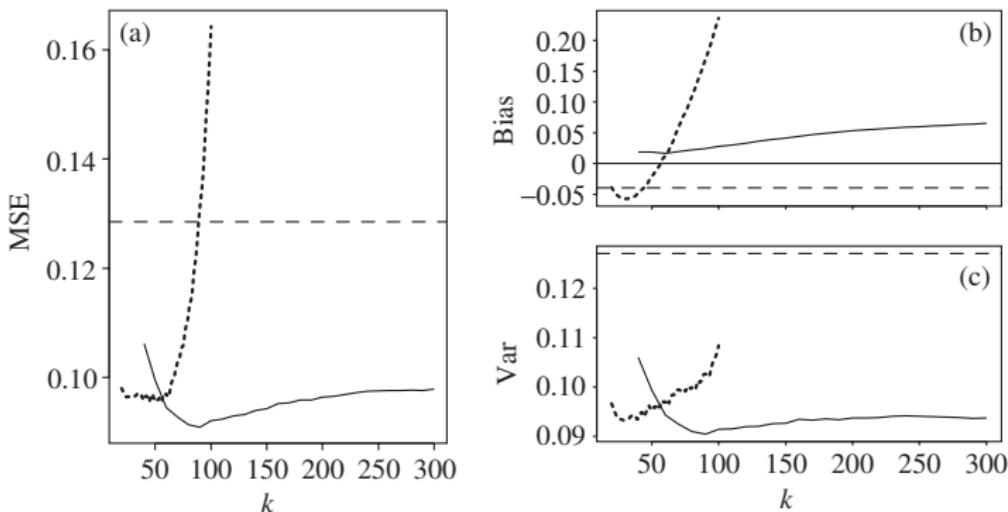


- The Hill estimator outperforms the GPD estimator (optimal k around 20–30) according to the variance for small k (number of order statistics)
- The biases are closer: the Hill (GPD) estimator tends to overestimate (underestimate) ξ .
- For the GPD method, the optimal u is around 100–150 exceedances.

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 $\text{VaR}_{0.99}$ estimator has a negative bias.
- The Hill $\text{VaR}_{0.99}$ estimator has a negative bias for small k but a rapidly growing positive bias for larger k .
- The GPD $\text{VaR}_{0.99}$ estimator has a positive bias which grows much more slowly.
- The GPD $\text{VaR}_{0.99}$ 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 $\text{VaR}_{0.99}$ estimator performs well for $20 \leq k \leq 75$ (we only use k values that lead to a quantile estimate beyond the effective threshold $X_{k,n}$) 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 iid from F_Z) \Rightarrow GPD-based estimates of $\text{VaR}_{\alpha}(Z)$ (see (12)) and $\text{ES}_{\alpha}(Z)$ (see (13)).

6 Multivariate models

- 6.1 Basics of multivariate modelling
- 6.2 Normal mixture distributions
- 6.3 Spherical and elliptical distributions
- 6.4 Dimension reduction techniques

6.1 Basics of multivariate modelling

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) F of \mathbf{X}* is

$$F(\mathbf{x}) = F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}) = \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d), \quad \mathbf{x} \in \mathbb{R}^d.$$

- The *jth margin* or *marginal df F_j of \mathbf{X}* is

$$\begin{aligned} F_j(x_j) &= \mathbb{P}(X_j \leq x_j) \\ &= \mathbb{P}(X_1 \leq \infty, \dots, X_{j-1} \leq \infty, X_j \leq x_j, X_{j+1} \leq \infty, \dots, X_d \leq \infty) \\ &= F(\infty, \dots, \infty, x_j, \infty, \dots, \infty), \quad x_j \in \mathbb{R}, \quad j \in \{1, \dots, d\}. \end{aligned}$$

(interpreted as a *limit*).

- Similarly for *k-dimensional margins*. Suppose we partition \mathbf{X} into $(\mathbf{X}'_1, \mathbf{X}'_2)',$ where $\mathbf{X}_1 = (X_1, \dots, X_k)'$ and $\mathbf{X}_2 = (X_{k+1}, \dots, X_d)',$ then the marginal distribution function of \mathbf{X}_1 is

$$F_{\mathbf{X}_1}(\mathbf{x}_1) = \mathbb{P}(\mathbf{X}_1 \leq \mathbf{x}_1) = F(x_1, \dots, x_k, \infty, \dots, \infty).$$

- F is absolutely continuous if

$$F(\mathbf{x}) \underset{(*)}{=} \int_{-\infty}^{x_d} \dots \int_{-\infty}^{x_1} f(z_1, \dots, z_d) dz_1 \dots dz_d = \int_{(-\infty, \mathbf{x}]} f(\mathbf{z}) d\mathbf{z}$$

for some $f \geq 0$ known as the *(joint) density of \mathbf{X} (or F)*. Similarly, the *jth marginal df F_j is absolutely continuous* if $F_j(x) = \int_{-\infty}^x f_j(z) dz$ for some $f_j \geq 0$ known as the *density of X_j (or F_j)*.

- In case f exists, $F_j(x_j) \underset{(*)}{=} \int_{-\infty}^{x_j} \int_{(-\infty, \infty)} f(\mathbf{z}) d\mathbf{z}_{-j} dz_j = \int_{-\infty}^{x_j} f_j(z_j) dz_j,$ so that $f_j(x_j)$ can be recovered from f via

$$\underbrace{\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty}}_{d-1\text{-many}} f(z_1, \dots, z_{j-1}, x_j, z_{j+1}, \dots, z_d) dz_1 \dots dz_{j-1} dz_{j+1} \dots dz_d.$$

- 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 Chapter 7).
- By **replacing integrals by sums**, one obtains similar formulas for the **discrete case**, in which the notion of densities is replaced by ***probability mass functions***.
- We sometimes work with the ***survival function* \bar{F} of \mathbf{X}** ,

$$\bar{F}(\mathbf{x}) = \bar{F}_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} > \mathbf{x}) = \mathbb{P}(X_1 > x_1, \dots, X_d > x_d), \quad \mathbf{x} \in \mathbb{R}^d,$$

with corresponding ***jth marginal survival function* \bar{F}_j**

$$\begin{aligned}\bar{F}_j(x_j) &= \mathbb{P}(X_j > x_j) \\ &= \bar{F}(-\infty, \dots, -\infty, x_j, -\infty, \dots, -\infty), \quad x_j \in \mathbb{R}, \quad j \in \{1, \dots, d\}.\end{aligned}$$

- Note that $\bar{F}(\mathbf{x}) \neq 1 - F(\mathbf{x})$ in general (unless $d = 1$).

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 $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2) \sim F$. The *conditional df of \mathbf{X}_2 given $\mathbf{X}_1 = \mathbf{x}_1$* is $F_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{x}_1) = \mathbb{P}(\mathbf{X}_2 \leq \mathbf{x}_2 | \mathbf{X}_1 = \mathbf{x}_1) = \mathbb{E}(I_{\{\mathbf{X}_2 \leq \mathbf{x}_2\}} | \mathbf{X}_1 = \mathbf{x}_1)$, where $\mathbb{E}(\cdot | \cdot)$ denotes conditional expectation (not discussed here).
- A useful identity for conditional dfs is

$$F(\mathbf{x}) = \int_{(-\infty, \mathbf{x}_1]} F_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{z}) dF_{\mathbf{X}_1}(\mathbf{z}); \quad (14)$$

see the appendix for a proof.

- ▶ If $\mathbf{x}_1 \rightarrow \infty$, then $F_{\mathbf{X}_2}(\mathbf{x}_2) = \int_{\mathbb{R}^d} F_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{z}) dF_{\mathbf{X}_1}(\mathbf{z})$.
- ▶ If F has a density f , 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 F has density f and f_{X_1} denotes the density of X_1 , then

$$\begin{aligned} f(\mathbf{x}_1, \mathbf{x}_2) &= \frac{\partial^2}{\partial \mathbf{x}_2 \partial \mathbf{x}_1} F(\mathbf{x}_1, \mathbf{x}_2) \stackrel{(14)}{=} \frac{\partial}{\partial \mathbf{x}_2} F_{X_2|X_1}(\mathbf{x}_2 | \mathbf{x}_1) f_{X_1}(\mathbf{x}_1) \\ &= f_{X_2|X_1}(\mathbf{x}_2 | \mathbf{x}_1) f_{X_1}(\mathbf{x}_1). \end{aligned}$$

We call

$$f_{X_2|X_1}(\mathbf{x}_2 | \mathbf{x}_1) = \frac{f(\mathbf{x}_1, \mathbf{x}_2)}{f_{X_1}(\mathbf{x}_1)}$$

the *conditional density of X_2 given $X_1 = \mathbf{x}_1$* . In this case, the conditional df $F_{X_2|X_1}(\mathbf{x}_2 | \mathbf{x}_1)$ is given by

$$F_{X_2|X_1}(\mathbf{x}_2 | \mathbf{x}_1) = \int_{-\infty}^{x_{k+1}} \cdots \int_{-\infty}^{x_d} f_{X_2|X_1}(z_{k+1}, \dots, z_d | \mathbf{x}_1) dz_{k+1} \dots dz_d.$$

- X_1, X_2 are *independent* if $F(\mathbf{x}_1, \mathbf{x}_2) = F_{X_1}(\mathbf{x}_1) F_{X_2}(\mathbf{x}_2)$ for all $\mathbf{x}_1, \mathbf{x}_2$.
- If F has density f , then X_1, X_2 are independent if $f(\mathbf{x}_1, \mathbf{x}_2) = f_{X_1}(\mathbf{x}_1) f_{X_2}(\mathbf{x}_2)$ for all $\mathbf{x}_1, \mathbf{x}_2$. In this case, $f_{X_2|X_1}(\mathbf{x}_2 | \mathbf{x}_1) = f_{X_2}(\mathbf{x}_2)$.

- The components X_1, \dots, X_d of \mathbf{X} are (*mutually*) *independent* if $F(\mathbf{x}) = \prod_{j=1}^d F_j(x_j)$ for all \mathbf{x} or, if F has density f , if $f(\mathbf{x}) = \prod_{j=1}^d f_j(x_j)$ for all \mathbf{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

$$\mathbb{E}\mathbf{X} = (\mathbb{E}X_1, \dots, \mathbb{E}X_d).$$

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 by

$$\text{cov}(\mathbf{X}) = \mathbb{E}((\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})').$$

If we write $\Sigma = \text{cov}(\mathbf{X})$, 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 $\sigma_{jj} = \text{var}(X_j)$, $j \in \{1, \dots, d\}$.

- X_1, X_2 independent $\not\Rightarrow \text{cov}(X_1, X_2) = 0$ (counter-examples can be constructed with **copulas**; see Chapter 7).
- The *cross covariance matrix between* two random vectors \mathbf{X}, \mathbf{Y} is defined by $\text{cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}((\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{Y} - \mathbb{E}\mathbf{Y})')$; 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 by the matrix $\text{corr}(\mathbf{X})$ with (i, j) th element

$$\text{corr}(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}(X_i) \text{var}(X_j)}}, \quad i, j \in \{1, \dots, d\},$$

which is in $[-1, 1]$ with $\text{corr}(X_i, X_j) = \pm 1$ if and only if $X_j \stackrel{\text{a.s.}}{=} aX_i + b$ for some $a \gtrless 0$ and $b \in \mathbb{R}$.

- **Some properties of $\mathbb{E}()$ and $\text{cov}()$:**

1) For all $A \in \mathbb{R}^{k \times d}$, $\mathbf{b} \in \mathbb{R}^k$:

► $\mathbb{E}(A\mathbf{X} + \mathbf{b}) = A\mathbb{E}\mathbf{X} + \mathbf{b} = A\boldsymbol{\mu} + \mathbf{b};$

- $\text{cov}(A\mathbf{X} + \mathbf{b}) = A \text{cov}(\mathbf{X})A' = A\Sigma A'$; if $k = 1$ ($A = \mathbf{a}'$),

$$\mathbf{a}'\Sigma\mathbf{a} = \text{cov}(\mathbf{a}'\mathbf{X}) = \text{var}(\mathbf{a}'\mathbf{X}) \geq 0, \quad \mathbf{a} \in \mathbb{R}^d, \quad (15)$$

i.e. covariance matrices are *positive semidefinite*.

- $\text{cov}(\mathbf{X}_1 + \mathbf{X}_2) = \text{cov}(\mathbf{X}_1) + \text{cov}(\mathbf{X}_2) + 2 \text{cov}(\mathbf{X}_1, \mathbf{X}_2)$

- 2) If Σ is a *positive definite matrix* (i.e. $\mathbf{a}'\Sigma\mathbf{a} > 0$ for all $\mathbf{a} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$), one can show that Σ is invertible.
- 3) A *symmetric, positive (semi)definite Σ* can be written as

$$\Sigma = AA' \quad \text{Cholesky decomposition} \quad (16)$$

for a lower triangular matrix A with $A_{jj} > 0$ ($A_{jj} \geq 0$) for all j . A is known as *Cholesky factor* (and also denoted by $\Sigma^{1/2}$).

- Properties of \mathbf{X} can often be shown with the *characteristic function (cf)*

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}(\exp(it'\mathbf{X})), \quad \mathbf{t} \in \mathbb{R}^d.$$

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 (Characterization of covariance matrices)

A symmetric matrix Σ is a covariance matrix if and only if it is symmetric and positive semidefinite.

Proof.

“ \Rightarrow ” As we have seen in (15), 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 $\text{cov}(A\mathbf{X}) = A \text{cov}(\mathbf{X}) A' = AA' = \Sigma$, i.e. Σ is a covariance matrix (namely that of $A\mathbf{X}$). \square

6.1.2 Standard estimators of covariance and correlation

- Assume $\mathbf{X}_1, \dots, \mathbf{X}_n \sim F$ (daily/weekly/monthly/yearly risk-factor changes) to be **serially uncorrelated** (i.e. multivariate white noise) with $\boldsymbol{\mu} := \mathbb{E}\mathbf{X}_1$, $\Sigma := \text{cov } \mathbf{X}_1$ and $P = \text{corr}(\mathbf{X}_1)$.

- Non-parametric method-of-moments-like estimators of μ , Σ , P are

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \quad (\text{sample mean})$$

$$S = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' \quad (\text{sample covariance matrix})$$

$$R = (R_{ij}) \text{ for } R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} \quad (\text{sample correlation matrix})$$

- Under joint normality (F multivariate normal), $\bar{\mathbf{X}}$, S and R are also MLEs. S is biased, but an unbiased version can be obtained by

$$S_n = \frac{n}{n-1} S.$$

- 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_n is unbiased since

$$\begin{aligned}
 \mathbb{E}S_n &= \frac{1}{n-1} \sum_{i=1}^n \mathbb{E}((\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})') \\
 &= \frac{1}{n-1} \sum_{i=1}^n \mathbb{E}(((\mathbf{X}_i - \boldsymbol{\mu}) - (\bar{\mathbf{X}} - \boldsymbol{\mu}))((\mathbf{X}_i - \boldsymbol{\mu}) - (\bar{\mathbf{X}} - \boldsymbol{\mu}))') \\
 &= \frac{1}{n-1} \sum_{i=1}^n \mathbb{E}((\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})' - (\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})') \\
 &= \frac{1}{n-1} \sum_{i=1}^n (\Sigma - \text{cov } \bar{\mathbf{X}}) \underset{\text{cov}(\bar{\mathbf{X}}) = \frac{\Sigma}{n}}{\stackrel{n}{=}} \frac{n}{n-1} \left(1 - \frac{1}{n}\right) \Sigma = \Sigma.
 \end{aligned}$$

- Further properties of $\bar{\mathbf{X}}, S, R$ depend on F .

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

$$\mathbf{X} \stackrel{\text{d}}{=} \boldsymbol{\mu} + A\mathbf{Z}, \quad (17)$$

where $\mathbf{Z} = (Z_1, \dots, Z_k)$, $Z_l \stackrel{\text{ind.}}{\sim} N(0, 1)$, $A \in \mathbb{R}^{d \times k}$, $\boldsymbol{\mu} \in \mathbb{R}^d$.

- $\mathbb{E}\mathbf{X} = \boldsymbol{\mu} + A\mathbb{E}\mathbf{Z} = \boldsymbol{\mu}$
- $\text{cov}(\mathbf{X}) = \text{cov}(\boldsymbol{\mu} + A\mathbf{Z}) = A \text{cov}(\mathbf{Z})A' = AA' =: \Sigma$

Proposition 6.3 (Cf of the multivariate normal distribution)

Let \mathbf{X} be as in (17) and $\Sigma = AA'$. Then the cf of \mathbf{X} is

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}(\exp(i\mathbf{t}'\mathbf{X})) = \exp\left(i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}\right), \quad \mathbf{t} \in \mathbb{R}^d.$$

Idea of proof. Using the fact that $\phi_Z(t) = \exp(-t^2/2)$ for $Z \sim N(0, 1)$ (see the appendix for a proof), we obtain that

$$\begin{aligned}\phi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}(\exp(i\mathbf{t}'(\boldsymbol{\mu} + A\mathbf{Z}))) \underset{\tilde{\mathbf{t}}' = \mathbf{t}'A}{=} \exp(i\mathbf{t}'\boldsymbol{\mu})\mathbb{E}(\exp(i\tilde{\mathbf{t}}'\mathbf{Z})) \\ &\stackrel{\text{ind.}}{=} \exp(i\mathbf{t}'\boldsymbol{\mu}) \prod_{j=1}^d \mathbb{E}(\exp(i(\tilde{t}_j Z_j))) = \exp\left(i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2} \sum_{j=1}^d \tilde{t}_j^2\right) \\ &= \exp\left(i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\tilde{\mathbf{t}}'\tilde{\mathbf{t}}\right) = \exp\left(i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'AA'\mathbf{t}\right) \\ &= \exp\left(i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right)\end{aligned}$$

□

- We see that the multivariate normal distribution is characterized by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, hence the notation $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be characterized by univariate normal distributions.

Proposition 6.4 (Characterization of $N_d(\mu, \Sigma)$)

$$\mathbf{X} \sim N_d(\mu, \Sigma) \iff \mathbf{a}' \mathbf{X} \sim N(\mathbf{a}' \mu, \mathbf{a}' \Sigma \mathbf{a}) \quad \forall \mathbf{a} \in \mathbb{R}^d.$$

Proof. “ \Rightarrow ” via uniqueness of cfs; “ \Leftarrow ” via Corollary A.10 □

Consequences:

- Margins: $\mathbf{X} \sim N_d(\mu, \Sigma) \stackrel{\mathbf{a} = e_j}{\not\Rightarrow} X_j \sim N(\mu_j, \sigma_{jj}^2), \quad j \in \{1, \dots, d\}.$
- Sums: $\mathbf{X} \sim N_d(\mu, \Sigma) \stackrel{\mathbf{a} = \mathbf{1}}{\Rightarrow} \sum_{j=1}^d X_j \sim N(\sum_{j=1}^d \mu_j, \sum_{i,j} \sigma_{ij}).$

Proposition 6.5 (Density)

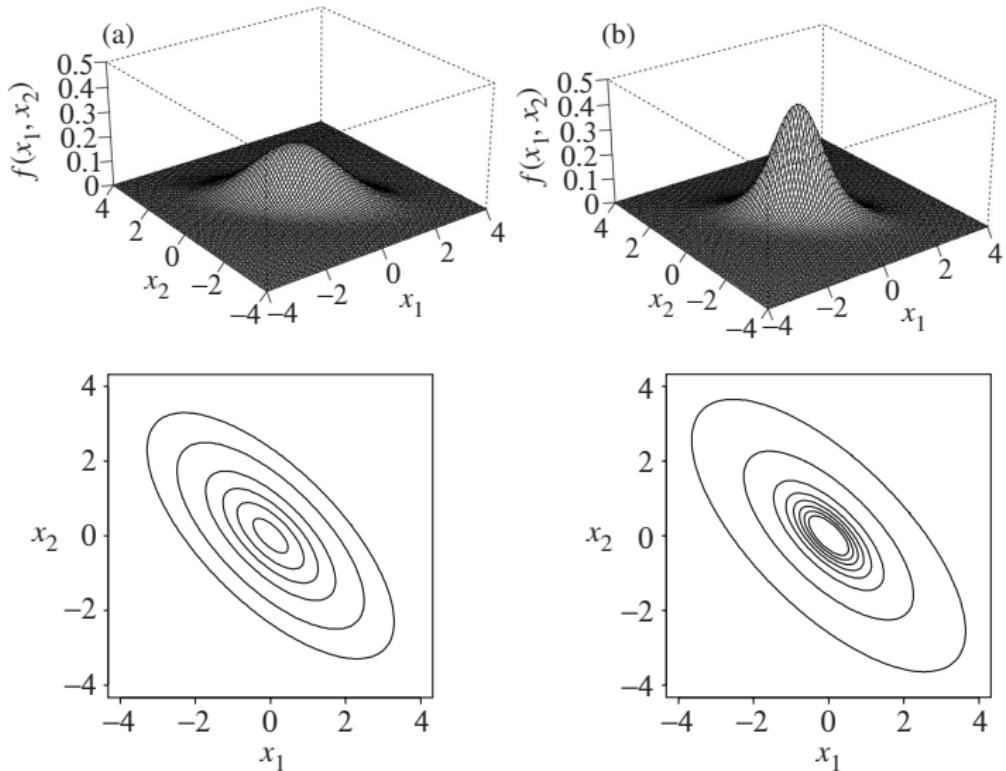
Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ with $\text{rank } A = d = k$ ($\Rightarrow \Sigma$ pos. definite, invertible).

Via the Density Transformation Theorem, it is an exercise to show that \mathbf{X} has density

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right), \quad \mathbf{x} \in \mathbb{R}^d.$$

Consequences:

- Sets of the form $S_c = \{\mathbf{x} \in \mathbb{R}^d : (\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c\}$, $c > 0$, 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 $(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})$, 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: $N_d(\mu, \Sigma)$ for $\mu = (0, 0)$, $\Sigma = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix}$; Right: $t_\nu(\mu, \frac{\nu-2}{\nu}\Sigma)$, $\nu = 4$,
(same mean and covariance matrix as on the left-hand side)

The definition of $N_d(\boldsymbol{\mu}, \Sigma)$ in terms of a stochastic representation ($\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + A\mathbf{Z}$) directly justifies the following sampling algorithm.

Algorithm 6.6 (Sampling $N_d(\boldsymbol{\mu}, \Sigma)$)

Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ with Σ symmetric and positive definite.

- 1) Compute the Cholesky factor A of Σ ; see, e.g. Press et al. (1992).
- 2) Generate $Z_j \stackrel{\text{ind.}}{\sim} N(0, 1)$, $j \in \{1, \dots, d\}$.
- 3) Return $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Z}$, where $\mathbf{Z} = (Z_1, \dots, Z_d)$.

Further useful properties of multivariate normal distributions

■ Linear combinations

If $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ and $B \in \mathbb{R}^{k \times d}$, $\mathbf{b} \in \mathbb{R}^k$, then

$$\begin{aligned} B\mathbf{X} + \mathbf{b} &= B(\boldsymbol{\mu} + A\mathbf{Z}) + \mathbf{b} = (B\boldsymbol{\mu} + \mathbf{b}) + BA\mathbf{Z} \\ &\sim N_k(B\boldsymbol{\mu} + \mathbf{b}, BA(BA)') = N_k(B\boldsymbol{\mu} + \mathbf{b}, B\Sigma B'). \end{aligned}$$

Special case (see variance-covariance method; or Proposition 6.4):
 $\mathbf{b}'\mathbf{X} \sim N(\mathbf{b}'\boldsymbol{\mu}, \mathbf{b}'\boldsymbol{\Sigma}\mathbf{b})$

- **Marginal dfs**

Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and write $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2)$, where $\mathbf{X}_1 \in \mathbb{R}^k$, $\mathbf{X}_2 \in \mathbb{R}^{d-k}$, and $\boldsymbol{\mu} = (\boldsymbol{\mu}'_1, \boldsymbol{\mu}'_2)$, $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$. 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, in the above.

- **Conditional distributions**

Let \mathbf{X} be as before and $\boldsymbol{\Sigma}$ be positive definite. One can show that

$$\mathbf{X}_2 | \mathbf{X}_1 = \mathbf{x}_1 \sim N_{d-k}(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}),$$

where $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)$ and $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$.

- **Quadratic forms**

Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ be 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,

$$(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Z}' \mathbf{Z} \sim \chi_d^2, \quad (18)$$

which is useful for (goodness-of-fit) testing of $N_d(\boldsymbol{\mu}, \Sigma)$; see later.

■ Convolutions

Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ and $\mathbf{Y} \sim N_d(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$ be independent. Via cfs it is then an exercise to show that

$$\mathbf{X} + \mathbf{Y} \sim N_d(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \Sigma + \tilde{\Sigma}).$$

6.1.4 Testing multivariate normality

- For testing univariate normality, all tests of Section 3.1.2 can be applied.
- Now consider multivariate normality. By Proposition 6.4,

$$\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{\text{ind.}}{\sim} N_d(\boldsymbol{\mu}, \Sigma) \Rightarrow \mathbf{a}' \mathbf{X}_1, \dots, \mathbf{a}' \mathbf{X}_n \stackrel{\text{ind.}}{\sim} N(\mathbf{a}' \boldsymbol{\mu}, \mathbf{a}' \Sigma \mathbf{a}).$$

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, (18) can be used to test joint normality.

- Multivariate Shapiro–Wilk
- Mardia's test

- ▶ According to (18), if $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ with Σ positive definite, then $(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_d^2$.
- ▶ Let $D_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})' S^{-1} (\mathbf{X}_i - \bar{\mathbf{X}})$ denote the *squared Mahalanobis distances* and $D_{ij} = (\mathbf{X}_i - \bar{\mathbf{X}})' S^{-1} (\mathbf{X}_j - \bar{\mathbf{X}})$ 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).

Example 6.7 (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.

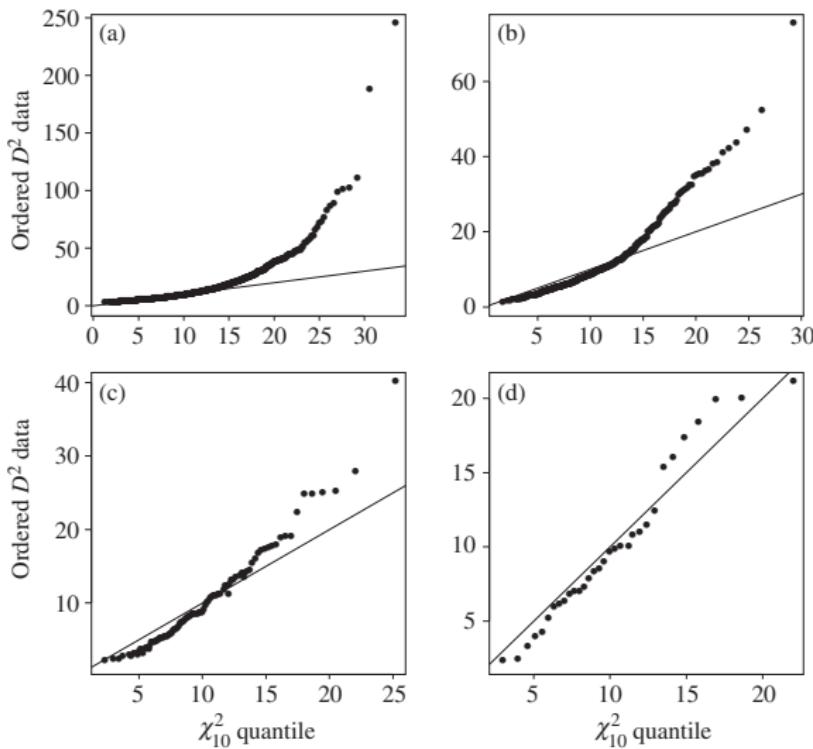
n	Daily 2020	Weekly 416	Monthly 96	Quarterly 32
b_{10}	9.31	9.91	21.10	50.10
$p\text{-value}$	0.00	0.00	0.00	0.02
k_{10}	242.45	177.04	142.65	120.83
$p\text{-value}$	0.00	0.00	0.00	0.44

- We also compare D_i^2 data to a χ_{10}^2 using a Q-Q plot; see the next page.

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:

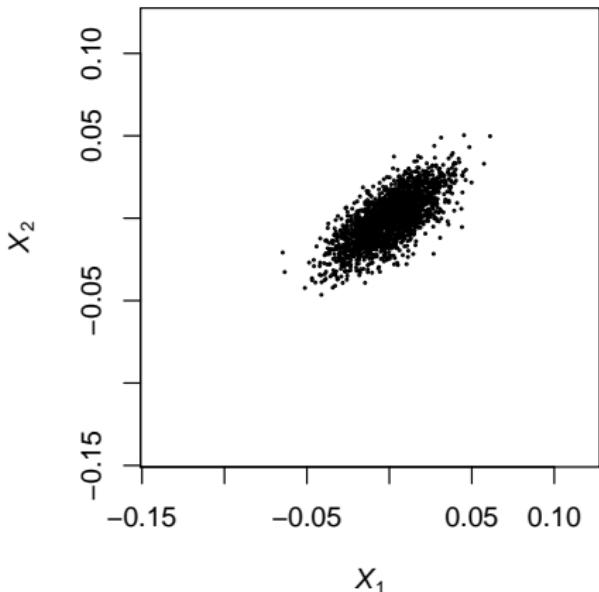
(a) daily data; (b) weekly data; (c) monthly data; and (d) quarterly data



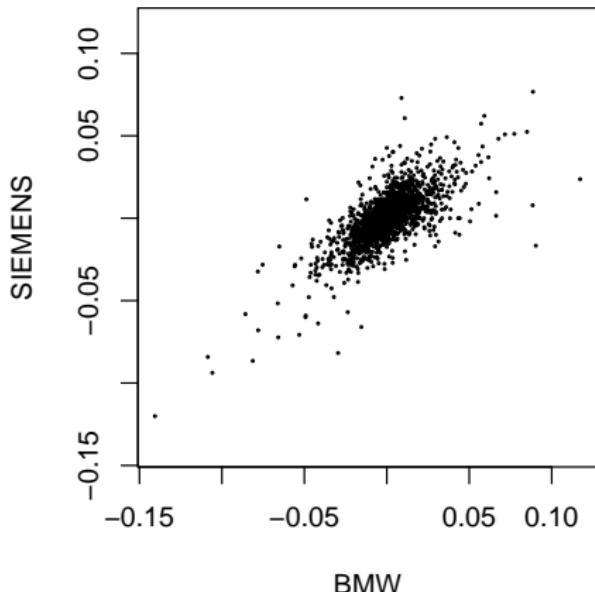
Example 6.8 (Simulated data vs BMW–Siemens)

Is the [BMW–Siemens data](#) (see Section 3.2.2) [jointly normal](#)?

Simulated data (fitted multivariate normal)

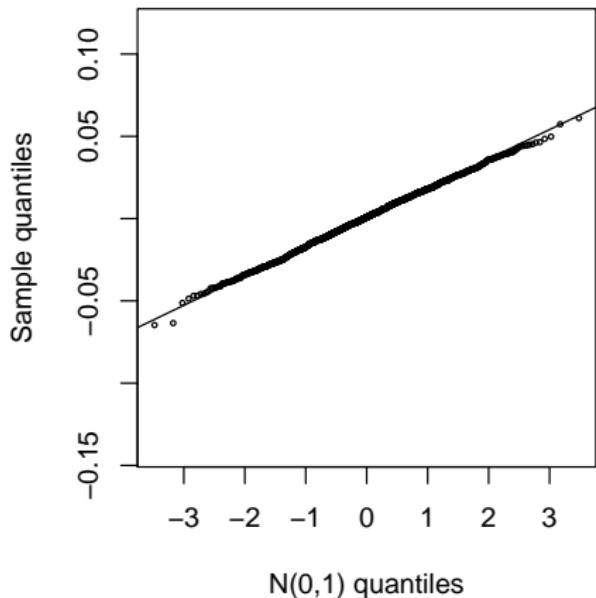


Real risk-factor changes

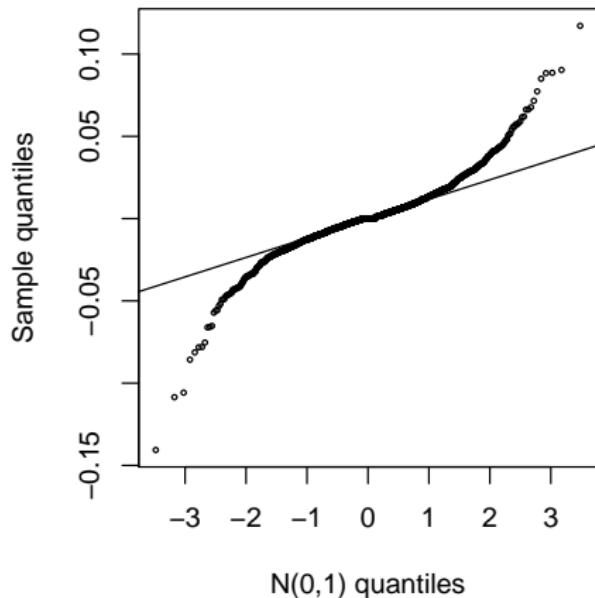


Considering the first margin only:

Q-Q plot for margin 1 (simulated data)

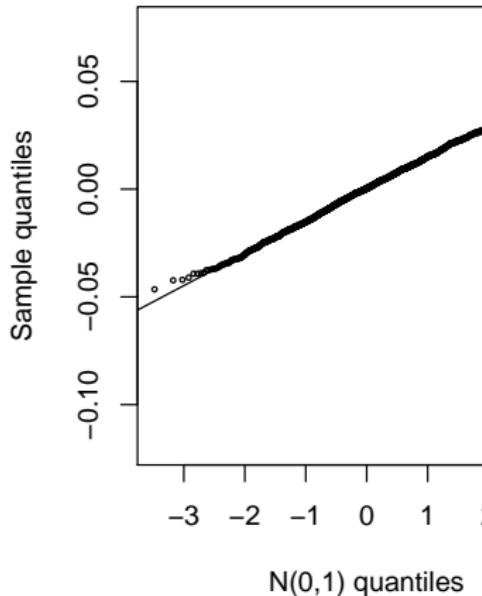


Q-Q plot for margin 1 (real data)

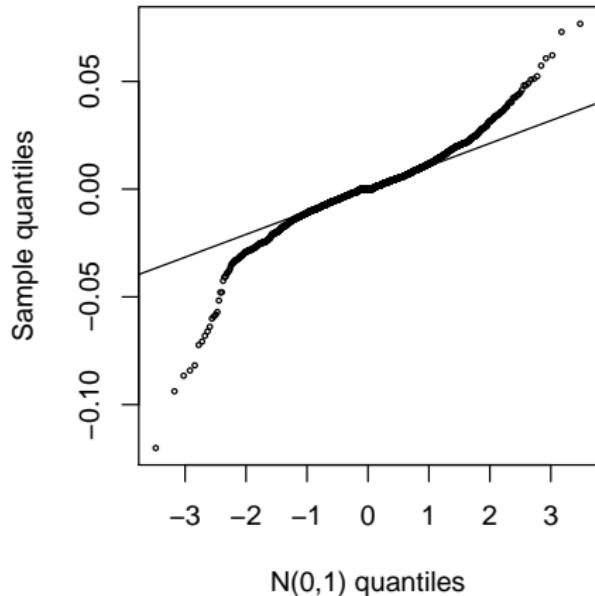


Considering the second margin only:

Q-Q plot for margin 2 (simulated data)

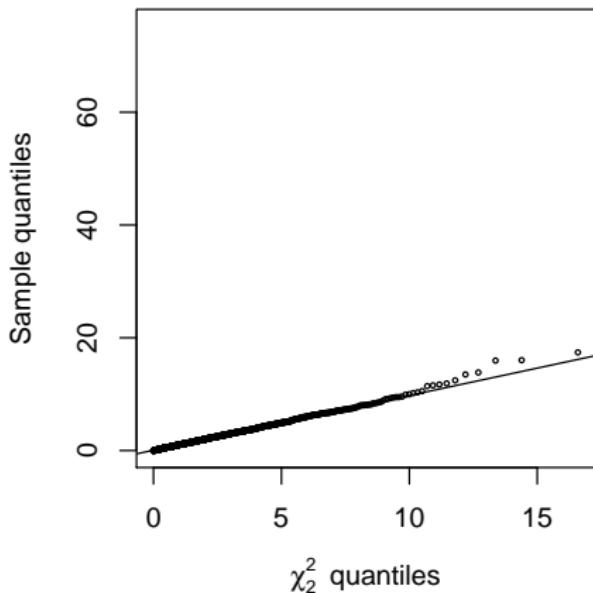


Q-Q plot for margin 2 (real data)

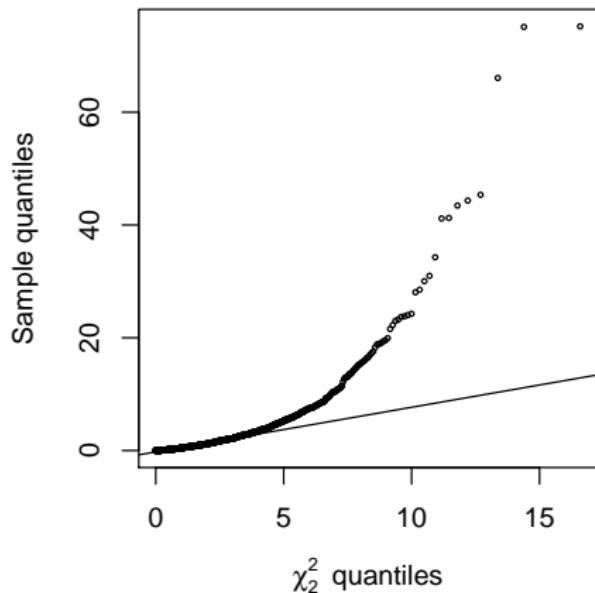


Q-Q plot of the simulated (left) or real (right) D_i^2 's against a χ_2^2 :

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 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 modelling 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 Chapter 7).
- 3) Very strong symmetry known as radial symmetry: X is called radially symmetric about μ if $X - \mu \stackrel{d}{=} \mu - X$. This is true for $N_d(\mu, \Sigma)$.

Short outlook:

- Normal variance mixture distributions can address 1) and 2) while sharing many of the desirable properties of $N_d(\mu, \Sigma)$.
- Normal mean-variance mixture distributions 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 6.9 (Multivariate normal variance mixtures)

The random vector X has a (multivariate) *normal variance mixture distribution* if

$$X \stackrel{d}{=} \mu + \sqrt{W} A Z, \quad (19)$$

where $Z \sim N_k(\mathbf{0}, I_k)$, $W \geq 0$ is a rv independent of Z , $A \in \mathbb{R}^{d \times k}$, and $\mu \in \mathbb{R}^d$. μ is called *location vector* and $\Sigma = AA'$ *scale* (or *dispersion matrix*).

Observe that $(X | W = w) \stackrel{d}{=} \mu + \sqrt{w} A Z = N_d(\mu, wAA') = N_d(\mu, w\Sigma)$; or $(X | W) \stackrel{d}{=} N_d(\mu, W\Sigma)$. W can be interpreted as a shock affecting the variances of all risk factors.

Properties of multivariate normal variance mixtures

Let $\mathbf{X} = \boldsymbol{\mu} + \sqrt{W}\mathbf{A}\mathbf{Z}$ and $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$. Assume that $\text{rank}(A) = d \leq k$ and that Σ is positive definite.

- If $\mathbb{E}\sqrt{W} < \infty$, then $\mathbb{E}(\mathbf{X}) \stackrel{\text{ind.}}{=} \boldsymbol{\mu} + \mathbb{E}(\sqrt{W})\mathbf{A}\mathbb{E}(\mathbf{Z}) = \boldsymbol{\mu} + \mathbf{0} = \boldsymbol{\mu} = \mathbb{E}\mathbf{Y}$
- If $\mathbb{E}W < \infty$, then

$$\begin{aligned}\text{cov}(\mathbf{X}) &= \text{cov}(\sqrt{W}\mathbf{A}\mathbf{Z}) = \mathbb{E}((\sqrt{W}\mathbf{A}\mathbf{Z})(\sqrt{W}\mathbf{A}\mathbf{Z})') \\ &\stackrel{\text{ind.}}{=} \mathbb{E}(W) \cdot \mathbb{E}(\mathbf{A}\mathbf{Z}\mathbf{Z}'\mathbf{A}') = \mathbb{E}(W) \cdot \mathbf{A}\mathbb{E}(\mathbf{Z}\mathbf{Z}')\mathbf{A}' \\ &= \mathbb{E}(W)\mathbf{A}\mathbf{I}_k\mathbf{A}' = \mathbb{E}(W)\Sigma_{\substack{\neq \\ \text{in general}}} \Sigma (= \text{cov}(\mathbf{Y}))\end{aligned}$$

- However, if they exist (i.e. if $\mathbb{E}W < \infty$), it is easy to check that $\text{corr}(\mathbf{X})$ and $\text{corr}(\mathbf{Y})$ are equal.

Lemma 6.10 (Independence in normal variance mixtures)

Let $\mathbf{X} = \boldsymbol{\mu} + \sqrt{W}\mathbf{Z}$ with $\mathbb{E}W < \infty$ (uncorrelated normal variance mixture). Then

$$X_i \text{ and } X_j \text{ are independent} \iff W \text{ is a.s. constant (i.e. } \mathbf{X} \sim \mathcal{N}_d).$$

See the appendix for a proof. Intuitively, W affects all components of \mathbf{X} and thus creates dependence (unless it is constant).

Recall: If $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$, then $\phi_{\mathbf{X}}(\mathbf{t}) = \exp(i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t})$.

Furthermore, $\mathbf{X} | W = w \sim \mathcal{N}_d(\boldsymbol{\mu}, w\Sigma)$

- **Characteristic function:** The cf of a multivariate normal variance mixtures is

$$\begin{aligned}\phi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}(\exp(i\mathbf{t}'\mathbf{X})) = \mathbb{E}(\mathbb{E}(\exp(i\mathbf{t}'\mathbf{X}) | W)) \\ &= \mathbb{E}(\exp(i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}W\mathbf{t}'\Sigma\mathbf{t})) = \exp(i\mathbf{t}'\boldsymbol{\mu})\mathbb{E}(\exp(-W\frac{1}{2}\mathbf{t}'\Sigma\mathbf{t})).\end{aligned}$$

- **LS transform:** The Laplace-Stieltjes transform of F_W is

$$\hat{F}_W(\theta) := \mathbb{E}(\exp(-\theta W)) = \int_0^\infty e^{-\theta w} dF_W(w).$$

Therefore, $\phi_X(\mathbf{t}) = \exp(i\mathbf{t}'\boldsymbol{\mu})\hat{F}_W(\frac{1}{2}\mathbf{t}'\Sigma\mathbf{t})$. 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

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \int_0^\infty f_{\mathbf{X}|W}(\mathbf{x} | w) dF_W(w) \\ &= \int_0^\infty \frac{1}{(2\pi)^{d/2} w^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2w}\right) dF_W(w). \end{aligned}$$

⇒ Only depends on \mathbf{x} through $(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$.

⇒ 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 a.s.

- **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 $\mathbf{Y} \sim M_k(B\boldsymbol{\mu} + \mathbf{b}, B\Sigma B', \hat{F}_W)$; this can be shown via cfs. If $\mathbf{a} \in \mathbb{R}^d$ ($\mathbf{b} = \mathbf{0}$, $B = \mathbf{a}' \in \mathbb{R}^{1 \times d}$), $\mathbf{a}'\mathbf{X} \sim M_1(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\Sigma\mathbf{a}, \hat{F}_W)$.
- **Sampling:**

Algorithm 6.11 (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 $AA' = \Sigma$).
- 4) Return $\mathbf{X} = \boldsymbol{\mu} + \sqrt{W}\mathbf{A}\mathbf{Z}$.

Example 6.12 ($t_d(\nu, \boldsymbol{\mu}, \Sigma)$ distribution)

For Step 2), generate $V \sim \chi_\nu^2$ and set $W = \frac{\nu}{V} \sim Ig(\nu/2, \nu/2)$; or $W = \frac{1}{V}$ with $V \sim \Gamma(\frac{\nu}{2}, \frac{\nu}{2})$ ($\Gamma(\alpha, \beta)$ density: $f(x) = \beta^\alpha x^{\alpha-1} e^{-\beta x} / \Gamma(\alpha)$).

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 $V \sim \Gamma(\nu/2, \nu/2)$.

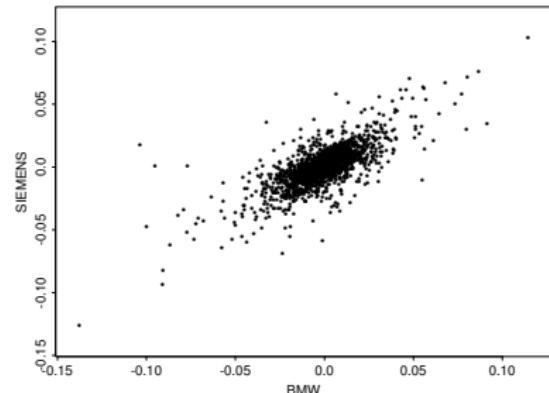
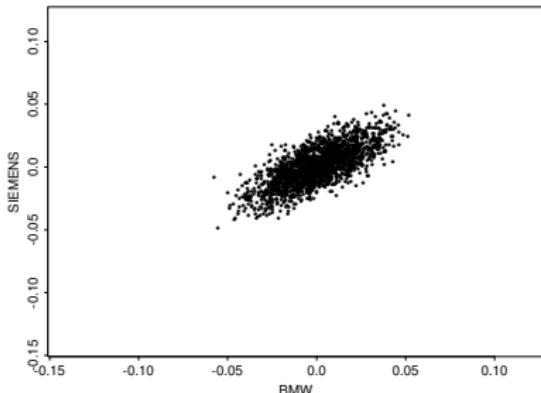
► $\mathbb{E}(W) = \frac{\nu}{\nu-2} \Rightarrow \text{cov}(\mathbf{X}) = \frac{\nu}{\nu-2} \boldsymbol{\Sigma}$. For finite variances/correlations, $\nu > 2$ is required. For finite mean, $\nu > 1$ is required.

- The density of the multivariate t distribution is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\Gamma((\nu + d)/2)}{\Gamma(\nu/2)(\nu\pi)^{d/2}|\Sigma|^{1/2}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+d}{2}},$$

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

$$\mathbf{X} \stackrel{\text{d}}{=} \mathbf{m}(W) + \sqrt{W} A \mathbf{Z}, \quad (20)$$

where

- $\mathbf{Z} \sim N_k(\mathbf{0}, I_k)$;
- $W \geq 0$ 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 **skewness**: Let $\Sigma = AA'$ and observe that $\mathbf{X} | W = w \sim N_d(\mathbf{m}(w), w\Sigma)$. In general, **they are no longer elliptical** (see later).

Example 6.13

- Suppose we have $\mathbf{m}(W) = \boldsymbol{\mu} + W\boldsymbol{\gamma}$. Since

$$\begin{aligned}\mathbb{E}(\mathbf{X} | W) &= \boldsymbol{\mu} + W\boldsymbol{\gamma}, \\ \text{cov}(\mathbf{X} | W) &= W\Sigma\end{aligned}$$

we have

$$\begin{aligned}\mathbb{E}\mathbf{X} &= \mathbb{E}(\mathbb{E}(\mathbf{X} | W)) = \boldsymbol{\mu} + \mathbb{E}(W)\boldsymbol{\gamma} \quad \text{if } \mathbb{E}W < \infty, \\ \text{cov}(\mathbf{X}) &= \mathbb{E}(\text{cov}(\mathbf{X} | W)) + \text{cov}(\mathbb{E}(\mathbf{X} | W)) \\ &= \mathbb{E}(W)\Sigma + \text{var}(W)\boldsymbol{\gamma}\boldsymbol{\gamma}' \quad \text{if } \mathbb{E}(W^2) < \infty.\end{aligned}$$

- 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 skewed normal mean-variance mixture distributions offer only a modest improvement.

We soon study elliptical distributions, a generalization of $M_d(\mu, \Sigma, \hat{F}_W)$.

6.3.1 Spherical distributions

Definition 6.14 (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 $UU' = U'U = I_d$)

$$\mathbf{Y} \stackrel{d}{=} U\mathbf{Y} \quad (\text{distributionally invariant under rotations and reflections})$$

Theorem 6.15 (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 (notation: $\mathbf{Y} \sim S_d(\psi)$ for ψ as below).
- 2) \exists a characteristic generator $\psi : [0, \infty) \rightarrow \mathbb{R}$, such that $\phi_{\mathbf{Y}}(\mathbf{t}) = \mathbb{E}(e^{it'\mathbf{Y}}) = \psi(\|\mathbf{t}\|^2)$, $\forall \mathbf{t} \in \mathbb{R}^d$.
- 3) For every $\mathbf{a} \in \mathbb{R}^d$, $\mathbf{a}'\mathbf{Y} \stackrel{d}{=} \|\mathbf{a}\|Y_1$ (lin. comb. are of the same type).
⇒ Subadditivity of VaR_{α} for jointly elliptical losses

Theorem 6.16 (Stochastic representation)

$\mathbf{Y} \sim S_d(\psi)$ if and only if $\mathbf{Y} \stackrel{d}{=} R\mathbf{S}$ for an independent radial part $R \geq 0$ and $\mathbf{S} \sim U(\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\})$.

- See the appendix for proofs for Theorems 6.15 and 6.16.
- If \mathbf{Y} has a density $f_{\mathbf{Y}}$, it satisfies $f_{\mathbf{Y}}(\mathbf{y}) = g(\|\mathbf{y}\|^2)$ for a function $g : [0, \infty) \rightarrow [0, \infty)$ referred to as density generator (i.e. $f_{\mathbf{Y}}$ is constant on spheres); see the appendix for a proof.

Corollary 6.17

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

$$(\|\mathbf{Y}\|, \frac{\mathbf{Y}}{\|\mathbf{Y}\|}) \stackrel{d}{=} (\|R\mathbf{S}\|, \frac{R\mathbf{S}}{\|R\mathbf{S}\|}) = (|R|\|\mathbf{S}\|, \frac{R\mathbf{S}}{|R|\|\mathbf{S}\|}) = (R, \mathbf{S}).$$

In particular, $\|\mathbf{Y}\|$ and $\mathbf{Y}/\|\mathbf{Y}\|$ are independent (\Rightarrow goodness-of-fit).

Example 6.18 (Standardized normal variance mixtures)

- $\mathbf{Y} \sim M_d(\mathbf{0}, I_d, \hat{F}_W)$ is spherical (recall: $\mathbf{Y} \stackrel{d}{=} \mathbf{0} + \sqrt{W}I_d\mathbf{Z}$) since

$$\begin{aligned}\phi_{\mathbf{Y}}(\mathbf{t}) &= \mathbb{E}(\exp(i\mathbf{t}'\sqrt{W}\mathbf{Z})) = \mathbb{E}_W(\mathbb{E}(\exp(i(\mathbf{t}\sqrt{W})'\mathbf{Z}) | W)) \\ &= \mathbb{E}(\exp(-\frac{1}{2}W\mathbf{t}'\mathbf{t})) = \hat{F}_W(\frac{1}{2}\mathbf{t}'\mathbf{t}) = \hat{F}_W(\frac{1}{2}\|\mathbf{t}\|^2),\end{aligned}$$

so $\mathbf{Y} \sim S_d(\psi)$ by Theorem 6.15 Part 2). We thus have $\psi(t) = \hat{F}_W(t/2)$.

- For $\mathbf{Y} \sim N_d(\mathbf{0}, I_d)$, $\psi(t) = \exp(-t/2)$. By Corollary 6.17, 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.

Example 6.19 (R , S , cov, corr)

- It follows from $\mathbf{Y} \sim N_d(\mathbf{0}, I_d)$ and $R^2 = \mathbf{Y}'\mathbf{Y} \sim \chi_d^2$ that

$$\mathbf{0} = \mathbb{E}\mathbf{Y} = \mathbb{E}R\mathbb{E}\mathbf{S} \stackrel{\text{Th. 6.16}}{\Rightarrow} \mathbb{E}\mathbf{S} = \mathbf{0},$$

$$I_d = \text{cov } \mathbf{Y} = \mathbb{E}(R^2) \text{cov } \mathbf{S} = d \text{cov } \mathbf{S} \stackrel{\text{Th. 6.16}}{\Rightarrow} \text{cov } \mathbf{S} = I_d/d. \quad (21)$$

- For $\mathbf{Y} \sim S_d(\psi)$ with $\mathbb{E}(R^2) < \infty$, it follows that

$$\text{cov } \mathbf{Y} = \mathbb{E}(R^2) \text{cov } \mathbf{S} = \frac{\mathbb{E}(R^2)}{d} I_d$$

and thus $\text{corr } \mathbf{Y} = I_d$.

- For $\mathbf{X} = \mu + A\mathbf{Y}$ 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 $\text{corr } \mathbf{X} = P$ (the correlation matrix corresponding to Σ).

Example 6.20 (t distribution)

For $\mathbf{Y} \sim t_d(\nu, \mathbf{0}, \mathbf{I}_d)$, $R^2 = \mathbf{Y}'\mathbf{Y} = W\mathbf{Z}'\mathbf{Z}$ for $\mathbf{Z} \sim \text{N}_d(\mathbf{0}, \mathbf{I}_d)$. Therefore,

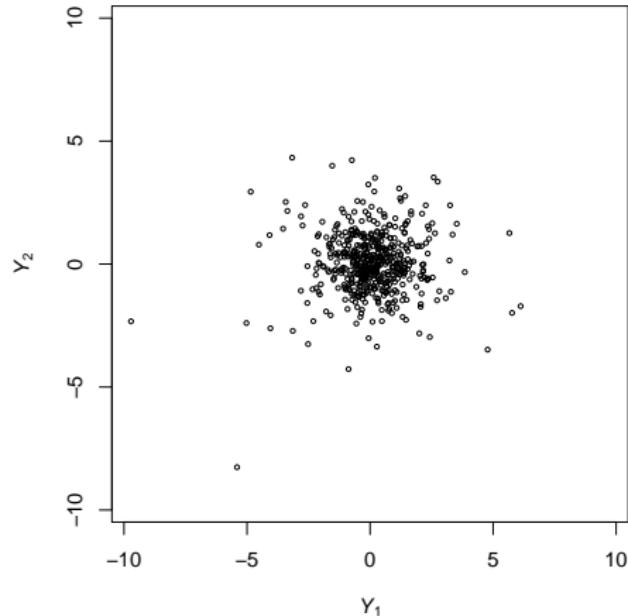
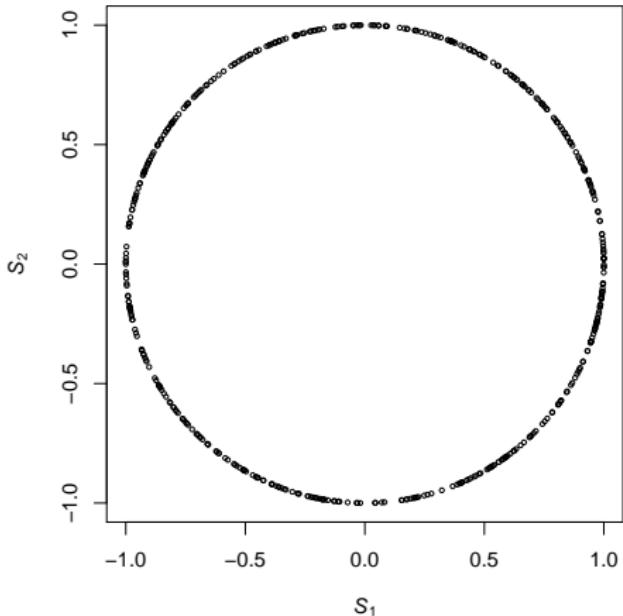
$$\frac{R^2}{d} = \frac{\mathbf{Z}'\mathbf{Z}/d}{(\nu/W)/\nu} = \frac{\chi_d^2/d}{\chi_\nu^2/\nu} \sim F(d, \nu)$$

and thus $\mathbb{E}(R^2/d) = \frac{\nu}{\nu-2}$.

- This, together with Example 6.19, implies that $\mathbf{X} \sim t_d(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ has $\text{cov } \mathbf{X} = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}$ and $\text{corr } \mathbf{X} = \mathbf{P}$ (which we already know from Section 6.2.1); note that in the univariate case $X \sim t(\nu, \mu, \sigma^2)$ and $\text{var}(X) = \frac{\nu}{\nu-2}\sigma^2$.
- We also see that we can use a Q-Q plot of the order statistics of $R^2/d = \|\mathbf{Y}\|^2/d$ versus the theoretical quantiles of a (hypothesized) $F(d, \nu)$ distribution to check the goodness-of-fit of the hypothesized t distribution (in any dimensions).
- See the appendix for the form of the density generator g .

Example 6.21 (Understanding spherical distributions)

$n = 500$ realizations of \mathbf{S} (left) and $\mathbf{Y} = \mathbf{R}\mathbf{S}$ (right) for $R \sim \sqrt{dF(d, \nu)}$, $d = 2$, $\nu = 4$ (as for the multivariate t distribution with $\nu = 4$).



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

$$\mathbf{X} \stackrel{\text{d}}{=} \boldsymbol{\mu} + A\mathbf{Y}, \quad (\text{multivariate affine transformation})$$

where $\mathbf{Y} \sim S_k(\psi)$, $A \in \mathbb{R}^{d \times k}$ (*scale matrix* $\Sigma = AA'$), and (*location vector*) $\boldsymbol{\mu} \in \mathbb{R}^d$.

- By Theorem 6.16, an elliptical random vector **admits the stochastic representation** $\mathbf{X} \stackrel{\text{d}}{=} \boldsymbol{\mu} + RAS$, with R and S as before.
- The **cf** of an elliptical random vector \mathbf{X} is $\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}(e^{i\mathbf{t}'\mathbf{X}}) = \mathbb{E}(e^{i\mathbf{t}'(\boldsymbol{\mu}+A\mathbf{Y})}) = e^{i\mathbf{t}'\boldsymbol{\mu}} \mathbb{E}(e^{i(A'\mathbf{t})'\mathbf{Y}}) = e^{i\mathbf{t}'\boldsymbol{\mu}} \psi(\mathbf{t}'\Sigma\mathbf{t})$. 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 $\mathbf{Y} = A^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim S_d(\psi)$.

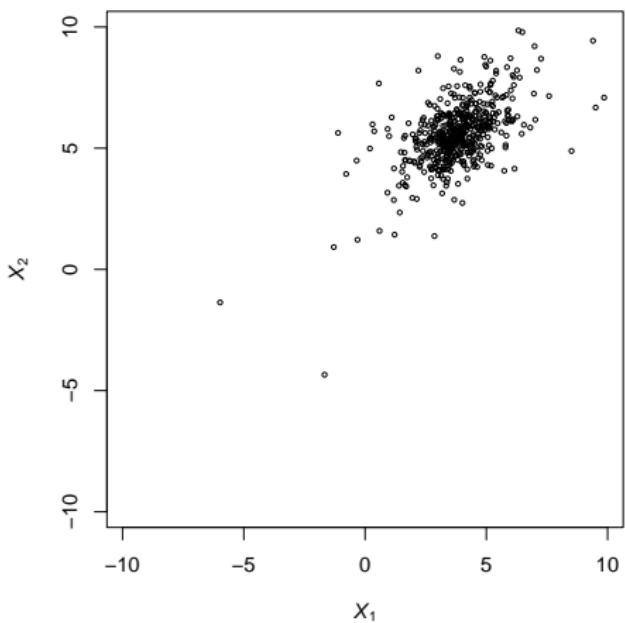
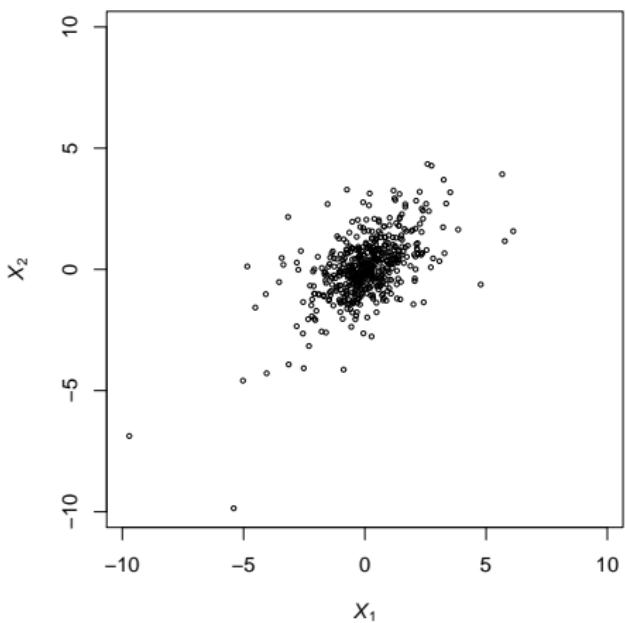
- Normal variance mixture distributions are elliptical (most useful examples) since $\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + \sqrt{W} A \mathbf{Z} = \boldsymbol{\mu} + \sqrt{W} \|\mathbf{Z}\| A \mathbf{Z} / \|\mathbf{Z}\| = \boldsymbol{\mu} + R A S$ with $R = \sqrt{W} \|\mathbf{Z}\|$ and $S = \mathbf{Z} / \|\mathbf{Z}\|$. By Corollary 6.17, 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 6.17 implies that

$$\left(\sqrt{(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})}, \frac{A^{-1}(\mathbf{X} - \boldsymbol{\mu})}{\sqrt{(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})}} \right) \stackrel{d}{=} (R, S), \quad (22)$$

which can be used for testing elliptical symmetry. One can also use the following result for testing.

Example 6.23 (Understanding elliptical distributions)

$n = 500$ realizations of $\mathbf{X} = R\mathbf{A}\mathbf{S}$ (left) and $\mathbf{X} = \boldsymbol{\mu} + R\mathbf{A}\mathbf{S}$ (right) for $R \sim \sqrt{dF(d, \nu)}$, $d = 2$, $\nu = 4$; based on the same samples as in Example 6.21.



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

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{\det \Sigma}} g((\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})),$$

which depends on \mathbf{x} only through $(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$, 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 $\mathbf{b} \in \mathbb{R}^k$,

$$B\mathbf{X} + \mathbf{b} \sim E_k(B\boldsymbol{\mu} + \mathbf{b}, B\Sigma B', \psi) \quad (\text{via cfs}).$$

If $\mathbf{a} \in \mathbb{R}^d$ (take $\mathbf{b} = \mathbf{0}$ and $B = \mathbf{a}' \in \mathbb{R}^{1 \times d}$),

$$\mathbf{a}'\mathbf{X} \sim E_1(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\Sigma\mathbf{a}, \psi) \quad (\text{as for } N(\boldsymbol{\mu}, \Sigma)). \quad (23)$$

From $\mathbf{a} = \mathbf{e}_j = (0, \dots, 0, 1, 0, \dots, 0)$ we see that all marginal distributions are of the same type.

- **Marginal dfs:** As for $N_d(\mu, \Sigma)$, it immediately follows that $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2)' \sim E_d(\mu, \Sigma, \psi)$ satisfies $\mathbf{X}_1 \sim E_k(\mu_1, \Sigma_{11}, \psi)$ and that $\mathbf{X}_2 \sim E_{d-k}(\mu_2, \Sigma_{22}, \psi)$; i.e. margins of elliptical distributions are elliptical.
- **Conditional distributions:** One can also show that conditional distributions of elliptical distributions are elliptical; see Embrechts et al. (2002). For $N_d(\mu, \Sigma)$ the characteristic generator remains the same.
- **Quadratic forms:** (22) implies that $(\mathbf{X} - \mu)' \Sigma^{-1} (\mathbf{X} - \mu) \stackrel{d}{=} R^2$. If $\mathbf{X} \sim N_d(\mu, \Sigma)$, $R^2 \sim \chi_d^2$; and if $\mathbf{X} \sim t_d(\nu, \mu, \Sigma)$, $R^2/d \sim F(d, \nu)$.
- **Convolutions:** Let $\mathbf{X} \sim E_d(\mu, \Sigma, \psi)$ and $\mathbf{Y} \sim E_d(\tilde{\mu}, c\Sigma, \tilde{\psi})$ be independent. Then $a\mathbf{X} + b\mathbf{Y}$ is elliptically distributed for $a, b \in \mathbb{R}$, $c > 0$.
- **Conditional correlations remain invariant** See Proposition A.11.

Many (but not all) nice properties of $N_d(\mu, \Sigma)$ are preserved. For estimating μ , Σ , P , see the appendix. The following result shows why elliptical distributions are known as the “Garden of Eden” of QRM.

Proposition 6.24 (Subadditivity of VaR in elliptical models)

Let $L_i = \boldsymbol{\lambda}'_i \mathbf{X}$, $\boldsymbol{\lambda}_i \in \mathbb{R}^d$, $i \in \{1, \dots, n\}$, with $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$. 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 $L = \boldsymbol{\lambda}' \mathbf{X} \stackrel{d}{=} \boldsymbol{\lambda}' \boldsymbol{\mu} + \boldsymbol{\lambda}' A \mathbf{Y}$ for $\mathbf{Y} \sim S_k(\psi)$. By Theorem 6.15 Part 3), $\boldsymbol{\lambda}' A \mathbf{Y} \stackrel{d}{=} \|\boldsymbol{\lambda}' A\| Y_1$, so $L \stackrel{d}{=} \boldsymbol{\lambda}' \boldsymbol{\mu} + \|\boldsymbol{\lambda}' A\| Y_1$ (all L_i 's are of the same type). By translation invariance and positive homogeneity,

$$\text{VaR}_\alpha(L) = \boldsymbol{\lambda}' \boldsymbol{\mu} + \|\boldsymbol{\lambda}' A\| \text{VaR}_\alpha(Y_1). \quad (24)$$

Applying (24) once to $L = \sum_{i=1}^n L_i = (\sum_{i=1}^n \boldsymbol{\lambda}_i)' \mathbf{X}$ and to each $L = L_i = \boldsymbol{\lambda}'_i \mathbf{X}$, $i \in \{1, \dots, n\}$, and using that $\text{VaR}_\alpha(Y_1) \geq 0$ for $\alpha \in [1/2, 1]$, we obtain $\text{VaR}_\alpha(\sum_{i=1}^n L_i) \stackrel{(24)}{=} \sum_{i=1}^n \boldsymbol{\lambda}'_i \boldsymbol{\mu} + \|\sum_{i=1}^n \boldsymbol{\lambda}'_i A\| \text{VaR}_\alpha(Y_1)$
 $\leq \sum_{i=1}^n \boldsymbol{\lambda}'_i \boldsymbol{\mu} + (\sum_{i=1}^n \|\boldsymbol{\lambda}'_i A\|) \text{VaR}_\alpha(Y_1) = \sum_{i=1}^n (\boldsymbol{\lambda}'_i \boldsymbol{\mu} + \|\boldsymbol{\lambda}'_i A\| \text{VaR}_\alpha(Y_1))$
 $\stackrel{(24)}{=} \sum_{i=1}^n \text{VaR}_\alpha(L_i)$. For $\boldsymbol{\lambda}_i = \mathbf{e}_i$, $\text{VaR}_\alpha(\sum_{i=1}^n X_i) \leq \sum_{i=1}^n \text{VaR}_\alpha(X_i)$. \square

6.4 Dimension reduction techniques

6.4.1 Factor models

Explain the variability of \mathbf{X} in terms of common factors.

Definition 6.25 (p -factor model)

\mathbf{X} follows a *p-factor model* if

$$\mathbf{X} = \mathbf{a} + B\mathbf{F} + \boldsymbol{\varepsilon}, \quad (25)$$

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 $\Omega := \text{cov}(\mathbf{F})$, (*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 $\Sigma := \text{cov}(\mathbf{X}) = B\Omega B' + \Upsilon$.
- 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 $\Sigma = B^*(B^*)' + \Upsilon$. Conversely, if $\text{cov}(\mathbf{X}) = BB' + \Upsilon$ 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 $\boldsymbol{\varepsilon}$.

- For a one-factor/equicorrelation example, see the appendix.

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, $t \in \{1, \dots, n\}$. Estimation of B, \mathbf{a} is accomplished by time series regression.
- 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.

6.4.3 Estimating macroeconomic factor models

There are two equivalent approaches.

Univariate regression

- Consider the (univariate) *time series regression* model

$$X_{t,j} = a_j + \mathbf{b}'_j \mathbf{F}_t + \varepsilon_{t,j}, \quad t \in \{1, \dots, n\}.$$

- 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 .

For the multivariate case, see the appendix.

6.4.4 Estimating fundamental factor models

- Consider the cross-sectional regression model $\mathbf{X}_t = B\mathbf{F}_t + \boldsymbol{\varepsilon}_t$ (B known; \mathbf{F}_t to be estimated; $\text{cov}(\boldsymbol{\varepsilon}) = \Upsilon$); note that a can be absorbed into \mathbf{F}_t . To obtain precision in estimating \mathbf{F}_t , we need $d \gg p$.
- First estimate \mathbf{F}_t via OLS by $\hat{\mathbf{F}}_t^{\text{OLS}} = (B'B)^{-1}B'\mathbf{X}_t$. This is the best linear unbiased estimator if the $\boldsymbol{\varepsilon}$ 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 $\hat{\boldsymbol{\varepsilon}}_t = \mathbf{X}_t - B\hat{\mathbf{F}}_t^{\text{OLS}}$, $t \in \{1, \dots, n\}$.
- Then estimate \mathbf{F}_t via $\hat{\mathbf{F}}_t = (B'\Upsilon^{-1}B)^{-1}B'\Upsilon^{-1}\mathbf{X}_t$.

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*

$$A = \Gamma \Lambda \Gamma'$$

where

- 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 $\Sigma = \Gamma \Lambda \Gamma'$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$ (positive semidefiniteness \Rightarrow all eigenvalues ≥ 0) and $\mathbf{Y} = \Gamma'(\mathbf{X} - \boldsymbol{\mu})$ (the so-called *principal component transform*). The j th component $\mathbf{Y}_j = \gamma_j'(\mathbf{X} - \boldsymbol{\mu})$ 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 $\text{cov}(\mathbf{Y}) = \Gamma'\Sigma\Gamma = \Gamma'\Gamma\Lambda\Gamma'\Gamma = \Lambda$, so the principal components are uncorrelated and $\text{var}(Y_j) = \lambda_j$, $j \in \{1, \dots, d\}$. The principal components are thus ordered by decreasing variance.
- 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. $\text{var}(\gamma_1' \mathbf{X}) = \max\{\text{var}(\mathbf{a}' \mathbf{X}) : \mathbf{a}'\mathbf{a} = 1\}$.
 - ▶ 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 $\mathbf{Y} = \boldsymbol{\Gamma}'(\mathbf{X} - \boldsymbol{\mu})$, we have

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}_1\mathbf{Y}_1 + \boldsymbol{\Gamma}_2\mathbf{Y}_2 =: \boldsymbol{\mu} + \boldsymbol{\Gamma}_1\mathbf{Y}_1 + \boldsymbol{\varepsilon}$$

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.
- In principle, the same can be applied to the sample covariance matrix to obtain the sample principal components; see the appendix.

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

7.6 A copulas-based proof of subadditivity of ES

7.1 Copulas

- We now look more closely at modelling the dependence among the components of a random vector $\mathbf{X} \sim F$ (risk-factor changes).
- **In short:** F “=” 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 F , 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,

$$C(1, \dots, 1, u_j, 1, \dots, 1) = u_j \text{ for all } u_j \in [0, 1] \text{ and } j \in \{1, \dots, d\}.$$

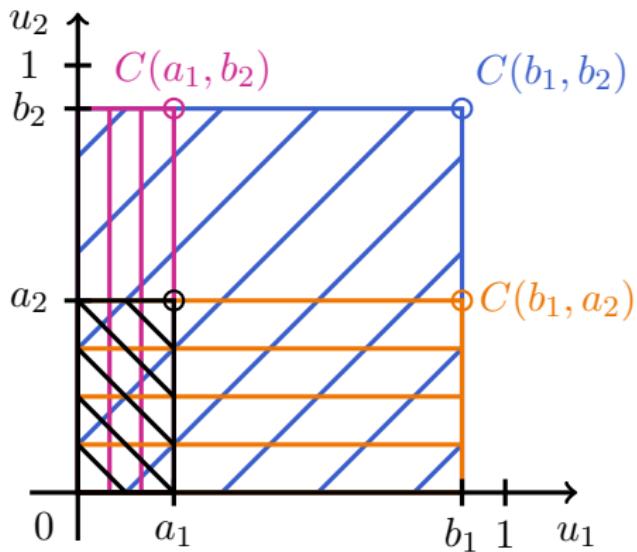
3) C is *d-increasing*, that is, for all $\mathbf{a}, \mathbf{b} \in [0, 1]^d$, $\mathbf{a} \leq \mathbf{b}$,

$$\Delta_{(\mathbf{a}, \mathbf{b})} C = \sum_{\mathbf{i} \in \{0, 1\}^d} (-1)^{\sum_{j=1}^d i_j} C(a_1^{i_1} b_1^{1-i_1}, \dots, a_d^{i_d} b_d^{1-i_d}) \geq 0.$$

Equivalently (if existent): density $c(\mathbf{u}) \geq 0$ for all $\mathbf{u} \in (0, 1)^d$.

2-increasingness explained in a picture:

$$\begin{aligned}\Delta_{(a,b]} C &= C(b_1, b_2) - C(b_1, a_2) - C(a_1, b_2) + C(a_1, a_2) \\ &= \mathbb{P}(\mathbf{U} \in (a, b]) \geq 0\end{aligned}$$



$\Rightarrow \Delta_{(a,b]} C$ is the probability of a random vector $\mathbf{U} \sim C$ to be in $(a, b]$.

Preliminaries

Lemma 7.2 (Probability transformation)

Let $X \sim F$, F continuous. Then $F(X) \sim U(0, 1)$.

Idea of the proof. $\mathbb{P}(F(X) \leq u) = \mathbb{P}(F^{-1}(F(X)) \leq F^{-1}(u)) = \mathbb{P}(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u$, $u \in [0, 1]$; more details in the appendix. \square

Note that F needs to be **continuous** (otherwise $F(X)$ would not reach all intervals $\subseteq [0, 1]$).

Lemma 7.3 (Quantile transformation)

Let $U \sim U(0, 1)$ and F be any df. Then $X = F^{-1}(U) \sim F$.

Proof. $\mathbb{P}(F^{-1}(U) \leq x) \stackrel{(GI5)}{=} \mathbb{P}(U \leq F(x)) = F(x)$, $x \in \mathbb{R}$. \square

Probability and quantile transformations are the key to all applications involving copulas. They allow us to go from \mathbb{R}^d to $[0, 1]^d$ and back.

Sklar's Theorem

Theorem 7.4 (Sklar's Theorem)

- 1) For any df F with margins F_1, \dots, F_d , there exists a copula C such that

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)), \quad \mathbf{x} \in \mathbb{R}^d. \quad (26)$$

C is uniquely defined on $\prod_{j=1}^d \text{ran } F_j$ and given by

$$C(u_1, \dots, u_d) = F(F_1^\leftarrow(u_1), \dots, F_d^\leftarrow(u_d)), \quad \mathbf{u} \in \prod_{j=1}^d \text{ran } F_j,$$

where $\text{ran } F_j = \{F_j(x) : x \in \mathbb{R}\}$ denotes the *range* of F_j .

- 2) Conversely, given any copula C and univariate dfs F_1, \dots, F_d , F defined by (26) is a df with margins F_1, \dots, F_d .

Proof.

- 1) **Proof for continuous F_1, \dots, F_d only.** Let $\mathbf{X} \sim F$ 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 $X_j = F_j^\leftarrow(F_j(X_j)) = F_j^\leftarrow(U_j)$, $j \in \{1, \dots, d\}$. Therefore,

$$\begin{aligned} F(\mathbf{x}) &= \mathbb{P}(X_j \leq x_j \ \forall j) = \mathbb{P}(F_j^\leftarrow(U_j) \leq x_j \ \forall j) \stackrel{(GI5)}{=} \mathbb{P}(U_j \leq F_j(x_j) \ \forall j) \\ &= C(F_1(x_1), \dots, F_d(x_d)), \quad \mathbf{x} \in \mathbb{R}^d. \end{aligned}$$

Hence C is a copula and satisfies (26).

(GI4) implies that $F_j(F_j^\leftarrow(u_j)) = u_j$ for all $u_j \in \text{ran } F_j$, so

$$\begin{aligned} C(u_1, \dots, u_d) &= C(F_1(F_1^\leftarrow(u_1)), \dots, F_d(F_d^\leftarrow(u_d))) \\ &\stackrel{(26)}{=} F(F_1^\leftarrow(u_1), \dots, F_d^\leftarrow(u_d)), \quad \mathbf{u} \in \prod_{j=1}^d \text{ran } F_j. \end{aligned}$$

2) For $\mathbf{U} \sim C$, define $\mathbf{X} = (F_1^\leftarrow(U_1), \dots, F_d^\leftarrow(U_d))$. Then

$$\begin{aligned}\mathbb{P}(\mathbf{X} \leq \mathbf{x}) &= \mathbb{P}(F_j^\leftarrow(U_j) \leq x_j \ \forall j) \stackrel{\text{(GI5)}}{=} \mathbb{P}(U_j \leq F_j(x_j) \ \forall j) \\ &= C(F_1(x_1), \dots, F_d(x_d)), \quad \mathbf{x} \in \mathbb{R}^d.\end{aligned}$$

Therefore, F defined by (26) is a df (that of \mathbf{X}), with margins F_1, \dots, F_d (obtained by the quantile transformation). \square

Example 7.5 (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 (26) (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 $\delta(u) = u^2$).

- A copula model for \mathbf{X} means $F(\mathbf{x}) = C(F_1(x_1), \dots, F_d(x_d))$ for some (parametric) copula C and (parametric) marginals F_1, \dots, F_d .
- \mathbf{X} (or F) with margins F_1, \dots, F_d has copula C if (26) holds.

Invariance principle

Lemma 7.6 (Core of the invariance principle)

Let $X_j \sim F_j$, F_j continuous, $j \in \{1, \dots, d\}$. Then

$$\mathbf{X} \text{ has copula } C \iff (F_1(X_1), \dots, F_d(X_d)) \sim C.$$

Proof. See the appendix. □

Theorem 7.7 (Invariance principle)

Let $\mathbf{X} \sim F$ 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

$$\begin{aligned} F_{T_j(X_j)}(x) &= \mathbb{P}(T_j(X_j) \leq x) = \mathbb{P}(T_j(X_j) < x) \stackrel{\text{(GI5)}}{=} \mathbb{P}(X_j < T_j^\leftarrow(x)) \\ &= \mathbb{P}(X_j \leq T_j^\leftarrow(x)) = F_j(T_j^\leftarrow(x)), \quad x \in \mathbb{R}. \end{aligned}$$

This implies that $\mathbb{P}(F_{T_j(X_j)}(T_j(X_j)) \leq u_j \forall j)$ equals

$$\mathbb{P}(F_j(T_j^\leftarrow(T_j(X_j))) \leq u_j \forall j) \stackrel{\text{(GI3)}}{=} \mathbb{P}(F_j(X_j) \leq u_j \forall j) \stackrel{\substack{\text{L.7.6} \\ \text{"only if"}}}{=} C(\mathbf{u}).$$

The claim follows from the if part (" \Leftarrow ") of Lemma 7.6. \square

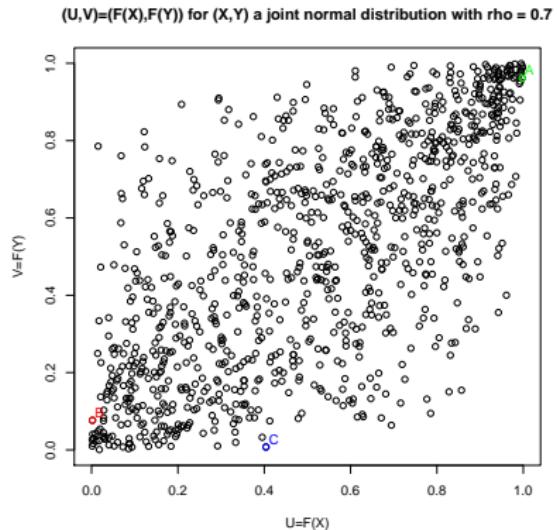
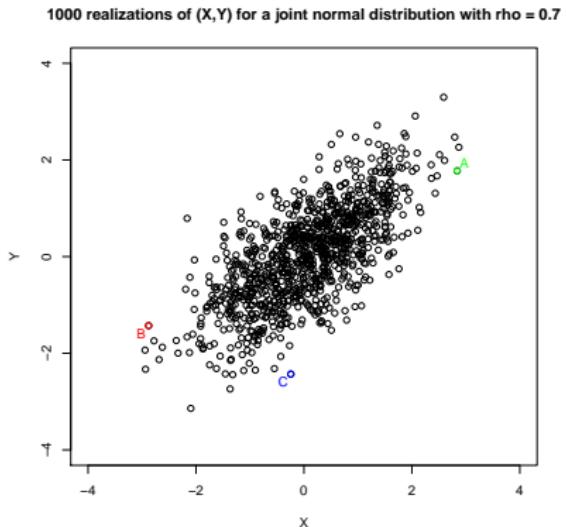
Interpretation of Sklar's Theorem (and the invariance principle)

- 1) Part 1) of Sklar's Theorem allows one to decompose any df F into its margins and a copula. This, together with the invariance principle, allows one to study dependence independently of the margins via the margin-free $\mathbf{U} = (F_1(X_1), \dots, F_d(X_d))$ instead of $\mathbf{X} = (X_1, \dots, X_d)$ (they both have the same copula!). This is interesting for statistical applications, e.g. parameter estimation or goodness-of-fit.
- 2) Part 2) allows one to construct flexible multivariate distributions for particular applications.

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.

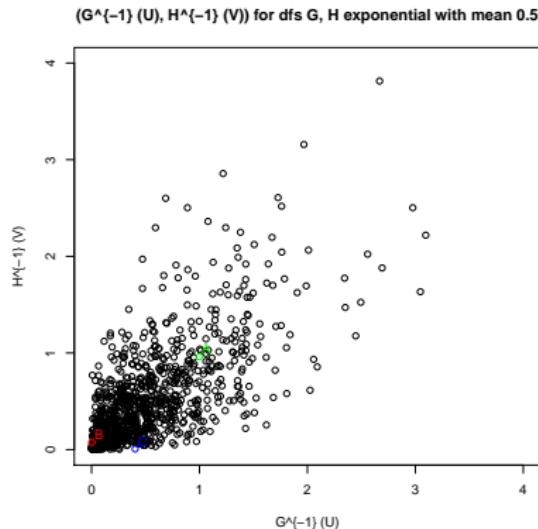
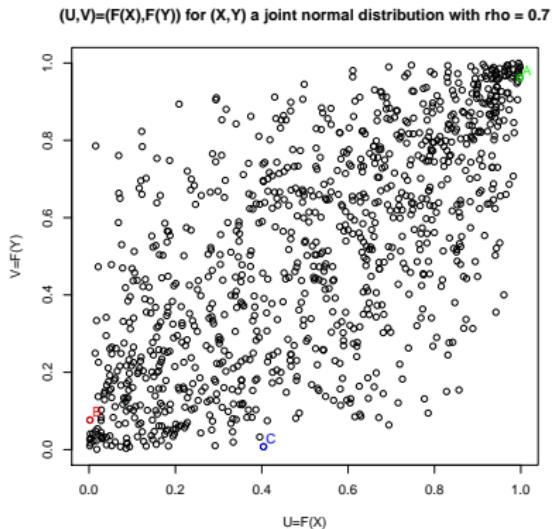
Right: Scatter plot of the corresponding Gauss copula (after applying the df Φ of $N(0, 1)$). Note how A, B, C change.



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.

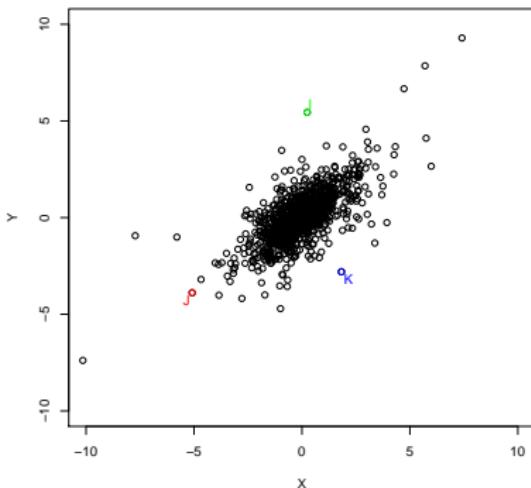


Visualizing the first part of Sklar's Theorem

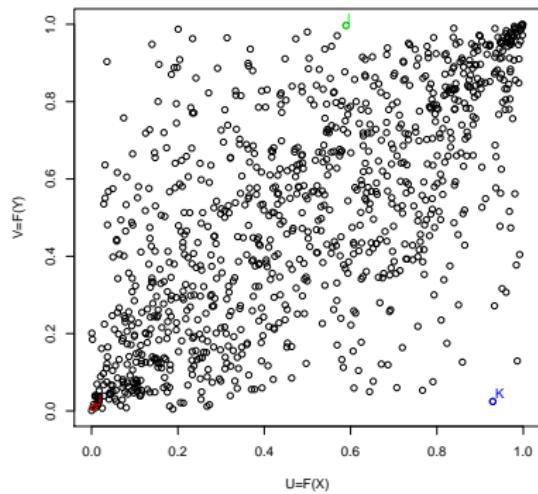
Left: Scatter plot of $n = 1000$ samples from $(X_1, X_2) \sim t_2(4, \mathbf{0}, P)$, where $P = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$. We mark three points **I**, **J**, **K**.

Right: Scatter plot of the corresponding t_4 copula (after applying the df t_4). Note how **A**, **B**, **C** change.

1000 realizations of (X, Y) for joint t-distribution with nu=4 and rho=0.7



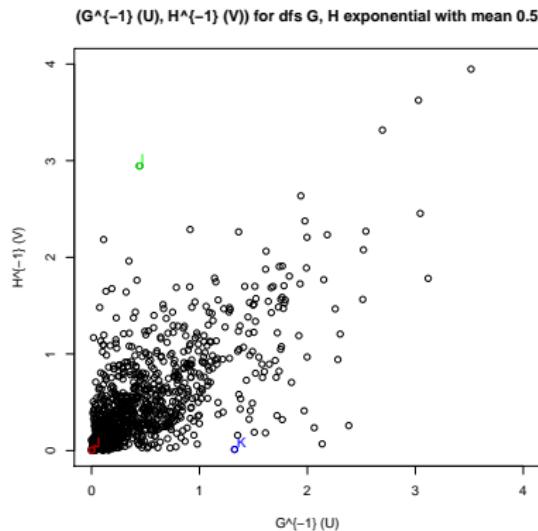
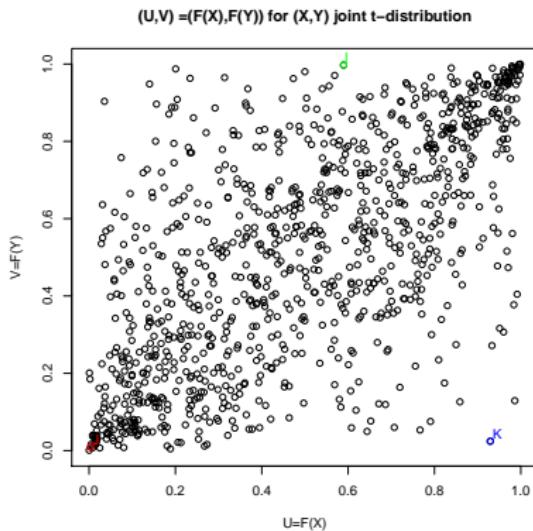
$(U, V) = (F(X), F(Y))$ for (X, Y) joint t-distribution



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–Höffding bounds

Theorem 7.8 (Fréchet–Höffding 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 $d = 2$.

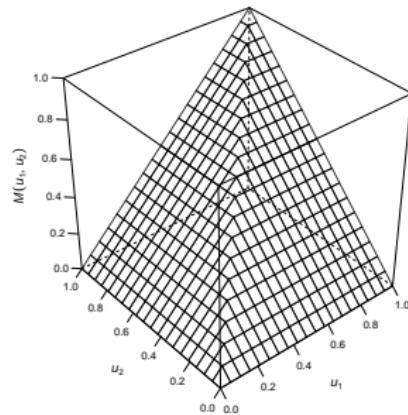
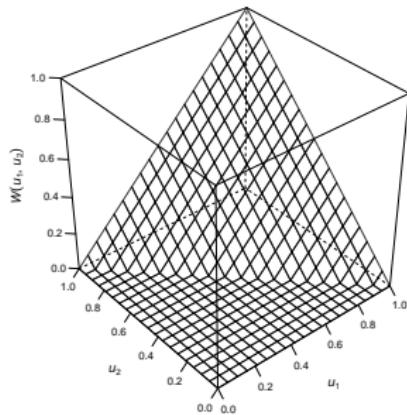
3) M is a copula for all $d \geq 2$.

Proof. See the appendix. □

■ It is easy to verify that, for $U \sim U(0, 1)$,

- ▶ $(U, \dots, U) \sim M$;
- ▶ $(U, 1 - U) \sim W$.

- Plot of W, M for $d = 2$ (compare with $(U, 1 - U) \sim W, (U, U) \sim M$)



- The Fréchet–Höffding bounds correspond to perfect dependence (negative for W ; positive for M); see Proposition 7.14 later.
- The Fréchet–Höffding bounds lead to bounds for any df F , via

$$\max\left\{\sum_{j=1}^d F_j(x_j) - d + 1, 0\right\} \leq F(\mathbf{x}) \leq \min_{1 \leq j \leq d} \{F_j(x_j)\}.$$

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 F 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 $C(F_1(x_1), \dots, F_d(x_d)) = F(\mathbf{x}) = \prod_{j=1}^d F_j(x_j)$ 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–Höffding 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–Höffding 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 $N(0, 1)$.

- Special cases: If $P = I_d$ then $C = \Pi$, and if $P = J_d = \mathbf{1}\mathbf{1}'$ then $C = M$.
If $d = 2$ and $\rho = P_{12} = -1$ then $C = W$.
- Sklar's Theorem \Rightarrow The density of $C(\mathbf{u}) = F(F_1^\leftarrow(u_1), \dots, F_d^\leftarrow(u_d))$ is

$$c(\mathbf{u}) = \frac{f(F_1^\leftarrow(u_1), \dots, F_d^\leftarrow(u_d))}{\prod_{j=1}^d f_j(F_j^\leftarrow(u_j))}, \quad \mathbf{u} \in (0, 1)^d.$$

In particular, the density of C_P^{Ga} is

$$c_P^{\text{Ga}}(\mathbf{u}) = \frac{1}{\sqrt{\det P}} \exp\left(-\frac{1}{2} \mathbf{x}'(P^{-1} - I_d)\mathbf{x}\right), \quad (27)$$

where $\mathbf{x} = (\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))$.

t copulas

- Consider (w.l.o.g.) $\mathbf{X} \sim t_d(\nu, \mathbf{0}, P)$. The *t copula* (family) is given by

$$\begin{aligned} C_{\nu, P}^t(\mathbf{u}) &= \mathbb{P}(t_\nu(X_1) \leq u_1, \dots, t_\nu(X_d) \leq u_d) \\ &= t_{\nu, P}(t_\nu^{-1}(u_1), \dots, t_\nu^{-1}(u_d)) \end{aligned}$$

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.

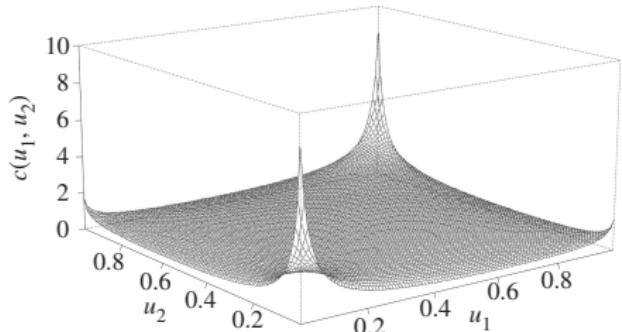
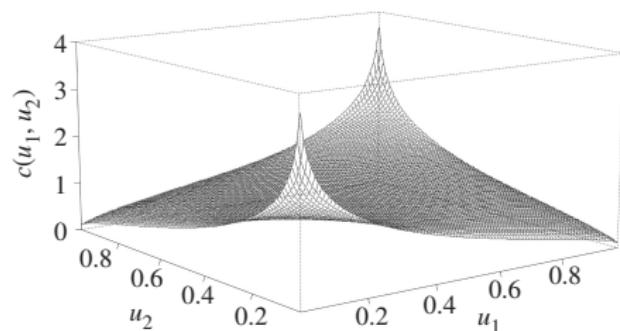
- Special cases: $P = J_d = \mathbf{1}\mathbf{1}'$ then $C = M$. However, if $P = I_d$ then $C \neq \Pi$ (unless $\nu = \infty$ in which case $C_{\nu,P}^t = C_P^{Ga}$). If $d = 2$ and $\rho = P_{12} = -1$ then $C = W$.
- Sklar's Theorem \Rightarrow The density of $C_{\nu,P}^t$ is

$$c_{\nu,P}^t(\mathbf{u}) = \frac{\Gamma((\nu + d)/2)}{\Gamma(\nu/2)\sqrt{\det P}} \left(\frac{\Gamma(\nu/2)}{\Gamma((\nu + 1)/2)} \right)^d \frac{(1 + \mathbf{x}'P^{-1}\mathbf{x}/\nu)^{-(\nu+d)/2}}{\prod_{j=1}^d (1 + x_j^2/\nu)^{-(\nu+1)/2}},$$

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:

- Modelling pairwise dependencies (comparably flexible)
- Density available
- Sampling (typically) simple

Drawbacks:

- Typically, C is not explicit
- Radially symmetric (so the same lower/upper tail behaviour)

Explicit copulas

Archimedean copulas are copulas of the form

$$C(\mathbf{u}) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)), \quad \mathbf{u} \in [0, 1]^d,$$

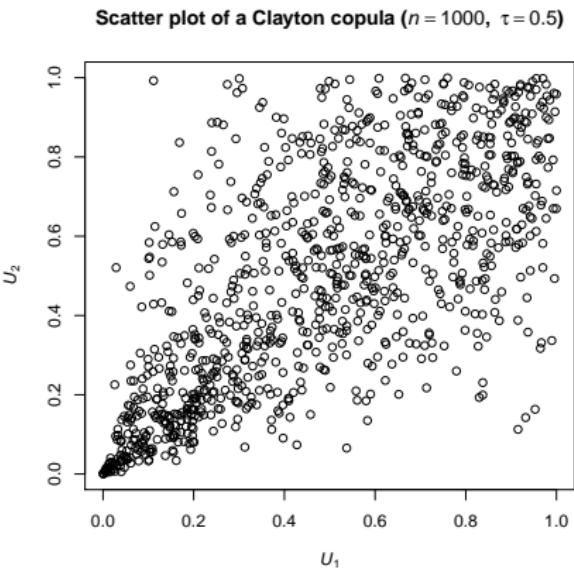
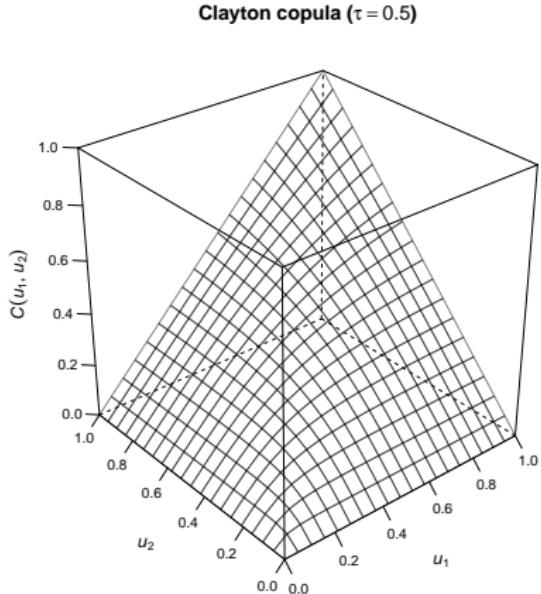
where the (*Archimedean*) generator $\psi : [0, \infty) \rightarrow [0, 1]$ is \downarrow on $[0, \inf\{t : \psi(t) = 0\}]$ and satisfies $\psi(0) = 1$, $\psi(\infty) = \lim_{t \rightarrow \infty} \psi(t) = 0$; we set $\psi^{-1}(0) = \inf\{t : \psi(t) = 0\}$. The set of all generators is denoted by Ψ . 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 $\theta \rightarrow \infty$, $C \rightarrow M$.

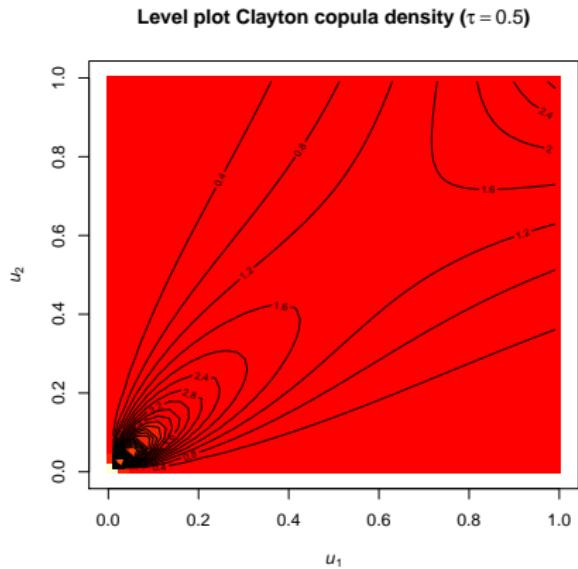
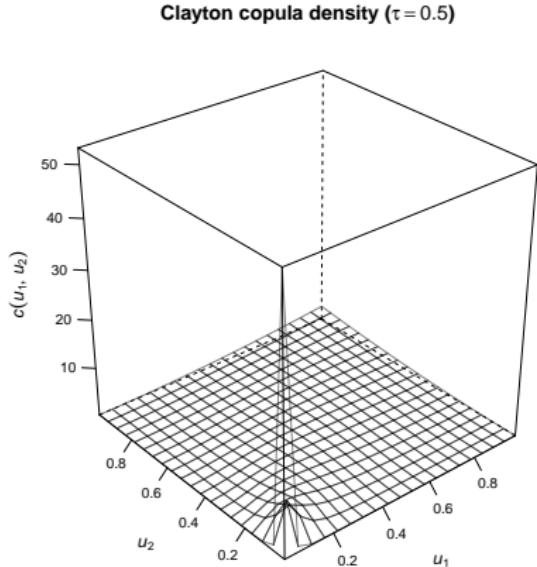
Left: Plot of a bivariate Clayton copula (Kendall's tau 0.5; see later).

Right: Corresponding scatter plot (sample size $n = 1000$)



Left: Plot of the corresponding density.

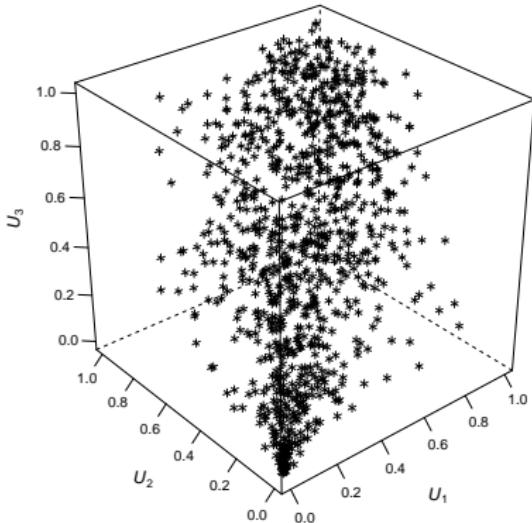
Right: Level plot of the density (with heat colors).



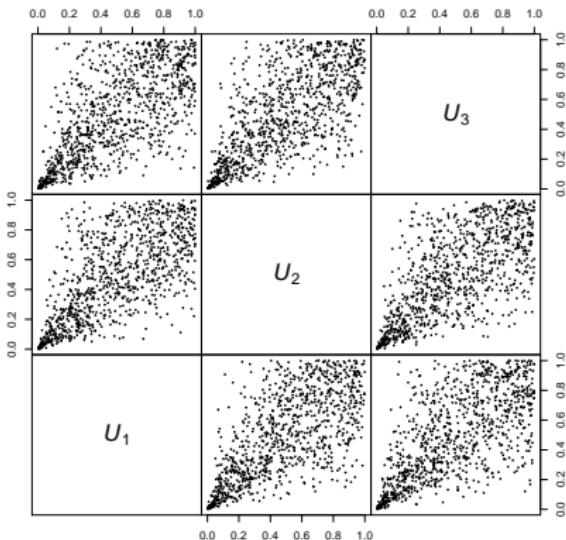
Left: Cloud plot of a trivariate Clayton copula (sample size $n = 1000$; Kendall's tau 0.5).

Right: Corresponding scatter plot matrix.

Clayton copula cloud plot ($n = 1000$, $\tau = 0.5$)

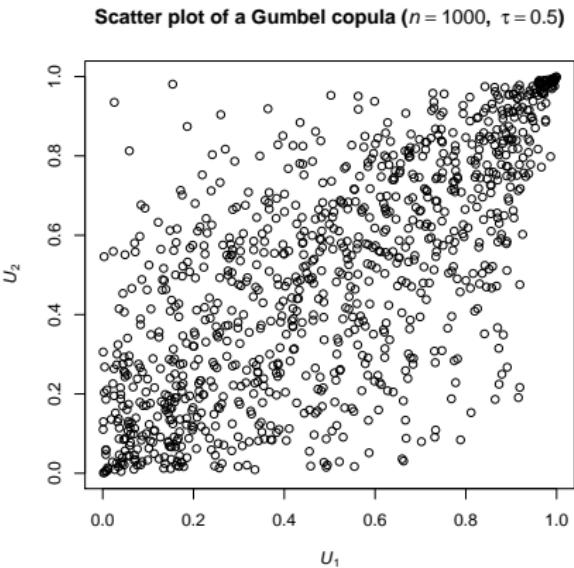
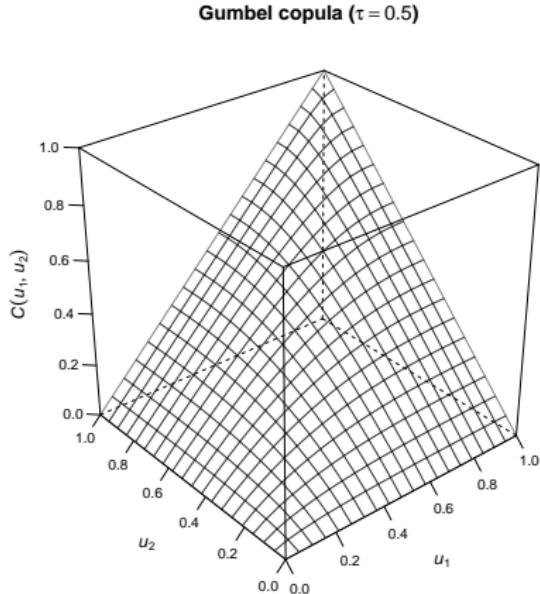


Scatter plot matrix of a Clayton copula ($n = 1000$, $\tau = 0.5$)



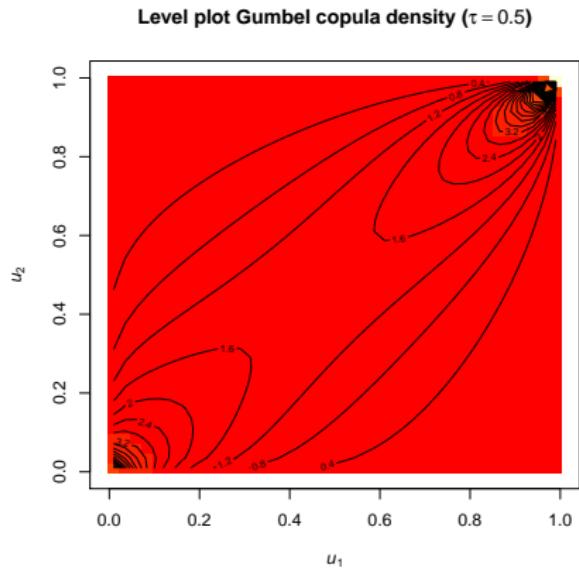
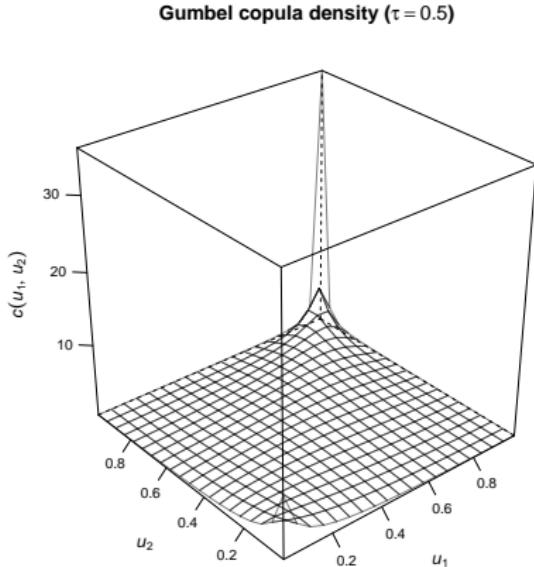
Left: Plot of a bivariate Gumbel copula (Kendall's tau 0.5).

Right: Corresponding scatter plot (sample size $n = 1000$)



Left: Plot of the corresponding density.

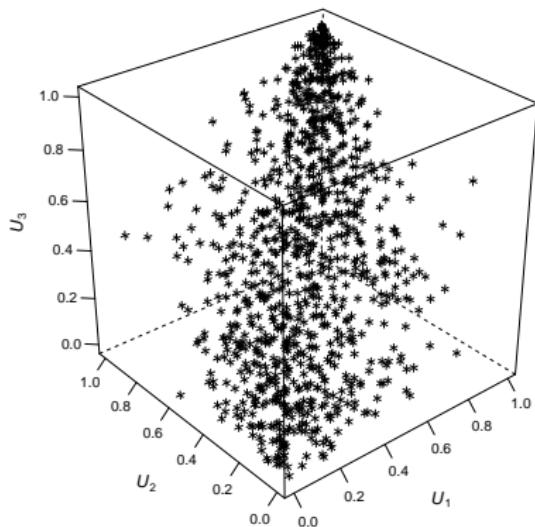
Right: Level plot of the density (with heat colors).



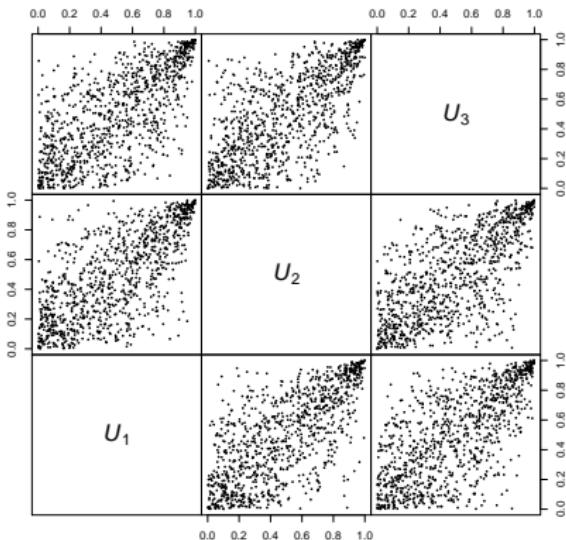
Left: Cloud plot of a trivariate Gumbel copula (sample size $n = 1000$; Kendall's tau 0.5).

Right: Corresponding scatter plot matrix.

Gumbel copula cloud plot ($n = 1000$, $\tau = 0.5$)



Scatter plot matrix of a Gumbel copula ($n = 1000$, $\tau = 0.5$)



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

Drawbacks:

- All margins of the same dimension are equal (symmetry or exchangeability; see later)
- Often used only with a small number of parameters (some extensions available, but still less than $d(d - 1)/2$)

7.1.3 Meta distributions

- *Fréchet class*: Class of all dfs F with given marginal dfs F_1, \dots, F_d ;
- *Meta- C models*: All dfs F with the same given copula C .
- **Example:** A *meta-Gauss model* is a multivariate df F with Gauss copula C and some margins F_1, \dots, F_d .

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 7.6.

Algorithm 7.9 (Simulation of implicit copulas)

- 1) Sample $\mathbf{X} \sim F$, where F 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 7.10

- Sampling *Gauss copulas* C_P^{Ga} :

- 1) Sample $\mathbf{X} \sim \text{N}_d(\mathbf{0}, P)$ ($\mathbf{X} \stackrel{\text{d}}{=} A\mathbf{Z}$ for $AA' = P$, $\mathbf{Z} \sim \text{N}_d(\mathbf{0}, I_d)$).
- 2) Return $\mathbf{U} = (\Phi(X_1), \dots, \Phi(X_d))$.

- Sampling *t_ν copulas* $C_{\nu,P}^t$:

- 1) Sample $\mathbf{X} \sim t_d(\nu, \mathbf{0}, P)$ ($\mathbf{X} \stackrel{\text{d}}{=} \sqrt{W}A\mathbf{Z}$ for $W = \frac{1}{V}$, $V \sim \Gamma(\frac{\nu}{2}, \frac{\nu}{2})$).
- 2) Return $\mathbf{U} = (t_\nu(X_1), \dots, t_\nu(X_d))$.

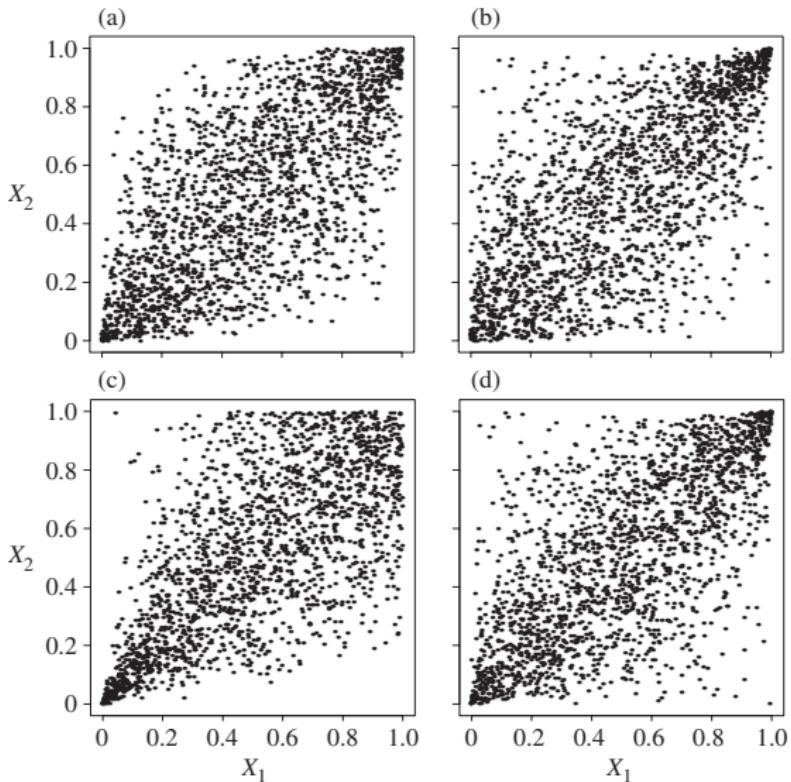
Sampling meta distributions

Meta- C distributions can be sampled via Sklar's Theorem, Part 2).

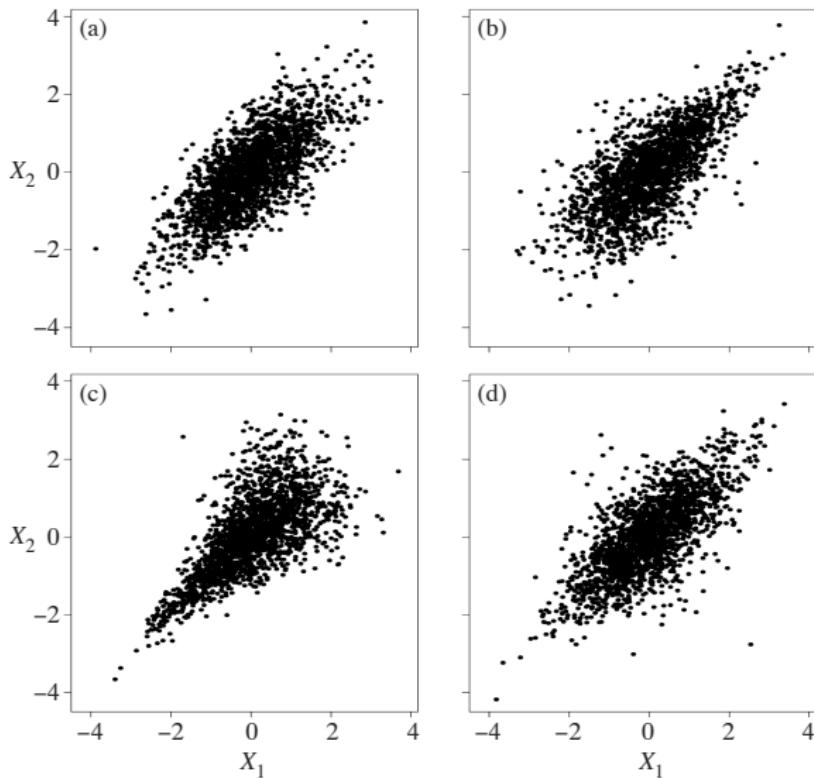
Algorithm 7.11 (Sampling meta- C models)

- 1) Sample $\mathbf{U} \sim C$.
- 2) Return $\mathbf{X} = (F_1^\leftarrow(U_1), \dots, F_d^\leftarrow(U_d))$ (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 $\approx 0.7!$



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 7.12 (Conditional distribution method)

If C is a d -dimensional copula and $\mathbf{U}' \sim \text{U}(0, 1)^d$, let

$$U_1 = U'_1,$$

$$U_2 = C^{\leftarrow}(U'_2 | U_1),$$

⋮

$$U_d = C^{\leftarrow}(U'_d | U_1, \dots, U_{d-1}).$$

Then $\mathbf{U} \sim C$.

This typically involves numerical root-finding and the following result.

Theorem 7.13 (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.

- For $d = 2$ one obtains that $C(u_2 | u_1) = D_1 C(u_1, u_2)$ for a.e. $u_1 \in [0, 1]$.
- For most well-known copula families, the conditional distribution method is neither simple to apply nor fast \Rightarrow Efficient sampling algorithms are typically family-specific.

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 .
- \hat{C} can be expressed as

$$\hat{C}(\mathbf{u}) = \sum_{J \subseteq \{1, \dots, d\}} (-1)^{|J|} C((1 - u_1)^{I_J(1)}, \dots, (1 - u_d)^{I_J(d)})$$

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

$$\bar{F}(\mathbf{x}) = \hat{C}(\bar{F}_1(x_1), \dots, \bar{F}_d(x_d)),$$

where $F(\mathbf{x}) = \mathbb{P}(\mathbf{X} > \mathbf{x})$ 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{F} and $\bar{F}_1, \dots, \bar{F}_d$ are not!

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 f of F satisfies

$$f(\mathbf{x}) = c(F_1(x_1), \dots, F_d(x_d)) \prod_{j=1}^d f_j(x_j) \quad (28)$$

As seen before, we can recover c via

$$c(\mathbf{u}) = \frac{f(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \cdots f_d(F_d^{-1}(u_d))}.$$

- It follows from (28) that the log-density splits into

$$\log f(\mathbf{x}) = \log c(F_1(x_1), \dots, F_d(x_d)) + \sum_{j=1}^d \log f_j(x_j).$$

which allows for a *two-stage estimation* (**marginal** and **copula parameters**); see Section 7.5.

Exchangeability

- \mathbf{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 and 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 approximate modelling homogeneous portfolios**.
- **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 7.14 (Perfect dependence)

- 1) $X_2 = T(X_1)$ a.s. with decreasing $T(x) = F_2^\leftarrow(1 - F_1(x))$ (*counter-monotone*) if and only if $C(u_1, u_2) = W(u_1, u_2)$, $u_1, u_2 \in [0, 1]$.
- 2) $X_j = T_j(X_1)$ a.s. with increasing $T_j(x) = F_j^\leftarrow(F_1(x))$, $j \in \{2, \dots, d\}$ (*comonotone*), if and only if $C(\mathbf{u}) = M(\mathbf{u})$, $\mathbf{u} \in [0, 1]^d$.

Proof. See the appendix. □

Proposition 7.15 (Comonotone additivity)

Let $\alpha \in (0, 1)$ and $X_j \sim F_j$, $j \in \{1, \dots, d\}$, be comontone. Then $F_{X_1 + \dots + X_d}^{\leftarrow}(\alpha) = F_1^{\leftarrow}(\alpha) + \dots + F_d^{\leftarrow}(\alpha)$; technical proof, see appendix.

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 7.16 (Höffding's formula)

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 F . Then

$$\text{cov}(X_1, X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (F(x_1, x_2) - F_1(x_1)F_2(x_2)) dx_1 dx_2.$$

Classical properties and drawbacks of linear correlation

Let X_1 and X_2 be two random variables with $\mathbb{E}(X_j^2) < \infty$, $j \in \{1, 2\}$.

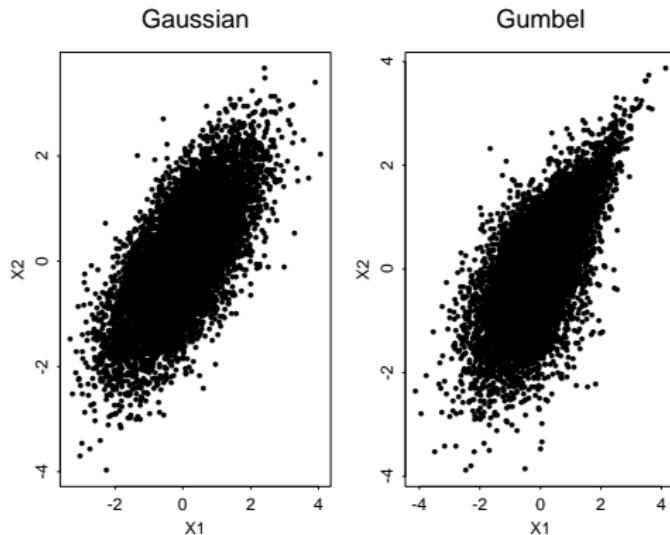
Note that ρ depends on the marginal distributions! In particular, second moments have to exist (not the case, e.g. for $X_1, X_2 \stackrel{\text{ind.}}{\sim} F(x) = 1 - x^{-3}$)

- $|\rho| \leq 1$. 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 $\rho = \pm 1$. This discards other strong functional dependence such as $X_2 = X_1^2$, for example.
- If X_1 and X_2 are independent, then $\rho = 0$. However, the converse is not true in general; see Example 7.17 below.
- ρ is invariant under strictly increasing linear transformations on $\text{ran } X_1 \times \text{ran } X_2$ but not invariant under strictly increasing functions in general. To see this, consider $(X_1, X_2) \sim N_2(\mathbf{0}, P)$ with $P_{12} = \rho$. Then $\rho(X_1, X_2) = \rho$, but $\rho(F_1(X_1), F_2(X_2)) = \frac{6}{\pi} \arcsin(\rho/2)$.

Correlation fallacies

Fallacy 1: F_1 , F_2 , and ρ uniquely determine F

This is true for bivariate elliptical distributions, but wrong in general. The following samples both have $N(0, 1)$ margins and correlation $\rho = 0.7$, yet come from different (copula) models:



Another example is this.

Example 7.17 (Uncorrelated $\not\Rightarrow$ independent)

- Consider the two risks

$$X_1 = Z \quad (\text{Profit \& Loss Country A}),$$

$$X_2 = ZV \quad (\text{Profit \& Loss Country B}),$$

where V, Z are independent with $Z \sim N(0, 1)$ and $\mathbb{P}(V = -1) = \mathbb{P}(V = 1) = 1/2$. 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).

- Consider $(X'_1, X'_2) \sim N_2(\mathbf{0}, I_2)$. Both (X'_1, X'_2) and (X_1, X_2) have $N(0, 1)$ margins and $\rho = 0$, but the copula of (X'_1, X'_2) is Π and the copula of (X_1, X_2) is the convex combination $C(\mathbf{u}) = \lambda M(\mathbf{u}) + (1 - \lambda)W(\mathbf{u})$ for $\lambda = 0.5$.

Fallacy 2: Given F_1, F_2 , any $\rho \in [-1, 1]$ is attainable

This is true for elliptically distributed (X_1, X_2) with $\mathbb{E}(R^2) < \infty$ (as then $\text{corr } \mathbf{X} = P$), but wrong in general:

- If F_1 and F_2 are not of the same type (no linearity), $\rho(X_1, X_2) = 1$ is not attainable (recall that $|\rho| = 1$ 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.).
- What is the attainable range then? Höffding's formula

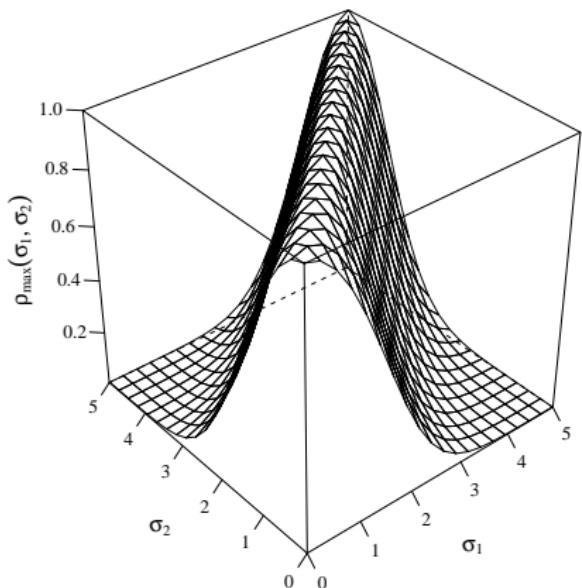
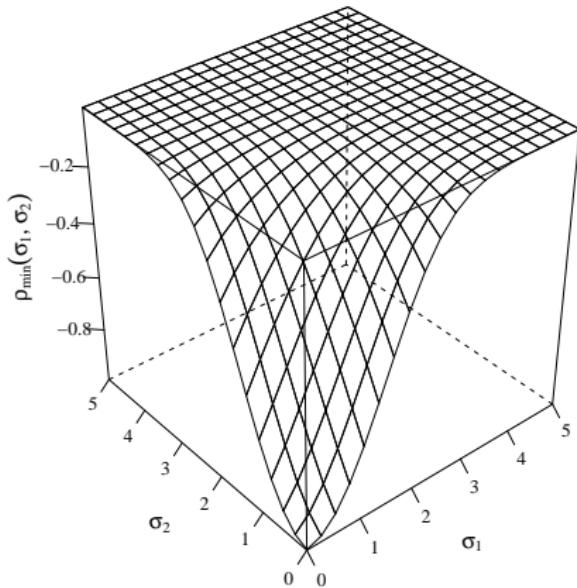
$$\text{cov}(X_1, X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (C(F_1(x_1), F_2(x_2)) - F_1(x_1)F_2(x_2)) dx_1 dx_2.$$

implies bounds on attainable ρ :

$$\rho \in [\rho_{\min}, \rho_{\max}] \quad (\rho_{\min} \text{ is attained for } C = W, \rho_{\max} \text{ for } C = M).$$

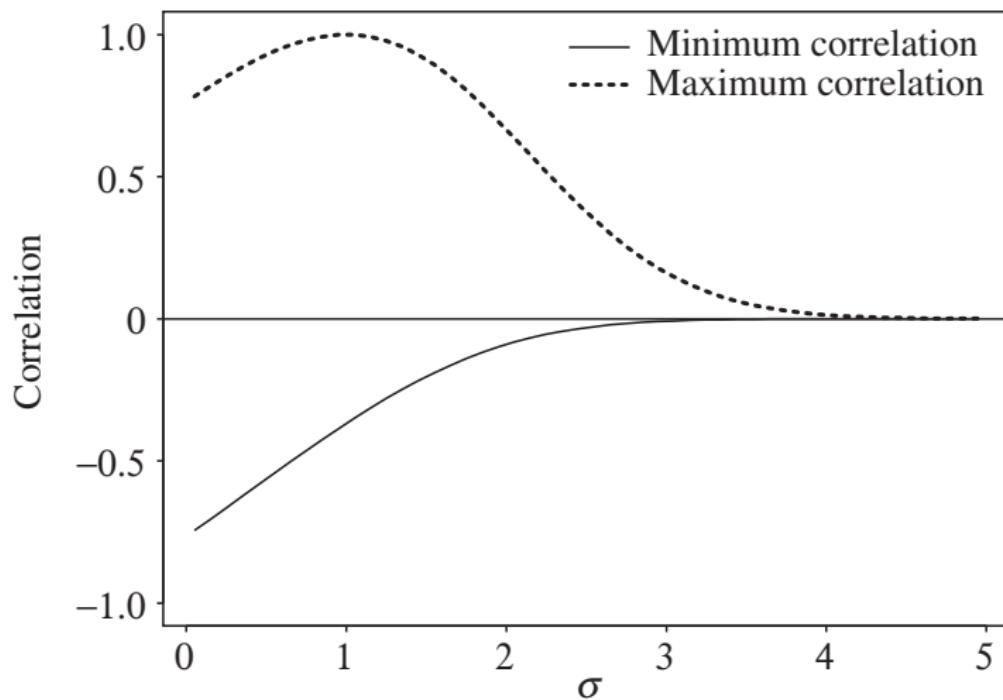
Example 7.18 (Bounds for a model with $\text{LN}(0, \sigma_j^2)$ margins)

Let $X_j \sim \text{LN}(0, \sigma_j^2)$, $j \in \{1, 2\}$. One can show that minimal (ρ_{\min} ; left) and maximal (ρ_{\max} ; right) correlations are given as follows.



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 :



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, so VaR_α is additive (by Proposition 7.15) and additivity provides the largest possible bound in this case as VaR_α is subadditive (by Proposition 6.24).
- Any superadditivity example $\text{VaR}_\alpha(X_1 + X_2) > \text{VaR}_\alpha(X_1) + \text{VaR}_\alpha(X_2)$ under comonotonicity (under comonotonicity, so maximal correlationqrm, the right-hand side is $\text{VaR}_\alpha(X_1 + X_2)$) serves as a counterexample; see Section 2.3.5.

7.2.3 Rank correlation

Rank correlation coefficients are...

- ... always defined;
- ... invariant under strictly increasing transformations of the random variables (hence only depend on the underlying copula).

Kendall's tau and Spearman's rho

Definition 7.19 (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

$$\begin{aligned}\rho_\tau &= \mathbb{E}(\text{sign}((X_1 - X'_1)(X_2 - X'_2))) \\ &= \mathbb{P}((X_1 - X'_1)(X_2 - X'_2) > 0) - \mathbb{P}((X_1 - X'_1)(X_2 - X'_2) < 0),\end{aligned}$$

where $\text{sign}(x) = I_{(0,\infty)}(x) - I_{(-\infty,0)}(x)$ (so -1 for $x < 0$, 0 for $x = 0$ and 1 for $x > 0$).

By definition, Kendall's tau is the probability of *concordance* minus the probability of *discordance*.

Proposition 7.20 (Formula for Kendall's tau)

Let $X_j \sim F_j$ with F_j continuous, $j \in \{1, 2\}$, and copula C . Then

$$\rho_\tau = 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1.$$

Proof. See the appendix. □

An estimator of ρ_τ is provided by the sample version of Kendall's tau

$$r_n^\tau = \frac{1}{\binom{n}{2}} \sum_{1 \leq i_1 < i_2 \leq n} \text{sign}((X_{i_1 1} - X_{i_2 1})(X_{i_1 2} - X_{i_2 2})). \quad (29)$$

Definition 7.21 (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 7.22 (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 Höffding's formula, 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 r_n^S is given by the sample correlation computed from componentwise (scaled) ranks (i.e. marginal empirical dfs) of the data.
- For $\kappa = \rho_\tau$ and $\kappa = \rho_S$, Embrechts et al. (2002) show that $\kappa = \pm 1$ if and only if X_1, X_2 are co-/countermonotonic.
- Fallacy 1 (F_1, F_2, ρ uniquely determine F) 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 2 (For F_1, F_2 , any $\rho \in [-1, 1]$ is attainable) is solved. Take $F(x_1, x_2) = \lambda \textcolor{brown}{M}(F_1(x_1), F_2(x_2)) + (1 - \lambda) \textcolor{brown}{W}(F_1(x_1), F_2(x_2))$. This is a model with $\rho_S = 2\lambda - 1$ and $\tau = -\lambda^2 + 3\lambda - 1$ (choose $\lambda \in [0, 1]$ as desired).
- 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 Fallacy 3 earlier.
- Also, in general, $\kappa = 0$ does not imply independence.
- Nevertheless, rank correlations are useful to summarize dependence, to parameterize copula families to make dependence comparable and for copula parameter calibration or estimation.

7.2.4 Coefficients of tail dependence

Goal: Measure *extremal dependence*, i.e. dependence in the *joint tails*.

Definition 7.23 (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

$$\lambda_l = \lim_{u \downarrow 0} \mathbb{P}(X_2 \leq F_2^\leftarrow(u) \mid X_1 \leq F_1^\leftarrow(u)),$$

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

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

$$\begin{aligned}\mathbb{P}(X_2 \leq F_2^\leftarrow(u) \mid X_1 \leq F_1^\leftarrow(u)) &= \frac{\mathbb{P}(X_1 \leq F_1^\leftarrow(u), X_2 \leq F_2^\leftarrow(u))}{\mathbb{P}(X_1 \leq F_1^\leftarrow(u))} \\ &= \frac{F(F_1^\leftarrow(u), F_2^\leftarrow(u))}{F_1(F_1^\leftarrow(u))} \stackrel{\text{Sklar}}{=} \frac{C(u, u)}{u}, \quad u \in (0, 1), \text{ so } \lambda_l = \lim_{u \downarrow 0} \frac{C(u, u)}{u}.\end{aligned}$$

- 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, $\lambda_l = 2 \lim_{u \downarrow 0} \mathbf{D}_1 C(u, u)$. By Theorem 7.13, $\lambda_l = 2 \lim_{u \downarrow 0} \mathbb{P}(U_2 \leq u \mid U_1 = u)$ for $(U_1, U_2) \sim C$. Combined with any continuous df F , and $(X_1, X_2) = (F_1^\leftarrow(U_1), F_2^\leftarrow(U_2))$, one has

$$\lambda_l = 2 \lim_{x \downarrow -\infty} \mathbb{P}(X_2 \leq x \mid X_1 = x) \stackrel{\substack{\text{if} \\ \text{density}}}{=} 2 \lim_{x \downarrow -\infty} \int_{-\infty}^x f_{X_2 \mid X_1=x}(x_2) dx_2. \tag{30}$$

- Similarly as above, for the upper tail-dependence coefficient,

$$\begin{aligned}\lambda_u &= \lim_{u \uparrow 1} \frac{1 - 2u + C(u, u)}{1 - u} = \lim_{u \downarrow 0} \frac{\hat{C}(u, u)}{u} \\ &= \lim_{u \uparrow 1} \frac{2(1 - u) - (1 - C(u, u))}{1 - u} = 2 - \lim_{u \uparrow 1} \frac{1 - C(u, u)}{1 - u}.\end{aligned}$$

- For all radially symmetric copulas (e.g. the bivariate $C_P^{G_a}$ and $C_{\nu, P}^t$ copulas), we have $\lambda_l = \lambda_u =: \lambda$.
- 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 $\mathbf{X} \stackrel{\text{d}}{=} \boldsymbol{\mu} + \sqrt{W}\mathbf{A}\mathbf{Z}$ ($\mathbf{X} \stackrel{\text{d}}{=} \mathbf{m}(W) + \sqrt{W}\mathbf{A}\mathbf{Z}$); e.g. Gauss, t copulas.

7.3.1 Tail dependence

Coefficients of tail dependence

Let (X_1, X_2) be distributed according to a normal variance mixture and assume (w.l.o.g.) that $\boldsymbol{\mu} = (0, 0)$ and $\mathbf{A}\mathbf{A}' = P = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. In this case, $F_1 = F_2$ and C is symmetric and radially symmetric. We thus obtain that

$$\lambda \stackrel{\text{radial}}{\underset{\text{symm.}}{=}} \lambda_l \stackrel{\text{symm.}}{\underset{(30)}{=}} 2 \lim_{x \downarrow -\infty} \mathbb{P}(X_2 \leq x \mid X_1 = x).$$

Example 7.24 (λ for the Gauss and t copula)

- Considering the bivariate $N(\mathbf{0}, P)$ density, one can show (via $f_{X_2|X_1}(x_2 \mid x_1) = \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_1}(x_1)}$) that $X_2 \mid X_1 = x \sim N(\rho x, 1 - \rho^2)$. This implies that

$\lambda = 2 \lim_{x \downarrow -\infty} \mathbb{P}(X_2 \leq x | X_1 = x) = 2 \lim_{x \downarrow -\infty} \Phi\left(\frac{x(1-\rho)}{\sqrt{1-\rho^2}}\right) = I_{\{\rho=1\}}$
 (essentially no tail dependence).

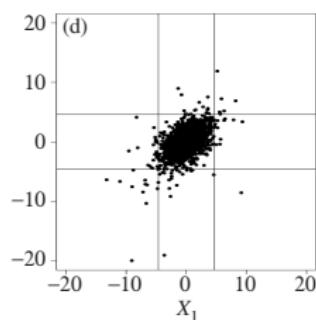
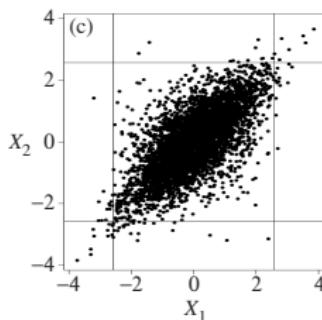
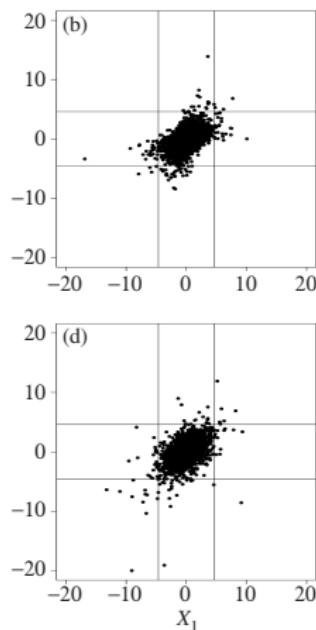
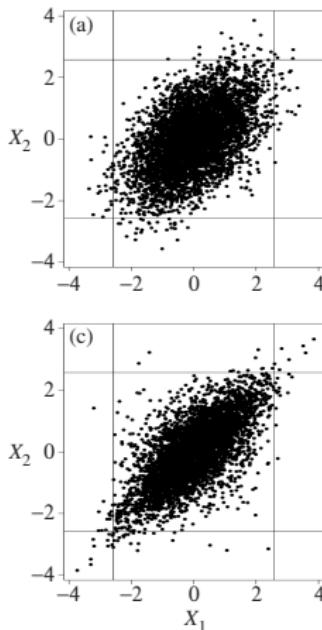
- For $C_{\nu, P}^t$, one can show that $X_2 | X_1 = x \sim t_{\nu+1}(\rho x, \frac{(1-\rho^2)(\nu+x^2)}{\nu+1})$ and thus $\mathbb{P}(X_2 \leq x | X_1 = x) = t_{\nu+1}\left(\frac{x-\rho x}{\sqrt{\frac{(1-\rho^2)(\nu+x^2)}{\nu+1}}}\right)$. Hence

$$\lambda = 2t_{\nu+1}\left(-\sqrt{\frac{(\nu+1)(1-\rho)}{1+\rho}}\right) \quad (\text{tail dependence}).$$

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



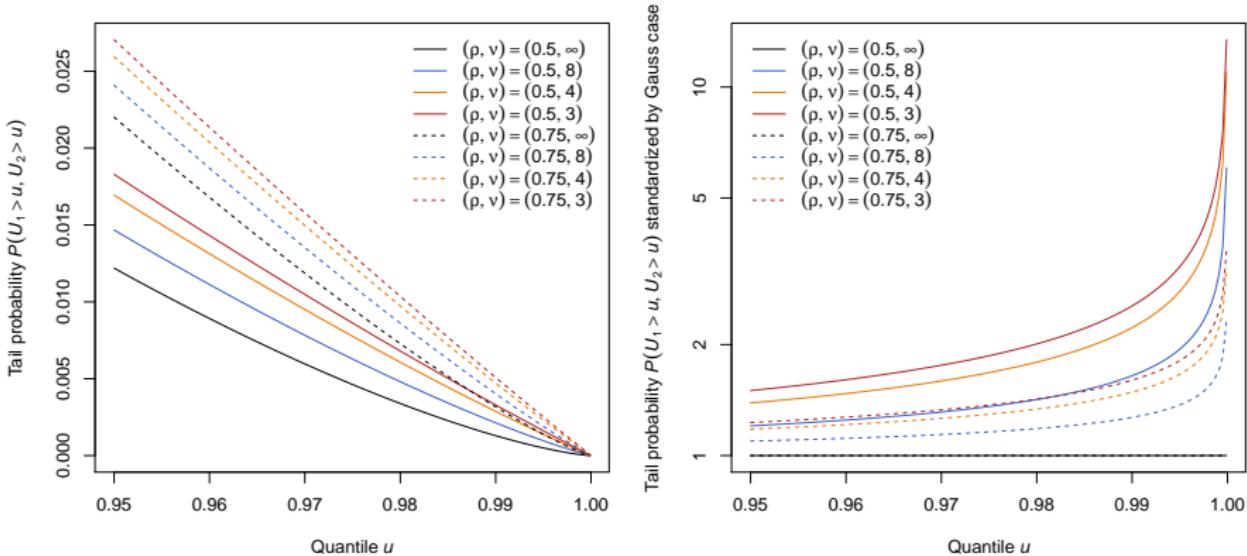
5000 samples from

- (a) $N_2(\mathbf{0}, P = (\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}))$, $\rho = 0.5$;
- (b) $C_\rho^{G_a}$ 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 the true 0.005- and 0.995-quantiles.

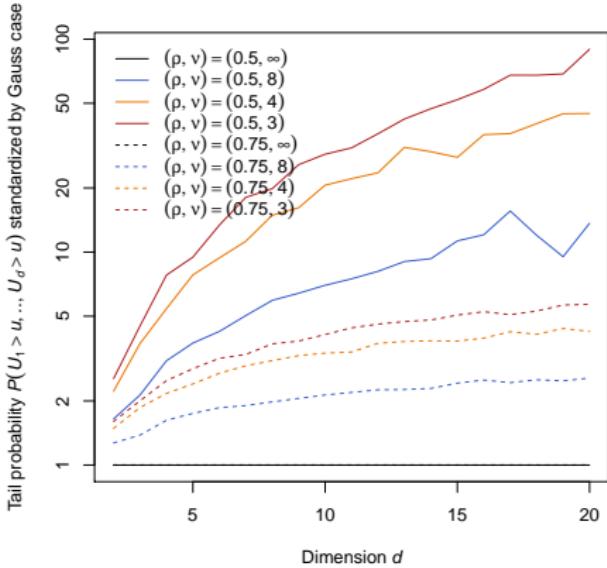
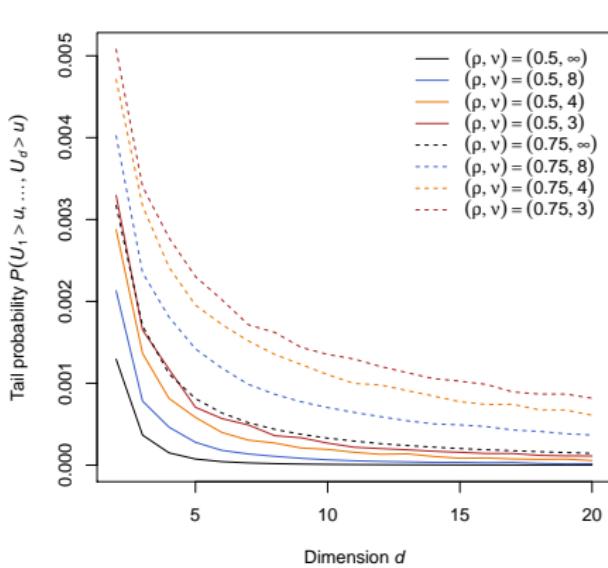
Note the different number of points in the bivariate tails (all models have the same Kendall's tau!)

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:** $u \mapsto \frac{\mathbb{P}(U_1 > u, U_2 > u)}{\mathbb{P}(V_1 > u, V_2 > u)} \stackrel{\text{radial}}{=} \stackrel{\text{symm.}}{=} \frac{C_{\nu, \rho}^t(u, u)}{C_{\rho}^{\text{Ga}}(u, u)}$

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:** $d \mapsto \frac{\mathbb{P}(U_1 > u, \dots, U_d > u)}{\mathbb{P}(V_1 > u, \dots, V_d > u)}$ $\stackrel{\text{radial}}{=}$ $\stackrel{\text{symm.}}{=}$ $\frac{C_{\nu, \rho}^t(u, \dots, u)}{C_{\rho}^{\text{Ga}}(u, \dots, u)}$ for $u = 0.99$.

Example 7.25 (Interpretation of joint tail probabilities)

- Consider 5 daily negative (log-)returns $\mathbf{X} = (X_1, \dots, X_5)$ with fixed margins and 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 negative returns lie above their $u = 0.99$ quantiles is

$$\begin{aligned}\mathbb{P}(X_1 > F_1^\leftarrow(u), \dots, X_5 > F_5^\leftarrow(u)) &= \mathbb{P}(U_1 > u, \dots, U_5 > u) \\ &\approx \frac{7.48 \times 10^{-5}}{\text{MC error}}.\end{aligned}$$

In the long run such an event will happen once every $1/7.48 \times 10^{-5} \approx 13\,369$ 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 !

7.3.2 Rank correlations

Proposition 7.26 (Spearman's rho for normal variance mixtures)

Let $\mathbf{X} \sim M_2(\mathbf{0}, P, \hat{F}_W)$ 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 $W, \tilde{W}, \bar{W} \stackrel{\text{ind.}}{\sim} F_W$ with Laplace–Stieltjes transform \hat{F}_W . For Gauss copulas, $\rho_S = \frac{6}{\pi} \arcsin(\frac{\rho}{2})$.

Proof. See the appendix. □

Proposition 7.27 (Kendall's tau for elliptical distributions)

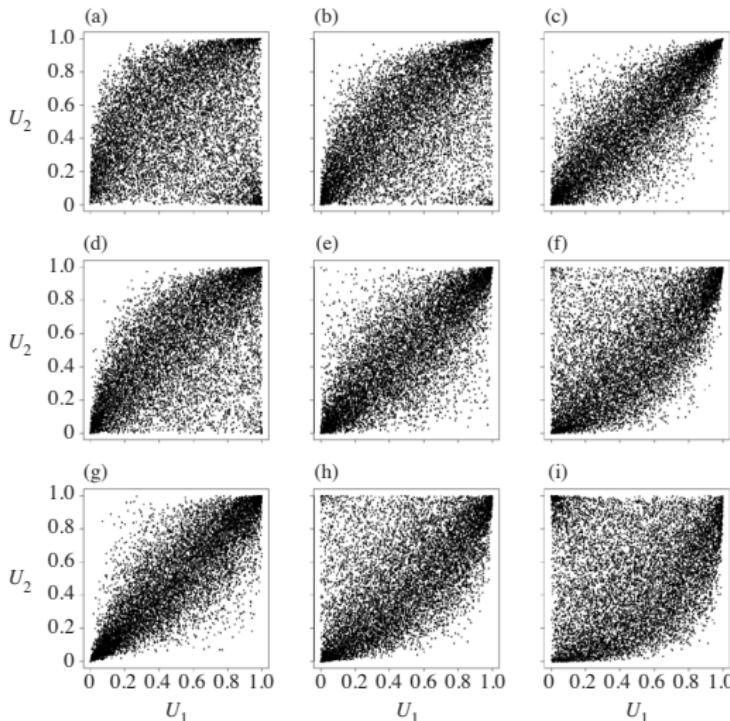
Let $\mathbf{X} \sim E_2(\mathbf{0}, P, \psi)$ with $\mathbb{P}(\mathbf{X} = \mathbf{0}) = 0$, $\rho = P_{12}$. Then $\rho_\tau = \frac{2}{\pi} \arcsin \rho$.

Proof. See the appendix. □

7.3.3 Skewed normal mixture copulas

- *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 a generalized hyperbolic distribution; see McNeil et al. (2015, Sections 6.2.3 and 7.3.3) for more details.
- It can be sampled as other implicit copulas; see Algorithm 7.9 (the evaluation of the margins requires numerical integration of a skewed *t* density).
- The main advantage of such a copula over $C_{\nu,P}^t$ is its radial asymmetry (e.g. for modelling $\lambda_l \neq \lambda_u$)

10 000 samples from $C_{\nu=5}^t$, $\rho=0.8$, $\gamma=0.8(I_{\{i<2\}}-I_{\{i>2\}}, I_{\{j>2\}}-I_{\{j<2\}})$:



- (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

- *Grouped normal mixture copulas* are copulas which attach together a set of normal mixture copulas.
- Let $\mathbf{Y} \sim N_d(\mathbf{0}, P)$ (so $\mathbf{Y} \stackrel{d}{=} A\mathbf{Z}$ as before). The *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 $(W_1, \dots, W_S) \sim M(IG(\frac{\nu_1}{2}, \frac{\nu_1}{2}), \dots, IG(\frac{\nu_S}{2}, \frac{\nu_S}{2}))$; see Demarta and McNeil (2005) for 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

Recall that 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 7.28 (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.

- For a strict and twice-continuously differentiable ψ , one can show that

$$\rho_\tau = 1 - 4 \int_0^\infty t(\psi'(t))^2 dt = 1 + 4 \int_0^1 \frac{\psi^{-1}(t)}{(\psi^{-1}(t))'} dt.$$

- If ψ is strict, $\lambda_l = 2 \lim_{t \rightarrow \infty} \frac{\psi'(2t)}{\psi'(t)}$ and $\lambda_u = 2 - 2 \lim_{t \downarrow 0} \frac{\psi'(2t)}{\psi'(t)}$ (as seen before).

- The most widely used one-parameter Archimedean copulas are:

Family	θ	$\psi(t)$	$V \sim F = \mathcal{LS}^{-1}(\psi)$
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	$[1, \infty)$	$\exp(-t^{1/\theta})$	$S(1/\theta, 1, \cos^\theta(\pi/(2\theta)), I_{\{\theta=1\}}; 1)$
J	$[1, \infty)$	$1 - (1 - \exp(-t))^{1/\theta}$	$\text{Sibuya}(1/\theta)$

Family	ρ_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 7.29 (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 $\psi \in \Psi_\infty$.

Bernstein's Theorem characterizes all $\psi \in \Psi_\infty$.

Theorem 7.30 (Bernstein (1928))

$\psi(0) = 1$, ψ c.m. if and only if $\psi(t) = \mathbb{E}(\exp(-tV))$ for $V \sim G$ with $V \geq 0$ and $G(0) = 0$.

We thus use the notation $\psi = \hat{G}$ and call all Archimedean copulas with $\psi \in \Psi_\infty$ *LT-Archimedean copulas*.

Proposition 7.31 (Stochastic representation, related properties)

Let $\psi \in \Psi_\infty$ with $V \sim G$ such that $\hat{G} = \psi$ and let $E_1, \dots, E_d \stackrel{\text{ind.}}{\sim} \text{Exp}(1)$ be independent of V . Then

- 1) The survival copula of $\mathbf{X} = (\frac{E_1}{V}, \dots, \frac{E_d}{V})$ is Archimedean (with ψ).
- 2) $\mathbf{U} = (\psi(X_1), \dots, \psi(X_d)) \sim \mathbf{C}$ and the U_j 's are conditionally independent given V with $\mathbb{P}(U_j \leq u | V = v) = \exp(-v\psi^{-1}(u))$.

Proof.

- 1) The joint survival function of \mathbf{X} is given by

$$\begin{aligned}\bar{F}(\mathbf{x}) &= \mathbb{P}(X_j > x_j \ \forall j) = \int_0^\infty \mathbb{P}(E_j/V > x_j \ \forall j | V = v) dG(v) \\ &= \int_0^\infty \mathbb{P}(E_j > vx_j \ \forall j) dG(v) = \int_0^\infty \prod_{j=1}^d \exp(-vx_j) dG(v) \\ &= \int_0^\infty \exp\left(-v \sum_{j=1}^d x_j\right) dG(v) = \psi\left(\sum_{j=1}^d x_j\right).\end{aligned}$$

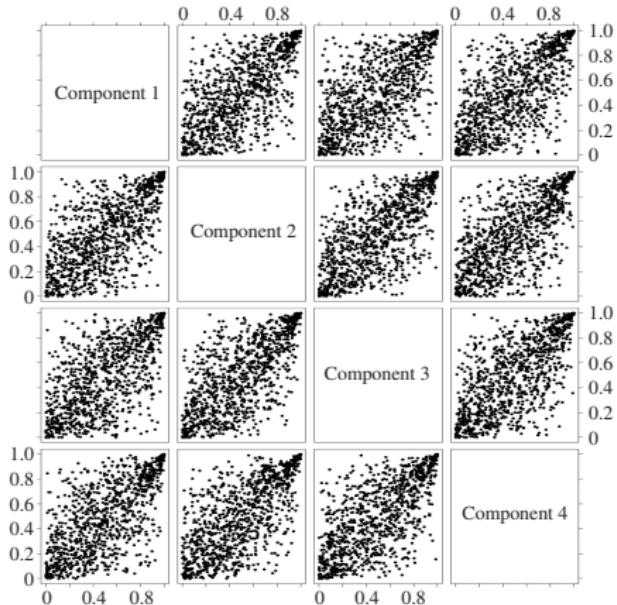
The j th marginal survival function is thus (set $x_k = 0 \ \forall k \neq j$)
 $\bar{F}_j(x_j) = \mathbb{P}(X_j > x_j) = \psi(x_j)$ (\downarrow and continuous) and therefore
 $\hat{C}(\mathbf{u}) = \bar{F}(\bar{F}_1^\leftarrow(u_1), \dots, \bar{F}_d^\leftarrow(u_d)) = \psi(\sum_{j=1}^d \psi^{-1}(u_j)).$

- 2) $\mathbb{P}(\mathbf{U} \leq \mathbf{u}) = \mathbb{P}(X_j > \psi^{-1}(u_j) \ \forall j) \stackrel{1)}{=} \psi(\sum_{j=1}^d \psi^{-1}(u_j)).$ Conditional independence is clear by construction and $\mathbb{P}(U_j \leq u | V = v) = \mathbb{P}(X_j > \psi^{-1}(u) | V = v) = \mathbb{P}(E_j > v\psi^{-1}(u)) = \exp(-v\psi^{-1}(u)).$ \square

Algorithm 7.32 (Marshall and Olkin (1988))

- 1) Sample $\mathbf{V} \sim \mathbf{G}$ (df corresponding to ψ).
- 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 ($\rho_\tau = 0.5$; $\lambda_u \approx 0.5858$)



- Various non-exchangeable extensions to Archimedean copulas exist.
- For fixed d , c.m. can be relaxed to d -monotonicity; see McNeil and Nešlehová (2009).

7.5 Fitting copulas to data

- Let $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_n$ be independent random vectors with df F , continuous margins F_1, \dots, F_d and copula C . We assume we have data $\mathbf{x}_1, \dots, \mathbf{x}_n$, interpreted as realizations of $\mathbf{X}_1, \dots, \mathbf{X}_n$; in what follows we work with the latter.

- Assume

- $F_j = F_j(\cdot; \theta_{0,j})$ for some $\theta_{0,j} \in \Theta_j$, $j \in \{1, \dots, d\}$;
($F_j(\cdot; \theta_j)$ continuous $\forall \theta_j \in \Theta_j$, $j \in \{1, \dots, d\}$)
- $C = C(\cdot; \theta_{0,C})$ for some $\theta_{0,C} \in \Theta_C$.

Thus F has the true but unknown parameter vector $\theta_0 = (\theta'_{0,C}, \theta'_{0,1}, \dots, \theta'_{0,d})'$ to be estimated.

- Here, we focus particularly on $\theta_{0,C}$. 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 estimating $\theta_{0,C}$ by solving $\rho_\tau(\theta_C) = r_n^\tau$ w.r.t. θ_C , i.e.

$$\hat{\theta}_{n,C}^{\text{IKTE}} = \rho_\tau^{-1}(r_n^\tau), \quad (\text{inversion of Kendall's tau estimator (IKTE)})$$

where $\rho_\tau(\cdot)$ denotes Kendall's tau as a function in θ and r_n^τ is the sample version of Kendall's tau (computed via (29) 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 $r_{n,j_1j_2}^\tau$ denotes the sample version of Kendall's tau for data pair (j_1, j_2) , then $\hat{P}_{n,j_1j_2}^{\text{IKTE}} = \sin(\frac{\pi}{2} r_{n,j_1j_2}^\tau)$; see Proposition 7.27.

For obtaining a proper correlation matrix P (positive semi-definite), see Higham (2002).

- ▶ For Gauss copulas, it is preferable to use Spearman's rho based on

$$\rho_S = \frac{6}{\pi} \arcsin \frac{\rho}{2} \approx \rho.$$

Prop.7.26

The latter approximation error is comparably small, so that the matrix of pairwise sample versions of Spearman's rho is an estimator for P .

- ▶ For t copulas, \hat{P}_n^{IKTE} can be used to estimate P and then ν can be estimated via its MLE based on \hat{P}_n^{IKTE} .

7.5.2 Forming a pseudo-sample from the copula

- X_1, \dots, 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 $\hat{U}_i = (\hat{U}_{i1}, \dots, \hat{U}_{id}) = (\hat{F}_1(X_{i1}), \dots, \hat{F}_d(X_{id}))$, $i \in \{1, \dots, n\}$, where \hat{F}_j denotes an estimator of F_j ; see Lemma 7.6. Note

that $\hat{U}_1, \dots, \hat{U}_n$ are typically neither independent (even if X_1, \dots, 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

$$\hat{U}_{ij} = \frac{n}{n+1} \hat{F}_{n,j}(X_{ij}) = \frac{R_{ij}}{n+1}, \quad (31)$$

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 (31) for estimating $\theta_{0,C}$ (to keep the error due to misspecification of the margins small).
- 3) EVT-based. Bodies are modelled empirically; tails semiparametrically via GPD.

7.5.3 Maximum likelihood estimation

The (classical) maximum likelihood estimator

- By Sklar's Theorem, the density of F is given by

$$f(\mathbf{x}; \boldsymbol{\theta}_0) = c(F_1(x_1; \boldsymbol{\theta}_{0,1}), \dots, F_d(x_d; \boldsymbol{\theta}_{0,d}); \boldsymbol{\theta}_{0,C}) \prod_{j=1}^d f_j(x_j; \boldsymbol{\theta}_{0,j}).$$

- The log-likelihood based on $\mathbf{X}_1, \dots, \mathbf{X}_n$ is thus

$$\begin{aligned}\ell(\boldsymbol{\theta}; \mathbf{X}_1, \dots, \mathbf{X}_n) &= \sum_{i=1}^n \ell(\boldsymbol{\theta}; \mathbf{X}_i) \\ &= \sum_{i=1}^n \ell_C(\boldsymbol{\theta}_C; F_1(X_{i1}; \boldsymbol{\theta}_1), \dots, F_d(X_{id}; \boldsymbol{\theta}_d)) + \sum_{i=1}^n \sum_{j=1}^d \ell_j(\boldsymbol{\theta}_j; X_{ij}),\end{aligned}$$

where

$$\ell_C(\boldsymbol{\theta}_C; u_1, \dots, u_d) = \log c(u_1, \dots, u_d; \boldsymbol{\theta}_C)$$

$$\ell_j(\boldsymbol{\theta}_j; x) = \log f_j(x; \boldsymbol{\theta}_j), \quad j \in \{1, \dots, d\}.$$

- The *maximum likelihood estimator (MLE)* of θ_0 is

$$\hat{\theta}_n^{\text{MLE}} = \underset{\theta \in \Theta}{\operatorname{arg\!sup}} \ell(\theta; \mathbf{X}_1, \dots, \mathbf{X}_n).$$

This optimization is typically done by numerical means. Note that this can be quite demanding, especially in high dimensions.

The inference functions for margins estimator

- 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

$$\hat{\theta}_{n,C}^{\text{IFME}} = \underset{\theta_C \in \Theta_C}{\operatorname{arg\!sup}} \ell(\theta_C, \hat{\theta}_{n,1}^{\text{MLE}}, \dots, \hat{\theta}_{n,d}^{\text{MLE}}; \mathbf{X}_1, \dots, \mathbf{X}_n).$$

The *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).
- $\hat{\theta}_n^{\text{IFME}}$ 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:

$\hat{\theta}_n^{\text{MLE}}$ solves $\left(\frac{\partial}{\partial \boldsymbol{\theta}_C} \ell, \frac{\partial}{\partial \boldsymbol{\theta}_1} \ell, \dots, \frac{\partial}{\partial \boldsymbol{\theta}_d} \ell \right) = \mathbf{0}$,

$\hat{\theta}_n^{\text{IFME}}$ 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 7.33 (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 $\theta_{0,C}$ (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 to $\hat{\theta}_n^{\text{IFME}}$, 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

$$\hat{\theta}_{n,C}^{\text{MPLE}} = \underset{\theta_C \in \Theta_C}{\operatorname{arg\!sup}} \sum_{i=1}^n \ell_C(\theta_C; \hat{U}_{i1}, \dots, \hat{U}_{id}) = \underset{\theta_C \in \Theta_C}{\operatorname{arg\!sup}} \sum_{i=1}^n \log c(\hat{U}_i; \theta_C).$$

- 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 7.34 (Fitting the Gauss copula)

- The (copula-related) log-likelihood ℓ_C is

$$\ell_C(P; \hat{U}_1, \dots, \hat{U}_n) = \sum_{i=1}^n \ell_C(P; \hat{U}_i) \stackrel{\text{Eq. (27)}}{=} \sum_{i=1}^n \log c_P^{\text{G}\alpha}(\hat{U}_i).$$

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 $O(d^2)$.

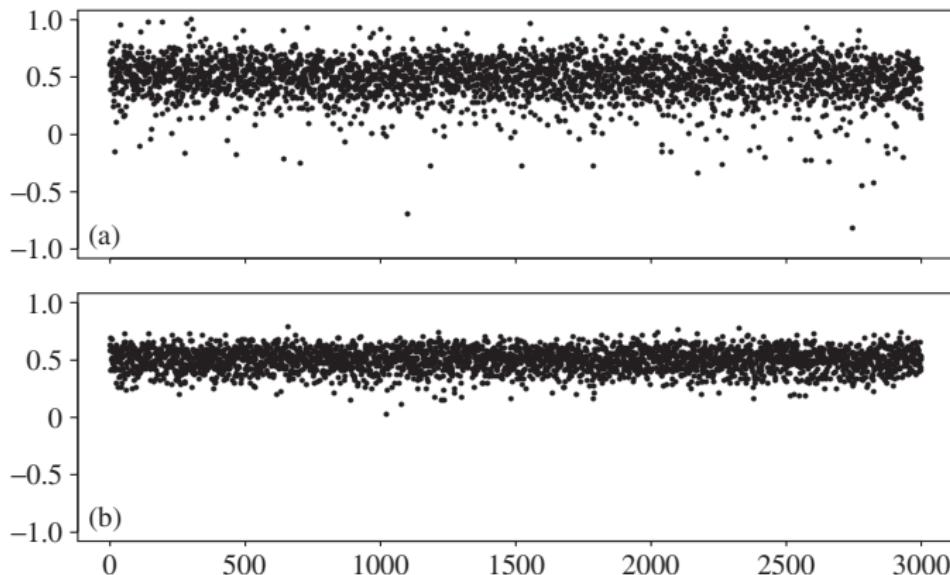
- Alternatively, use pairwise inversion of Spearman's rho or Kendall's tau.

Example 7.35 (Fitting the t copula)

- For small d , maximize the likelihood over all correlation matrices (as for the Gauss copula case) and the d.o.f. ν .
- For moderate/larger d , do:
 - Estimate P via pairwise inversion of Kendall's tau (see above).
 - Plug \hat{P} into the likelihood and maximize it w.r.t. ν to obtain $\hat{\nu}_n$.

Example 7.36 (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 $\rho = 0.5$). Shall we estimate ρ via the sample correlation (estimates are shown in (a)) or via inversion of Kendall's tau (shown in (b))? The variance of the latter is smaller!



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 . See the appendix for a graphical approach.

7.6 A copulas-based proof of subadditivity of ES

Proposition 7.37 (Subadditivity of ES)

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

Proof.

- Let $L = F_L^\leftarrow(U)$ and $Y = I_{\{U>\alpha\}} \sim \text{B}(1, 1 - \alpha)$ for $U \sim \text{U}(0, 1)$.
- Then $\text{ES}_\alpha(L) = \frac{1}{1-\alpha} \int_\alpha^1 F_L^\leftarrow(u) du = \frac{1}{1-\alpha} \int_0^1 F_L^\leftarrow(u) I_{\{u>\alpha\}} \cdot 1 du = \frac{1}{1-\alpha} \mathbb{E}(F_L^\leftarrow(U) I_{\{U>\alpha\}}) = \frac{1}{1-\alpha} \mathbb{E}(LY).$
- L and Y are comontone. Hence, for any other $\tilde{Y} \sim \text{B}(1, 1 - \alpha)$,

$$\mathbb{E}(L\tilde{Y}) = \text{cov}(L, \tilde{Y}) + \mathbb{E}(L)\mathbb{E}(\tilde{Y}) \stackrel{\text{Höffding}}{\leq} \text{cov}(L, Y) + \mathbb{E}(L)\mathbb{E}(Y) = \mathbb{E}(LY)$$

and thus $\text{ES}_\alpha(L) = \sup_{\{\tilde{Y} \sim \text{B}(1, 1 - \alpha)\}} \mathbb{E}(L\tilde{Y}) / (1 - \alpha)$. □

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.
- Berndt, A., Douglas, R., Duffie, D., Ferguson, F., and Schranz, D. (2008), Measuring Default Risk Premia from Default Swap rates and EDF's, Preprint, Tepper School of Business, Carnegie Mellon University.
- Bernstein, S. N. (1928), Sur les fonctions absolument monotones, *Acta Mathematica*, 52, 1–66.
- 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, 2nd, New York: Springer.
- Chavez-Demoulin, V., Embrechts, P., and Hofert, M. (2014), An extreme value approach for modeling operational risk losses depending on covariates, *Journal of Risk and Insurance*, to appear.
- CME SPAN: Standard Portfolio Analysis of Risk (2010), www.cmegroup.com/cme-span Chicago Mercantile Exchange.
- Crosbie, P. J. and Bohn, J. R. (2002), Modeling default risk, Technical document, Moody's/KMV, New York.
- D'Agostino, R. B. and Stephens, M. A. (1986), Goodness-of-fit techniques, Dekker.
- 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, Berlin: 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: 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. T. Rachev, Elsevier, 331–385.

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:10.1017/asb.2013.6.

Embrechts, P. and Wang, R. (2015), Seven Proofs for the Subadditivity of Expected Shortfall, *Dependence Modeling*, 3(1), 126–140.

- Fang, K.-T., Kotz, S., and Ng, K.-W. (1990), Symmetric Multivariate and Related Distributions, London: Chapman & Hall.
- Föllmer, H. and Schied, A. (2002), Convex measures of risk and trading constraints, *Finance and Stochastics*, 6, 429–447.
- Frey, R. and McNeil, A. J. (2001), Modelling dependent defaults, ETH E-Collection, No. 273, ETH Zurich.
- Frey, R., McNeil, A. J., and Nyfeler, M. (2001), Copulas and Credit Models, *Risk*, 14(10), 111–114.
- Frye, J. (2000), Depressing recoveries.
- Genest, C. and Rivest, L. (1993), Statistical inference procedures for bivariate Archimedean copulas, *Journal of the American Statistical Association*, 88, 1034–1043.
- Genest, C., Ghoudi, K., and Rivest, L. (1995), A semi-parametric estimation procedure of dependence parameters in multivariate families of distributions, *Biometrika*, 82, 543–552.

- 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.
- Glasserman, P. and Li, J. (2005), Importance sampling for portfolio credit risk, *Management Science*, 51, 1643–1656.
- 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.
- Hamilton, R., Varma, S., Ou, S., and Cantor, R. (2005), Default and Recovery Rates of Corporate CBond Issuers, Technical report, Moody's Investor Service.

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 Scherer, M. (2011), CDO pricing with nested Archimedean copulas, *Quantitative Finance*, 11(5), 775–787, doi:<http://dx.doi.org/10.1080/14697680903508479>.
- Huang, J.-Z., Lanfranconi, M., Patel, N., and Pospisil, L. (2012), Modeling credit correlations: an overview of the Moody's Analytics GCorr Model, tech. rep., Moody's Analytics.
- Jarrow, R. A. and Turnbull, S. M. (1995), Pricing Derivatives on Financial Securities Subject to Credit Risk, *The Journal of Finance*, 50(1), 53–85.
- 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.
- 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.
- 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).

- Lando, D. and Skodeberg, T. (2002), Analyzing Rating Transitions and Rating Drift with Continuous Observations, *Journal of Banking and Finance*, 26, 423–444.
- Leadbetter, M. R. (1991), On a basis for Peaks over Threshold modeling, *Statistics and 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. (2003), Kendall's tau for elliptical distributions, *Credit Risk: Measurement, Evaluation and Management*, ed. by G. Bol et al., Heidelberg: Physica-Verlag (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, 834–841.
- McNeil, A. J. and Nešlehová, J. (2009), Multivariate Archimedean copulas, d -monotone functions and ℓ_1 -norm symmetric distributions, *Annals of Statistics*, 37(5b), 3059–3097.

- McNeil, A. J., Frey, R., and Embrechts, P. (2005), Quantitative Risk Management: Concepts, Techniques and Tools, Princeton: Princeton University Press.
- McNeil, A. J., Frey, R., and Embrechts, P. (2015), Quantitative Risk Management: Concepts, Techniques and Tools, 2nd, Princeton: Princeton University Press.
- Ou, S. (2013), Annual Default Study: Corporate Default and recovery Rates, 1920-2012, Technical document, Moody's Investor Service.
- Petrov, V. V. (1995), Limit Theorems of Probability Theory, Oxford: Oxford University Press.
- 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.
- 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/201008shreve.html.
- Sibuya, M. (1960), Bivariate extreme statistics, *Annals of the Institute of Statistical Mathematics*, 11, 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.
- Sun, Z., Munves, D., and Hamilton, D. (2012), Public Firm Expected Default Frequency (EDF) Credit Measures: Methodology, Performance and Model Extensions, Technical document, Moody's Analytics.
- 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.

A Appendix

- A.1 Risk in perspective
- A.2 Basics concepts in risk management
- A.3 Empirical properties of financial data
- A.4 Financial time series
- A.5 Extreme value theory
- A.6 Multivariate models
- A.7 Copulas and dependence
- A.8 Aggregate risk

A.1 Risk in perspective

Background information

- A *bond* is an instrument of indebtedness. The 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).
- *Netting* refers to the compensation of long versus short positions on the same underlying.
- A *derivative* is a financial instrument *derived from an underlying asset*, e.g. stocks, bonds, commodities, currencies, interest rates etc. Examples:
 - ▶ *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).
- A *credit default swap (CDS)* is a credit derivative which allows the (protection) buyer (who pays premiums) to transfer credit risk inherent in a reference entity to a seller (investor; pays in case of default).
- A *CDS spread* is the annual 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.
- The *Fundamental Theorems of Asset Pricing*:
 - ▶ A (model for) a market is *arbitrage free* if and only if there exists a risk-neutral probability measure Q equivalent to \mathbb{P} ;
 - ▶ A market is *complete* (i.e. every contingent claim can be replicated) if and only if Q is unique.

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.

A.2 Basics concepts in risk management

Background information

- A *balance sheet* is a financial statement showing *assets* (investments) and *liabilities* (obligations; show how funds have been raised)
- (X_t) is a (discrete) *martingale* with respect to the filtration (\mathcal{F}_t) if
 - ▶ $X_t \in \mathcal{F}_t$ for all $t \in \mathbb{N}_0$ (*adapted*);
 - ▶ $\mathbb{E}X_t < \infty$ for all $t \in \mathbb{N}_0$;
 - ▶ $\mathbb{E}(X_{t+1} | \mathcal{F}_t) = X_t$ for all $t \in \mathbb{N}_0$.

Physical (\mathbb{P}) vs risk-neutral (\mathbb{Q}) measure: An example

- 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**.

- Here, \mathbb{Q} is determined by specifying a q such that

$$\frac{1}{1+r}(1 \cdot (1 - q) + R \cdot q) = V_0.$$

This implies $q = 0.03$ which is greater than $p = 0.01$; the larger value reflects the risk premium.

Expected shortfall under continuity

If F_L is continuous, we know that

$$\begin{aligned} F_{L|L > \text{VaR}_\alpha(L)}(x) &= \mathbb{P}(L \leq x | L > \text{VaR}_\alpha(L)) \\ &= \frac{\mathbb{P}(L \leq x, L > \text{VaR}_\alpha(L))}{\mathbb{P}(L > \text{VaR}_\alpha(L))} = \frac{\mathbb{P}(\text{VaR}_\alpha(L) < L \leq x)}{\mathbb{P}(L > \text{VaR}_\alpha(L))} \\ &= \frac{F_L(x) - F_L(\text{VaR}_\alpha(L))}{1 - F_L(\text{VaR}_\alpha(L))} I_{\{x \geq \text{VaR}_\alpha(L)\}} \\ &\stackrel{(GI4)}{=} \frac{F_L(x) - \alpha}{1 - \alpha} I_{\{x \geq \text{VaR}_\alpha(L)\}} \quad \text{for all } \alpha \in (0, 1) \end{aligned}$$

and thus, since $dF_{L|L > \text{VaR}_\alpha(L)}(x) = dF_L(x)/(1 - \alpha)$,

$$\begin{aligned} \mathbb{E}(L | L > \text{VaR}_\alpha(L)) &= \int_{\text{VaR}_\alpha(L)}^{\infty} x dF_{L|L > \text{VaR}_\alpha(L)}(x) = \frac{1}{1 - \alpha} \int_{\text{VaR}_\alpha(L)}^{\infty} x dF_L(x) \\ &= \frac{1}{1 - \alpha} \int_{\alpha}^1 \text{VaR}_z(L) dz = \text{ES}_\alpha(L), \end{aligned}$$

where we used the substitution $x = \text{VaR}_z(L)$ ($F_L(x) = z$, $dF_L(x) = dz$).

Elicitability explained in words

We follow Kou and Peng (2014, Sections 1 and 2.2) and McNeil et al. (2005, Chapter 9).

- Computing a (one-period ahead) risk measure $\varrho(L) =: \varrho(F_L)$ is a point forecasting problem because F_L is unknown and one has to find an estimate \hat{F}_L of it and forecast the unknown true $\varrho(F_L)$ via the point forecast $\varrho(\hat{F}_L)$.
- As different \hat{F}_L can be used to forecast the risk measure, it is desirable to be able to evaluate which of them gives a better point forecast.
- Suppose we want to forecast L (or F_L) by a point y . The *forecasting error* is

$$\mathbb{E}(S(y, L)) = \int_{\mathbb{R}} S(y, l) dF_L(l),$$

where $S(y, l)$ is a *scoring* (i.e. forecasting objective) function.

- Two point forecasting methods can be compared via their forecasting errors. For a given S , the **optimal point forecast** is

$$\varrho^*(F_L) = \operatorname{arginf}_y \mathbb{E}(S(y, L)) \quad (\text{minimizing the forecast error}).$$

For example, for $S(y, l) = (y - l)^2$ and $S(y, l) = |y - l|$, the optimal point forecasts are the mean and median of F_L , respectively.

- **Elicitable risk measures** (or: statistical functionals) are risk measures ϱ which minimize $\mathbb{E}(S(y, L))$ of some scoring function S ; hence that S can be used to compare different point forecasting procedures for ϱ (“the smaller the forecasting error, the better” makes sense).
- If ϱ is not elicitable, one cannot find such an S and thus the minimization of the forecasting error does not yield the true value $\varrho(F_L)$ for any S . Hence, for two competing point forecasts of $\varrho(F_L)$, one cannot tell which performs the best by comparing their forecasting error, no matter what S is used.

The (nonparametric) bootstrap

- Suppose $X_1, \dots, X_n \stackrel{\text{ind.}}{\sim} F$ (F unknown) and we are interested in estimating $\theta = \theta(F)$. If we could, we would sample from F and estimate θ by $\hat{\theta}_n = \theta(\hat{F}_n)$. But even then, we would only get a point estimate $\hat{\theta}_n$ of θ ; what is the distribution of $\hat{\theta}_n$, what is $\text{var}(\hat{\theta}_n)$?
- The (nonparametric) bootstrap treats \hat{F}_n as the true df F (first approximation error) and samples from that (i.e. resamples X_1, \dots, X_n) to approximate the distribution of $\hat{\theta}_n$ (2nd approximation error).

Algorithm A.1 (Nonparametric bootstrap)

- 1) Fix a large $B \in \mathbb{N}$.
- 2) For $b \in \{1, \dots, B\}$, do:
 - 2.1) Randomly sample $X_{b,1}, \dots, X_{b,n}$ from X_1, \dots, X_n with replacement.

- 2.2) Compute the bootstrap estimator $\hat{\theta}_{b,n} = \theta(\hat{F}_{b,n})$ where $\hat{F}_{b,n}(x) = \frac{1}{n} \sum_{i=1}^n I_{\{X_{b,i} \leq x\}}$ (informally: compute $\hat{\theta}_{b,n}$ from $X_{b,1}, \dots, X_{b,n}$).
- 3) Use the *bootstrap sample* $\hat{\theta}_{b,n}$, $b \in \{1, \dots, B\}$, to approximate the distribution of $\hat{\theta}_n$.
- Examples for Step 3):
 - ▶ $\hat{\mu}_{B,n} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_{b,n}$ approximates $\mathbb{E}(\hat{\theta}_n)$ (similarly for $\text{var}(\hat{\theta}_n)$; the empirical df based on the $\hat{\theta}_{b,n}$'s approximates the df of $\hat{\theta}_n$).
 - ▶ An approximate/bootstrapped *$(1 - \beta)$ -confidence interval* for θ is
- $$\left[\hat{\theta}_{\left(\lceil \frac{\beta}{2} B \rceil\right),n}, \hat{\theta}_{\left(\lceil (1 - \frac{\beta}{2}) B \rceil\right),n} \right]$$
- where $\hat{\theta}_{(1),n} \leq \dots \leq \hat{\theta}_{(B),n}$.
- Advantages: Applicable if F or the df of θ is *unknown*; applicable if n is *small* (unlike the CLT); applicable if F is *skewed* (CIs based on the CLT are always centered around the sample mean); *easy to implement*.

A.3 Empirical properties of financial data

Non-normality and heavy tails

Justification for P-P and Q-Q plots:

- 1) Glivenko–Cantelli: $\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow[n \uparrow \infty]{\text{a.s.}} 0$
- 2) $\hat{F}_n(x) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} F(x) \quad \forall x \in C(F) \Leftrightarrow \hat{F}_n^\leftarrow(u) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} F^\leftarrow(u) \quad \forall u \in C(F^\leftarrow)$;
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 , $x_{(i)} = \hat{F}_n^\leftarrow(i/n) \approx \hat{F}_n^\leftarrow(p_i) \approx F^\leftarrow(p_i)$ (a justification for both P-P and Q-Q plots).

A.4 Financial time series

Conditional expectations

Definition A.2 (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) $\mathbb{E}(\mathbf{Y} I_G) = \int_G \mathbf{Y} d\mathbb{P} = \int_G \mathbf{X} d\mathbb{P} = \mathbb{E}(\mathbf{X} I_G)$ for all $G \in \mathcal{G}$

is called *conditional expectation of \mathbf{X} given \mathcal{G}* and denoted by $\mathbb{E}(\mathbf{X} | \mathcal{G})$.

$\mathbb{P}(A | \mathcal{G}) = \mathbb{E}(I_A | \mathcal{G})$ is called *conditional probability of A given \mathcal{G}* .

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

Lemma A.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 A.2 Part 3) to $\mathbb{E}(\mathbb{E}(X | \mathcal{G}) | \mathcal{F})$ and then to $\mathbb{E}(X | \mathcal{G})$ implies that $\mathbb{E}(\mathbb{E}[\mathbb{E}[X | \mathcal{G}] | \mathcal{F}]I_G) = \mathbb{E}(\mathbb{E}[X | \mathcal{G}]I_G) = \mathbb{E}(XI_G)$. \square

On partial autocorrelation in stationary time series

For introducing it, we need some tools.

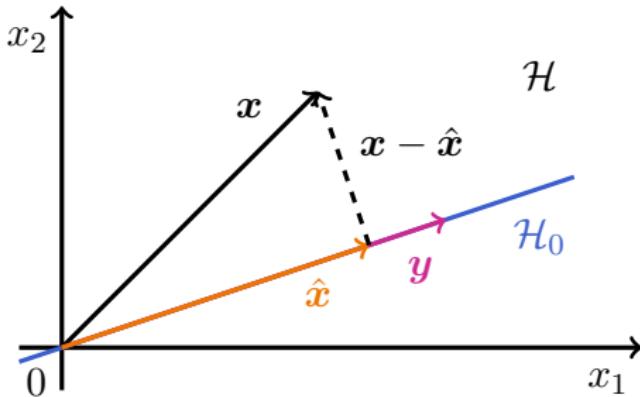
- *Hilbert's Projection Theorem* (see Brockwell and Davis (1991, p. 51)):
If \mathcal{H}_0 is a closed subspace of the Hilbert space \mathcal{H} and $x \in \mathcal{H}$, then:
 - i) There exists a unique $\hat{x} \in \mathcal{H}_0$: $\|x - \hat{x}\| = \inf_{y \in \mathcal{H}_0} \|x - y\|$;
 - ii) $\hat{x} \in \mathcal{H}_0$, $\|x - \hat{x}\| = \inf_{y \in \mathcal{H}_0} \|x - y\|$ 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 A.4

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



- **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

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

$$\begin{aligned}
&\Leftrightarrow \underbrace{\langle X_n - \hat{X}_n, X_l \rangle}_{= \mathbb{E}((X_n - \sum_{k=1}^{n-1} \phi_{n-1,k} X_{n-k}) X_l)} = 0, \quad \forall l \in \{1, \dots, n-1\} \\
&= \mathbb{E}(X_n X_l) - \sum_{k=1}^{n-1} \phi_{n-1,k} \mathbb{E}(X_{n-k} X_l) \\
&\Leftrightarrow \gamma(n-l) = \sum_{k=1}^{n-1} \gamma(n-k-l) \phi_{n-1,k} \\
&\Leftrightarrow \underset{\text{station.}}{\gamma(h)} = \sum_{k=1}^{n-1} \gamma(h-k) \phi_{n-1,k}, \quad \forall h \in \{1, \dots, n-1\} \\
&\Leftrightarrow \Gamma_{n-1} \phi_{n-1} = \gamma_{n-1}, \quad (\text{Yule-Walker equations})
\end{aligned}$$

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

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

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. $\phi_{n,n}$. \square

Definition A.5 (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

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

- PACF for MA(1): Let $\theta = \theta_1$.

$$\rho(h) = \begin{cases} 1, & \text{if } h = 0, \\ \frac{\theta}{1+\theta^2}, & \text{if } |h| = 1, \\ 0, & \text{if } |h| > 1. \end{cases}$$

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.

- PACF for AR(p): 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,

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

Proof idea of Theorem 4.10.

“ \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 4.9 $\Rightarrow \varepsilon_t/\phi(B)$ 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 $|z| \leq 1$. □

Properties of (Q)MLEs

- We consider two situations: The model which has been fitted...
 - 1) ... 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 iid 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 $\text{GARCH}(p, q)$ model with Gaussian innovations,

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

where

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

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

$$\bar{I}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^n \left(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)' \right) \quad \text{and} \quad \bar{J}(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{t=1}^n \frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2},$$

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.

The SARIMA model

$(X_t)_{t \in \mathbb{Z}}$ is a $\text{SARIMA}(p, d, q) \times (\tilde{p}, \tilde{d}, \tilde{q})_s$ (Seasonal; Integrated) 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.

A.5 Extreme value theory

The convergence to types theorem

Theorem A.6 (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). □

The Gumbel MDA

Theorem A.7 (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 $d_n = F^\leftarrow(1 - 1/n)$, $n \in \mathbb{N}$.

Derivation of the Hill estimator

Let e be the mean excess function for $\log X$. Using partial integration ($\int H dG = [HG] - \int G dH$), we obtain

$$\begin{aligned} e(\log u) &= \mathbb{E}(\log X - \log u \mid \log X > \log u) \\ &= \frac{1}{\bar{F}(u)} \int_u^\infty (\log x - \log u) dF(x) = -\frac{1}{\bar{F}(u)} \int_u^\infty \log\left(\frac{x}{u}\right) d\bar{F}(x) \\ &= -\frac{1}{\bar{F}(u)} \left(\underbrace{\left[\log\left(\frac{x}{u}\right) \bar{F}(x) \right]_u^\infty}_{=0} - \int_u^\infty \bar{F}(x) \frac{1}{x} dx \right) \\ &= \frac{1}{\bar{F}(u)} \int_u^\infty \frac{\bar{F}(x)}{x} dx = \frac{1}{\bar{F}(u)} \int_u^\infty x^{-\alpha-1} L(x) dx. \end{aligned}$$

For u sufficiently large, $L(x) \approx L(u)$, $x \geq u$ (by 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 and k sufficiently small, replace $e(\cdot)$ by $e_n(\cdot)$ and use $u = X_{k,n}$. We obtain that

$$\begin{aligned}\frac{1}{\alpha} \approx e_n(\log X_{k,n}) &= \frac{\sum_{i=1}^n (\log X_i - \log X_{k,n}) I_{\{\log X_i > \log X_{k,n}\}}}{\sum_{i=1}^n I_{\{\log X_i > \log X_{k,n}\}}} \\ &= \frac{\sum_{i=1}^{k-1} (\log X_{i,n} - \log X_{k,n})}{k-1} = \frac{1}{k-1} \sum_{i=1}^{k-1} \log X_{i,n} - \log X_{k,n}\end{aligned}$$

The standard form of the estimator is typically written with the average taken over the largest k (instead of $k-1$) terms.

Non-iid 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 iid generalized Pareto distributed.
- If extremal clustering is present ($\theta < 1$; e.g. GARCH 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 the following section for more details on threshold exceedances.

Point process models

So far: loss size distribution. Now: loss frequency distribution

Threshold exceedances for strict white noise

- Consider a strict white noise $(X_i)_{i \in \mathbb{N}}$ (iid 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.

1) 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

$$N(A) = \sum_{i=1}^n I_{\{Y_i \in A\}}.$$

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 $\Lambda(A) = \int_A \lambda(x) dx$.

2) Asymptotic behaviour of the point process of exceedances

- For $n \in \mathbb{N}$ and $i \in \{1, \dots, n\}$ let $Y_{i,n} = \frac{i}{n} I_{\{X_i > u_n(x)\}}$. The **point process of exceedances over u_n** is the process $N_n(\cdot)$ with state space $\mathcal{X} = (0, 1]$ given by

$$N_n(A) = \sum_{i=1}^n I_{\{Y_{i,n} \in A\}}, \quad A \subseteq \mathcal{X}.$$

- 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, $\mathbb{E}N_n(A) \xrightarrow{n \uparrow \infty} \mathbb{E}N(A) = \Lambda(A) = (t_2 - t_1)\lambda(x)$. 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 .

3) 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 iid data over u are separated by iid exponential waiting times.

The POT model

- Putting the pieces together, we obtain an asymptotic model for threshold exceedances in regularly spaced iid 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 *iid* 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))

1) 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 $N(A) = \sum_{i=1}^n I_{\{(i/n, X_i) \in A\}}$ 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.

2) 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

$$\underbrace{N_u!}_{\substack{\text{ordered} \\ \text{sample prob. of } N_u \text{ samples}}} \underbrace{\frac{e^{-\Lambda(A)} \Lambda(A)^{N_u}}{N_u!}}_{\substack{\text{prob. of } N_u \text{ samples} \\ \text{density of } \tilde{X}_i}} \prod_{i=1}^{N_u} \underbrace{\frac{\lambda(\tilde{X}_i)}{\Lambda(A)}}_{\text{density of } \tilde{X}_i} = e^{-\Lambda(A)} \prod_{i=1}^{N_u} \lambda(\tilde{X}_i) = e^{-\tau_u} \prod_{i=1}^{N_u} \lambda(\tilde{X}_i).$$

- 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

$$\begin{aligned}&= -\tau_u + \sum_{i=1}^{N_u} \log \lambda(\tilde{X}_i) = -\tau_u + N_u \log \tau_u + \overbrace{\sum_{i=1}^{N_u} (\log \lambda(\tilde{X}_i) - \log \tau_u)}^{\text{= } \log g_{\xi, \beta}(\tilde{X}_i - u)} \\ &= \ell_{\text{Poi}}(\tau_u; N_u) - N_u \log(T) + \ell_{\text{GPD}}(\xi, \beta; \tilde{X}_1 - u, \dots, \tilde{X}_{N_u} - u),\end{aligned}\tag{32}$$

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 $\sigma = \tau_u^\xi \beta$.

3) Advantages of the POT model formulation

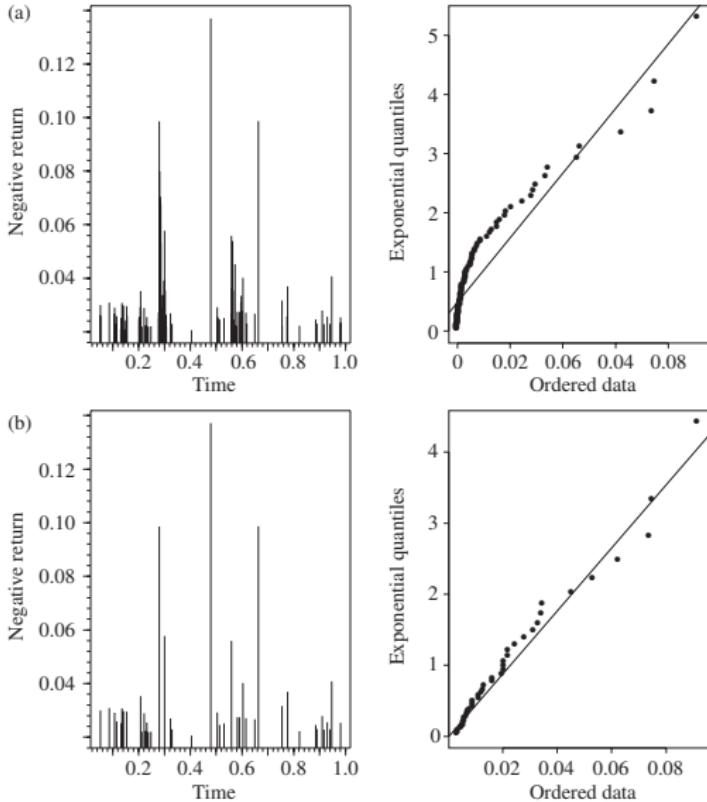
- One advantage of the two-dimensional Poisson point process model is that ξ , μ and σ do not depend on u (unlike β in the GPD model).
 - ⇒ In practice, we would expect the estimated parameters of the Poisson model to be roughly stable over a range of high thresholds.
- The intensity λ is thus often 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. (2014).

4) 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.

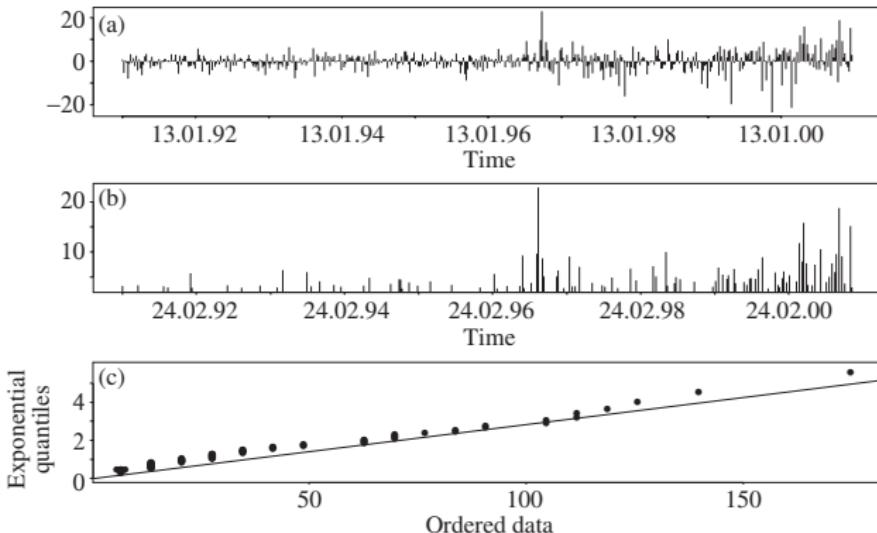
- *Declustering* may circumvent the problem. One identifies clusters (ad hoc; not easy) of exceedances and then applies the POT model to cluster maxima only.
- A possible declustering algorithm is the *runs method*. A run size r is fixed and two successive exceedances are said to belong to two different 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 modelling of cluster formation, we cannot make more dynamic statements about the intensity of occurrence of exceedances.

Example A.8 (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 (which contradicts a homogeneous Poisson process)

- Using the log-likelihood (32), 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 5.19.
- 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.

A.6 Multivariate models

Conditional distributions and independence

Proof of (14). We have

$$\begin{aligned} & \int_{(-\infty, x_1]} F_{X_2|X_1}(x_2 | z) dF_{X_1}(z) \\ &= \int_{\mathbb{R}^d} I_{\{z \leq x_1\}} \mathbb{E}(I_{\{X_2 \leq x_2\}} | X_1 = z) dF_{X_1}(z) \\ &= \mathbb{E}(I_{\{X_1 \leq x_1\}} \mathbb{E}(I_{\{X_2 \leq x_2\}} | X_1)) = \mathbb{E}(\mathbb{E}(I_{\{X_1 \leq x_1, X_2 \leq x_2\}} | X_1)) \\ &\stackrel{\substack{\text{tower} \\ \text{property}}}{=} \mathbb{E}(I_{\{X_1 \leq x_1, X_2 \leq x_2\}}) = F(x), \end{aligned}$$

where the second-last equality holds by the **tower property** of conditional expectations. □

The multivariate normal distribution

Proof of the form of the cf of $N(0, 1)$; see Proposition 6.3. The rv $Z \sim N(0, 1)$ 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 $\phi_Z(t)$ of Z is given by

$$\phi_Z(t) = \int_{-\infty}^{\infty} (\cos(tx) + i \sin(tx)) \varphi(x) dx = \int_{-\infty}^{\infty} \cos(tx) \varphi(x) dx.$$

Hence,

$$\phi'_Z(t) = \int_{-\infty}^{\infty} \sin(tx)(-x) \varphi(x) dx \stackrel{\text{ii)}}{=} \int_{-\infty}^{\infty} \sin(tx) \varphi'(x) dx = -t\phi_Z(t).$$

We also know that $\phi_Z(0) = 1$. This initial value problem has the unique solution $\phi_Z(t) = \exp(-t^2/2)$. □

Theorem A.9 (Cramér–Wold)

Let \mathbf{X}, \mathbf{X}_n , $n \in \mathbb{N}$, be random vectors. Then

$$\mathbf{X}_n \xrightarrow[n \uparrow \infty]{d} \mathbf{X} \iff \mathbf{a}' \mathbf{X}_n \xrightarrow[n \uparrow \infty]{d} \mathbf{a}' \mathbf{X} \quad \forall \mathbf{a} \in \mathbb{R}^d$$

Proof.

" \Rightarrow " This follows from the Continuous Mapping Theorem with the map $g(\mathbf{x}) = \mathbf{a}' \mathbf{x}$.

" \Leftarrow " Note that $\phi_{\mathbf{X}_n}(\mathbf{t}) = \mathbb{E}(\exp(i \cdot \mathbf{1} \cdot \mathbf{t}' \mathbf{X}_n)) = \phi_{\mathbf{t}' \mathbf{X}_n}(1) \xrightarrow[n \uparrow \infty]{} \phi_{\mathbf{t}' \mathbf{X}}(1) = \phi_{\mathbf{X}}(\mathbf{t})$ for all \mathbf{t} and apply the Lévy Continuity Theorem. \square

Corollary A.10

Let \mathbf{X}, \mathbf{Y} be two random vectors. Then

$$\mathbf{X} \stackrel{d}{=} \mathbf{Y} \iff \mathbf{a}' \mathbf{X} \stackrel{d}{=} \mathbf{a}' \mathbf{Y} \quad \forall \mathbf{a} \in \mathbb{R}^d.$$

Properties of multivariate normal variance mixtures

Proof of Lemma 6.10. W.l.o.g. assume $\mu = \mathbf{0}$.

$$\begin{aligned} \Rightarrow \mathbb{E}|X_i| \mathbb{E}|X_j| &\stackrel{\text{ind.}}{=} \mathbb{E}(|X_i||X_j|) = \mathbb{E}(W|Z_i||Z_j|) \stackrel{\text{ind.}}{=} \mathbb{E}(W) \mathbb{E}|Z_i| \mathbb{E}|Z_j| \\ &\geq \mathbb{E}(\sqrt{W})^2 \mathbb{E}|Z_i| \mathbb{E}|Z_j| \stackrel{\text{ind.}}{=} \mathbb{E}|\sqrt{W}Z_i| \mathbb{E}|\sqrt{W}Z_j| = \mathbb{E}|X_i| \mathbb{E}|X_j| \end{aligned}$$

Jensen

⇒ 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. □

Spherical distributions

Proof of Theorem 6.15.

1) \Rightarrow 2): $\phi_{\mathbf{Y}}(\mathbf{t}) = \phi_{U\mathbf{Y}}(\mathbf{t}) = \phi_{\mathbf{Y}}(U'\mathbf{t})$ 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} ⇒ 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 $\phi_{\mathbf{a}'\mathbf{Y}}(t) = \phi_{\mathbf{Y}}(t\mathbf{a}) \stackrel{2)}{=}$

$$\psi(t^2\|\mathbf{a}\|^2) = \psi((t\|\mathbf{a}\|)^2) \stackrel{(*)}{=} \phi_{Y_1}(t\|\mathbf{a}\|) = \phi_{\|\mathbf{a}\|Y_1}(t)$$

$3) \Rightarrow 1)$: $\phi_{U\mathbf{Y}}(\mathbf{t}) = \mathbb{E}(\exp(i(U'\mathbf{t})'\mathbf{Y})) \stackrel{U'\mathbf{t}:=\mathbf{a}}{=} \mathbb{E}(\exp(i\mathbf{a}'\mathbf{Y})) \stackrel{3)}{=} \mathbb{E}(\exp(i\|\mathbf{a}\|Y_1))$
 $= \mathbb{E}(\exp(i\|\mathbf{t}\|Y_1)) \stackrel{3)}{=} \mathbb{E}(\exp(i\mathbf{t}'\mathbf{Y})) = \phi_{\mathbf{Y}}(\mathbf{t})$ \square

Proof of Theorem 6.16. Let Ω_d be the characteristic generator of S .

“ \Rightarrow ” $\mathbf{Y} \sim S_d(\psi) \Rightarrow \phi_{\mathbf{Y}}(\|\mathbf{t}\|\mathbf{u}) \stackrel{2)}{=} \psi(\|\mathbf{t}\|^2\mathbf{u}'\mathbf{u}) = \psi(\|\mathbf{t}\|^2)$ for all $\mathbf{u} \in \mathbb{R}^d$:
 $\|\mathbf{u}\| = 1$. Replacing \mathbf{u} by \mathbf{S} and integrating leads to $\psi(\|\mathbf{t}\|^2) = \mathbb{E}_{\mathbf{S}}(\phi_{\mathbf{Y}}(\|\mathbf{t}\|\mathbf{S})) = \mathbb{E}_{\mathbf{S}}(\mathbb{E}_{\mathbf{Y}}(e^{i\|\mathbf{t}\|S'\mathbf{Y}})) \stackrel{\text{Fubini}}{=} \mathbb{E}_{\mathbf{Y}}(\mathbb{E}_{\mathbf{S}}(e^{i\|\mathbf{t}\|S'\mathbf{Y}})) = \mathbb{E}_{\mathbf{Y}}(\phi_{\mathbf{S}}(\|\mathbf{t}\|\mathbf{Y})) \stackrel{2)}{=} \mathbb{E}_{\mathbf{Y}}(\Omega_d(\|\mathbf{t}\|^2\mathbf{Y}'\mathbf{Y}))$. We thus obtain that

$$\begin{aligned} \phi_{\mathbf{Y}}(\mathbf{t}) &\stackrel{2)}{=} \psi(\|\mathbf{t}\|^2) \stackrel{\mathbf{R}:=\|\mathbf{Y}\|}{=} \mathbb{E}_R(\Omega_d(\|\mathbf{t}\|^2R^2)) = \int_0^\infty \Omega_d(\|\mathbf{t}\|^2r^2) dF_R(r) \\ &\stackrel{2)}{=} \int_0^\infty \phi_{\mathbf{S}}(r\mathbf{t}) dF_R(r) = \phi_{R\mathbf{S}}(\mathbf{t}) \text{ for all } \mathbf{t} \in \mathbb{R}^d. \end{aligned}$$

“ \Leftarrow ” Let $\mathbf{Z} \sim N_d(\mathbf{0}, I_d)$. Since \mathbf{Z} is spherical and $\|\mathbf{Z}/\|\mathbf{Z}\|\| = \|\mathbf{Z}\|/\|\mathbf{Z}\| =$

$1, \mathbf{S} \stackrel{\text{d}}{=} \mathbf{Z}/\|\mathbf{Z}\|$. As such, \mathbf{S} itself is spherical, since $U\mathbf{S} \stackrel{\text{d}}{=} U\mathbf{Z}/\|\mathbf{Z}\| \stackrel{\text{d}}{=} \mathbf{Z}/\|\mathbf{Z}\| \stackrel{\text{d}}{=} \mathbf{S}$ for any orthogonal $U \in \mathbb{R}^{d \times d}$. Theorem 6.15 Part 2) implies that $\phi_{\mathbf{S}}(\mathbf{t}) = \Omega_d(\|\mathbf{t}\|^2)$, so $\phi_{R\mathbf{S}}(\mathbf{t}) = \mathbb{E}(\exp(i\mathbf{t}' R\mathbf{S})) = \mathbb{E}_R(\mathbb{E}(\exp(i\mathbf{t}' R\mathbf{S}) | R)) = \mathbb{E}_R(\phi_{\mathbf{S}}(R\mathbf{t})) = \mathbb{E}_R(\Omega_d(R^2\|\mathbf{t}\|^2))$, which is a function in $\|\mathbf{t}\|^2$ and thus, by 2), $R\mathbf{S}$ is spherical. \square

Density of $\mathbf{Y} \sim S_d(\psi)$ constant on spheres

If \mathbf{Y} admits a density $f_{\mathbf{Y}}$, then $f_{\mathbf{Y}}(\mathbf{y})$ is constant on hyperspheres in \mathbb{R}^d . The *inversion formula* $f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\mathbf{t}' \mathbf{y}} \phi_{\mathbf{Y}}(\mathbf{t}) d\mathbf{t}$ and Theorem 6.15 Part 2) show that for any orthogonal U ,

$$\begin{aligned} f_{\mathbf{Y}}(U\mathbf{y}) &\stackrel{\text{inv.}}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i(U'\mathbf{t})' \mathbf{y}} \phi_{\mathbf{Y}}(\mathbf{t}) d\mathbf{t} \\ &\stackrel{\text{subs.}}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-is' \mathbf{y}} \phi_{\mathbf{Y}}(Us) ds \\ &\stackrel{2)}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-is' \mathbf{y}} \psi((Us)' Us) ds \end{aligned}$$

$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-is' \mathbf{y}} \psi(s' s) ds \underset{\text{backwards}}{=} \cdots = f_{\mathbf{Y}}(\mathbf{y}).$$

This implies that $f_{\mathbf{Y}}(\mathbf{y}) = g(\|\mathbf{y}\|^2)$ for a function $g : [0, \infty) \rightarrow [0, \infty)$, the density generator. For $\mathbf{Y} \sim t_d(\nu, \mathbf{0}, I_d)$, one can show that $g(x) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)(\pi\nu)^{d/2}} (1 + \frac{x}{\nu})^{-(\nu+d)/2}$.

Elliptical distributions

Proposition A.11

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 $\mathbb{P}((\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \geq c) > 0$,

$$\text{corr}(\mathbf{X} | (\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \geq c) = \text{corr}(\mathbf{X}).$$

Proof. $\mathbf{X} | ((\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \geq c) \stackrel{\text{d}}{=} \boldsymbol{\mu} + RAS | (R^2 \geq c) \stackrel{\text{ind.}}{=} \boldsymbol{\mu} + \tilde{R}AS$ (22)
where $\tilde{R} \stackrel{\text{d}}{=} (R | R^2 \geq c)$. A (and thus Σ) remains the same. \square

Estimating scale and correlation

- Suppose $\mathbf{X}_1, \dots, \mathbf{X}_n \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$. How can we estimate $\boldsymbol{\mu}$, Σ 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 $D_i = \sqrt{(\mathbf{X}_i - \hat{\boldsymbol{\mu}})' \hat{\Sigma}^{-1} (\mathbf{X}_i - \hat{\boldsymbol{\mu}})}$ (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 A.12 (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 $D_i = \sqrt{(\mathbf{X}_i - \hat{\mu}^{[k]})' \hat{\Sigma}^{[k]-1} (\mathbf{X}_i - \hat{\mu}^{[k]})}$.

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 $I_{x \leq a} + (a/x)I_{x > a}$ for some value a).

2.3) Update:

$$\hat{\Sigma}^{[k+1]} = \frac{1}{n} \sum_{i=1}^n w_2(D_i^2) (\mathbf{X}_i - \hat{\mu}^{[k]}) (\mathbf{X}_i - \hat{\mu}^{[k]})',$$

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 $k + 1$.

Factor models

Example A.13 (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 $\Sigma = BB' + \Upsilon$ 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}

$$F = \frac{\sqrt{\rho}}{1 + \rho(d - 1)} \sum_{j=1}^d \mathbf{X}_j + \sqrt{\frac{1 - \rho}{1 + \rho(d - 1)}} Y, \quad \varepsilon_j = X_j - \sqrt{\rho}F.$$

We have $\mathbb{E}(F) = 0$, $\text{var}(F) = 1$, so $\mathbf{X} = \mathbf{0} + BF + \varepsilon = \sqrt{\rho}\mathbf{1}F + \varepsilon$.

- The requirements of Definition 6.25 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 $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 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 $F, Z_1, \dots, Z_d \stackrel{\text{ind.}}{\sim} N(0, 1)$.

Multivariate regression

- Here, construct large matrices:

$$X = \underbrace{\begin{pmatrix} \mathbf{X}'_1 \\ \vdots \\ \mathbf{X}'_n \end{pmatrix}}_{n \times d}, \quad F = \underbrace{\begin{pmatrix} 1 & \mathbf{F}'_1 \\ \vdots & \vdots \\ 1 & \mathbf{F}'_n \end{pmatrix}}_{n \times (p+1)}, \quad \tilde{B} = \underbrace{\begin{pmatrix} \mathbf{a}' \\ B' \end{pmatrix}}_{(p+1) \times d}, \quad E = \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}'_1 \\ \vdots \\ \boldsymbol{\varepsilon}'_n \end{pmatrix}}_{n \times d}.$$

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 $\hat{\tilde{B}} = (F'F)^{-1}F'X$.

- Now examine the conditions of Definition 6.25: Do the errors vectors ε_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 .

Sample principal components

- Assume $\mathbf{X}_1, \dots, \mathbf{X}_n$ with identical distribution, unknown mean vector μ and covariance matrix Σ with the spectral decomposition $\Sigma = \Gamma \Lambda \Gamma'$ as before.
- Estimate μ by $\bar{\mathbf{X}}$ and Σ by $S_x = \frac{1}{n} \sum_{t=1}^n (\mathbf{X}_t - \bar{\mathbf{X}})(\mathbf{X}_t - \bar{\mathbf{X}})'$.
- Apply the spectral decomposition to S_x to get $S_x = GLG'$, 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” $\mathbf{Y}_t = G'(\mathbf{X}_t - \bar{\mathbf{X}})$, $t \in \{1, \dots, n\}$. The j th component $Y_{t,j} = \mathbf{g}'_j(\mathbf{X}_t - \bar{\mathbf{X}})$ is the *jth sample principal component at time t* (\mathbf{g}_j is the j th column of G).
- The rotated vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ have sample covariance matrix L :

$$\begin{aligned} S_y &= \frac{1}{n} \sum_{t=1}^n (\mathbf{Y}_t - \bar{\mathbf{Y}})(\mathbf{Y}_t - \bar{\mathbf{Y}})' = \frac{1}{n} \sum_{t=1}^n \mathbf{Y}_t \mathbf{Y}_t' \\ &= \frac{1}{n} \sum_{t=1}^n G'(\mathbf{X}_t - \bar{\mathbf{X}})(\mathbf{X}_t - \bar{\mathbf{X}})'G = G'S_xG = L. \end{aligned}$$

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 \mathbf{Y}_t to calibrate an approximate factor model. We assume our data are realizations 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 ε_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.

A.7 Copulas and dependence

An example

Let $C(\mathbf{u}) = \lambda C_1(\mathbf{u}) + (1 - \lambda)C_2(\mathbf{u})$ for copulas C_1, C_2 and $\lambda \in [0, 1]$ (convex combination). Then C is again a copula since:

1) Analytical proof:

- 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.

- 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.

- $\Delta_{(a,b]} C = \lambda \Delta_{(a,b]} C_1 + (1 - \lambda) \Delta_{(a,b]} C_2 \geq 0$, so C is d -increasing.

2) Stochastic proof:

Let $\mathbf{U}_k \sim C_k$, $k \in \{1, 2\}$ and let $X \sim B(1, \lambda)$, 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 that

$$\begin{aligned}\mathbb{P}(\mathbf{U} \leq \mathbf{u}) &= \mathbb{P}(\mathbf{U} \leq \mathbf{u}, X = 1) + \mathbb{P}(\mathbf{U} \leq \mathbf{u}, X = 0) \\ &= \mathbb{P}(\mathbf{U}_1 \leq \mathbf{u}, X = 1) + \mathbb{P}(\mathbf{U}_2 \leq \mathbf{u}, X = 0) \\ &= \mathbb{P}(\mathbf{U}_1 \leq \mathbf{u}) \mathbb{P}(X = 1) + \mathbb{P}(\mathbf{U}_2 \leq \mathbf{u}) \mathbb{P}(X = 0) \\ &= C_1(\mathbf{u}) \lambda + C_2(\mathbf{u}) (1 - \lambda) = C(\mathbf{u}).\end{aligned}$$

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.

Basic properties

Lemma A.14

For any copula C , $|C(\mathbf{b}) - C(\mathbf{a})| \leq \sum_{j=1}^d |b_j - a_j|$ for all $\mathbf{a}, \mathbf{b} \in [0, 1]^d$.

Proof. Using a telescoping sum expansion and the triangle inequality, we obtain

$$\begin{aligned}|C(\mathbf{b}) - C(\mathbf{a})| &\leq \sum_{j=1}^d |C(b_1, \dots, b_{j-1}, b_j, a_{j+1}, \dots, a_d) \\&\quad - C(b_1, \dots, b_{j-1}, a_j, a_{j+1}, \dots, a_d)|.\end{aligned}$$

W.l.o.g. let $\mathbf{a} \leq \mathbf{b}$. By d -increasingness, $C \nearrow$ component-wise, so omit $|\cdot|$. Since, 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

Generalized inverses

$T \nearrow$ 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 A.15 (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) $T^\leftarrow(y) = -\infty$ if and only if $T(x) \geq y$ for all $x \in \mathbb{R}$. Similarly,
 $T^\leftarrow(y) = \infty$ if and only if $T(x) < y$ for all $x \in \mathbb{R}$.

(GI2) $T^\leftarrow \nearrow$. If $T^\leftarrow(y) \in (-\infty, \infty)$, T^\leftarrow is left-continuous at y and
admits a limit from the right at y .

(GI3) $T^\leftarrow(T(x)) \leq x$. If $T \uparrow$, then $T^\leftarrow(T(x)) = x$.

(GI4) Let T be right-continuous and $\text{ran } T$ denote the *range of T* , i.e.
 $\text{ran } T = \{T(x) : x \in \mathbb{R}\}$. $T^\leftarrow(y) < \infty$ implies $T(T^\leftarrow(y)) \geq y$.
Furthermore, $y \in \text{ran } T \cup \{\inf T, \sup T\}$ implies $T(T^\leftarrow(y)) = y$.

Moreover, if $y < \inf T$ then $T(T^\leftarrow(y)) > y$ and if $y > \sup T$ then $T(T^\leftarrow(y)) < y$.

- (GI5) $T(x) \geq y$ implies $x \geq T^\leftarrow(y)$. The other implication holds if T is right-continuous. Furthermore, $T(x) < y$ implies $x \leq T^\leftarrow(y)$.
- (GI6) $(T^\leftarrow(y-), T^\leftarrow(y+)) \subseteq \{x \in \mathbb{R} : T(x) = y\} \subseteq [T^\leftarrow(y-), T^\leftarrow(y+)]$, where $T^\leftarrow(y-) = \lim_{z \uparrow y} T^\leftarrow(z)$ and $T^\leftarrow(y+) = \lim_{z \downarrow y} T^\leftarrow(z)$.
- (GI7) T is continuous if and only if $T^\leftarrow \uparrow$ on $[\inf T, \sup T]$.
 $T \uparrow$ if and only if T^\leftarrow is continuous on $\text{ran } T$.
- (GI8) If T_1 and T_2 are right-continuous transformations with properties as T , then $(T_1 \circ T_2)^\leftarrow = T_2^\leftarrow \circ T_1^\leftarrow$.

Proof. See Embrechts and Hofert (2013a). □

Proof of Lemma 7.2. 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 $F^\leftarrow \uparrow$ on $[\inf F, \sup F] = [0, 1]$. Thus

$$\begin{aligned} \mathbb{P}(F(X) \leq u) &\stackrel{\text{(GI7)}}{=} \mathbb{P}(F^\leftarrow(F(X)) \leq F^\leftarrow(u)) \stackrel{\text{(GI3)}}{=} \mathbb{P}(X \leq F^\leftarrow(u)) \\ &= F(F^\leftarrow(u)) \stackrel{\text{(GI4)}}{=} u, \quad u \in [0, 1], \end{aligned}$$

where (GI3) applies since $F \uparrow$ on $\text{ran } X$. □

Proof of Lemma 7.6.

$$\begin{aligned} \text{"\Rightarrow"} \quad \mathbb{P}(F_j(X_j) \leq u_j \forall j) &\stackrel{\text{cont.}}{=} \mathbb{P}(F_j(X_j) < u_j \forall j) \stackrel{\text{(GI5)}}{=} \mathbb{P}(X_j < F_j^\leftarrow(u_j) \forall j) \\ &\stackrel{\text{cont.}}{=} \mathbb{P}(X_j \leq F_j^\leftarrow(u_j) \forall j) = F(F_1^\leftarrow(u_1), \dots, F_d^\leftarrow(u_d)) \stackrel{\text{Sklar}}{=} C(\mathbf{u}). \end{aligned}$$

“\$\Leftarrow\$” Since $F_j \uparrow$ on $\text{ran } X_j$, $j \in \{1, \dots, d\}$,

$$\begin{aligned} F(\mathbf{x}) &\stackrel{\text{(GI3)}}{=} \mathbb{P}(F_j^\leftarrow(F_j(X_j)) \leq x_j \forall j) \stackrel{\text{(GI5)}}{=} \mathbb{P}(F_j(X_j) \leq F_j(x_j) \forall j) \\ &\stackrel{\text{ass.}}{=} C(F_1(x_1), \dots, F_d(x_d)) \stackrel{\text{Sklar}}{\Rightarrow} \mathbf{X} \text{ has copula } C \quad \square \end{aligned}$$

Proof of Theorem 7.8.

- 1) ■ By Lemma A.14, $1 - C(\mathbf{u}) = C(\mathbf{1}) - C(\mathbf{u}) \leq \sum_{j=1}^d (1 - u_j) = d - \sum_{j=1}^d u_j$, 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

$$\begin{aligned}
 & \Delta_{(\frac{1}{2}, 1]} W \\
 &= \sum_{i \in \{0,1\}^d} (-1)^{\sum_{j=1}^d i_j} W\left(\frac{1}{2}^{i_1}, \dots, \frac{1}{2}^{i_d}\right) \\
 &= \max\{1 + 1 + 1 + \dots + 1 - d + 1, 0\} \quad (i_j = 0 \ \forall j) \\
 &\quad - d \max\{\frac{1}{2} + 1 + 1 + \dots + 1 - d + 1, 0\} \quad (\exists! j : i_j = 1)
 \end{aligned}$$

$$\begin{aligned}
& + \binom{d}{2} \max\left\{\frac{1}{2} + \frac{1}{2} + 1 + \cdots + 1 - d + 1, 0\right\} \quad (\exists! \text{ two } j : i_j = 1) \\
& - \cdots + (-1)^d \max\left\{\frac{1}{2} + \cdots + \frac{1}{2} - d + 1, 0\right\} \quad (i_j = 1 \ \forall j) \\
& = 1 - \frac{d}{2} < 0 \quad \text{for } d \geq 3.
\end{aligned}$$

3) M is a copula for all $d \geq 2$ since $(U, \dots, U) \sim M$ for $U \sim \text{U}(0, 1)$. \square

Extreme value and Marshall–Olkin 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

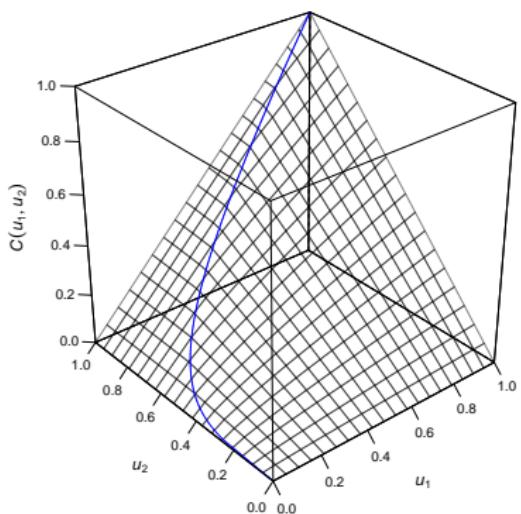
$$C(\mathbf{u}) = \left(\prod_{j=1}^d u_j \right)^{A\left(\frac{\log u_1}{\log \Pi(\mathbf{u})}, \dots, \frac{\log u_d}{\log \Pi(\mathbf{u})}\right)}$$

for a *Pickands dependence function* A ; see Ressel (2013) for a characterization of A .

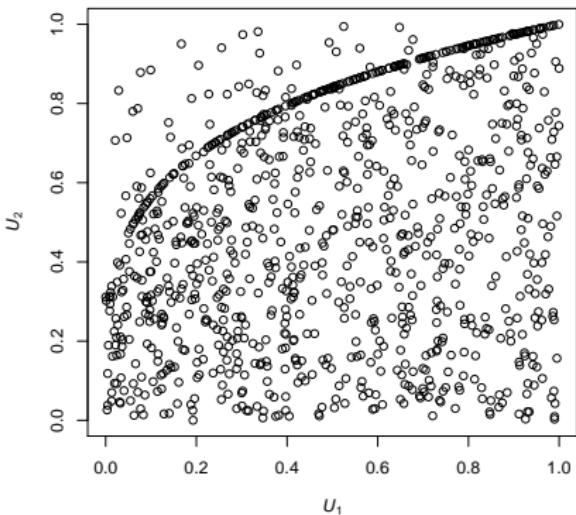
- Examples: Gumbel copula, Marshall-Olkin copulas.
- For more details, see Jaworski et al. (2010, Chapter 6).

Another class of copulas is given by $C(u_1, u_2) = \min\{u_1 u_2^{1-\alpha_2}, u_1^{1-\alpha_1} u_2\}$, $\alpha_1, \alpha_2 \in [0, 1]$. Such copulas are called *Marshall–Olkin copulas* and one of their characteristics is a *singular component* (set of Lebesgue measure 0 in which (U_1, U_2) take values with a positive probability).

MO copula with singular component ($\alpha_1 = 0.2$, $\alpha_2 = 0.8$, $\tau = 0.19$)



Scatter plot MO copula ($n = 1000$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, $\tau = 0.19$)



Perfect dependence

Proof of Proposition 7.14. Consider Part 1); Part 2) works similarly.

“ \Rightarrow ” By assumption, $\mathbb{P}(X_2 \leq x)$ equals $\mathbb{P}(F_2^\leftarrow(1 - F_1(X_1)) \leq x)$ $\stackrel{(GI5)}{=}$ $\mathbb{P}(1 - F_1(X_1) \leq F_2(x)) = F_2(x)$. If (X_1, X_2) has copula C , then

$$\begin{aligned} C(\mathbf{u}) &\stackrel{\text{L.7.6}}{=} \mathbb{P}(F_1(X_1) \leq u_1, F_2(F_2^\leftarrow(1 - F_1(X_1))) \leq u_2) \\ &\stackrel{(GI4)}{=} \mathbb{P}(F_1(X_1) \leq u_1, 1 - F_1(X_1) \leq u_2) \\ &= \mathbb{P}(1 - u_2 < U \leq u_1) = W(u_1, u_2) \quad \text{for } U \sim U(0, 1). \end{aligned}$$

“ \Leftarrow ” $W(u_1, u_2) = 0$ for all $u_1, u_2 \in [0, 1]$ such that $u_1 + u_2 - 1 < 0$, so W puts no mass below the secondary diagonal. Similarly 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 $\mathbb{P}(X_2 = F_2^\leftarrow(1 - F_1(X_1))) = \mathbb{P}(F_2(X_2) = F_2(F_2^\leftarrow(1 - F_1(X_1))))$ $\stackrel{(GI4)}{=} \mathbb{P}(F_2(X_2) = 1 - F_1(X_1)) = \mathbb{P}(U_2 = 1 - U_1) = 1$. \square

Proof of Proposition 7.15. Consider $T(u) = F_1^\leftarrow(u) + \cdots + F_d^\leftarrow(u)$, left-continuous and let $U \sim U(0, 1)$. We first show that $F_{T(U)}^\leftarrow(u) = T(u)$, for all $u \in [0, 1]$.

$$1) \quad T \text{ left-continuous} \Rightarrow T(\textcolor{brown}{u}) \leq x \Leftrightarrow \textcolor{brown}{u} \leq u_x := \sup\{u : T(u) \leq x\}$$

$$2) \quad 1) \Rightarrow \{T(\textcolor{brown}{U}) \leq x\} = \{\textcolor{brown}{U} \leq u_x\} \Rightarrow F_{T(U)}(x) = F_U(u_x) = u_x.$$

$$\Rightarrow F_{T(U)}^\leftarrow(u) \leq \textcolor{brown}{x} \stackrel{(\text{GI5})}{\Leftrightarrow} F_{T(U)}(x) \geq u \stackrel{2)}{\Leftrightarrow} u_x \geq u \stackrel{1)}{\Leftrightarrow} T(u) \leq \textcolor{brown}{x}$$

Choosing $\textcolor{brown}{x} = T(u)$ and $\textcolor{brown}{x} = F_{T(U)}^\leftarrow(u)$ in the last line, we see that $F_{T(U)}^\leftarrow(u) = T(u)$. Now Proposition 7.14 2) implies that

$$(X_1, \dots, X_d) \stackrel{d}{=} (F_1^\leftarrow(U), \dots, F_d^\leftarrow(U)),$$

so that

$$F_{\sum_{j=1}^d X_j}(x) = \mathbb{P}\left(\sum_{j=1}^d X_j \leq x\right) = \mathbb{P}\left(\sum_{j=1}^d F_j^\leftarrow(U) \leq x\right) = \mathbb{P}(T(U) \leq x)$$

$$\text{and thus } F_{\sum_{j=1}^d X_j}^\leftarrow(\alpha) = F_{T(U)}^\leftarrow(\alpha) = T(\alpha) = \sum_{j=1}^d F_j^\leftarrow(\alpha). \quad \square$$

Linear correlation

Proof of Proposition 7.16. Let (X'_1, X'_2) be an iid-copy of (X_1, X_2) . Consider

$$\begin{aligned} & 2 \operatorname{cov}(X_1, X_2) \\ &= \mathbb{E}((X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)) + \mathbb{E}((X'_1 - \mathbb{E}X'_1)(X'_2 - \mathbb{E}X'_2)) \\ &= \mathbb{E}\left(((X_1 - \mathbb{E}X_1) - (X'_1 - \mathbb{E}X'_1)) \cdot ((X_2 - \mathbb{E}X_2) - (X'_2 - \mathbb{E}X'_2))\right) \\ &\stackrel{\text{check}}{=} \mathbb{E}((X_1 - X'_1)(X_2 - X'_2)). \end{aligned}$$

With $b - a = \int_{-\infty}^{\infty} (I_{\{a \leq x\}} - I_{\{b \leq x\}}) dx$ for all $a, b \in \mathbb{R}$, we obtain that

$$\begin{aligned} & 2 \operatorname{cov}(X_1, X_2) \\ &= \mathbb{E}\left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (I_{\{X'_1 \leq x_1\}} - I_{\{X_1 \leq x_1\}})(I_{\{X'_2 \leq x_2\}} - I_{\{X_2 \leq x_2\}}) dx_1 dx_2\right] \\ &\stackrel{\text{Fubini}}{=} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{E}(\dots) dx_1 dx_2 \stackrel{\substack{\text{multiply} \\ \text{ind.}}}{=} 2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (F(x_1, x_2) - F_1(x_1)F_2(x_2)) dx_1 dx_2. \end{aligned}$$



Rank correlation

To overcome (some) of the deficiencies of ρ , Scarsini (1984) introduced:

Definition A.16 (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 (or measure of concordance)* 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 A.17 (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 $\kappa(T_1(X_1), T_2(X_2)) = \kappa(C) = \kappa(X_1, X_2)$.



Kendall's tau and Spearman's rho

Proof of Proposition 7.20. Let (X'_1, X'_2) be an independent copy of (X_1, X_2) . Then

$$\begin{aligned}\rho_\tau &= \mathbb{P}((X_1 - X'_1)(X_2 - X'_2) > 0) - \mathbb{P}((X_1 - X'_1)(X_2 - X'_2) < 0) \\ &= 2\underbrace{\mathbb{P}((X_1 - X'_1)(X_2 - X'_2) > 0)}_{=2\mathbb{P}(X_1 < X'_1, X_2 < X'_2)} - 1 = 4\mathbb{P}(U_1 \leq U'_1, U_2 \leq U'_2) - 1 \\ &= 4 \int_0^1 \int_0^1 \mathbb{P}(U_1 \leq u_1, U_2 \leq u_2) dC(u_1, u_2) - 1 \quad \square\end{aligned}$$

For computing ρ_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). One can also show that for any bivariate copulas C , \tilde{C} , $\int_{[0,1]^2} C(\mathbf{u}) d\tilde{C}(\mathbf{u}) = \int_{[0,1]^2} \tilde{C}(\mathbf{u}) dC(\mathbf{u})$.

Rank correlation for elliptical copulas

Lemma A.18

Let $\mathbf{X} \sim E_2(\mathbf{0}, \Sigma, \psi)$ with $\mathbb{P}(\mathbf{X} = \mathbf{0}) = 0$ and $\rho = P_{12} = \text{corr}(\Sigma)_{12}$.

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 $AA' = P$. Then $\mathbf{Y} \stackrel{d}{=} RAU \stackrel{d}{=} RA \begin{pmatrix} \cos \Theta \\ \sin \Theta \end{pmatrix}$, $\Theta \sim U(-\pi, \pi)$ 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 $\mathbb{P}(X_1 > 0, X_2 > 0) = \mathbb{P}(Y_1 > 0, Y_2 > 0) = \mathbb{P}(\cos \Theta > 0, \sin(\varphi + \Theta) > 0) = \mathbb{P}(\Theta \in (-\frac{\pi}{2}, \frac{\pi}{2}), \varphi + \Theta \in (0, \pi)) = \mathbb{P}(\Theta \in (-\frac{\pi}{2}, \frac{\pi}{2}), \Theta \in (-\varphi, \pi - \varphi)) = \mathbb{P}(\Theta \in (-\varphi, \frac{\pi}{2})) = (\frac{\pi}{2} - (-\varphi))/(2\pi)$. \square

Lemma A.19 (Representation of Spearman's rho)

Let $(U_1, U_2) \sim C$ and $\tilde{U}_1, \bar{U}_2 \stackrel{\text{ind.}}{\sim} \text{U}(0, 1)$ 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

Proof of Proposition 7.26. $\mathbf{X} \stackrel{\text{d}}{=} \sqrt{W}\mathbf{Z}$ for $\mathbf{Z} \sim \text{N}_2(\mathbf{0}, P)$. Let $\tilde{Z}, \bar{Z} \sim \text{N}(0, 1)$ and assume $\mathbf{Z}, \tilde{Z}, \bar{Z}, W, \tilde{W}$ and \bar{W} are all independent. Let

$$\tilde{X} = \sqrt{\tilde{W}}\tilde{Z}, \quad \bar{X} = \sqrt{\bar{W}}\bar{Z},$$

$$Y_1 = X_1 - \tilde{X} = \sqrt{W}Z_1 - \sqrt{\tilde{W}}\tilde{Z},$$

$$Y_2 = X_2 - \bar{X} = \sqrt{W}Z_2 - \sqrt{\bar{W}}\bar{Z}.$$

$$\rho_S(X_1, X_2) \stackrel{\substack{\text{L.A.19} \\ \Phi^{-1}}}{=} 12\mathbb{P}(X_1 \leq \tilde{X}_1, X_2 \leq \bar{X}_2) - 3$$

$$\begin{aligned}
&= 6\mathbb{P}((X_1 - \tilde{X}_1)(X_2 - \bar{X}_2) > 0) - 3 \\
&= 3(2\mathbb{E}(\mathbb{P}(Y_1 Y_2 > 0 | W, \tilde{W}, \bar{W})) - 1) \\
&= 3(4\mathbb{E}(\mathbb{P}(Y_1 > 0, Y_2 > 0 | W, \tilde{W}, \bar{W})) - 1).
\end{aligned}$$

Now note that $\mathbf{Y} | W, \tilde{W}, \bar{W} \sim N_2(\mathbf{0}, (\begin{smallmatrix} W+\tilde{W} & W\rho \\ W\rho & W+\bar{W} \end{smallmatrix}))$ with $\rho(Y_1, Y_2) = \frac{W\rho}{\sqrt{(W+\tilde{W})(W+\bar{W})}}$. Apply Lemma A.18 to see that

$$\rho_S(X_1, X_2) = 3\left(4\mathbb{E}\left(\frac{1}{4} + \frac{\arcsin \rho}{2\pi}\right) - 1\right) = \frac{12}{2\pi}\mathbb{E}(\arcsin \rho(Y_1, Y_2)).$$

For Gauss copulas, $F_W(x) = I_{\{x \geq 1\}}$, thus $W = \tilde{W} = \bar{W} = 1$ a.s. and the result follows. \square

Proof of Proposition 7.27.

- Let (X'_1, X'_2) be an independent copy of (X_1, X_2) . We have already seen that $\rho_\tau = 2\mathbb{P}((X_1 - X'_1)(X_2 - X'_2) > 0) - 1$.
- With $\mathbf{X} \stackrel{d}{=} R\mathbf{A}\mathbf{U}$ and $\mathbf{X}' \stackrel{d}{=} R'\mathbf{A}\mathbf{U}'(\stackrel{d}{=} -\mathbf{X}')$ 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 $\phi_{\mathbf{Y}}(\mathbf{t}) \stackrel{\text{ind.}}{=} \phi_{\mathbf{X}}(\mathbf{t})\phi_{-\mathbf{X}'}(\mathbf{t}) = \phi_{\mathbf{X}}(\mathbf{t})^2$,
hence $\mathbf{Y} \sim \text{E}_2(\mathbf{0}, P, \psi^2)$.

- We thus obtain that

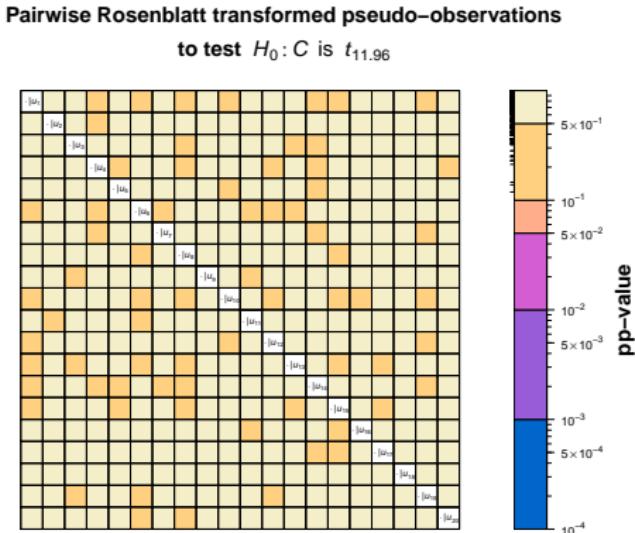
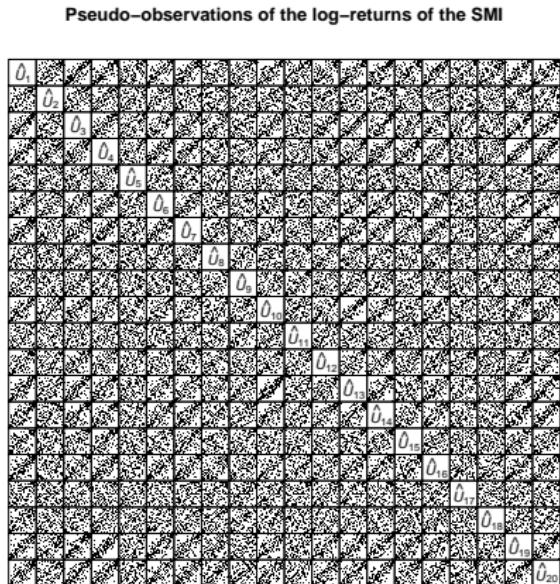
$$\begin{aligned}\rho_{\tau} &= 2\mathbb{P}(Y_1 Y_2 > 0) - 1 = 2(\mathbb{P}(Y_1 > 0, Y_2 > 0) + \mathbb{P}(Y_1 < 0, Y_2 < 0)) - 1 \\ &= 4\mathbb{P}(\mathbf{Y} > \mathbf{0}) - 1 \stackrel{\substack{\text{cont.} \\ \text{L.A.18}}}{=} \frac{2}{\pi} \arcsin \rho.\end{aligned}$$

□

For a generalization to componentwise n.d. \mathbf{X} , see Lindskog et al. (2003).

Goodness-of-fit

A graphical goodness-of-fit approach by Hofert and Mächler (2014) based on daily log-returns of the SMI from 2011-09-09 to 2012-03-28.

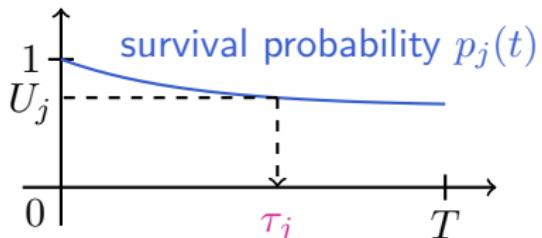


How intensity-/copula-based default models work

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 (1960): $\lambda_U = 0$)
- 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.

A.8 Aggregate risk

We now present an elementary proof for subadditivity of ES. We start with some auxiliary results.

Lemma A.20

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

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

For the following result let

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

Lemma A.21 (Properties of $I_{\{L > F_L^\leftarrow(\alpha)\}}^{(\alpha)}$)

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

Proof.

1) If $\mathbb{P}(L = F_L^\leftarrow(\alpha)) = 0$ we are done, so consider $\mathbb{P}(L = F_L^\leftarrow(\alpha)) > 0$.

On the set of all $\omega \in \Omega$ such that $L(\omega) > F_L^\leftarrow(\alpha)$, we are again done.

Now consider all $\omega \in \Omega$ such that $L(\omega) = F_L^\leftarrow(\alpha)$. Then $I_{\{L > F_L^\leftarrow(\alpha)\}}^{(\alpha)} =$

$\frac{1 - \alpha - \bar{F}_L(F_L^\leftarrow(\alpha))}{\mathbb{P}(L = F_L^\leftarrow(\alpha))}$. By definition, $F_L(F_L^\leftarrow(\alpha)) \geq \alpha$, so $\bar{F}_L(F_L^\leftarrow(\alpha)) \leq$

$1 - \alpha$, thus $I_{\{L > F_L^\leftarrow(\alpha)\}}^{(\alpha)} \geq 0$. Also, $F_L(F_L^\leftarrow(\alpha)-) < \alpha$, so $I_{\{L > F_L^\leftarrow(\alpha)\}}^{(\alpha)}$

$$= \frac{1 - \alpha - (1 - F_L(F_L^\leftarrow(\alpha)))}{\mathbb{P}(L = F_L^\leftarrow(\alpha))} = \frac{F_L(F_L^\leftarrow(\alpha)) - \alpha}{F_L(F_L^\leftarrow(\alpha)) - F_L(F_L^\leftarrow(\alpha)-)} < 1.$$

2) We have

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

Consider $\mathbb{P}(L = q) = 0$. Since $q = F_L^\leftarrow(\alpha)$, Lemma A.20 implies that $\bar{F}_L(q) = 1 - F_L(F_L^\leftarrow(\alpha)) = 1 - \alpha$. Thus $\mathbb{E}(I_{\{L>q\}}^{(\alpha)}) = 1 - \alpha$. \square

Lemma A.22 (Representation of ES_α in terms of $I_{\{L>F_L^\leftarrow(\alpha)\}}^{(\alpha)}$)

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

Proof.

- If $\mathbb{P}(L = F_L^\leftarrow(\alpha)) = 0$, Lemma A.20 implies that $\bar{F}_L(F_L^\leftarrow(\alpha)) = 1 - \alpha$. By Proposition ?? ?? and since $\mathbb{P}(L = F_L^\leftarrow(\alpha)) = 0$,

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

$$= \frac{\mathbb{E}(LI_{\{L > F_L^\leftarrow(\alpha)\}})}{1 - \alpha} = \frac{\mathbb{E}(LI_{\{L > F_L^\leftarrow(\alpha)\}}^{(\alpha)})}{1 - \alpha}.$$

- If $\mathbb{P}(L = F_L^\leftarrow(\alpha)) > 0$, $\mathbb{E}(LI_{\{L > F_L^\leftarrow(\alpha)\}}^{(\alpha)})$ equals

$$\mathbb{E}(LI_{\{L > F_L^\leftarrow(\alpha)\}}) + \frac{1 - \alpha - \bar{F}_L(F_L^\leftarrow(\alpha))}{\mathbb{P}(L = F_L^\leftarrow(\alpha))} \underbrace{\mathbb{E}(LI_{\{L = F_L^\leftarrow(\alpha)\}})}_{= \mathbb{E}(F_L^\leftarrow(\alpha)I_{\{L = F_L^\leftarrow(\alpha)\}})} = F_L^\leftarrow(\alpha)\mathbb{P}(L = F_L^\leftarrow(\alpha))$$

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

Proposition A.23 (Subadditivity of ES)

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

Proof. It suffices to show that

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

By Lemma A.22, this equals

$$\begin{aligned}
 & \left(\sum_{j=1}^2 \mathbb{E}(L_j I_{\{L_j > F_{L_j}^{\leftarrow}(\alpha)\}}^{(\alpha)}) \right) - \mathbb{E}((L_1 + L_2) I_{\{L_1 + L_2 > F_{L_1 + L_2}^{\leftarrow}(\alpha)\}}^{(\alpha)}) \\
 & \stackrel{\text{Linearity}}{=} \sum_{j=1}^2 \mathbb{E}(L_j (I_{\{L_j > F_{L_j}^{\leftarrow}(\alpha)\}}^{(\alpha)} - I_{\{L_1 + L_2 > F_{L_1 + L_2}^{\leftarrow}(\alpha)\}}^{(\alpha)})). \tag{33}
 \end{aligned}$$

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

In both cases, we make the expectations in (33) smaller by replacing L_j by $F_{L_j}^{\leftarrow}(\alpha)$. Hence

$$(33) \geq \sum_{j=1}^2 F_{L_j}^{\leftarrow}(\alpha) \underbrace{\mathbb{E}(I_{\{L_j > F_{L_j}^{\leftarrow}(\alpha)\}}^{(\alpha)} - I_{\{L_1 + L_2 > F_{L_1 + L_2}^{\leftarrow}(\alpha)\}}^{(\alpha)})}_{\substack{= (1-\alpha) - (1-\alpha) = 0 \\ \text{Lem. A.21 2)}} \geq 0. \quad \square$$