

Quantitative Risk Management

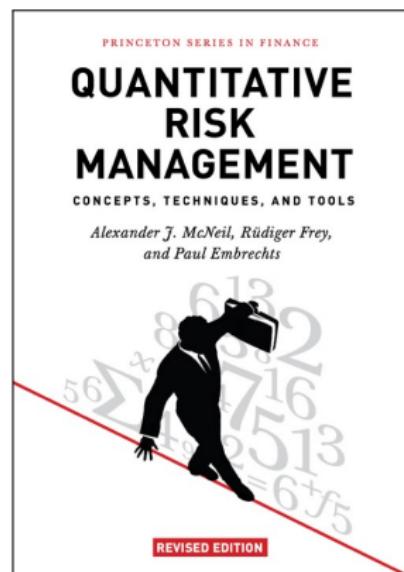
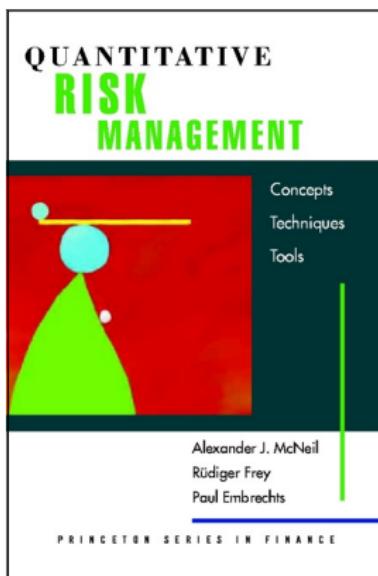
<http://www.qrmtutorial.org>

Last update: 2018-02-28

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**
- 8 Aggregate risk**
- 9 Market Risk**
- 10 Credit risk**

11 Portfolio credit risk management

14 Multivariate time series

17 Introduction to counterparty risk

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 randomness

1.1.1 Risk and randomness

- We will mostly model situations in which an investor holds today an asset with an uncertain future value.
- We use probabilistic notions (random variables, random vectors, distributions, stochastic processes) and statistical tools. In particular, we assume to work on a *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$; see Kolmogorov (1933).

1.1.2 Financial Risk

There are various types of risks. We focus on (those affected by regulation):

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 $X = (X_1, \dots, X_d)$. 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 comply with 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 Weatherstone 4¹⁵ 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](#) (see Basel Committee on Banking Supervision (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 (exchanges);
 - 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 them. 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, (more) capital and (higher) 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 (single customers and as a whole (systemic risk)) 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 others 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.
- 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 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 be beneficial, since a company may have better access to capital markets than individual investors.
- ▶ 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

The risks of a firm (here: bank) can be understood from its balance sheet:

Assets Investments of the firm		Liabilities Obligations from fundraising	
Cash (and central bank balance)	£10M	Customer deposits	£80M
Securities - bonds, stocks, derivatives	£50M	Bonds issued - senior bond issues	£25M
Loans and mortgages - corporates	£100M	- subordinated bond issues	£15M
- retail and smaller clients		Short-term borrowing	£30M
- government		Reserves (for losses on loans)	£20M
Other assets - property	£20M	Debt (sum of above)	£170M
- 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 to shareholders).
- Regulatory capital*
 - Capital required according to regulatory rules;

- For European insurance firms: Minimum (MCR) and solvency capital requirements (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 set up a general mathematical model for (changes in) value caused by financial risks. To this end 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 Δt .*

- The *change* in value of the portfolio is 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*.
- 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*, 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+1} 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} . We will thus also write $L_{t+1} = L(\mathbf{X}_{t+1})$, where $L(\mathbf{x}) = -(f(t+1, \mathbf{Z}_t + \mathbf{x}) - f(t, \mathbf{Z}_t))$ is known as *loss operator*.

- If f is differentiable, its first-order (Taylor) approximation ($f(\mathbf{y}) \approx f(\mathbf{y}_0) + \nabla f(\mathbf{y}_0)(\mathbf{y} - \mathbf{y}_0)$ for $\mathbf{y} = (t+1, Z_{t,1} + X_{t+1,1}, \dots, Z_{t,d} + X_{t+1,d}$ and $\mathbf{y}_0 = (t, Z_{t,1}, \dots, Z_{t,d})$) 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

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

which is non-linear in $X_{t+1,j}$ (here: $L(x) = - \sum_{j=1}^d \tilde{w}_{t,j} (e^{x_j} - 1)$).

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

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

- Note that $L_{t+1}^{\Delta} = -(c_t + \mathbf{b}_t' \mathbf{X}_{t+1})$ for $c_t = 0$ and $\mathbf{b}_t = \tilde{\mathbf{w}}_t$.
- If $\boldsymbol{\mu} = \mathbb{E} \mathbf{X}_{t+1}$ and $\Sigma = \text{cov } \mathbf{X}_{t+1}$ are known, then expectation and variance of the (linearized) one-period ahead loss are

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

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

- If \mathbf{X}_{t+1} is multivariate normal, then $L_{t+1}^{\Delta} \sim N(-\tilde{\mathbf{w}}_t' \boldsymbol{\mu}, \tilde{\mathbf{w}}_t' \Sigma \tilde{\mathbf{w}}_t)$.

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

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

This implies that the mapping f (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 1d** (as opposed to 1y), 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 $L_{t+1} \Rightarrow$ 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*. 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 (see later) 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 $\boldsymbol{X}_{t+1} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (e.g. if (\boldsymbol{Z}_t) is a Brownian motion, (S_t) a geometric Brownian motion)

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

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

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

Drawbacks:

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

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

2) Historical simulation

Idea: Estimate $F_{L_{t+1}}$ by its *empirical distribution function (edf)*

$$\hat{F}_{L_{t+1},n}(x) = \frac{1}{n} \sum_{i=1}^n I_{\{L_{t-i+1} \leq x\}}, \quad x \in \mathbb{R}, \quad (4)$$

based on

$$L_k = L(\mathbf{X}_k) = -(f(t+1, \mathbf{Z}_t + \mathbf{X}_k) - f(t, \mathbf{Z}_t)), \quad (5)$$

$k \in \{t-n+1, \dots, t\}$. L_{t-n+1}, \dots, 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 past losses considered ("driving a car by looking in the back mirror")

3) Monte Carlo method

Idea: Take any model for \mathbf{X}_{t+1} , simulate \mathbf{X}_{t+1} , compute the corresponding losses as in (5) and estimate $F_{L_{t+1}}$ (typically via edf as in (4)).

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 mapping f ; 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 risk measurement approaches grouped into three categories:

1) Notional-amount approach

- oldest approach; “standardized approaches” of Basel II (e.g. OpRisk)
- *risk of a portfolio* = summed notional 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 loss distribution over some predetermined time horizon Δt .
- Examples: variance, value-at-risk, expected shortfall (see later)
- Advantages:
 - ▶ Makes sense on all levels (from single portfolios to the overall position of a financial institution).
 - ▶ Loss distributions reflect netting and diversification.

- 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 $\mathcal{X} = \{x_1, \dots, x_n\}$ denote the risk-factor changes (*scenarios*) with corresponding weights $w = (w_1, \dots, w_n)$, the risk is $\psi_{\mathcal{X}, w} = \max_{1 \leq i \leq n} \{w_i L(x_i)\}$, where $L(x)$ denotes the loss the portfolio

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

- Mathematical interpretation:
 - ▶ 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).

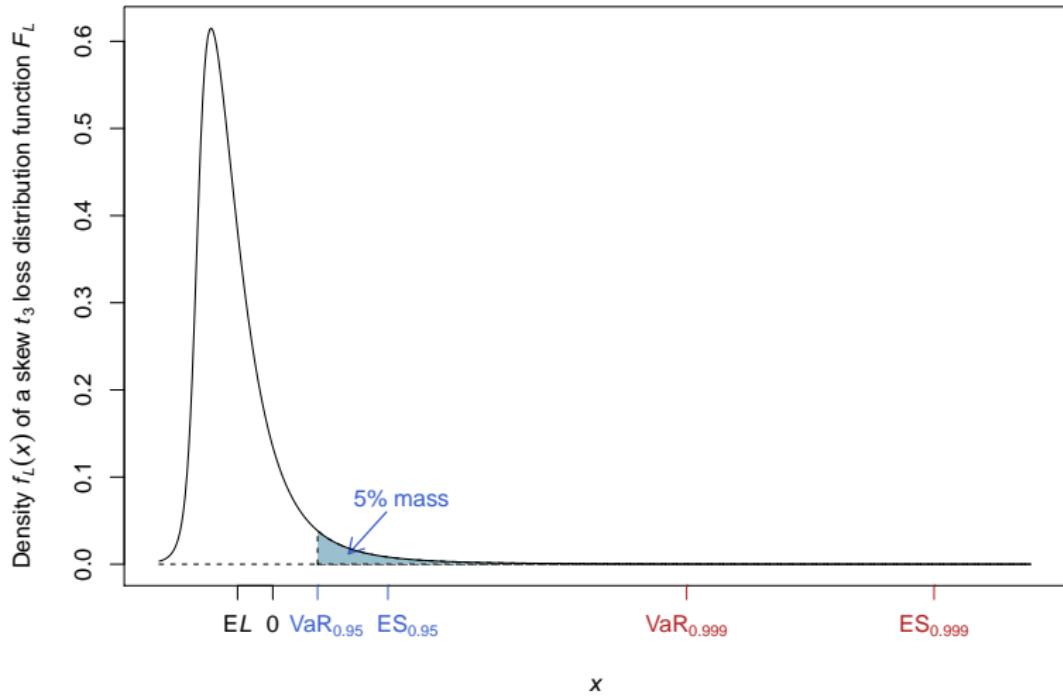
Drawbacks: ► Determining scenarios and weights.

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

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

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

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

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

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

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

- Δt should reflect the time period over which the portfolio is held (unchanged) (e.g. insurance contracts: $\Delta t = 1 \text{ y}$)
- 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 (considerable model/liquidity risk).

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)

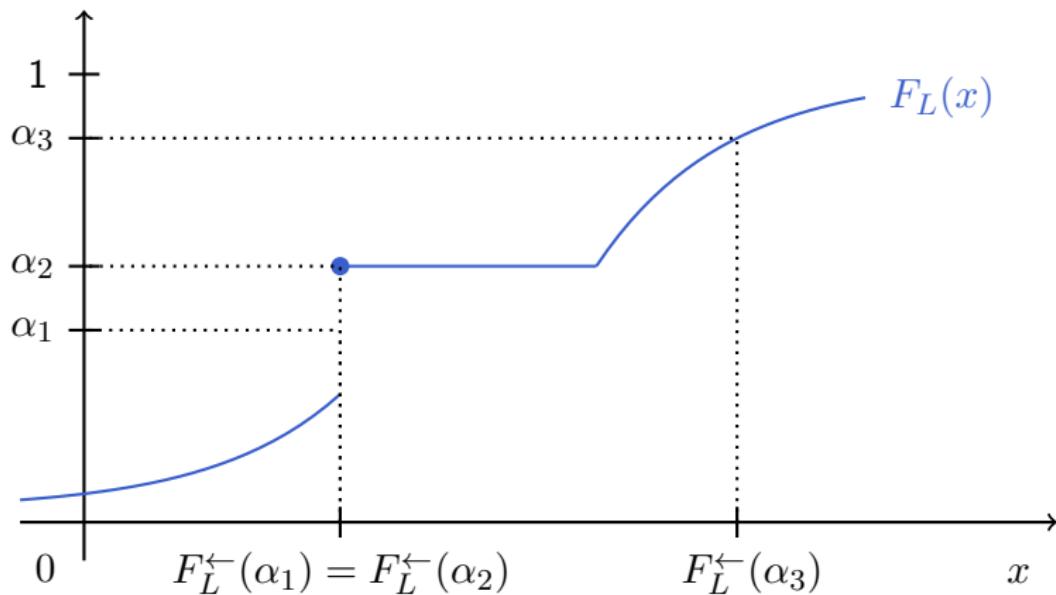
For any increasing function $T : \mathbb{R} \rightarrow \mathbb{R}$, with $T(-\infty) = \lim_{x \downarrow -\infty} T(x)$ and $T(\infty) = \lim_{x \uparrow \infty} T(x)$, the *generalized inverse* $T^\leftarrow : \mathbb{R} \rightarrow \bar{\mathbb{R}} = [-\infty, \infty]$ of T is defined by

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

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 (often, not always) similar to T^{-1} ; see Proposition A.15.

How to determine F_L^\leftarrow from F_L :



⇒ Flat parts of F_L correspond to jumps of F_L^\leftarrow ;
Jumps of F_L correspond to flat parts of F_L^\leftarrow .

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 (similarly but more involved for Basel III), the daily risk capital formula is

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

- $VaR_{\alpha}^{s,10}$ 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; VaR}_{0.999}\text{-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 $VaR_{0.99}^{t,10}$ 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 $\Delta t = 1 \text{ year}$ with $\alpha = 0.995$. Let V_t denote equity capital. The insurer wants to determine the minimum amount of extra capital x_0 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, or standard deviation, 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

Let $x_+ = \max\{x, 0\}$.

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

$$\text{ES}_\alpha = \text{ES}_\alpha(L) = \frac{1}{1 - \alpha} \int_\alpha^1 \text{VaR}_u(L) \, du. \quad (6)$$

- ES_α is the **average over VaR_u** for all $u \geq \alpha \Rightarrow \text{ES}_\alpha \geq \text{VaR}_\alpha$.
- 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 (the variance of estimators is typically larger; larger sample size required).
- $\text{ES}_\alpha(L) < \infty$ requires $\mathbb{E}(L_+) < \infty$.
- Subadditivity and elicibility (see the appendix). One can show:
 - ▶ In contrast to VaR_α , ES_α is **subadditive** (more 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 (i.e. minimizes some expected functional (scoring function); see Gneiting (2011). This can be used for backtesting, comparing risk measures).

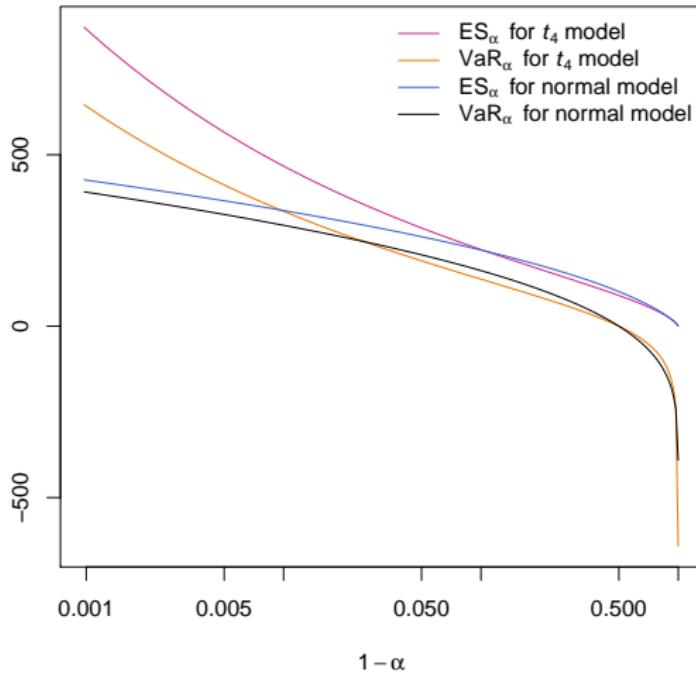
Example 2.9 (A comparison of VaR and ES for stock returns)

- Consider Example 2.2 with a 1-stock portfolio and $V_t = S_t = 10\,000$. 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
 - $X_{t+1} \sim N(0, \sigma^2) \Rightarrow L_{t+1}^\Delta \sim N(0, S_t^2 \sigma^2)$;
 - $X_{t+1} \sim t_\nu(0, \sigma^2 \frac{\nu-2}{\nu})$, $\nu > 2$ (so that $\text{var } X_{t+1} = \sigma^2$, too). Then

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

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

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



⇒ 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 $\varphi(x) = \Phi'(x) = \exp(-x^2/2)/\sqrt{2\pi}$. Since $x\varphi(x) = -\varphi'(x)$,

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

This implies (see soon) 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 VaR and ES can be huge (for large α as required by Basel II).

Under continuity, expected shortfall equals *conditional tail expectation*.

Proposition 2.11 ($\text{ES}_\alpha(L)$ under continuity)

If F_L is continuous, $\text{ES}_\alpha(L) = \mathbb{E}(L | L > \text{VaR}_\alpha(L))$.

Proof. If F_L is continuous, $F_L(\text{VaR}_\alpha(L)) = F_L(F_L^\leftarrow(\alpha)) = \alpha$ and thus, for all $x \geq \text{VaR}_\alpha(L)$,

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

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 substituted $x = \text{VaR}_z(L) = F_L^\leftarrow(z)$ (so $F_L(x) = z$, $dF_L(x) = dz$). □

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 net asset values** V_{t+1} : $\tilde{\varrho}(V_{t+1})$ denotes the **additional capital 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 shifting 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 (that is, a *superadditive*) 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): For $L_1 = \dots = L_n = L$, subadditivity implies $\varrho(nL) \leq n\varrho(L)$, but there is no diversification, so equality should hold.

Criticism: If $\lambda > 1$ is large, liquidity risk plays a role and one should rather have $\varrho(\lambda L) > \lambda \varrho(L)$ (also to penalize risk concentration), but this contradicts subadditivity. This has led to *convex risk measures*, i.e. monotone, translation invariant 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 2.12 (Coherent risk measure)

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

Coherent risk measures are convex. The converse is not true in general (but for positive homogeneous risk measures ϱ).

Example 2.13 (Coherence of 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 $\mathbb{E}_{\mathbb{P}}(\cdot)$). For subadditivity, note that

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

Remark 2.14

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.

Theorem 2.15 (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. \square

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.

Let's have a look at examples for 1), 2) and 4); for 3), see later.

Example 2.16 (Skewed loss distributions)

Consider two independent losses of the form

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

Set $\alpha = 0.99$. Then $\text{VaR}_\alpha(L_j) = -5$, $j \in \{1, 2\}$. The loss $L_1 + L_2$ is

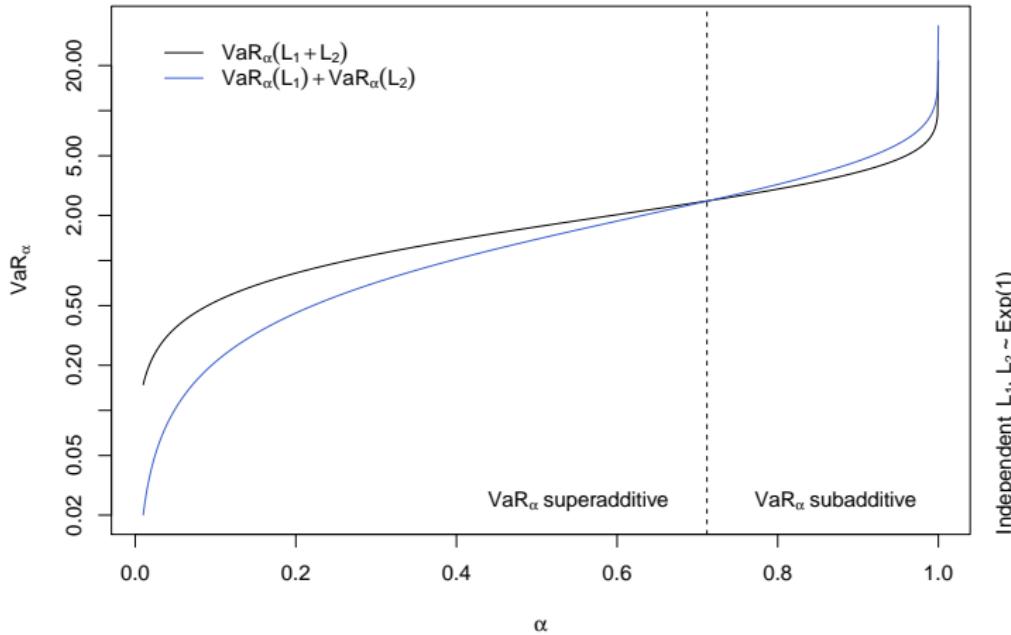
$$L_1 + L_2 = \begin{cases} -10, & \text{with prob. } (1-p)^2 = 0.982081, \\ 95, & \text{with prob. } 2p(1-p) = 0.017838, \\ 200, & \text{with prob. } p^2 = 0.000081. \end{cases}$$

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

- One can show: VaR_α is superadditive if and only if $(1-p)^d < \alpha \leq 1-p$.
- From the money lender's perspective, the losses could come from **two independently defaultable zero-coupon bonds** (maturity $T = 1$ y, nominal/face value 100, paid interest of 5%, default probability $p = 0.009$, no recovery).

Example 2.17 (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 $\iff \alpha < 0.71$.



One can show that (independently of λ) VaR_α is superadditive if and only if $(1 - \alpha)(1 - 2 \log(1 - \alpha)) > 1$.

Example 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)

- Note that $L_1 \stackrel{\text{a.s.}}{=} L_2$ does not lead to the largest $\text{VaR}_\alpha(L_1 + L_2)$ since

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

so “only” equality (whereas all above scenarios have produced “ $>$ ”).

- VaR_α is subadditive and thus coherent for all elliptical distributions (strictly including the multivariate normal and t) when $\alpha \geq 1/2$; see later.

3 Empirical properties of financial data

3.1 Stylized facts of financial return series

3.2 Multivariate stylized facts

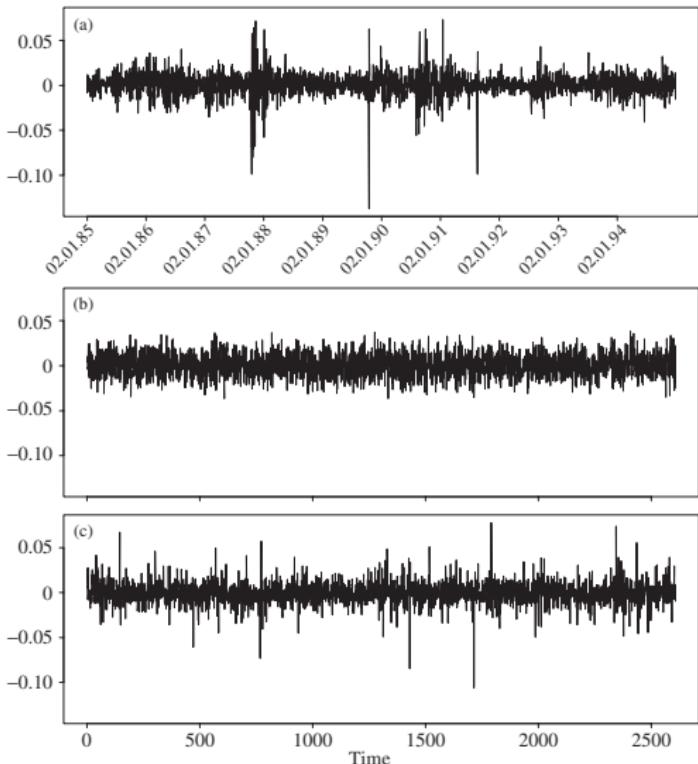
3.1 Stylized facts of financial return series

- The **stylized facts** are a collection of **empirical observations** and **related inferences**, which apply to many time series of risk-factor changes (e.g. log-returns on equities, indices, exchange rates, commodity prices).
- The best-known stylized facts **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 sparseness**; non-stationarity).
- Consider **discrete-time risk-factor changes** $X_t = Z_t - Z_{t-1}$ for a log-price or rate $Z_t = \log S_t$. In this case

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

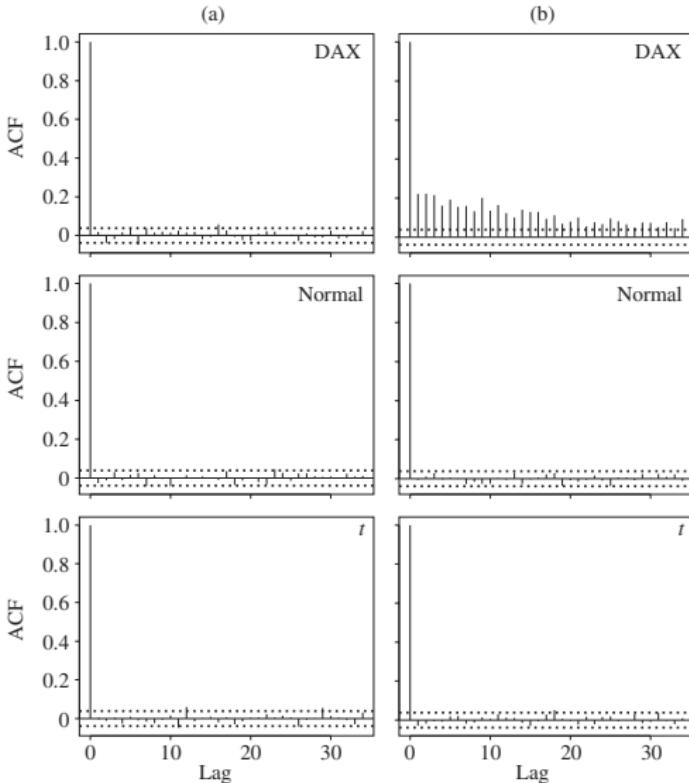
the former is often called a **(log-)return**, the latter a **simple return**.

3.1.1 Volatility Clustering



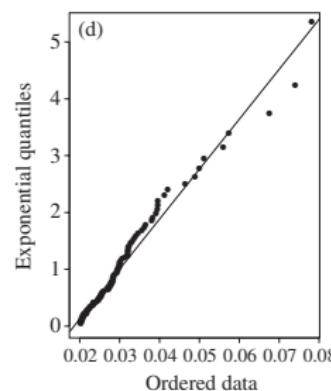
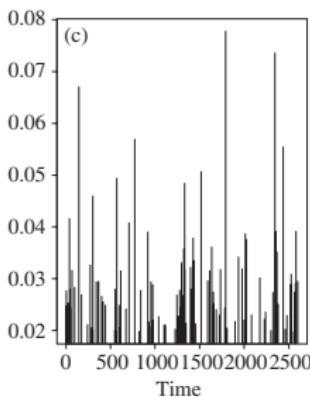
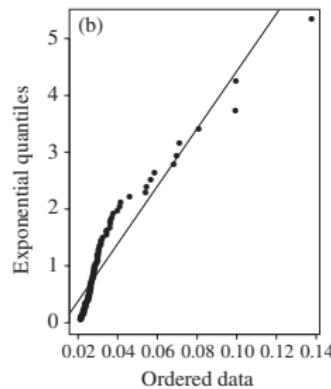
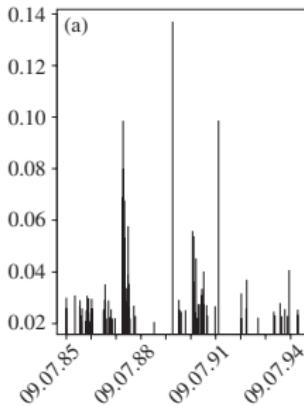
- (a) Log-returns for the DAX index from 1985-01-02 to 1994-12-30 ($n = 2608$).
(b) Simulated iid data from a fitted normal with $\hat{\mu} = \bar{X}_n$, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ show too few extremes.
(c) Simulated iid data from a fitted $t_{3.8}$. Better range of values but still no volatility clustering (= tendency for extreme returns to be followed by extreme returns).

Estimated *autocorrelation function (ACF)* $\rho(h) = \text{corr}(X_0, X_h)$, $h \in \mathbb{Z}$



- Estimated
 - (a) ACF of $(X_t)_{t \in \mathbb{Z}}$
 - (b) ACF of $(|X_t|)_{t \in \mathbb{Z}}$
- Non-zero ACF at lag 1 implies a tendency for a return to be followed by a return of equal sign; not the case here.
⇒ Predicted return ≈ 0
- For iid process $(X_t)_{t \in \mathbb{Z}}$ $\rho_X(h) = \rho_{|X|}(h) = I_{\{h=0\}}$; not the case here; see also Ljung–Box tests to confirm.
- $(Z_t)_{t \in \mathbb{Z}}$ not a random walk
 $(S_t)_{t \in \mathbb{Z}}$ not GBM.

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**);
- The **DAX** data shows shorter and longer waiting times than the iid data, so **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, biased by n)
 - ▶ 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 $T_n = \frac{n}{6} \left(\hat{\beta}^2 + \frac{1}{4}(\hat{\kappa} - 3)^2 \right) \underset{n \text{ large}}{\sim} H_0 \chi^2_2$.

Graphical tests

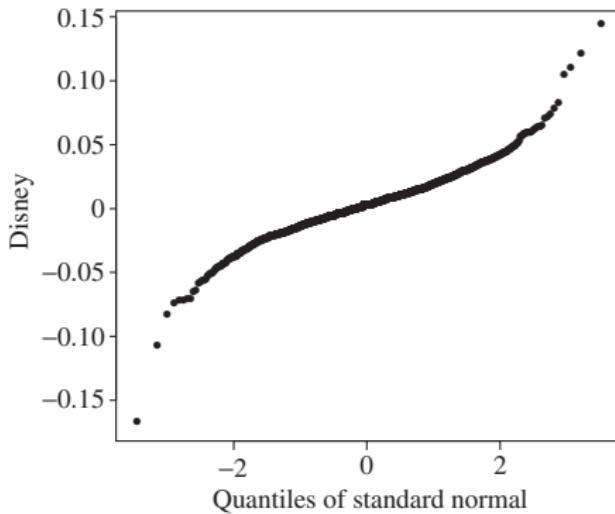
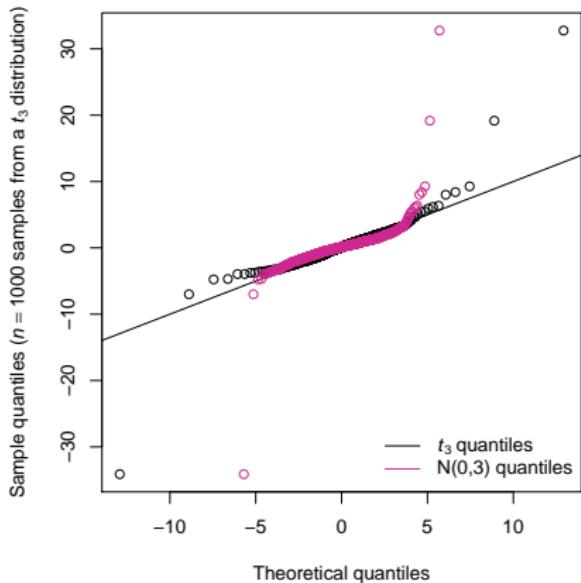
- We can also graphically test whether $X_1, \dots, X_n \sim F$ for some df F based on realizations of iid X_1, \dots, X_n .
- Let $X_{(1)} \leq \dots \leq X_{(n)}$ denote the corresponding order statistics and note that

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I_{\{X_i \leq x\}} = \frac{1}{n} \sum_{i=1}^n I_{\{X_{(i)} \leq x\}}, \quad x \in \mathbb{R},$$

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:** For $p_i \underset{n > 10}{=} \frac{i-1/2}{n} \approx \frac{i}{n}$, plot $\{(p_i, F(X_{(i)})) : i = 1, \dots, n\}$. If $F \approx \hat{F}_n$, $F(X_{(i)}) \approx p_i$, so the points lie on a line with slope 1.
 - ▶ **Q-Q plot:** Plot $\{(F^\leftarrow(p_i), X_{(i)}) : i = 1, \dots, n\}$ (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 **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 are 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).

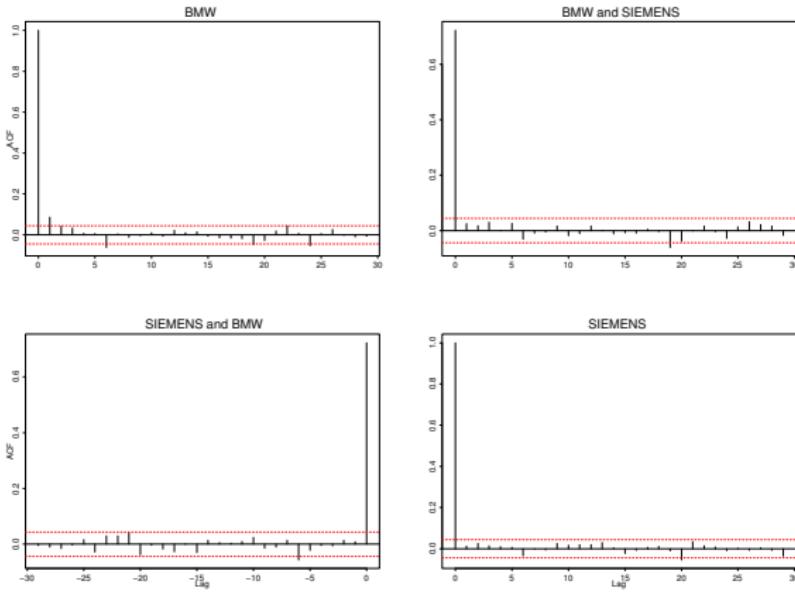
3.2 Multivariate stylized facts

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

3.2.1 Correlation between series

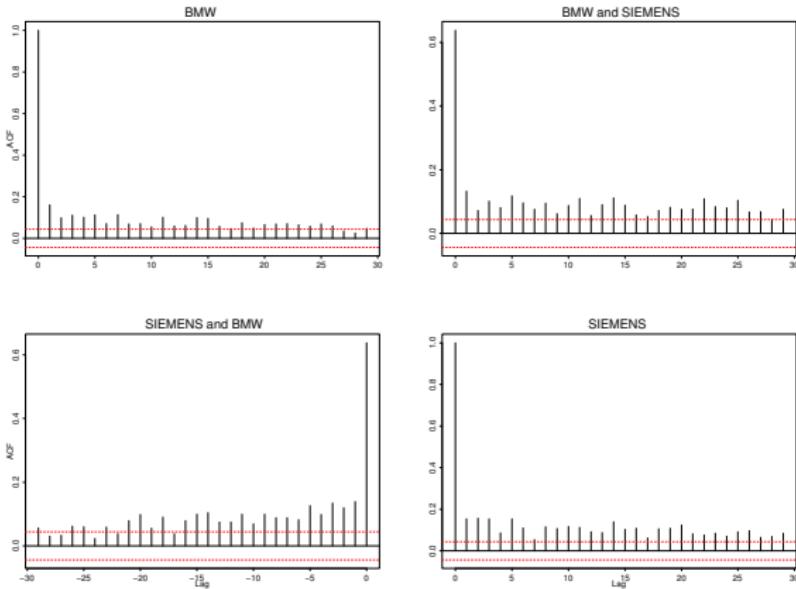
- By (U1), the returns of stock A at t and $t+h$ show little (auto)correlation. The returns of stock A at t and stock B at $t+h$, $h > 0$, also show little cross-correlation. However, Stock A and stock B on day t may be correlated due to factors that affect the whole market (*contemporaneous dependence*).
- These 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, so returns of large magnitude in A at t may be followed by returns of large magnitude in A and B at $t+h$.

Estimated correlations between/within series:



Based on 2000 values from period 1985-01-23 to 1994-09-22. Little autocorrelation, little crosscorrelation (at different lags), contemporaneous correlation.

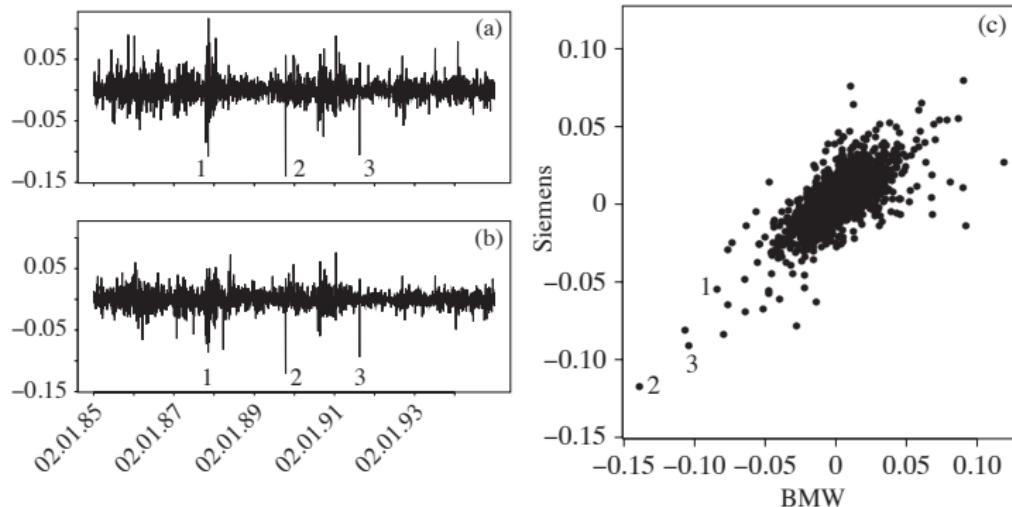
Estimated correlations between/within series of absolute values:



Autocorrelation of absolute returns (indication of volatility clustering).
Common to more than one stock.

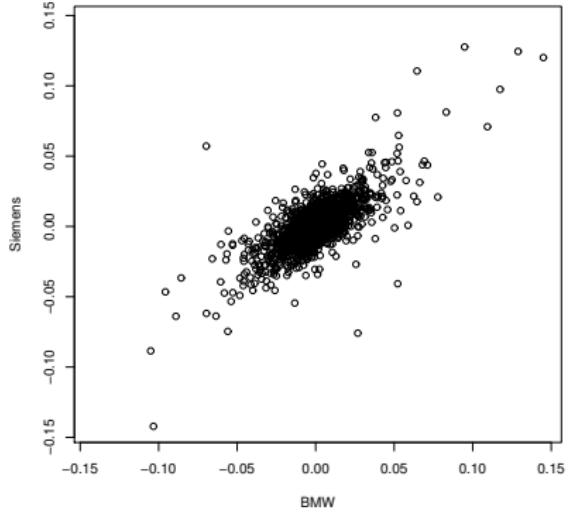
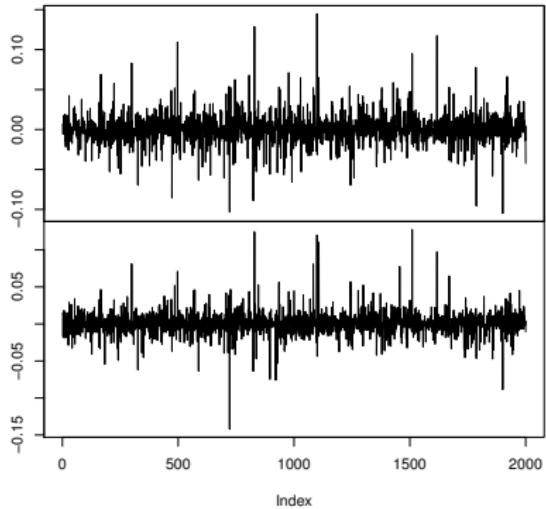
3.2.2 Tail dependence (for quantifying joint extremes)

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



In volatile/extreme(ly bad) periods, dependence seems stronger (1: 1987-10-19 Black Monday (DJ drop by 22%); 2: 1989-10-16 Monday demonstrations (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 (not so easy to infer from empirical correlations due to estimation error in small samples);
- (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 the idea that $(X_t)_{t \in \mathbb{Z}}$ behaves similarly in any time period.
- 2)
 - Strict stationarity $\not\Rightarrow$ stationarity (unless also $\mathbb{E}(X_t^2)$ exists).
 - Stationarity $\not\Rightarrow$ strict stationarity ($\mathbb{E}(|X_t|^p)$, $p > 2$, could change).
- 3) If $(X_t)_{t \in \mathbb{Z}}$ is stationary, $\gamma(0, t - s) = \gamma(s, t) = \gamma(t, s) = \gamma(0, s - t)$, so $\gamma(t, s)$ only depends on the lag $h = |t - s|$. We can thus define $\gamma(h) := \gamma(0, |h|)$, $h \in \mathbb{Z}$.

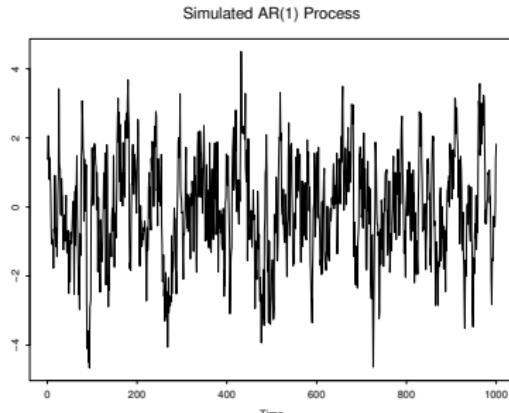
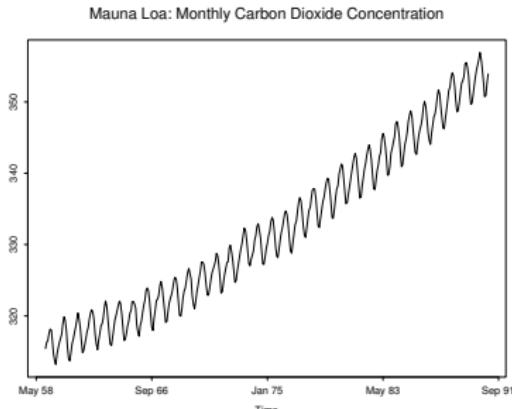
Autocorrelation in stationary time series

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

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

Stationary?



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 AR(p) processes similarly to how the ACF is used for 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 $\mu(t) = 0$, $\gamma(0) = \text{var}(X_t) = \sigma^2$, $(X_t)_{t \in \mathbb{Z}}$ is denoted by $(\varepsilon_t)_{t \in \mathbb{Z}} \sim \text{WN}(0, \sigma^2)$.
- 2) $(X_t)_{t \in \mathbb{Z}}$ is a *strict white noise* process if $(X_t)_{t \in \mathbb{Z}}$ is an iid sequence of rvs with $\gamma(0) = \text{var}(X_t) = \sigma^2 < \infty$. If $\mu(t) = 0$, we write $(Z_t)_{t \in \mathbb{Z}} \sim \text{SWN}(0, \sigma^2)$.

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. $(\mathcal{F}_t)_{t \in \mathbb{Z}}$ (typically the natural filtration $\mathcal{F}_t = \sigma(\{X_s : s \leq t\})$) if $\mathbb{E}|X_t| < \infty$, $t \in \mathbb{Z}$, $(X_t)_{t \in \mathbb{Z}}$ is adapted to $(\mathcal{F}_t)_{t \in \mathbb{Z}}$; and $\mathbb{E}(X_{t+1} | \mathcal{F}_t) = 0$ 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 $t \in \mathbb{Z}$,

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}. \quad (7)$$

$(X_t)_{t \in \mathbb{Z}}$ 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 (7) 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 $(X_t)_{t \in \mathbb{Z}}$ 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)

Let $X_t = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k}$ with $\sum_{k=0}^{\infty} |\psi_k| < \infty$. This process is *stationary* with ACF given by

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

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*), $(X_t)_{t \in \mathbb{Z}}$ is *invertible*, i.e. we can recover ε_t from $(X_s)_{s \leq t}$ (via $\varepsilon_t = \phi(B)X_t/\theta(B)$), so $\varepsilon_t \in \mathcal{F}_t = \sigma(\{X_s : s \leq t\})$.
- 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) $\text{MA}(q) = \text{ARMA}(0, q)$: $X_t = \varepsilon_t + \sum_{k=1}^q \theta_k \varepsilon_{t-k} \stackrel{\theta_0 := 1}{=} \sum_{k=0}^q \theta_k \varepsilon_{t-k}$

\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 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) (**I**ntegrated) 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** (**S**easonal) 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 a matrix W depending on ρ ; see McNeil et al. (2015, Theorem 4.13).

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

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

where $q_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$. This interval is usually shown in correlogram.

If more than 5% of $\hat{\rho}(k)$, $k \in \{1, \dots, h\}$, lie outside, 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. minimal 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

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

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

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

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 $\text{var}(X_t | \mathcal{F}_{t-1}) = \mathbb{E}((\sigma_t Z_t)^2 | \mathcal{F}_{t-1}) = \sigma_t^2 \mathbb{E}(Z_t^2 | \mathcal{F}_{t-1}) = \sigma_t^2 \mathbb{E}(Z_t^2) = \sigma_t^2$.
- \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.

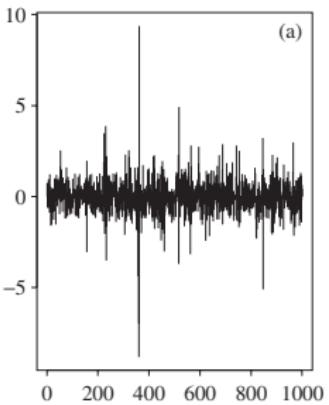
Example 4.17 (ARCH(1))

- 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, $\text{var}(X_t) = \alpha_0 / (1 - \alpha_1)$.

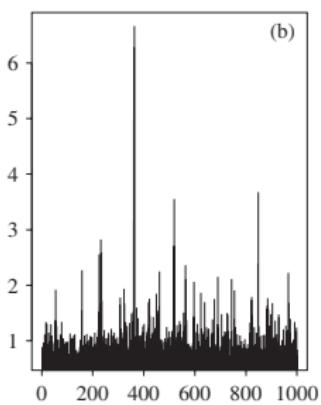
Proof of necessity. $X_t^2 = \sigma_t^2 Z_t^2 = (\alpha_0 + \alpha_1 X_{t-1}^2) Z_t^2 \Rightarrow \sigma_X^2 = \mathbb{E}(X_t^2) = \alpha_0 + \alpha_1 \mathbb{E}(X_{t-1}^2 Z_t^2) = \alpha_0 + \alpha_1 \sigma_X^2 \Rightarrow \sigma_X^2 = \frac{\alpha_0}{1 - \alpha_1}, \alpha_1 < 1$. \square

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

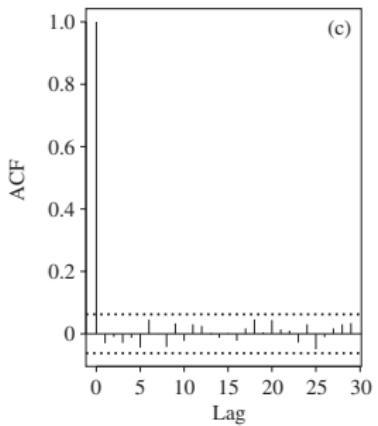
- Provided that $\mathbb{E}(Z_t^4) < \infty$ and $\alpha_1 < (\mathbb{E}(Z_t^4))^{-1/2}$, one can show that $\kappa(X_t) = \frac{\mathbb{E}(X_t^4)}{\mathbb{E}(X_t^2)^2} = \frac{\kappa(Z_t)(1 - \alpha_1^2)}{(1 - \alpha_1^2 \kappa(Z_t))}$. If $\kappa(Z_t) > 1$, $\kappa(X_t) > \kappa(Z_t)$. For Gaussian 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 $X_t^2 - \frac{\alpha_0}{1 - \alpha_1} = \alpha_1(X_{t-1}^2 - \frac{\alpha_0}{1 - \alpha_1}) + \varepsilon_t$.



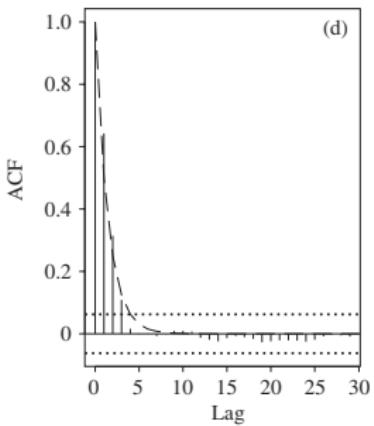
(a)



(b)



(c)



(d)

- a) $n = 1000$ realizations of an $\text{ARCH}(1)$ 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 $(X_t^2)_{t \in \mathbb{Z}}$ ($\text{AR}(1)$); 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

$$X_t = \sigma_t Z_t,$$

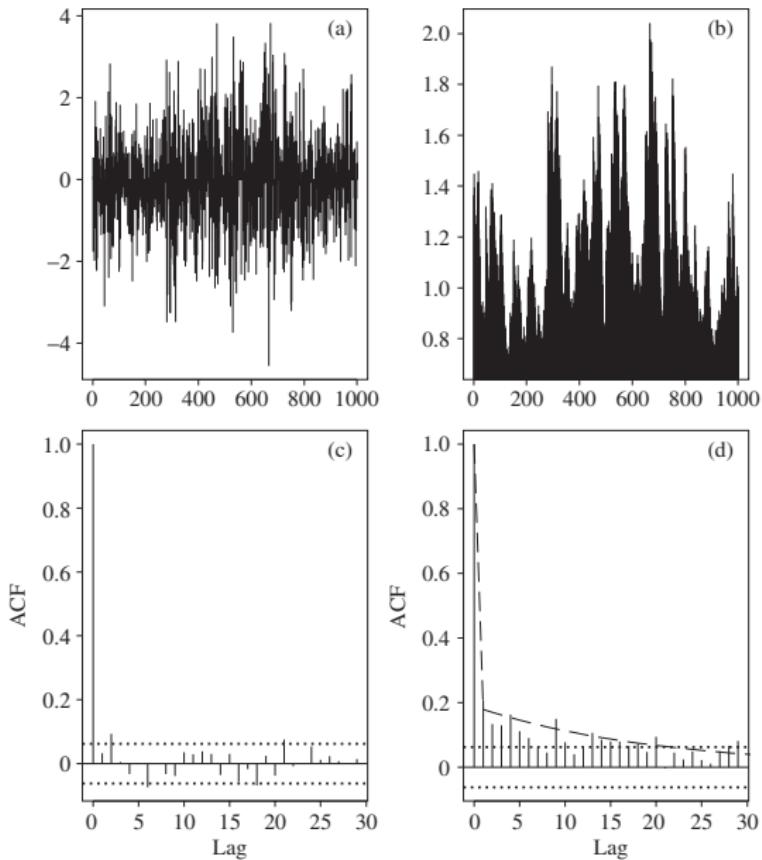
$$\sigma_t^2 = \alpha_0 + \sum_{k=1}^p \alpha_k X_{t-k}^2 + \sum_{k=1}^q \beta_k \sigma_{t-k}^2,$$

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, $\text{var}(X_t) = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}$.
- GARCH(1, 1) is typically leptokurtic:
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

$$X_t = \mu_t + \varepsilon_t \quad \text{for} \quad \varepsilon_t = \sigma_t Z_t \quad (\text{so } X_t = \mu_t + \sigma_t Z_t),$$

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

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

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, we check residuals. 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

$$\hat{Z}_t = \hat{\varepsilon}_t / \hat{\sigma}_t, \quad \hat{\sigma}_t^2 = \hat{\alpha}_0 + \sum_{k=1}^{p_2} \hat{\alpha}_k \hat{\varepsilon}_{t-k}^2 + \sum_{k=1}^{q_2} \hat{\beta}_k \hat{\sigma}_{t-k}^2; \quad (8)$$

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.

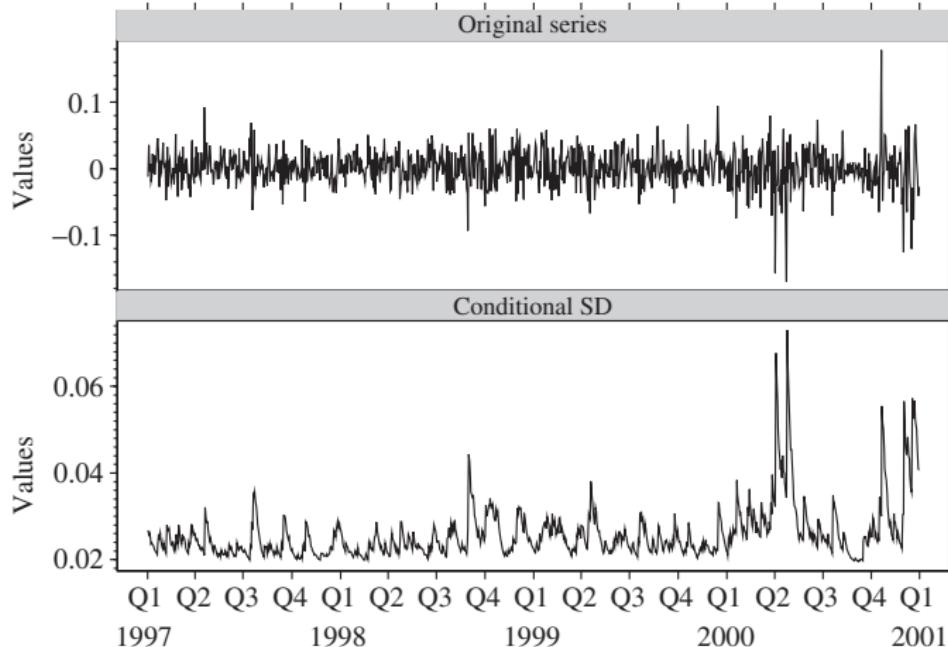
Drawbacks:

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

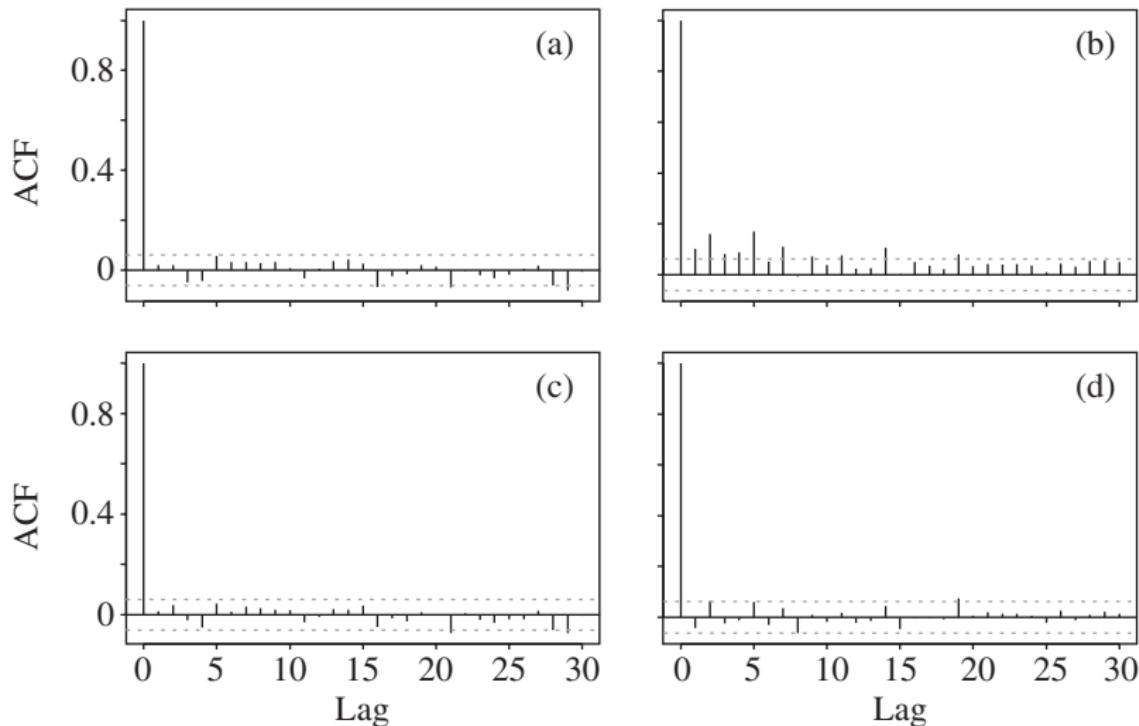
Example 4.21 (GARCH model for Microsoft log-returns)

- 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).

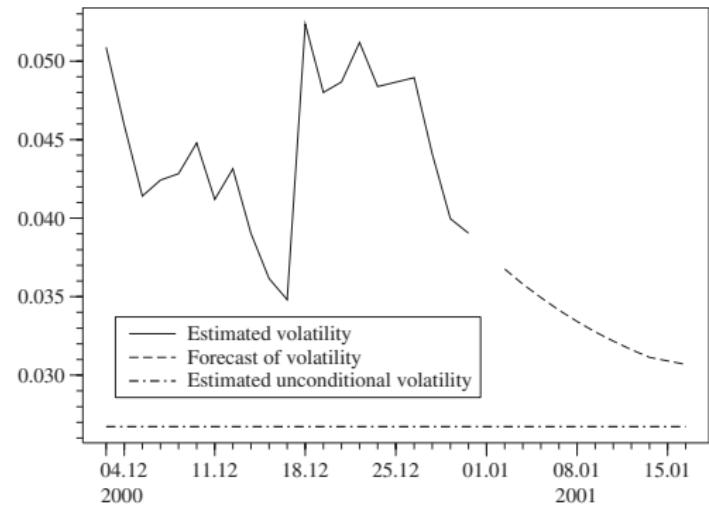
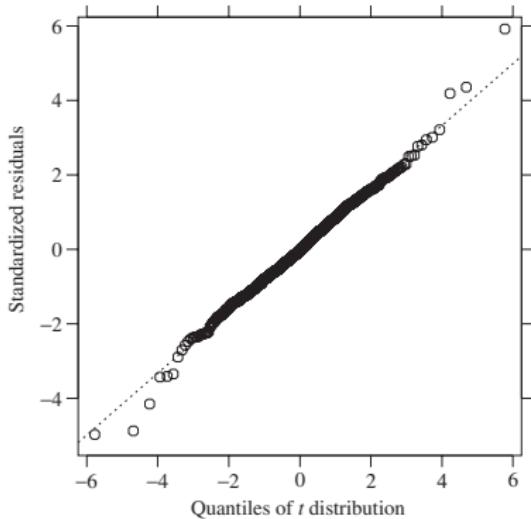
Microsoft log-returns 1997–2000: Data (top) and estimated volatility (bottom) from a GJR-GARCH(1, 1).



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 compute $P_t \sigma_{t+h}$, $h \geq 1$, a forecast of volatility based on these data.
- 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 $(X_{t+h} - \mu_{t+h})^2$.
- We consider two approaches: (1) calculating conditional expectations (optimal squared error forecasts) using model of GARCH type; (2) the more ad hoc exponentially weighted moving average (EWMA) approach.

Conditional expectation

The general procedure becomes clear from 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 (8).

Exponentially weighted moving averages

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

$$P_0 Y_1 = 0, \quad P_t Y_{t+1} = \alpha Y_t + (1 - \alpha) P_{t-1} Y_t, \quad t \geq 1. \quad (9)$$

With $Y_{t+1} = (X_{t+1} - \mu_{t+1})^2$ one obtains

$$P_t (X_{t+1} - \mu_{t+1})^2 = \alpha (X_t - \mu_t)^2 + (1 - \alpha) P_{t-1} (X_t - \mu_t)^2. \quad (10)$$

- Since $\sigma_{t+1}^2 = \mathbb{E}((X_{t+1} - \mu_{t+1})^2 | \mathcal{F}_t)$, we can use (10) 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 small (e.g. RiskMetrics: $\alpha = 0.06$); $\hat{\mu}_t$ is usually set to 0 (see Chapter 3).

Estimators of VaR_α and ES_α

- Suppose we now want to estimate VaR_α^t , ES_α^t , risk measures based on $F_{X_{t+1}|\mathcal{F}_t}$. If $Z_t \stackrel{\text{ind.}}{\sim} F_Z$, the \mathcal{F}_t -measurability of μ_{t+1} and σ_{t+1} , and $X_t = \mu_t + \sigma_t Z_t$ imply 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 have estimated σ_{t+1} (and μ_{t+1} ; often taken as 0) it remains 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 $\text{ES}_\alpha(Z)$.
 - 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 (worst-case losses). Let $(X_i)_{i \in \mathbb{N}} \stackrel{\text{ind.}}{\sim} F$ (can be relaxed to a strictly stationary time series) and F continuous. Then the *block maximum* is given by

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

One can show that $M_n \xrightarrow[n \rightarrow \infty]{\text{a.s.}} x_F$ (similar as in the SLLN; due to monotone convergence to a constant) where

$$x_F := \sup\{x \in \mathbb{R} : F(x) < 1\} = F^\leftarrow(1) \leq \infty$$

denotes the *right endpoint* of F .

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 ($F \in \text{MDA}(H)$).

The convergence to types theorem (see the appendix) guarantees that H is determined up to location/scale, i.e. H specifies a *unique type* of distribution.

Question: What does H look like?

Definition 5.2 (Generalized extreme value (GEV) distribution)

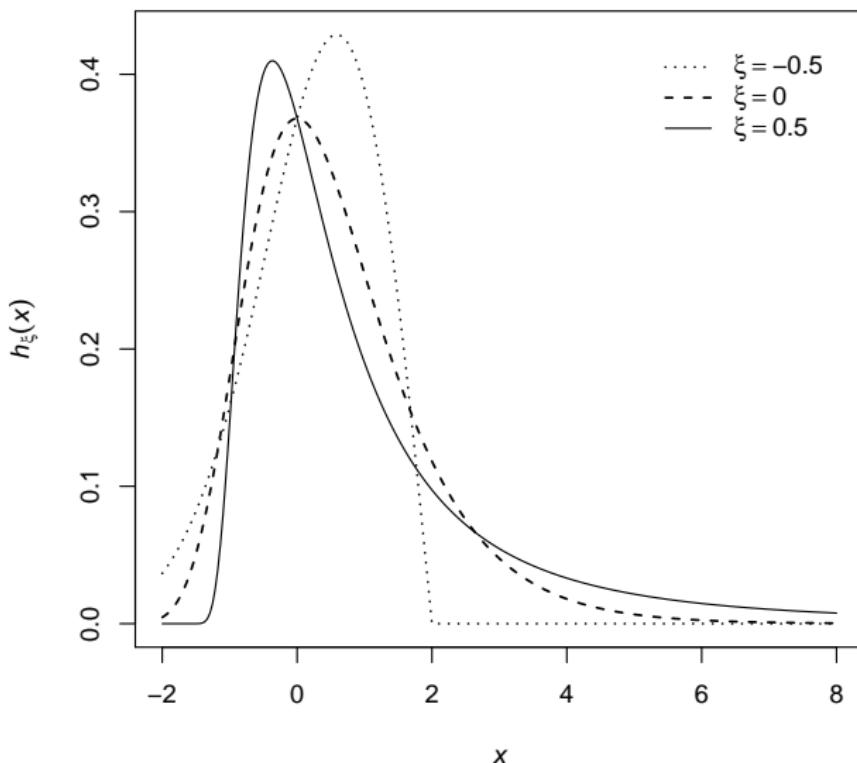
The (standard) *generalized extreme value (GEV) distribution* is given by

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

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 of iid random variables converge in distribution to a non-degenerate limit, the limiting distribution must be a location-scale transformed GEV distribution (that is, of GEV type).
- One can always choose normalizing sequences $(c_n) > 0$, (d_n) such that H_ξ appears in standard form (although from a statistical point of view, $(c_n) > 0$, (d_n) are simply estimated).
- All commonly encountered continuous distributions are in the MDA of some 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 $\xi \in \mathbb{R}$ and μ, σ can be estimated. But how can we characterize/determine ξ ? All $F \in \text{MDA}(H_\xi)$, $\xi > 0$, allow for a characterization based on:

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))

$F \in \text{MDA}(H_\xi)$ for $\xi > 0$ 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?

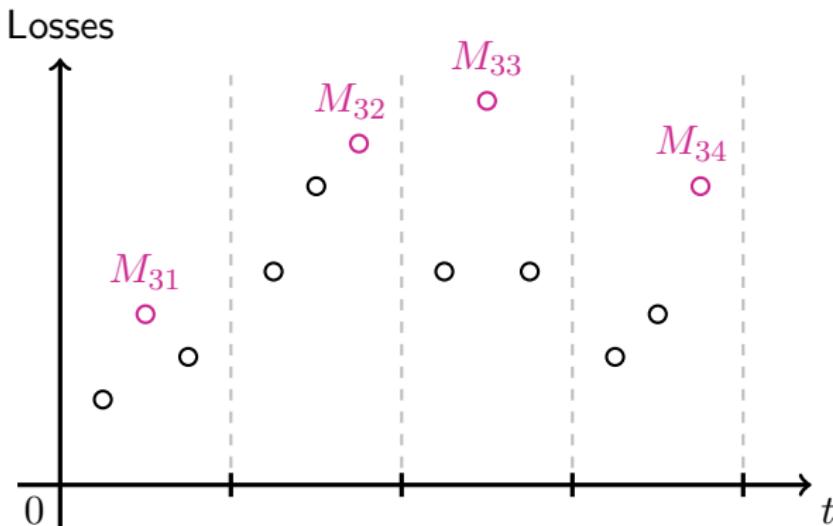
- Let $(X_k)_{k \in \mathbb{Z}}$ denote a strictly stationary time series with stationary distribution $X_k \sim F$, $k \in \mathbb{Z}$.

- 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 and H_ξ are of the same type, 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 $MDA(H_\xi)$, $\xi > 0$);
 - ▶ 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 $(X_i)_{i \in \mathbb{N}} \stackrel{\text{ind.}}{\sim} F \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$, where F is unknown. By Fisher–Tippett–Gnedenko Theorem,

$$\mathbb{P}(M_n \leq x) = \mathbb{P}((M_n - d_n)/c_n \leq (x - d_n)/c_n) \underset{n \text{ large}}{\approx} H_{\xi, \mu=d_n, \sigma=c_n}(x).$$

- 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

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

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 for finding 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 return periods

(Approximately) $M_n \sim H_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}$, so $\bar{H}_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}(r) = \mathbb{P}(M_n > r) = 1/k$ can be used to estimate the...

1) ... *k n-block return level* $r_{n,k}$, that is, the (smallest) r which is expected to be exceeded (at most) in one out of every k blocks of size n .

- e.g., 10 year return level $r_{260,10} = (\text{smallest})$ level exceeded in (at most) one out of every 10 years (where 260d \approx 1y)
- $r_{n,k} = H_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}^{\leftarrow}(1 - 1/k) = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}((-\log(1 - 1/k))^{-\hat{\xi}} - 1)$

2) ... *return period* $k_{n,u}$ of the event $\{M_n > u\}$, that is, the smallest number of n -blocks for which we expect to see at least one n -block exceeding u .

- $k_{n,u} = 1/\bar{H}_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}(u) = 1/(1 - \exp(-(1 + \hat{\xi}(x - \hat{\mu})/\hat{\sigma})^{-1/\hat{\xi}}))$
- $k_{n,u}$ satisfies $r_{n,k_{n,u}} = u$

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? $1 - H_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}(\text{"previous maxima"})$

- Was a risk-factor change as on 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

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

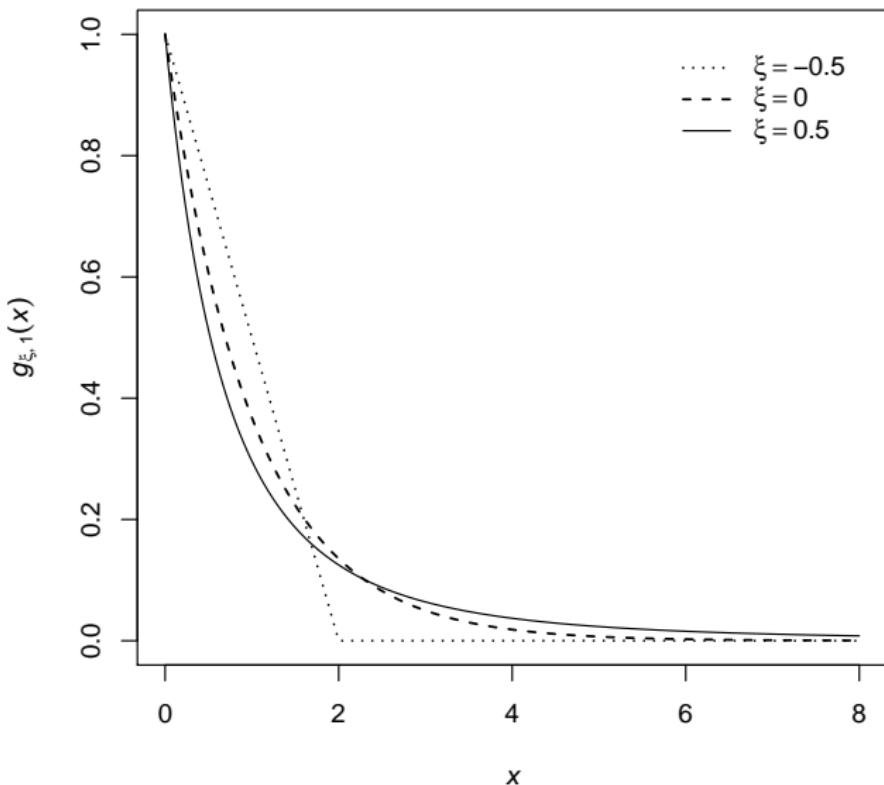
where $\beta > 0$, and the support is $x \geq 0$ when $\xi \geq 0$ and $x \in [0, -\beta/\xi]$ when $\xi < 0$.

- The parameterization is continuous in ξ .
- 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 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 $\mathbb{E}X = \beta/(1 - \xi)$).
- $G_{\xi,\beta} \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$, (same ξ)
- 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!).

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 is the distribution of the excess loss $X - u$ over u , given that $X > u$. $e(u)$ is the mean of F_u as a function of u .
- One can show the useful formulas $e(u) = \int_0^{x_F-u} \bar{F}_u(x) dx = \frac{\int_u^{x_F} \bar{F}(x) dx}{\bar{F}(u)}$.
- For continuous $X \sim F$ with $\mathbb{E}|X| < \infty$, the following formula holds:

$$\text{ES}_\alpha(X) = e(\text{VaR}_\alpha(X)) + \text{VaR}_\alpha(X), \quad \alpha \in (0, 1) \quad (11)$$

Example 5.14 (F_u , $e(u)$ for $\text{Exp}(\lambda)$, $G_{\xi,\beta}$)

- 1) If F is $\text{Exp}(\lambda)$, then $F_u(x) = 1 - e^{-\lambda x}$ (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 $F_u(x) = G_{\xi,\beta+\xi u}(x)$ (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 $\xi/(1 - \xi)$ 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 $F \in \text{MDA}(H_\xi)$, $\xi \in \mathbb{R}$.

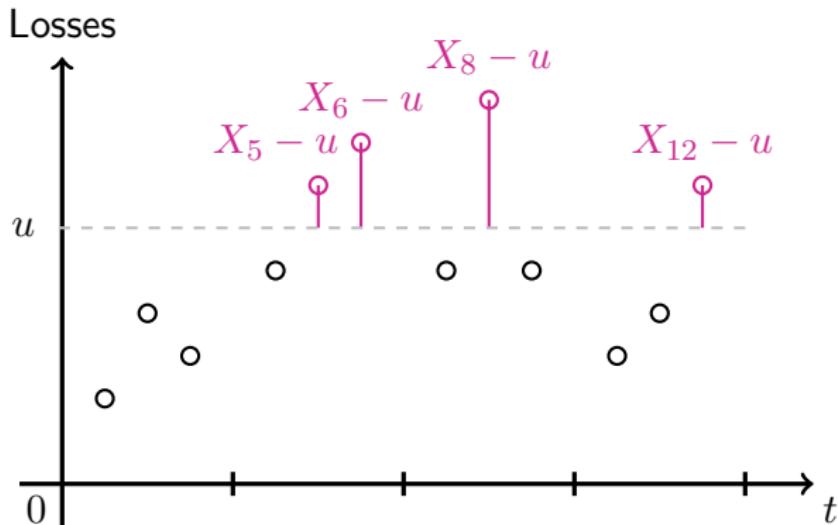
Proof. Non-trivial; see, e.g. Pickands (1975) and Balkema and de Haan (1974). \square

Interpretation

- The GPD is the canonical df for excess losses over high u . This leads to the peaks-over-threshold method for modeling excess losses.
- 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 when 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 peaks-over-threshold (POT) 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 *independent* 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

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

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

$$e(v) = \mathbb{E}(G_{\xi, \beta + \xi(v-u)}) = \frac{\beta + \xi(v-u)}{1 - \xi} = \frac{\xi}{1 - \xi}v + \frac{\beta - \xi u}{1 - \xi}, \quad v \in [u, x_F], \quad (12)$$

where $x_F = \infty$ if $\xi \in [0, 1)$ and $x_F = u - \beta/\xi$ 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)

For $X_1, \dots, X_n > 0$, the *sample mean excess function* is defined by

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

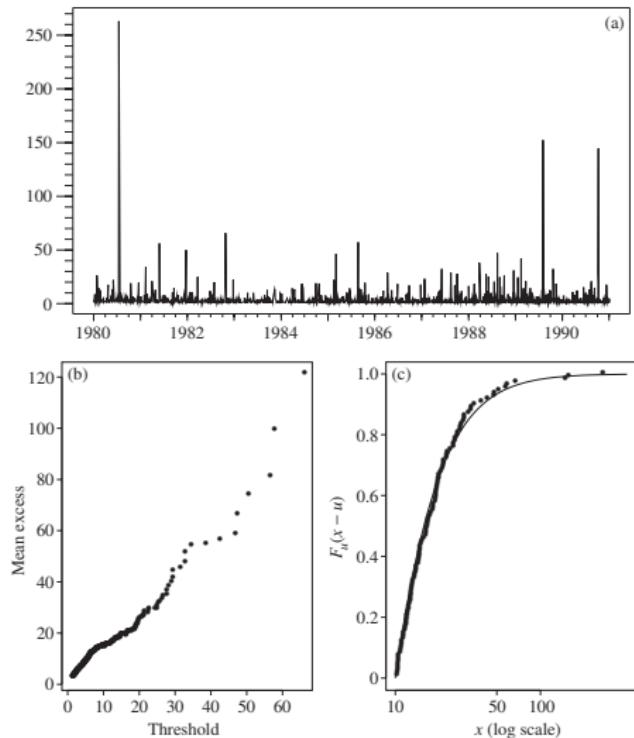
The *mean excess plot* is the plot of $\{(X_{(i)}, e_n(X_{(i)})) : 1 \leq i \leq n-1\}$, where $X_{(i)}$ denotes the i th order statistic.

- 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 .

Example 5.18 (Danish fire loss data)

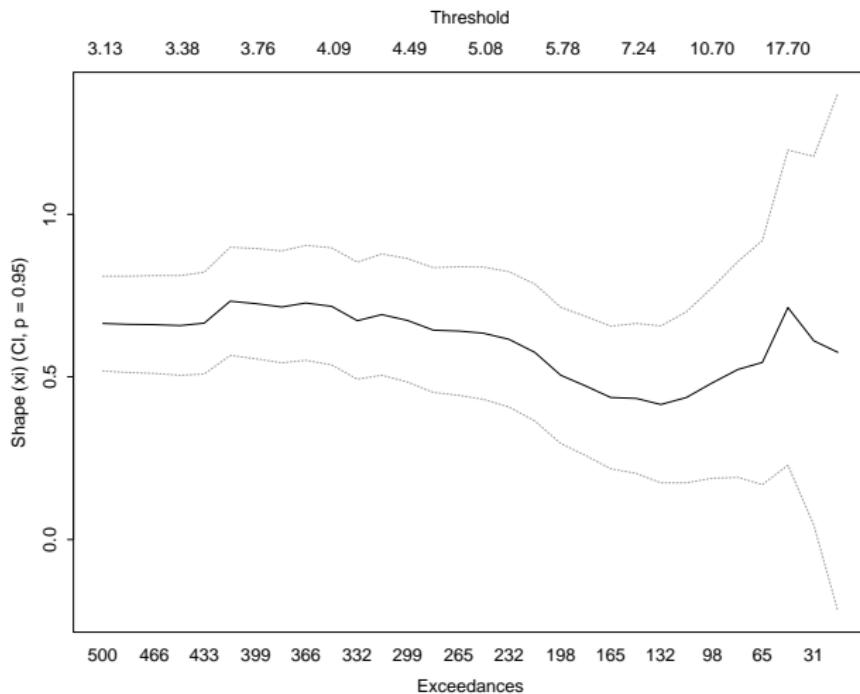
- 2156 fire insurance losses over 1M Danish kroner from 1980-01-03 to 1990-12-31; combined loss for a building and its contents, in some cases also a loss of business earnings. The losses are inflation adjusted to reflect values as of 1985.
- The mean excess function shows a “kink” below 10; “straightening 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 (12) 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) $\hat{F}_{u,n}(x - u)$, $G_{\hat{\xi},\hat{\beta}}(x - u)$



⇒ 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 $u = 10$).

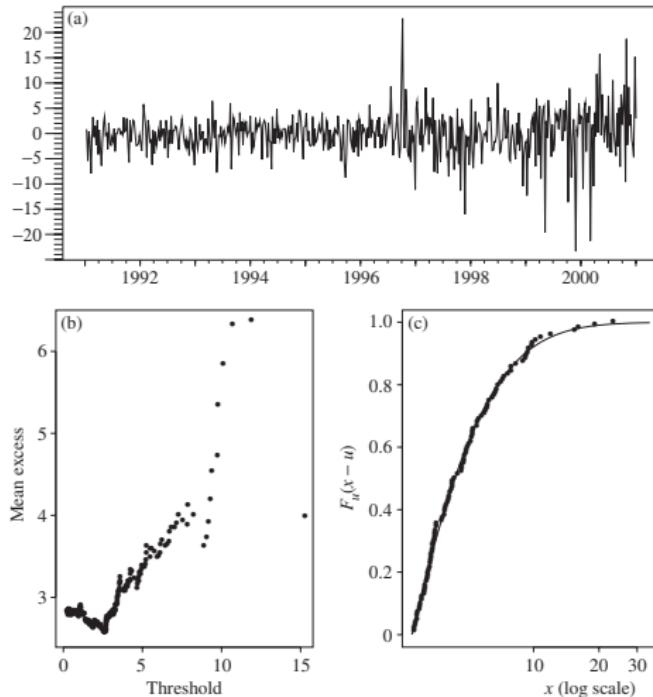
Example 5.19 (AT&T weekly loss data)

- 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): $\hat{F}_{u,n}(x - u)$, $G_{\hat{\xi}, \hat{\beta}}(x - u)$.



⇒ 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:

$$\begin{aligned}\bar{F}(x) &= \mathbb{P}(X > x) = \mathbb{P}(X > u)\mathbb{P}(X > x | X > u) \\ &= \bar{F}(u)\mathbb{P}(X - u > x - u | X > u) = \bar{F}(u)\bar{F}_u(x - u) \\ &= \bar{F}(u)\left(1 + \xi \frac{x - u}{\beta}\right)^{-1/\xi}, \quad x \geq u.\end{aligned}\tag{13}$$

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

$$\text{VaR}_\alpha = F^\leftarrow(\alpha) = u + \frac{\beta}{\xi} \left(\left(\frac{1-\alpha}{\bar{F}(u)} \right)^{-\xi} - 1 \right), \tag{14}$$

$$\text{ES}_\alpha = \frac{\text{VaR}_\alpha}{1-\xi} + \frac{\beta - \xi u}{1-\xi}, \quad \xi < 1. \tag{15}$$

The formula for ES_α can also be obtained from $e(\cdot)$ via (11) and (12).

- $\bar{F}(x)$, 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*

$$\hat{F}(x) = \frac{N_u}{n} \left(1 + \hat{\xi} \frac{x - u}{\hat{\beta}} \right)^{-1/\hat{\xi}}, \quad x \geq u \quad (\text{see (13)});$$

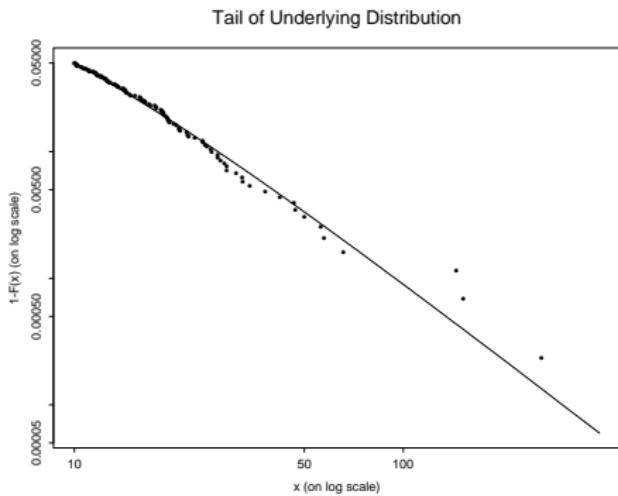
also known as the *Smith estimator* (note that it is only valid for $x \geq u$). It faces a **bias-variance tradeoff**: If u is increased, the bias of parametrically estimating $\bar{F}_u(x - u)$ decreases, but the variance of it and the nonparametrically estimated $\bar{F}(u)$ increases.

- Similarly, semi-parametric GPD-based $\widehat{\text{VaR}}_\alpha$, $\widehat{\text{ES}}_\alpha$ for $\alpha \geq 1 - N_u/n$ can be obtained from (14), (15).

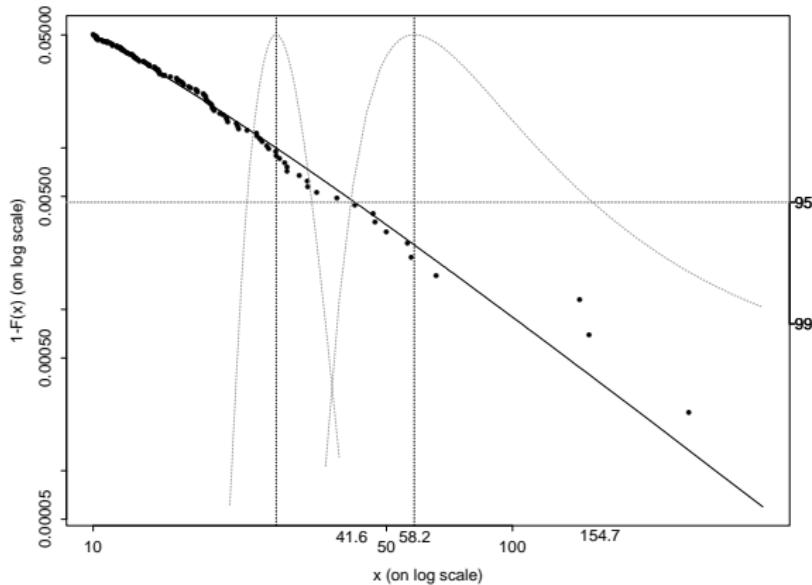
- Confidence intervals for $\bar{F}(x)$, $x \geq u$, VaR_α , ES_α can be obtained likelihood-based (neglecting the uncertainty in N_u/n): Reparametrize the GPD model in terms of $\phi = g(\xi, \beta, N_u/n)$ and construct a confidence interval for ϕ based on the likelihood ratio test.

Example 5.20 (Danish fire loss data (continued))

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

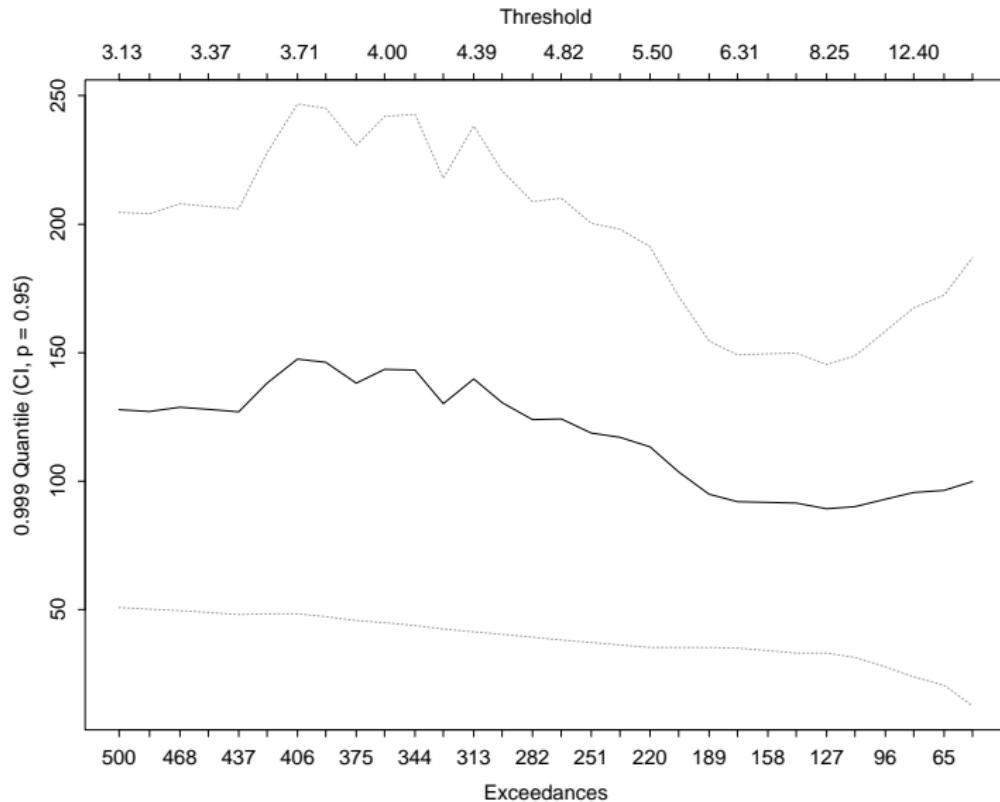


Here are $\hat{F}(x)$, $x \geq u$, $\widehat{\text{VaR}}_{0.99}$, $\widehat{\text{ES}}_{0.99}$ including confidence intervals.

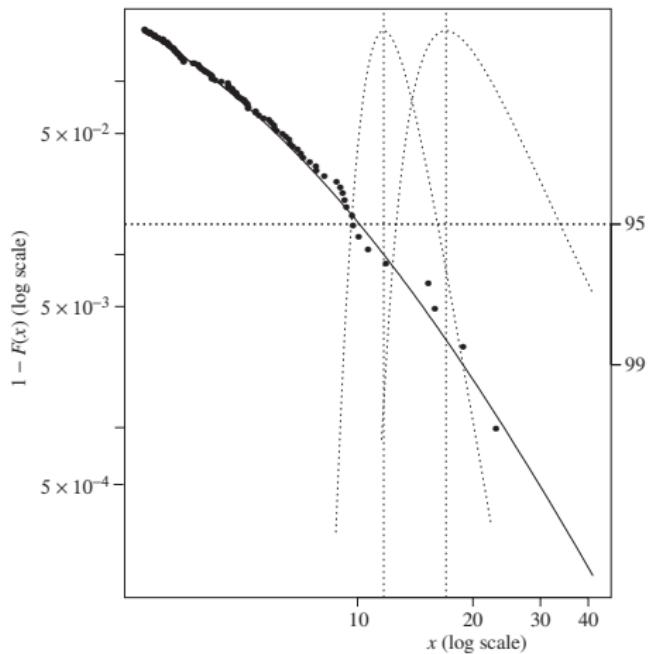


Log-log scale often helpful: If $\bar{F}(x) = x^{-\alpha} L(x)$, $\log \bar{F}(x) = -\alpha \log(x) + \log L(x)$ which is approximately linear in $\log x$.

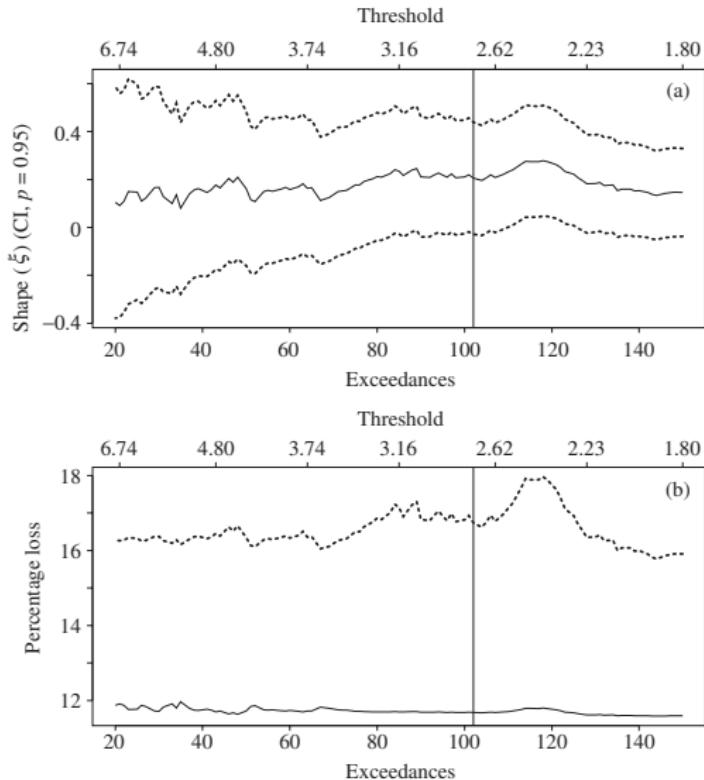
It is important to check the sensitivity of \hat{F} (or $\widehat{\text{VaR}}_\alpha$, $\widehat{\text{ES}}_\alpha$) w.r.t. u .



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



- Fitted GPD model as in Example 5.19.
- Plot of $\widehat{F}(x)$.
- 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)

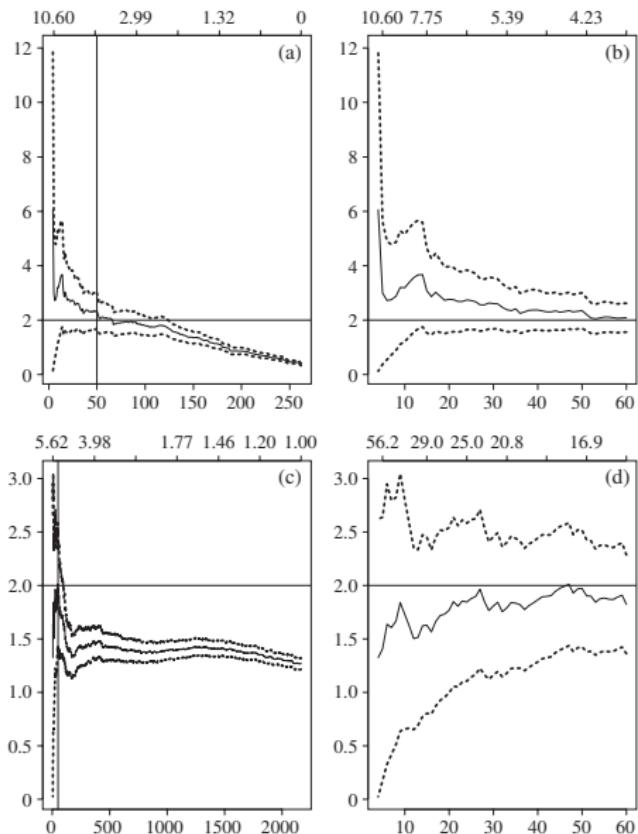
5.2.4 The Hill estimator

- 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

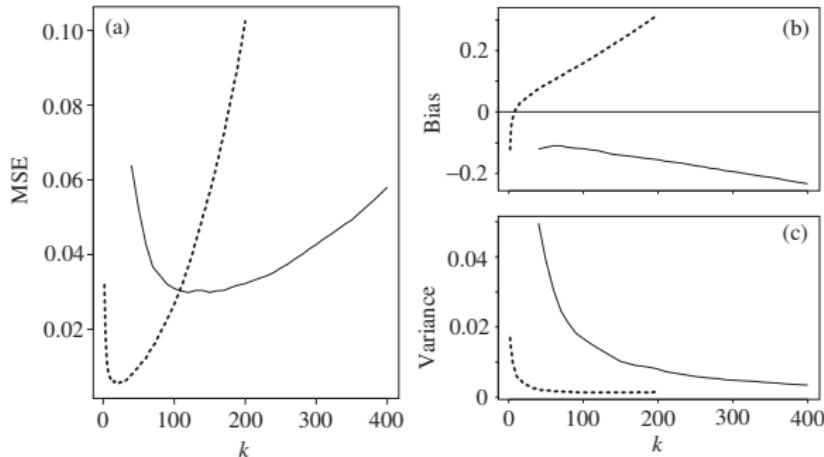
$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \theta)^2) = \mathbb{E}((\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}(\hat{\theta}) - \theta)^2) \\ &= \mathbb{E}((\hat{\theta} - \mathbb{E}[\hat{\theta}])^2) + \mathbb{E}(2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}(\hat{\theta})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

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)

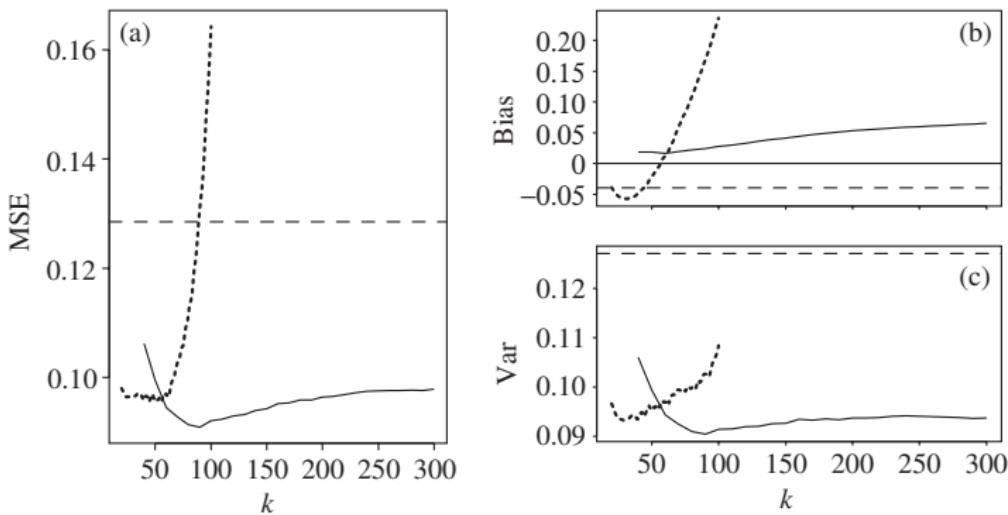


- The Hill estimator outperforms the GPD estimator (optimal k around 20–30) according to the variance for small k (number of order statistics)
- The biases are closer: the Hill (GPD) estimator tends to overestimate (underestimate) ξ .
- For the GPD method, the optimal u is around 100–150 exceedances.

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 VaR_{0.99} estimator has a negative bias.
- The Hill VaR_{0.99} estimator has a negative bias for small k but a rapidly growing positive bias for larger k .
- The GPD VaR_{0.99} estimator has a positive bias which grows much more slowly.
- The GPD VaR_{0.99} estimator attains lowest MSE for a value of k around 100, and the MSE is very robust to the choice of k (because of the slow growth of the bias) ⇒ Choice of u less critical
- The Hill VaR_{0.99} estimator performs well for $20 \leq k \leq 75$ (we only use k values that lead to a quantile estimate beyond the effective threshold $X_{k,n}$) 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 (14)) and $\text{ES}_{\alpha}(Z)$ (see (15)).

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 *j th margin F_j of F* or *j th marginal df F_j of \mathbf{X}* is

$$F_j(x_j) = \mathbb{P}(X_j \leq x_j)$$

$$= \mathbb{P}(X_1 \leq \infty, \dots, X_{j-1} \leq \infty, X_j \leq x_j, X_{j+1} \leq \infty, \dots, X_d \leq \infty)$$

$$= F(\infty, \dots, \infty, x_j, \infty, \dots, \infty), \quad x_j \in \mathbb{R}, \quad j \in \{1, \dots, d\}.$$

(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

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

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

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

- 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 singular **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 \mathbf{X} in the form of a joint df, survival function or density, implicitly describes the dependence of X_1, \dots, X_d . 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

$$F(\mathbf{x}) = \int_{(-\infty, \mathbf{x}_1]} F_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{z}) dF_{\mathbf{X}_1}(\mathbf{z}); \quad (16)$$

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

$$\begin{aligned} f(\mathbf{x}_1, \mathbf{x}_2) &= \frac{\partial^2}{\partial \mathbf{x}_2 \partial \mathbf{x}_1} F(\mathbf{x}_1, \mathbf{x}_2) \stackrel{(16)}{=} \frac{\partial}{\partial \mathbf{x}_2} F_{X_2|X_1}(\mathbf{x}_2 | \mathbf{x}_1) f_{X_1}(\mathbf{x}_1) \\ &= f_{X_2|X_1}(\mathbf{x}_2 | \mathbf{x}_1) f_{X_1}(\mathbf{x}_1). \end{aligned}$$

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

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

- 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} (if F has density f , then X_1, \dots, X_d are independent 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-example: $X_1 \sim U(-1, 1)$, $X_2 = X_1^2 \Rightarrow \text{cov}(X_1, X_2) = \mathbb{E}(X_1^3) - 0 \cdot \mathbb{E}(X_1^2) = 0$).
- 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 $a \neq 0$ and $b \in \mathbb{R}$.

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

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 $k = 1$ ($A = \mathbf{a}'$),
 $\mathbf{a}'\Sigma\mathbf{a} = \text{cov}(\mathbf{a}'\mathbf{X}) = \text{var}(\mathbf{a}'\mathbf{X}) \geq 0, \quad \mathbf{a} \in \mathbb{R}^d,$ (17)

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

$$\Sigma = AA' \quad \text{Cholesky decomposition} \quad (18)$$

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

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

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

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 positive semidefinite.

Proof.

“ \Rightarrow ” As we have seen in (17), 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) are serially uncorrelated (i.e. multivariate white noise) with $\mu := \mathbb{E}\mathbf{X}_1$, $\Sigma := \text{cov } \mathbf{X}_1$ and $P = \text{corr}(\mathbf{X}_1)$.

- Standard 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

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

- Further properties of \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

$$\mathbf{X} \stackrel{\text{d}}{=} \boldsymbol{\mu} + A\mathbf{Z}, \quad (19)$$

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 (19) and $\Sigma = AA'$. Then the cf of \mathbf{X} is

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

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

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

□

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

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

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

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

□

Consequences:

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

Proposition 6.5 (Density)

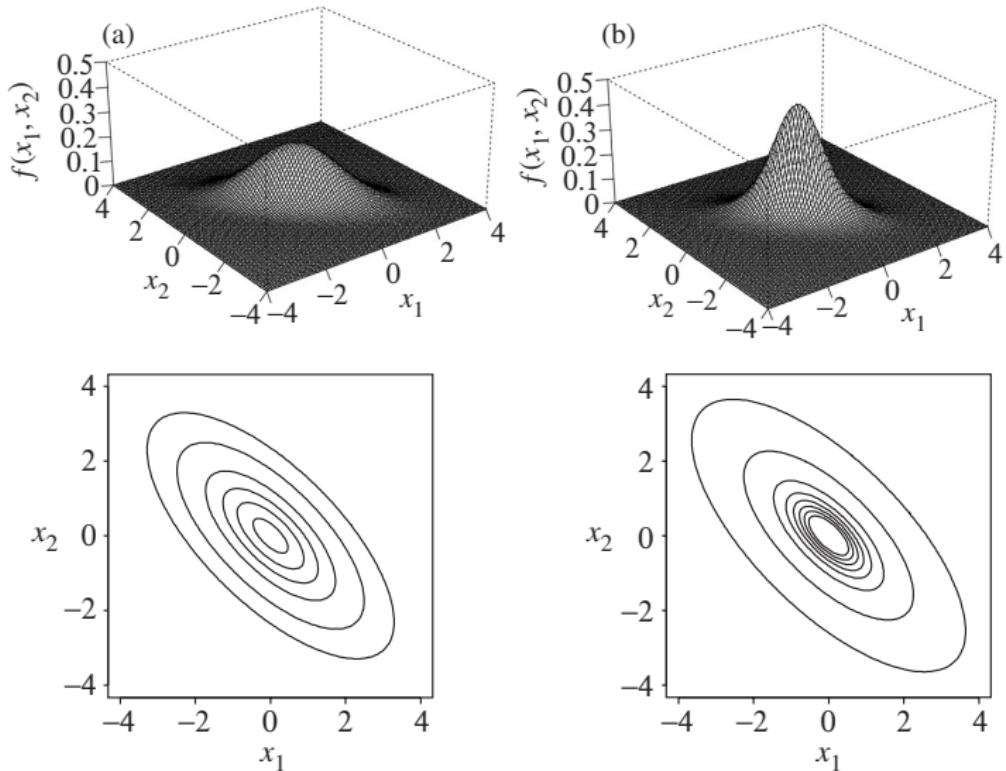
Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ with $\text{rank } A = k = d$ ($\Rightarrow \Sigma$ pos. definite, invertible).

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: $N_d(\boldsymbol{\mu}, \Sigma)$ for $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix}$; Right: $t_\nu(\boldsymbol{\mu}, \frac{\nu-2}{\nu}\Sigma)$, $\nu = 4$,
(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

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

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

- **Marginal dfs**

Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ 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 $\boldsymbol{\mu} = (\boldsymbol{\mu}'_1, \boldsymbol{\mu}'_2)$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then

$$\mathbf{X}_1 \sim N_k(\boldsymbol{\mu}_1, \Sigma_{11}) \quad \text{and} \quad \mathbf{X}_2 \sim N_{d-k}(\boldsymbol{\mu}_2, \Sigma_{22}).$$

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 Σ be positive definite. One can show that

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

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

- **Quadratic forms**

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

$$(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Z}' \mathbf{Z} \sim \chi_d^2, \quad (20)$$

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

■ Convolutions

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

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

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,

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

This can be tested statistically (for some \mathbf{a}) with various goodness-of-fit tests (e.g. Q-Q plots) used 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, (20) can be used to test joint normality (see Mardia's test below).

- Multivariate Shapiro–Wilk
- Mardia's test

- ▶ According to (20), if $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ with Σ positive definite, then $(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_d^2$ (can approx. be used in a Q-Q plot).
- ▶ 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.

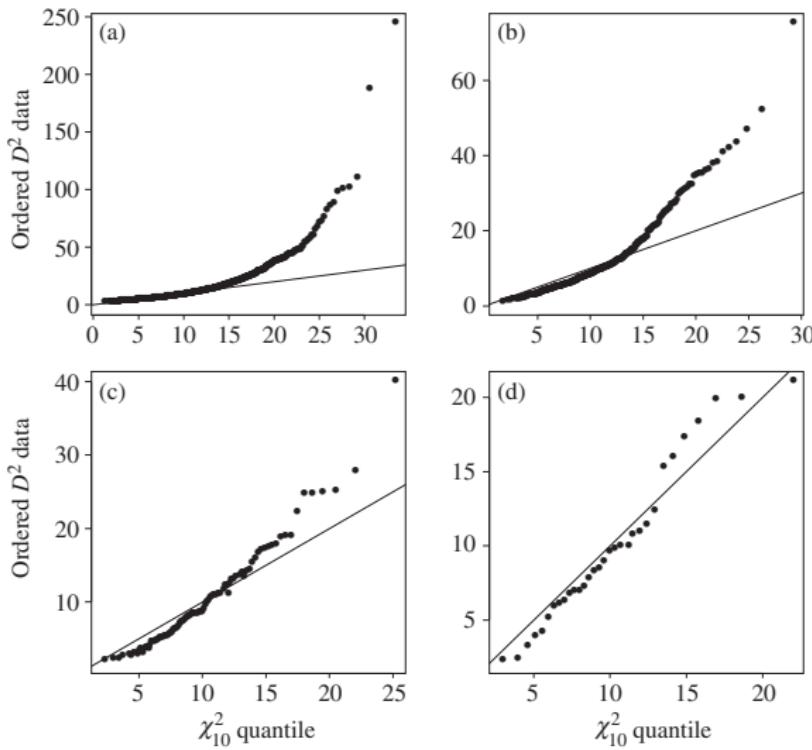
n	Daily	Weekly	Monthly	Quarterly
	2020	416	96	32
b_{10}	9.31	9.91	21.10	50.10
p-value	0.00	0.00	0.00	0.02
k_{10}	242.45	177.04	142.65	120.83
p-value	0.00	0.00	0.00	0.44

- We can also compare D_i^2 data to a χ_{10}^2 graphically using a Q-Q plot.

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:

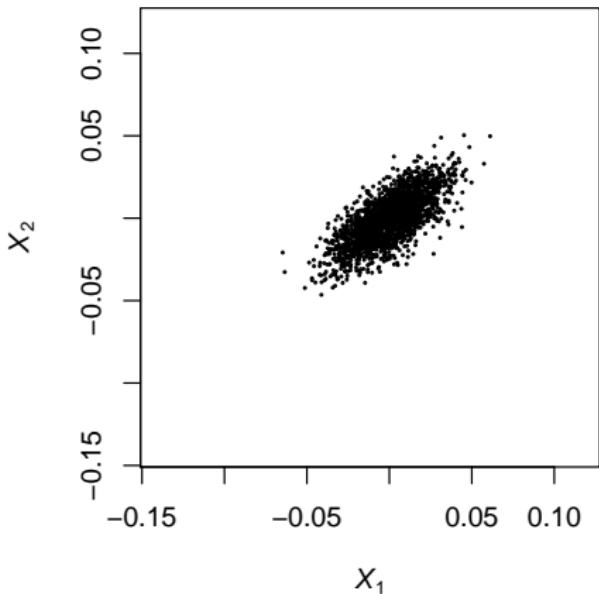
(a) daily data; (b) weekly data; (c) monthly data; and (d) quarterly data



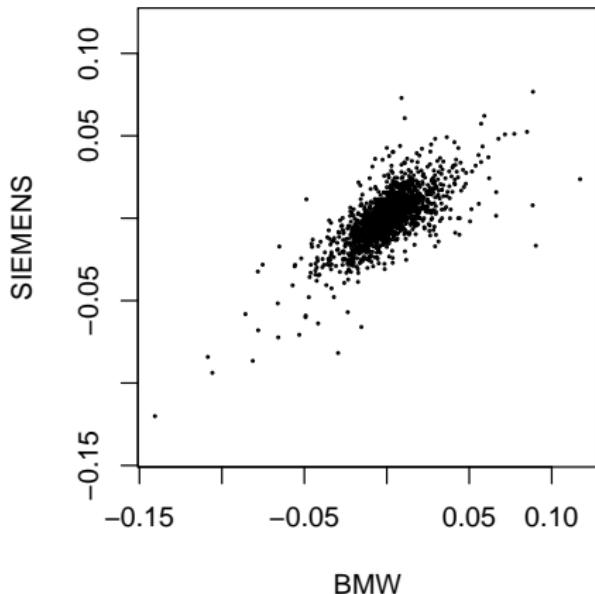
Example 6.8 (Simulated data vs BMW–Siemens)

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

Simulated data (fitted multivariate normal)

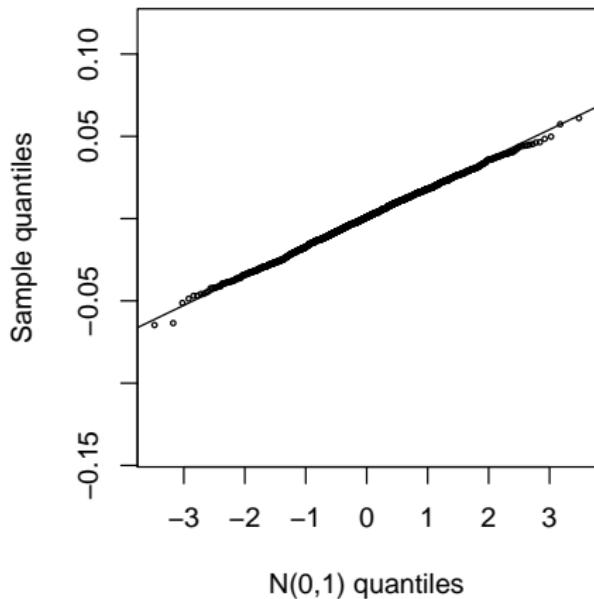


Real risk-factor changes

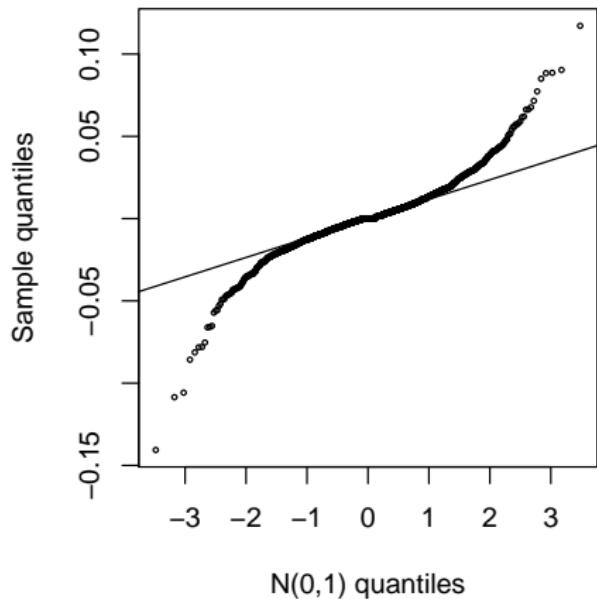


Considering the first margin only:

Q-Q plot for margin 1 (simulated data)

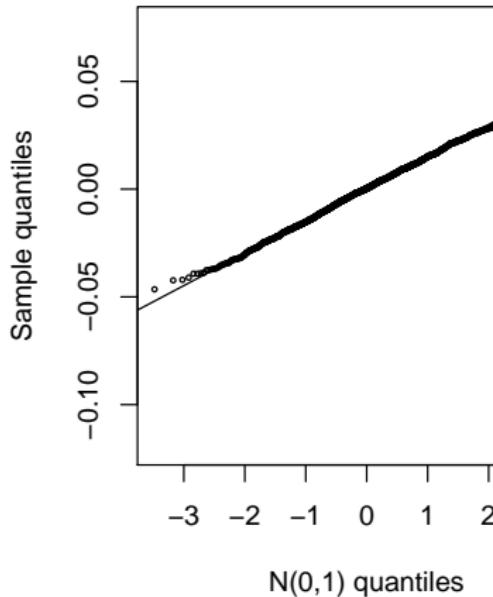


Q-Q plot for margin 1 (real data)

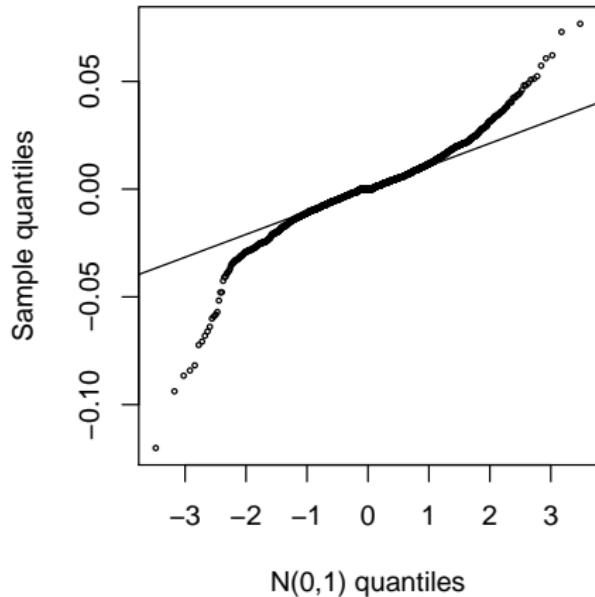


Considering the second margin only:

Q-Q plot for margin 2 (simulated data)

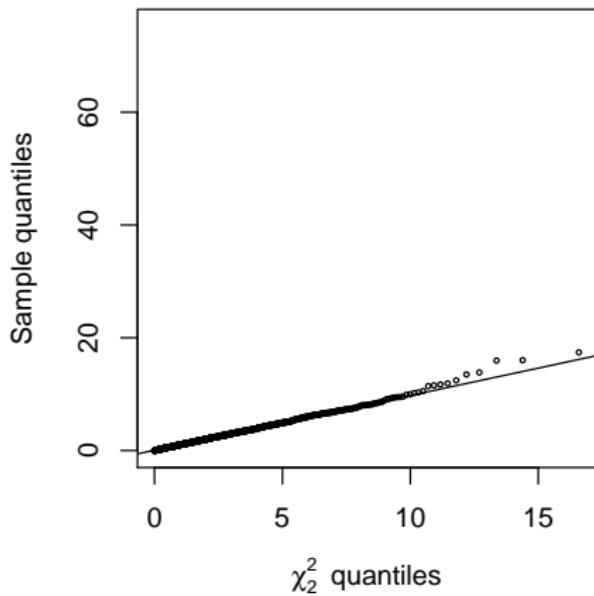


Q-Q plot for margin 2 (real data)

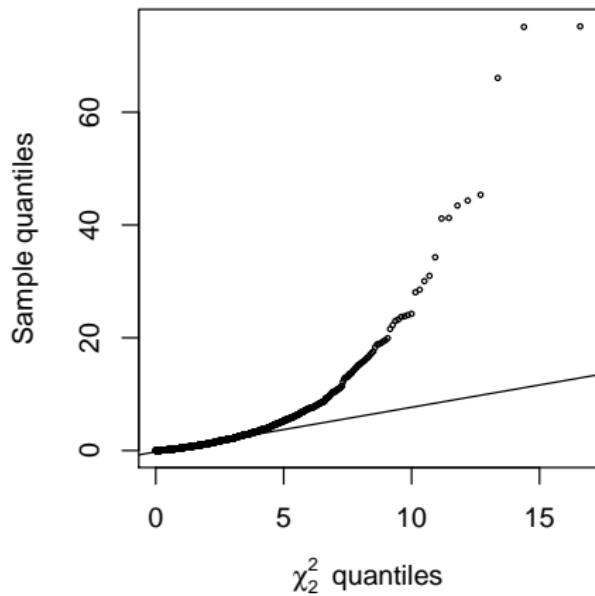


Q-Q plot of the simulated (left) or real (right) D_i^2 's against a χ_2^2 :

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 mixtures (or, more general, elliptical distributions can address 1) and 2) while sharing many of the desirable properties of $N_d(\mu, \Sigma)$.
- Normal mean-variance mixtures 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 possibly 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

$$X \stackrel{d}{=} \mu + \sqrt{W} A Z, \quad (21)$$

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

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

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

- However, if they exist (i.e. if $\mathbb{E}W < \infty$) $\text{corr}(\mathbf{X}) = \text{corr}(\mathbf{Y})$ since

$$\begin{aligned}\text{corr}(X_i, X_j) &= \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}(X_i) \text{var}(X_j)}} = \frac{\mathbb{E}(W)\Sigma_{ij}}{\sqrt{\mathbb{E}(W)\Sigma_{ii}\mathbb{E}(W)\Sigma_{jj}}} \\ &= \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}} = \text{corr}(Y_i, Y_j), \quad i, j \in \{1, \dots, d\}.\end{aligned}$$

Lemma 6.10 (Independence in normal variance mixtures)

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

X_i and X_j are independent $\iff W$ is a.s. constant (i.e. \mathbf{X} is normal).

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

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

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

- **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. (see stoch. representation).

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

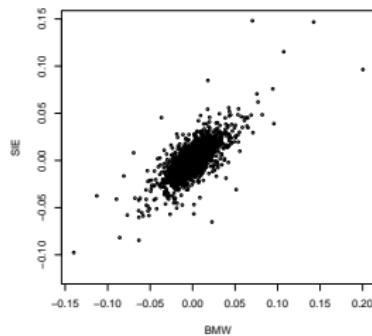
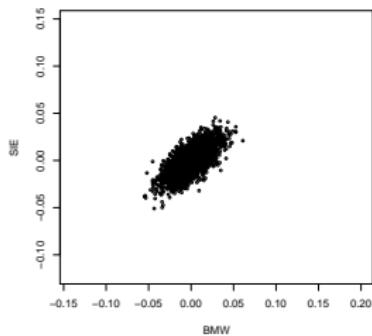
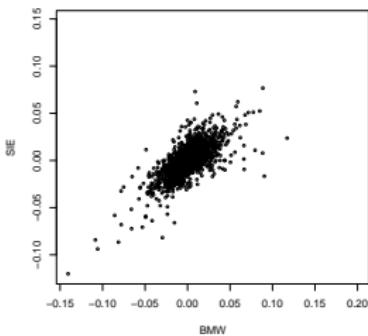
► $\mathbb{E}(W) = \frac{\nu}{\nu-2} \Rightarrow \text{cov } (\mathbf{X}) = \frac{\nu}{\nu-2} \boldsymbol{\Sigma}$. For finite variances/correlations, $\nu > 2$ is required. For finite mean, $\nu > 1$ is required.

- 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

$$\mathbf{X} \stackrel{d}{=} \mathbf{m}(W) + \sqrt{W} A \mathbf{Z}, \quad (22)$$

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 $\mathbf{X} | W = w \sim N_d(\boldsymbol{m}(w), w\Sigma)$. In general, **they are no longer elliptical** (see later).

Example 6.13

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

$$\mathbb{E}(\mathbf{X} | W) = \boldsymbol{\mu} + W\boldsymbol{\gamma},$$

$$\text{cov}(\mathbf{X} | W) = W\Sigma$$

we have

$$\mathbb{E}\mathbf{X} = \mathbb{E}(\mathbb{E}(\mathbf{X} | W)) = \boldsymbol{\mu} + \mathbb{E}(W)\boldsymbol{\gamma} \quad \text{if } \mathbb{E}W < \infty,$$

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

- 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 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 $UU' = U'U = I_d$)

$\mathbf{Y} \stackrel{d}{=} U\mathbf{Y}$ (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 $\phi_{\mathbf{Y}}(\mathbf{t}) = \mathbb{E}(e^{it'Y}) = \psi(\|\mathbf{t}\|^2)$, $\forall \mathbf{t} \in \mathbb{R}^d$.
- 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)

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

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

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

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

$$I_d = \text{cov } \mathbf{Y} = \text{cov}_{\text{Th. 6.16}}(R\mathbf{S}) = \mathbb{E}(R^2) \text{cov } \mathbf{S} = d \text{cov } \mathbf{S} \Rightarrow \text{cov } \mathbf{S} = I_d/d. \quad (23)$$

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

$$\text{cov } \mathbf{Y} = \text{cov}_{\text{Th. 6.16}}(R\mathbf{S}) = \mathbb{E}(R^2) \text{cov } \mathbf{S} = \frac{\mathbb{E}(R^2)}{d} I_d$$

and thus $\text{corr } \mathbf{Y} = \frac{(\mathbb{E}(R^2)/d)I_d}{\sqrt{(\mathbb{E}(R^2)/d)(\mathbb{E}(R^2)/d)}} = I_d$.

- For $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Y}$ with $\mathbb{E}(R^2) < \infty$ and Cholesky factor A of a covariance matrix Σ , we have $\text{cov } \mathbf{X} = \frac{\mathbb{E}(R^2)}{d} \Sigma$ and $\text{corr } \mathbf{X} = P$ (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,

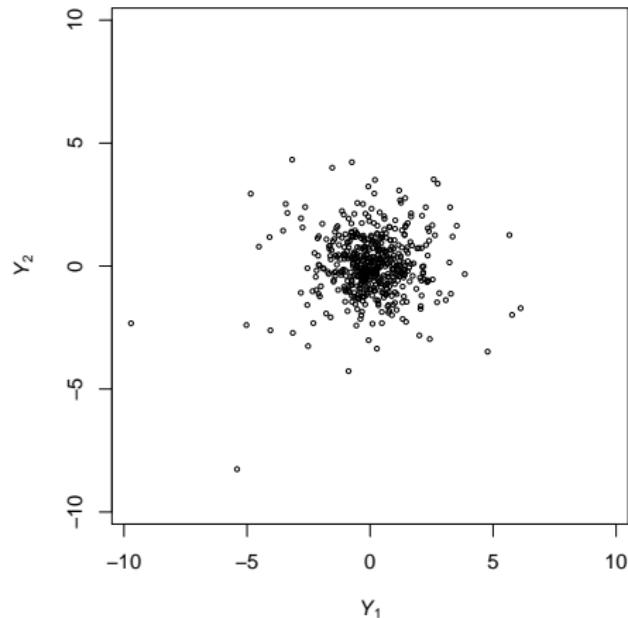
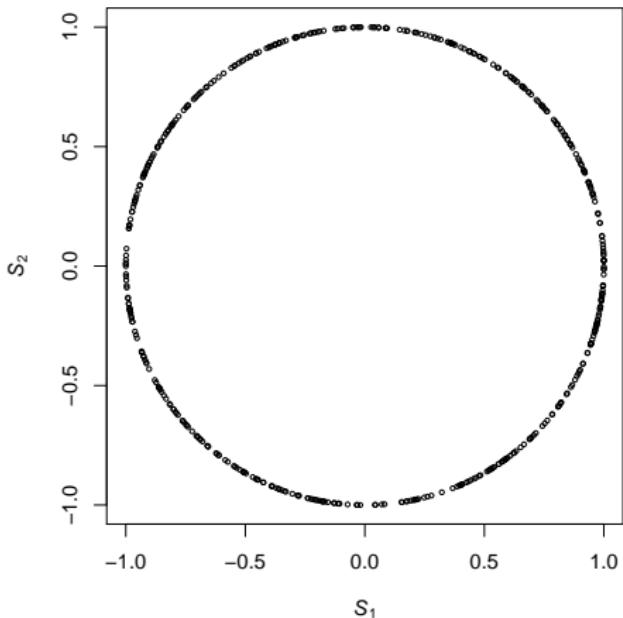
$$\frac{R^2}{d} = \frac{\mathbf{Z}'\mathbf{Z}/d}{(\nu/W)/\nu} = \frac{\chi_d^2/d}{\chi_\nu^2/\nu} \sim F(d, \nu)$$

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

- This, together with Example 6.19, implies that $\mathbf{X} \sim t_d(\nu, \mu, \Sigma)$ has $\text{cov } \mathbf{X} = \frac{\nu}{\nu-2}\Sigma$ and $\text{corr } \mathbf{X} = P$ (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

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

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** $\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + R\mathbf{S}$, with R and \mathbf{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 $\mathbf{Y} = A^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim S_d(\psi)$.

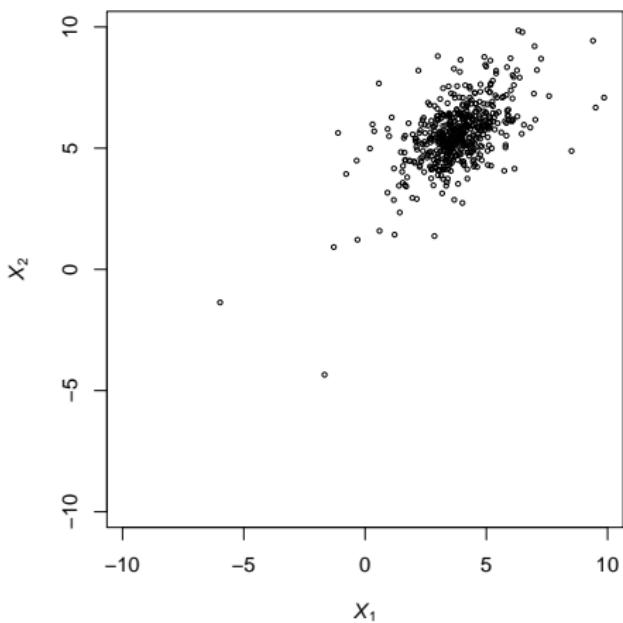
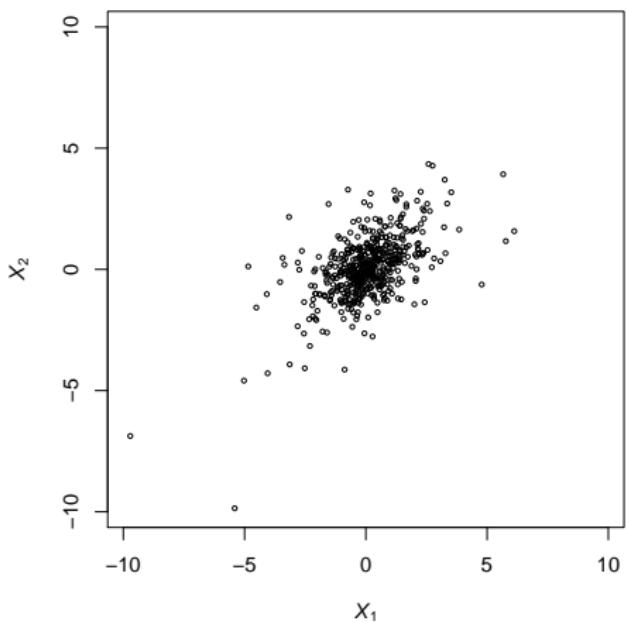
- Normal variance mixture distributions are elliptical (most useful examples) since $\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + \sqrt{W} A \mathbf{Z} = \boldsymbol{\mu} + \sqrt{W} \|\mathbf{Z}\| A \mathbf{Z} / \|\mathbf{Z}\| = \boldsymbol{\mu} + RAS$ with $R = \sqrt{W} \|\mathbf{Z}\|$ and $S = \mathbf{Z}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

$$\left(\sqrt{(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})}, \frac{A^{-1}(\mathbf{X} - \boldsymbol{\mu})}{\sqrt{(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})}} \right) \stackrel{d}{=} (R, S), \quad (24)$$

which can be used for testing elliptical symmetry.

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 $B = \mathbf{a}' \in \mathbb{R}^{1 \times d}$),

$$\mathbf{a}'\mathbf{X} \sim E_1(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\Sigma\mathbf{a}, \psi) \quad (\text{as for } N(\boldsymbol{\mu}, \Sigma)). \quad (25)$$

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:** (24) 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. 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,

$$\text{VaR}_\alpha(L) = \boldsymbol{\lambda}' \boldsymbol{\mu} + \|\boldsymbol{\lambda}' A\| \text{VaR}_\alpha(Y_1). \quad (26)$$

Applying (26) 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 $\text{VaR}_\alpha(\sum_{i=1}^n L_i) \stackrel{(26)}{=} \sum_{i=1}^n \boldsymbol{\lambda}'_i \boldsymbol{\mu} + \|\sum_{i=1}^n \boldsymbol{\lambda}'_i A\| \text{VaR}_\alpha(Y_1)$
 $\leq \sum_{i=1}^n \boldsymbol{\lambda}'_i \boldsymbol{\mu} + (\sum_{i=1}^n \|\boldsymbol{\lambda}'_i A\|) \text{VaR}_\alpha(Y_1) = \sum_{i=1}^n (\boldsymbol{\lambda}'_i \boldsymbol{\mu} + \|\boldsymbol{\lambda}'_i A\| \text{VaR}_\alpha(Y_1))$
 $\stackrel{(26)}{=} \sum_{i=1}^n \text{VaR}_\alpha(L_i)$. 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

$$\mathbf{X} = \mathbf{a} + B\mathbf{F} + \boldsymbol{\varepsilon}, \quad (27)$$

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

This is achieved by [time series regression](#).

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 .

Models can also be estimated simultaneously using [multivariate regression](#); see McNeil et al. (2015).

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 $Y_j = \gamma'_j(\mathbf{X} - \boldsymbol{\mu})$ 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 $\mathbf{Y} = \Gamma'(\mathbf{X} - \boldsymbol{\mu})$, we have

$$\mathbf{X} = \boldsymbol{\mu} + \Gamma\mathbf{Y} = \boldsymbol{\mu} + \Gamma_1\mathbf{Y}_1 + \Gamma_2\mathbf{Y}_2 =: \boldsymbol{\mu} + \Gamma_1\mathbf{Y}_1 + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}_1 \in \mathbb{R}^k$ contains the first k principal components. This is reminiscent of the basic factor model.

- Although $\varepsilon_1, \dots, \varepsilon_d$ will tend to have small variances, the assumptions of the factor model are generally violated (since they need not have a diagonal covariance matrix and need not be uncorrelated with \mathbf{Y}_1). Nevertheless, principal components are often interpreted as factors.
- 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

7.1 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,

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

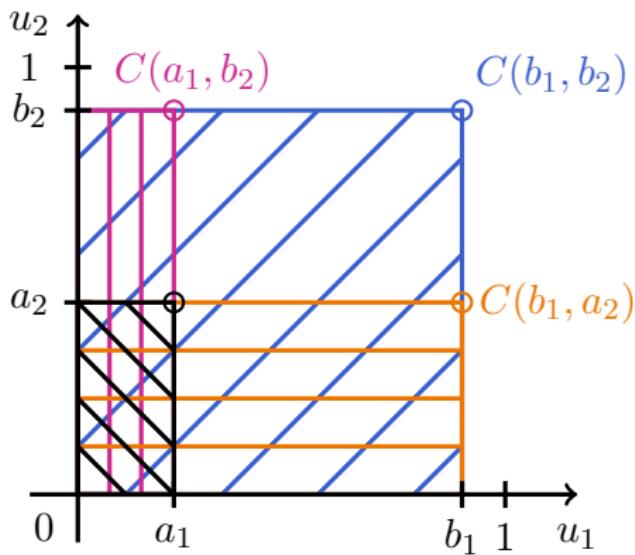
3) C is *d-increasing*, that is, for all $\mathbf{a}, \mathbf{b} \in [0, 1]^d$, $\mathbf{a} \leq \mathbf{b}$,

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

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

2-increasingness explained in a picture:

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



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

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)), \quad \mathbf{x} \in \mathbb{R}^d. \quad (28)$$

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 (28) 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 (28).

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

$$\begin{aligned} C(u_1, \dots, u_d) &= C(F_1(F_1^\leftarrow(u_1)), \dots, F_d(F_d^\leftarrow(u_d))) \\ &\stackrel{(28)}{=} F(F_1^\leftarrow(u_1), \dots, F_d^\leftarrow(u_d)), \quad \mathbf{u} \in \prod_{j=1}^d \text{ran } F_j. \end{aligned}$$

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 (28) 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 (28) (e.g. $C(u_1, u_2) = u_1 u_2$ 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 .
- We say that \mathbf{X} (or F) has copula C if (28) 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. □

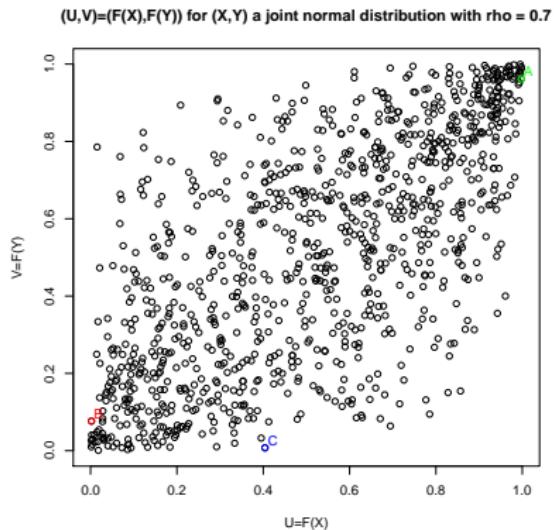
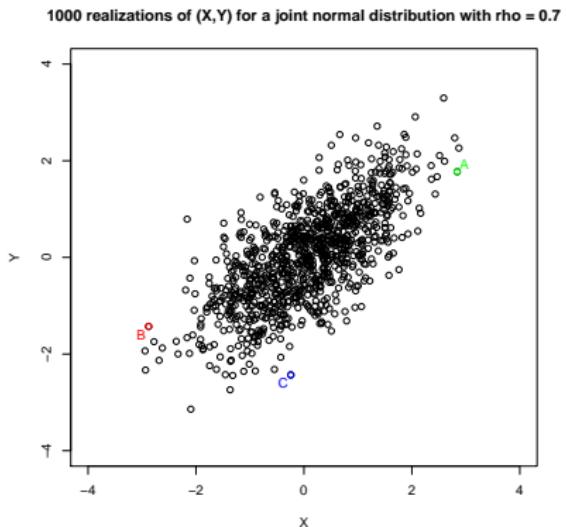
Interpretation of Sklar's Theorem (and the invariance principle)

- 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 Part 1) 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.

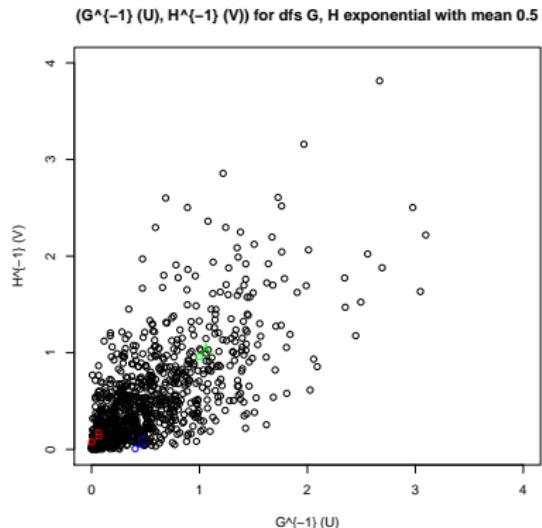
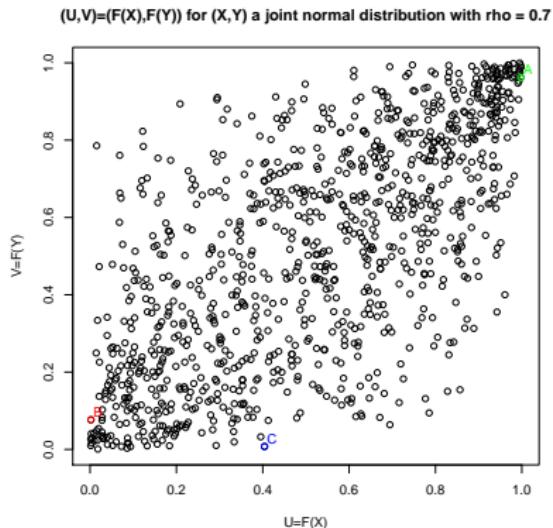
Right: Scatter plot of the corresponding Gauss copula (after applying the df Φ of $N(0, 1)$). Note how A, B, C change.



Visualizing Part 2) of Sklar's Theorem

Left: Same Gauss copula scatter plot as before. Apply marginal $\text{Exp}(2)$ -quantile functions ($F_j^{-1}(u) = -\log(1-u)/2$, $j \in \{1, 2\}$).

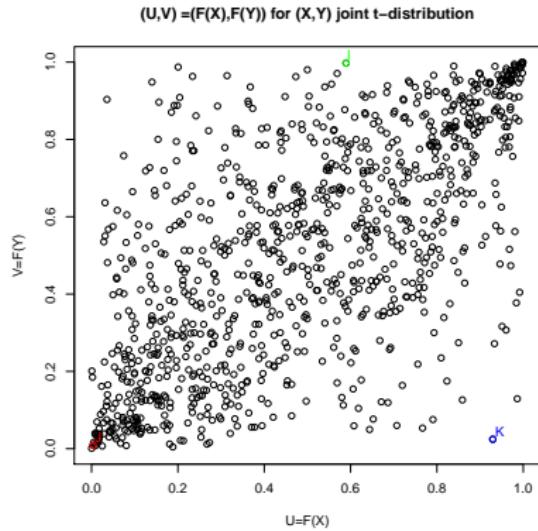
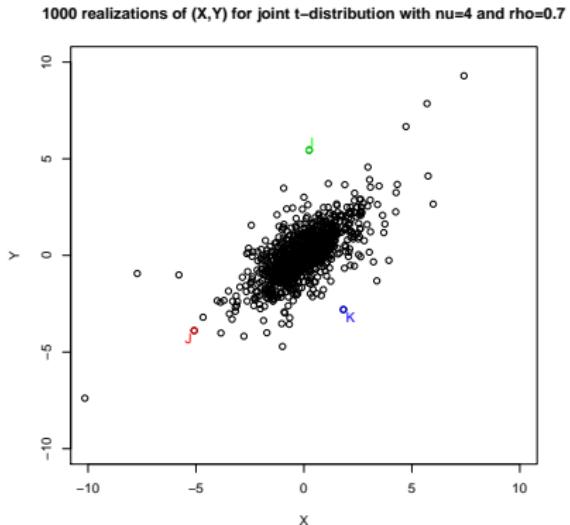
Right: The corresponding transformed random variates. Again, note the three points A, B, C.



Visualizing Part 1) 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**.

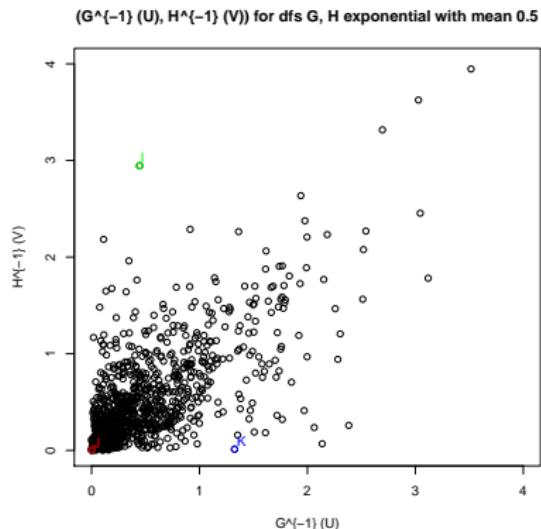
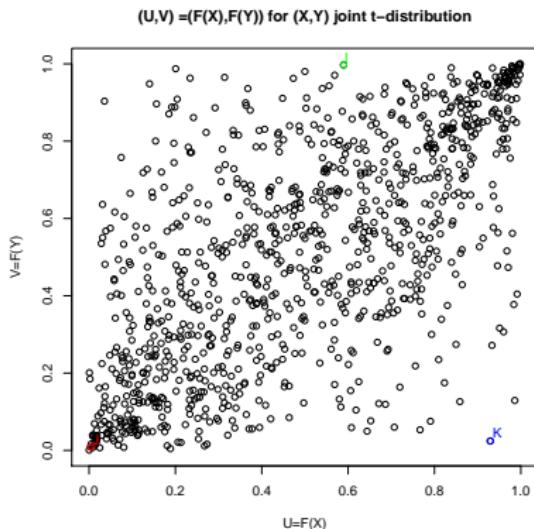
Right: Scatter plot of the corresponding t_4 copula (after applying the df t_4). Note how **A**, **B**, **C** change.



Visualizing Part 2) 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 $d = 2$.

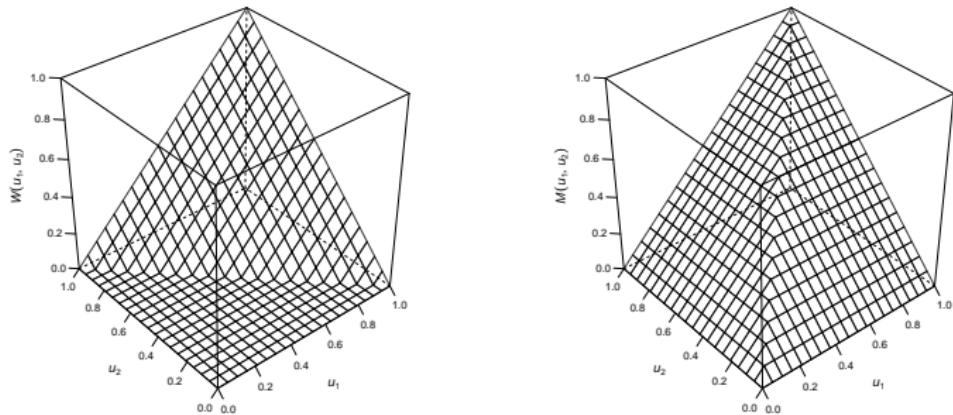
3) M is a copula for all $d \geq 2$.

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–Hoeffding bounds correspond to perfect dependence (negative for W ; positive for M); see Proposition 7.14 later.
- The Fréchet–Hoeffding 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 $C(\mathbf{u}) = \Pi(\mathbf{u})$ (replace x_j by $F_j^\leftarrow(u_j)$ and apply (GI4)). Therefore, X_1, \dots, X_d are independent if and only if their copula is Π ; the density is thus $c(\mathbf{u}) = 1$, $\mathbf{u} \in [0, 1]^d$.
- W is the *countermonotonicity copula*. It is the df of $(U, 1 - U)$. It can be shown that if X_1, X_2 are perfectly negatively dependent (X_2 is a.s. a strictly decreasing function of X_1), their copula is W .

- M is the *comonotonicity copula*. It is the df of (U, \dots, U) . It can be shown that if X_1, \dots, X_d are perfectly positively dependent (X_2, \dots, X_{d-1} are a.s. strictly increasing functions of 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 are the **Gauss copula** and the ***t* copula** (stemming from normal variance mixtures).

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

$$c_P^{\text{Ga}}(\mathbf{u}) = \frac{1}{\sqrt{\det P}} \exp\left(-\frac{1}{2} \mathbf{x}'(P^{-1} - I_d)\mathbf{x}\right), \quad (29)$$

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.

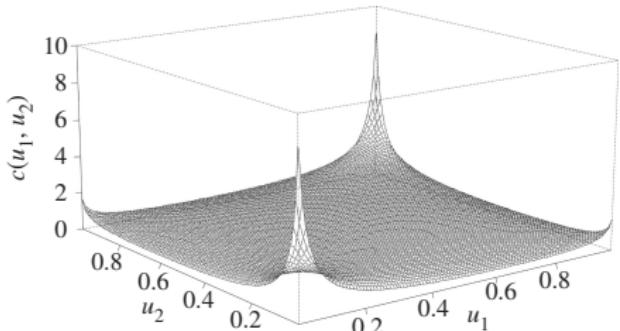
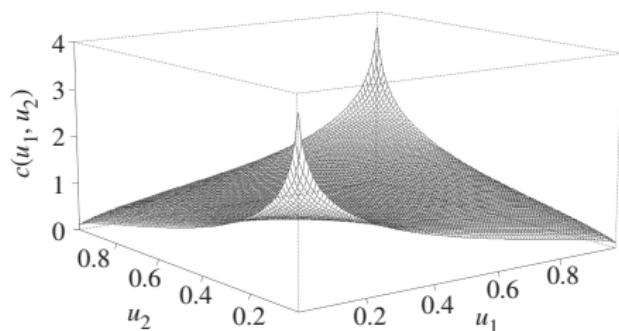
- Special cases: $P = J_d = \mathbf{1}\mathbf{1}'$ then $C = M$. However, if $P = I_d$ then $C \neq \Pi$ (unless $\nu = \infty$ in which case $C_{\nu,P}^t = C_P^{Ga}$). If $d = 2$ and $\rho = P_{12} = -1$ then $C = W$.
- Sklar's Theorem \Rightarrow The density of $C_{\nu,P}^t$ is

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

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:

Advantages:

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

Drawbacks:

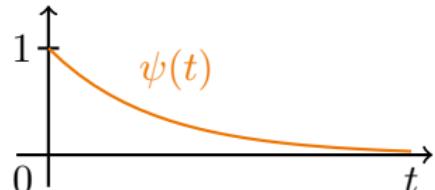
- Typically, C is not explicit
- Radially symmetric (so the same lower/upper tail behaviour)

Explicit copulas

Archimedean copulas are copulas of the form

$$C(\mathbf{u}) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d))$$

where ψ is the *(Archimedean) generator*.

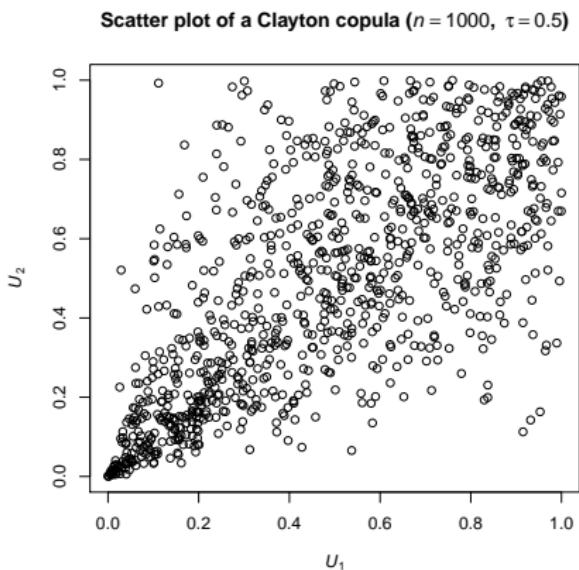
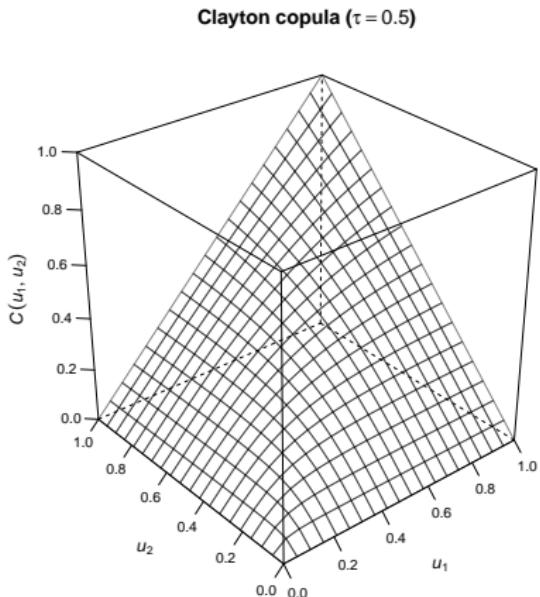


- $\psi : [0, \infty) \rightarrow [0, 1]$ is \downarrow on $[0, \inf\{t : \psi(t) = 0\}]$ and satisfies $\psi(0) = 1$, $\psi(\infty) = \lim_{t \rightarrow \infty} \psi(t) = 0$.
- We set $\psi^{-1}(0) = \inf\{t : \psi(t) = 0\}$.
- The set of all generators is denoted by Ψ .
- Not every generator $\psi \in \Psi$ generates indeed a proper copula (there are conditions, e.g. complete monotonicity, i.e. derivatives alternating in sign).
- If $\psi(t) > 0$, $t \in [0, \infty)$, we call ψ strict.

Clayton copulas are obtained for $\psi(t) = (1+t)^{-1/\theta}$, $t \in [0, \infty)$, $\theta \in (0, \infty)$.
For $\theta \downarrow 0$, $C \rightarrow \Pi$; and for $\theta \uparrow \infty$, $C \rightarrow M$.

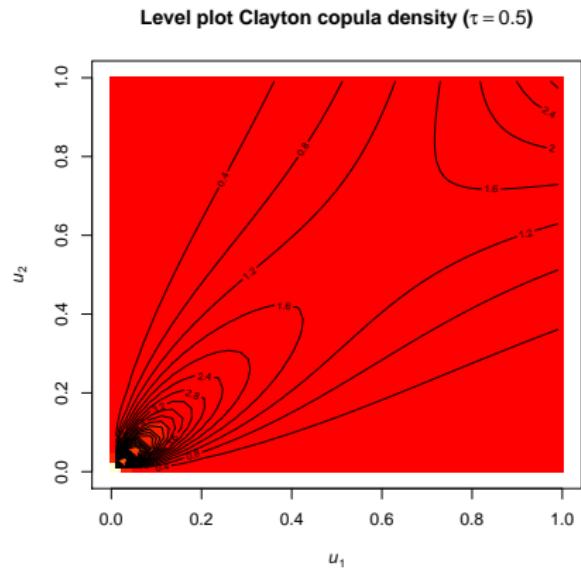
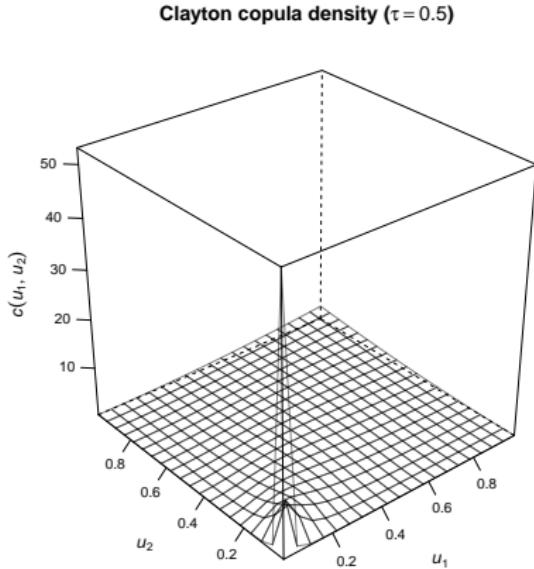
Left: Plot of a bivariate Clayton copula (Kendall's tau 0.5; see later).

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



Left: Plot of the corresponding density.

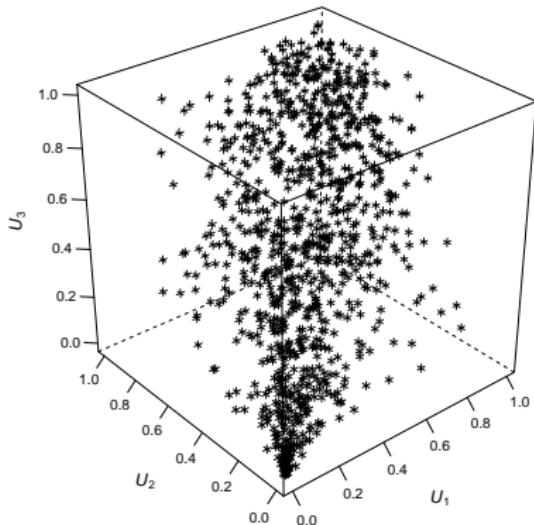
Right: Level plot of the density (with heat colors).



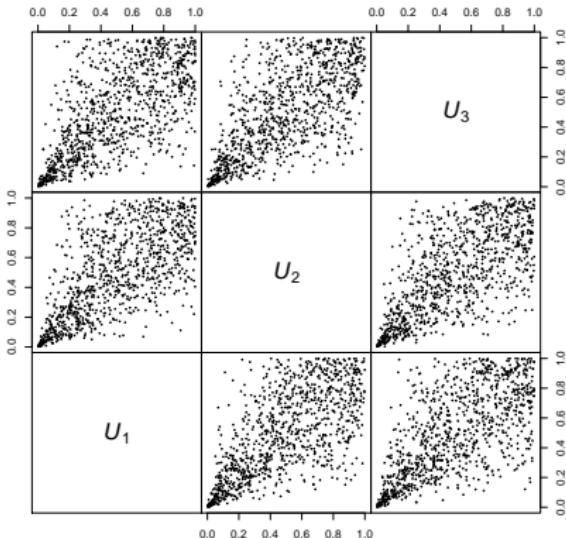
Left: Cloud plot of a trivariate Clayton copula (sample size $n = 1000$; Kendall's tau 0.5).

Right: Corresponding scatter plot matrix.

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



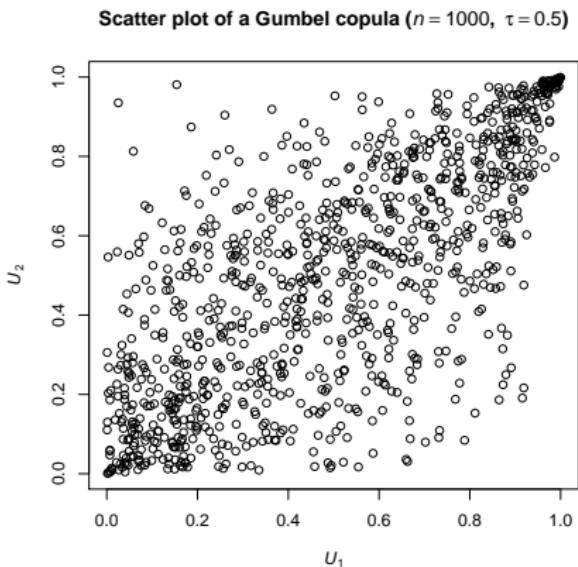
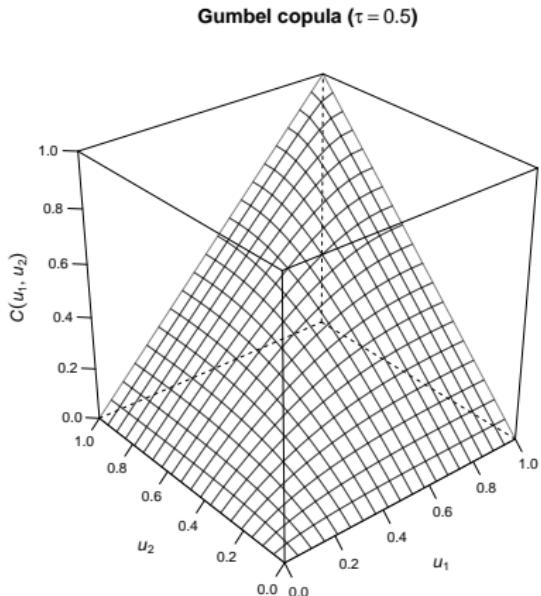
Scatter plot matrix of a Clayton copula ($n = 1000$, $\tau = 0.5$)



Gumbel copulas are obtained for $\psi(t) = \exp(-t^{1/\theta})$, $t \in [0, \infty)$, $\theta \in [1, \infty)$. For $\theta = 1$, $C = \Pi$; and for $\theta \rightarrow \infty$, $C \rightarrow M$.

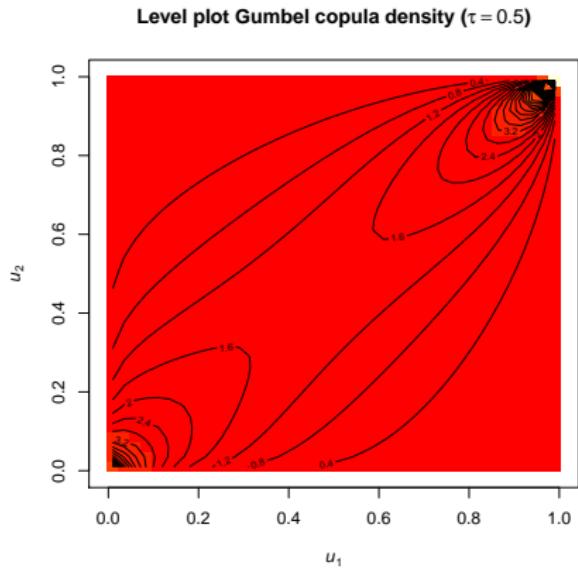
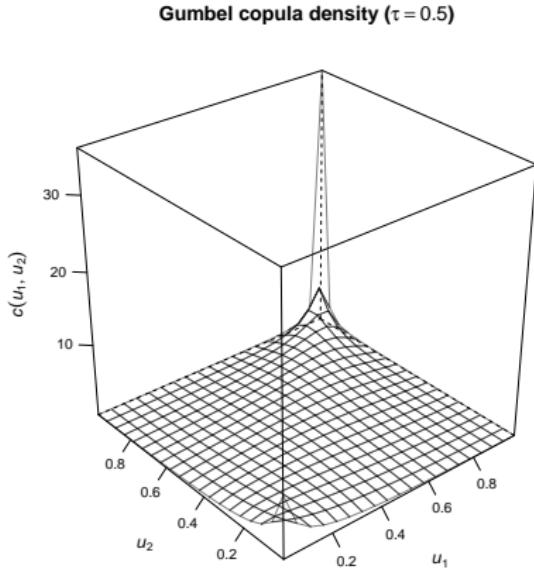
Left: Plot of a bivariate Gumbel copula (Kendall's tau 0.5).

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



Left: Plot of the corresponding density.

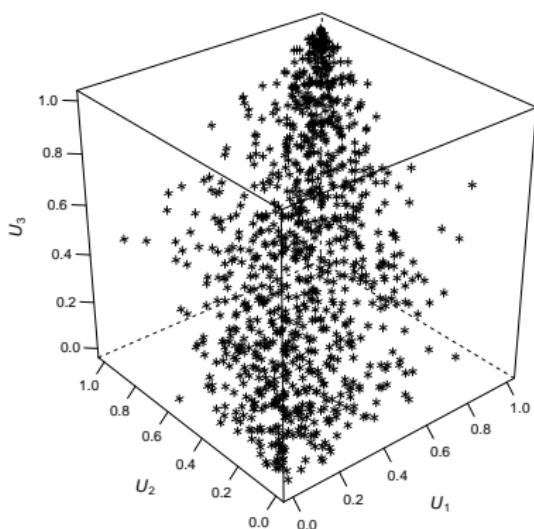
Right: Level plot of the density (with heat colors).



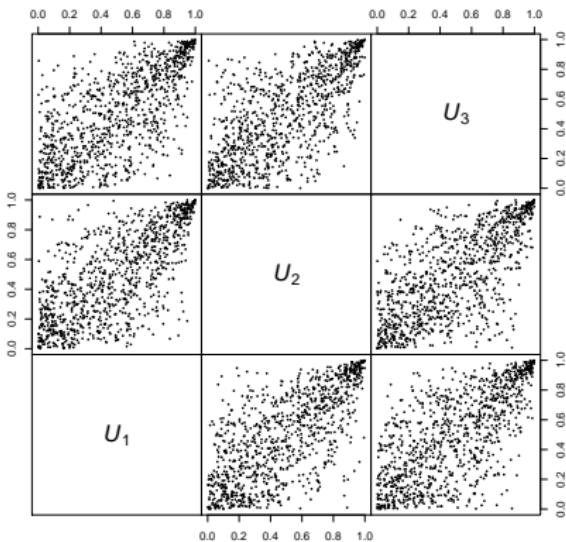
Left: Cloud plot of a trivariate Gumbel copula (sample size $n = 1000$; Kendall's tau 0.5).

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:

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- t model is a multivariate df F with t 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 $\mathbf{X} \sim N_d(\mathbf{0}, P)$ ($\mathbf{X} \stackrel{d}{=} A\mathbf{Z}$ for $AA' = P$, $\mathbf{Z} \sim N_d(\mathbf{0}, I_d)$).
- 2) Return $\mathbf{U} = (\Phi(X_1), \dots, \Phi(X_d))$.

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

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

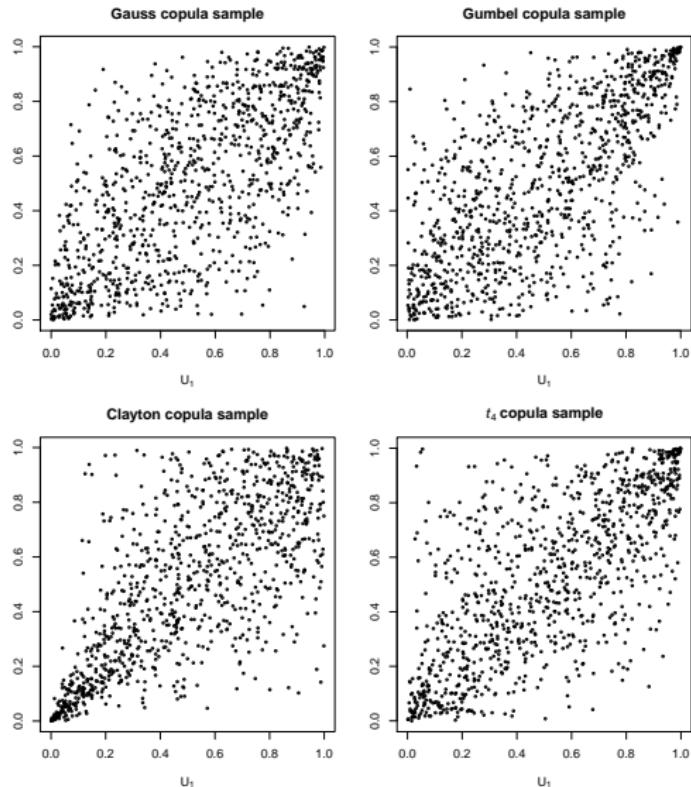
Sampling meta distributions

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

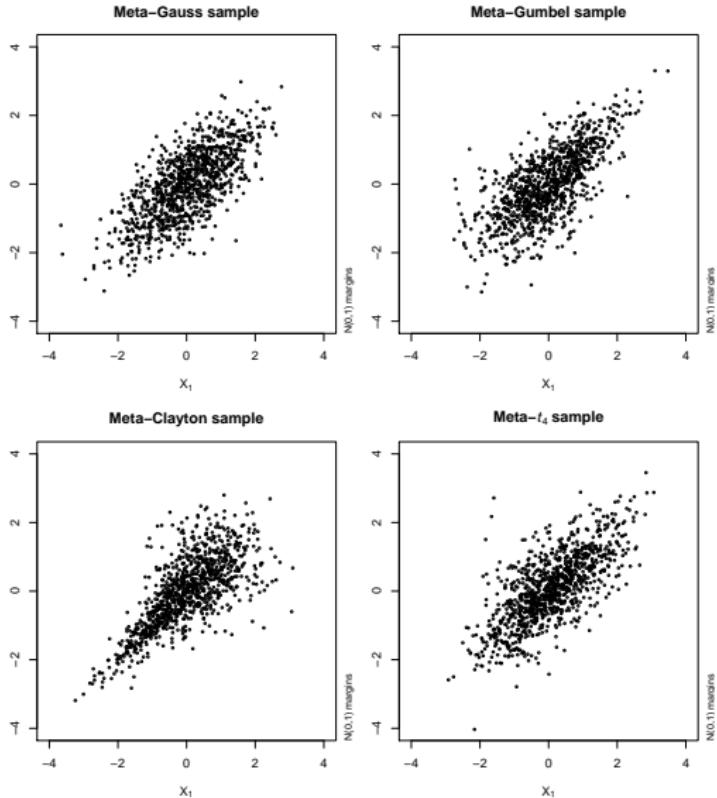
Algorithm 7.11 (Sampling meta- C models)

- 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 $\mathbf{U}' \sim \text{U}(0, 1)^d$ then $\mathbf{U} \sim C$, where

$$U_1 = U'_1,$$

$$U_2 = C_{2|1}^\leftarrow(U'_2 | U_1),$$

$$U_3 = C_{3|1,2}^\leftarrow(U'_3 | U_1, U_2),$$

$$\vdots$$

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

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. u_1, \dots, u_{d-1} . For a.e. $u_1, \dots, u_{j-1} \in [0, 1]$,

$$C_{j|1,\dots,j-1}(u_j | u_1, \dots, u_{j-1}) = \frac{D_{j-1,\dots,1} C^{(1,\dots,j)}(u_1, \dots, u_j)}{D_{j-1,\dots,1} C^{(1,\dots,j-1)}(u_1, \dots, u_{j-1})},$$

where $C^{(1,\dots,j)}(u_1, \dots, u_j) = C(u_1, \dots, u_j, 1, \dots, 1)$ and $D_{j-1,\dots,1}$ is the differential operator w.r.t. u_1, \dots, u_{j-1} .

Note: $C_{2|1}(u_2 | u_1) = \frac{D_1 C(u_1, u_2)}{1} = D_1 C(u_1, u_2)$ which also follows from

$$\begin{aligned} & \lim_{h \downarrow 0} \frac{C(u_1 + h, u_2) - C(u_1, u_2)}{h} \\ &= \lim_{h \downarrow 0} \frac{\mathbb{P}(U_1 \leq u_1 + h, U_2 \leq u_2) - \mathbb{P}(U_1 \leq u_1, U_2 \leq u_2)}{h} \\ &= \lim_{h \downarrow 0} \frac{\mathbb{P}(U_2 \leq u_2, u_1 < U_1 \leq u_1 + h)}{\mathbb{P}(u_1 < U_1 \leq u_1 + h)} = \lim_{h \downarrow 0} \mathbb{P}(U_2 \leq u_2 | u_1 < U_1 \leq u_1 + h). \end{aligned}$$

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 $d = 2$, $\hat{C}(u_1, u_2) = 1 - (1 - u_1) - (1 - u_2) + C(1 - u_1, 1 - u_2) = -1 + u_1 + u_2 + C(1 - u_1, 1 - u_2)$. We can also verify this directly by noting that $\hat{C}(u_1, u_2)$ equals

$$\begin{aligned}\mathbb{P}(1 - U_1 \leq u_1, 1 - U_2 \leq u_2) &= \mathbb{P}(U_1 > 1 - u_1, U_2 > 1 - u_2) \\&= \mathbb{P}(U_1 > 1 - u_1) - \mathbb{P}(U_1 > 1 - u_1, U_2 \leq 1 - u_2) \\&= 1 - (1 - u_1) - (\mathbb{P}(U_2 \leq 1 - u_2) - \mathbb{P}(U_1 \leq 1 - u_1, U_2 \leq 1 - u_2)) \\&= u_1 - (1 - u_2 - C(1 - u_1, 1 - u_2)).\end{aligned}$$

- If C admits a density, $\hat{c}(\mathbf{u}) = c(1 - \mathbf{u})$.
- 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 $\bar{F}(\mathbf{x}) = \mathbb{P}(\mathbf{X} > \mathbf{x})$ with corresponding marginal survival functions $\bar{F}_1, \dots, \bar{F}_d$ (with $\bar{F}_j(x) = \mathbb{P}(X_j > x)$). Hence survival copulas combine marginal to joint survival functions.

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 $\text{U}(0, 1)$ margins. This holds if only if $C(u_1, \dots, u_d) = C(u_{\pi(1)}, \dots, u_{\pi(d)})$ for all possible permutations of arguments, i.e. if C is *symmetric*.
- Exchangeable/symmetric copulas are useful for 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, $d = 2$

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

$$f(\mathbf{x}) = c(F_1(x_1), \dots, F_d(x_d)) \prod_{j=1}^d f_j(x_j). \quad (30)$$

This implies

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

- It follows from (30) 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.

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 $\mathbb{E}(X_j^2) < \infty$, $j \in \{1, 2\}$.

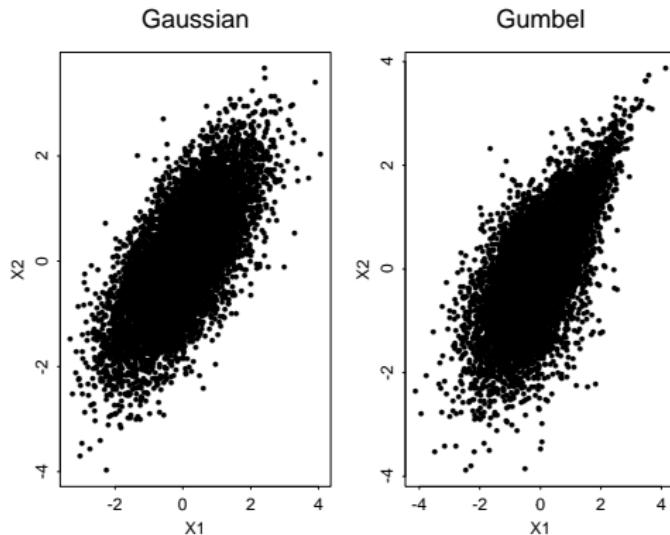
Note that ρ depends on the marginal distributions! In particular, second moments have to exist (not the case, e.g. for $X_1, X_2 \stackrel{\text{ind.}}{\sim} F(x) = 1 - x^{-3}$)

- $|\rho| \leq 1$. 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 $(X_1, X_2) \sim N_2(\mathbf{0}, P)$. Then $\rho(X_1, X_2) = P_{12}$, but $\rho(F_1(X_1), F_2(X_2)) = \frac{6}{\pi} \arcsin(P_{12}/2)$.

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 $\rho(X_1, X_2) = \text{cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) = \mathbb{E}(V) \mathbb{E}(Z^2) = 0$, 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

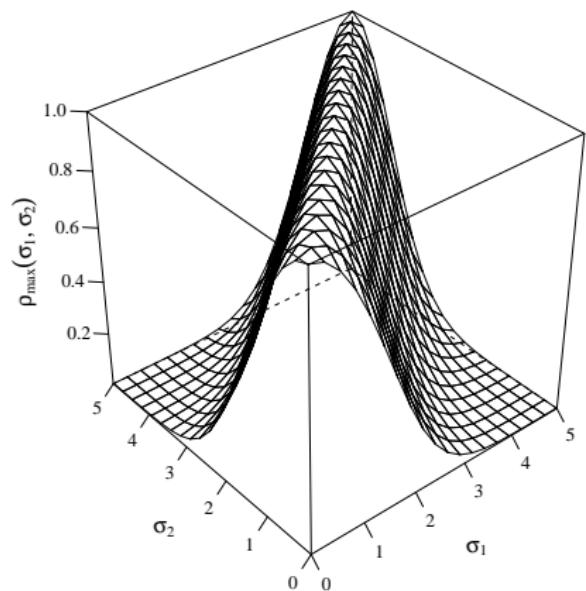
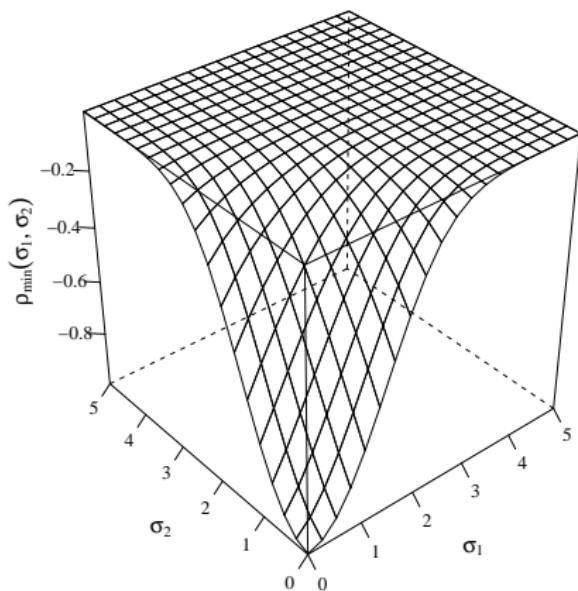
$$\text{cov}(X_1, X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (C(F_1(x_1), F_2(x_2)) - F_1(x_1)F_2(x_2)) dx_1 dx_2.$$

implies bounds on attainable ρ :

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

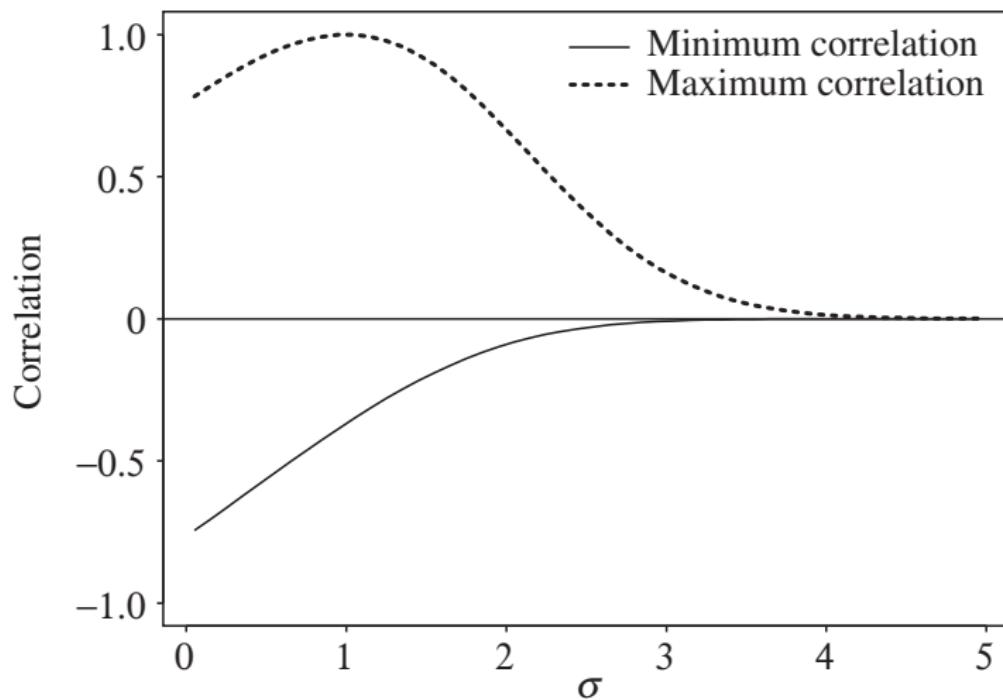
Example 7.18 (Bounds for a model with $\text{LN}(0, \sigma_j^2)$ margins)

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 $\text{VaR}_\alpha(X_1 + X_2) > \text{VaR}_\alpha(X_1) + \text{VaR}_\alpha(X_2)$ serves as a counterexample as the right-hand side under comonotonicity (so maximal correlation) only equals $\text{VaR}_\alpha(X_1 + X_2)$; 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* ($\mathbb{P}((X_1 - X'_1)(X_2 - X'_2) > 0)$; probability of two independent points from F to have a positive slope) *minus the probability of discordance* ($\mathbb{P}((X_1 - X'_1)(X_2 - X'_2) < 0)$; probability of two independent points from F to have a negative slope).

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

$$\rho_\tau = 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1 = 4 \mathbb{E}(C(U_1, U_2)) - 1,$$

where $(U_1, U_2) \sim C$.

Proof. See the appendix. □

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

$$r_n^\tau = \frac{1}{\binom{n}{2}} \sum_{1 \leq i_1 < i_2 \leq n} \text{sign}((X_{i_1 1} - X_{i_2 1})(X_{i_1 2} - X_{i_2 2})). \quad (31)$$

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

$$\rho_S = 12 \int_0^1 \int_0^1 C(u_1, u_2) du_1 du_2 - 3 = 12 \mathbb{E}(C(U'_1, U'_2)) - 3,$$

where $(U'_1, U'_2) \sim \Pi$.

Proof. By Höffding's formula, we have $\rho_S(X_1, X_2) = \rho(F_1(X_1), F_2(X_2)) = 12 \int_0^1 \int_0^1 (C(u_1, u_2) - u_1 u_2) du_1 du_2 = 12 \int_0^1 \int_0^1 C(u_1, u_2) du_1 du_2 - 3$.

□

- 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. In general, $\kappa = 0$ does not imply independence.

- 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 when ρ is replaced by ρ_τ or ρ_S . 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 $\rho_\tau = \rho_S = 2\lambda - 1$ (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.
- 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

$$\begin{aligned} \mathbb{P}(X_2 \leq F_2^\leftarrow(u) \mid X_1 \leq F_1^\leftarrow(u)) &= \frac{\mathbb{P}(X_1 \leq F_1^\leftarrow(u), X_2 \leq F_2^\leftarrow(u))}{\mathbb{P}(X_1 \leq F_1^\leftarrow(u))} \\ &= \frac{F(F_1^\leftarrow(u), F_2^\leftarrow(u))}{F_1(F_1^\leftarrow(u))} \stackrel{\text{Sklar}}{=} \stackrel{\text{(GI4)}}{=} \frac{C(u, u)}{u}, \quad u \in (0, 1), \text{ so } \lambda_l = \lim_{u \downarrow 0} \frac{C(u, u)}{u}. \end{aligned}$$

- 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 $\lambda_l = \lim_{u \downarrow 0} (D_1 C(u, u) + D_2 C(u, u))$ (Chain Rule). If C is symmetric, $\lambda_l = 2 \lim_{u \downarrow 0} D_1 C(u, u) \stackrel{\text{Th. 7.13}}{=} 2 \lim_{u \downarrow 0} C_{2|1}(u \mid u) = 2 \lim_{u \downarrow 0} \mathbb{P}(U_2 \leq u \mid U_1 = u)$ 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

$$\lambda_l = 2 \lim_{x \downarrow -\infty} \mathbb{P}(X_2 \leq x \mid X_1 = x). \quad (32)$$

- 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

$$\lambda \stackrel{\text{radial}}{\underset{\text{symm.}}{=}} \lambda_l \stackrel{\text{symm.}}{\underset{(32)}{=}} 2 \lim_{x \downarrow -\infty} \mathbb{P}(X_2 \leq x \mid X_1 = x).$$

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

$$\lambda = 2 \lim_{x \downarrow -\infty} \mathbb{P}(X_2 \leq x \mid X_1 = x) = 2 \lim_{x \downarrow -\infty} \Phi\left(\frac{x(1-\rho)}{\sqrt{1-\rho^2}}\right) = I_{\{\rho=1\}}.$$

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

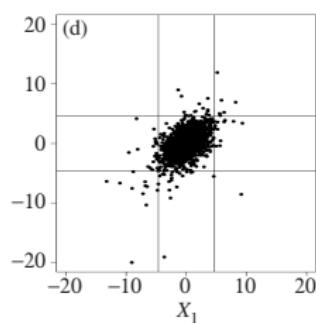
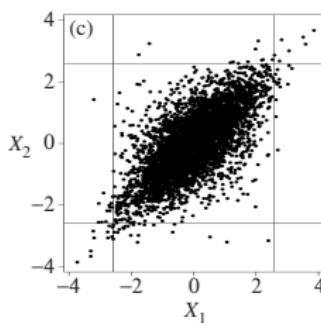
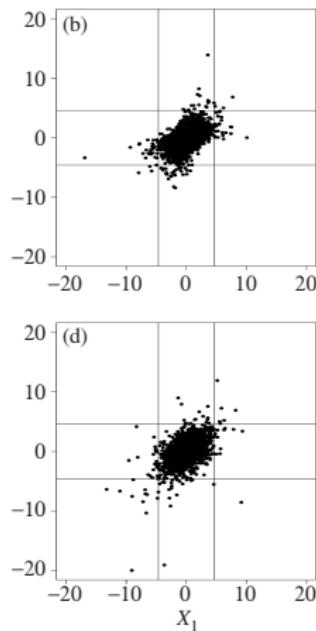
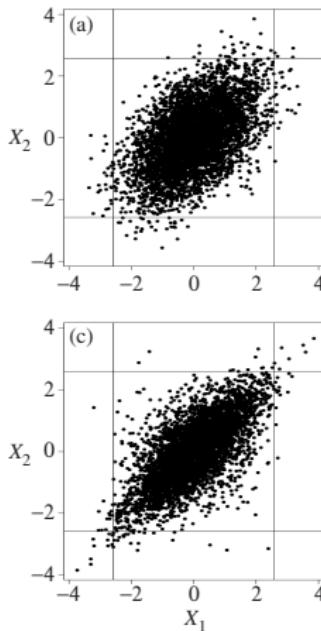
$$\lambda = 2t_{\nu+1}\left(-\sqrt{\frac{(\nu+1)(1-\rho)}{1+\rho}}\right) \quad (\text{tail dependence; } \lambda \uparrow \text{ in } \rho \uparrow \text{ and } \nu \downarrow).$$

- λ values for various ν, ρ :

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

If W has a power tail, $\lambda > 0$, otherwise $\lambda = 0$.

Joint quantile exceedance probabilities



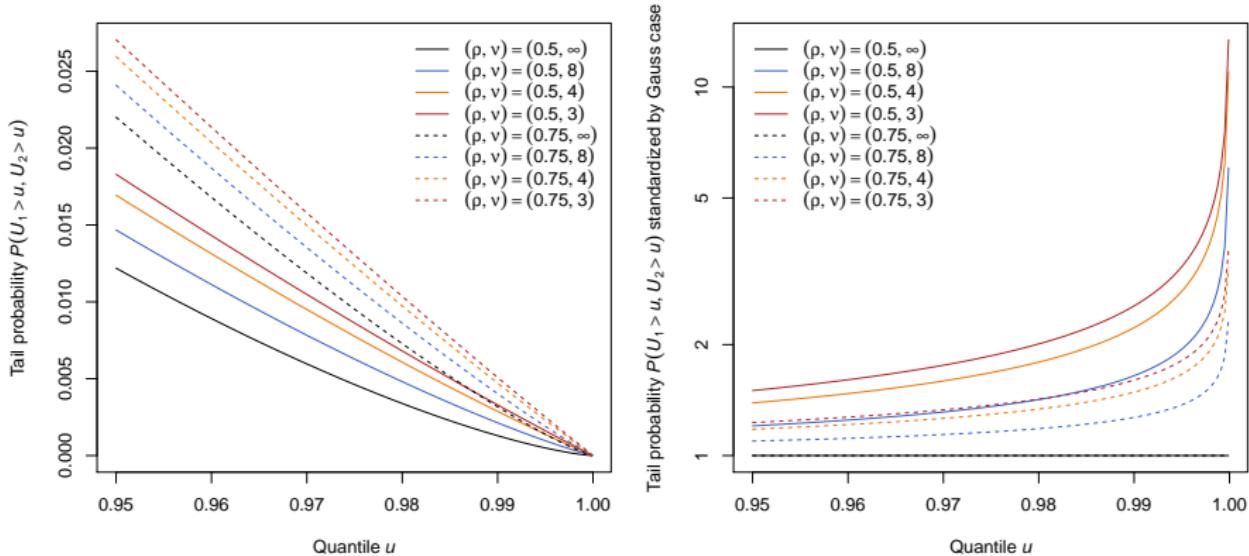
5000 samples from

- (a) $N_2(\mathbf{0}, P = (\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}))$, $\rho = 0.5$;
- (b) C_ρ^{Ga} with t_4 margins (same dependence as in (a));
- (c) $C_{4,\rho}^t$ with $N(0, 1)$ margins;
- (d) $t_2(4, \mathbf{0}, P)$ (same dependence as in (c)).

Lines denote the true 0.005- and 0.995-quantiles.

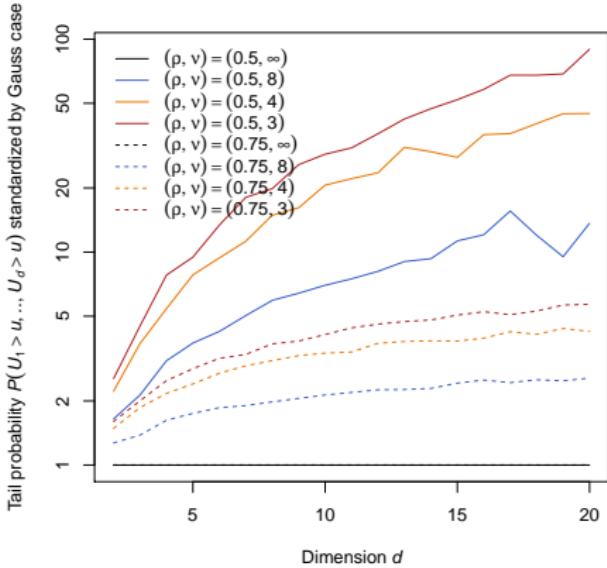
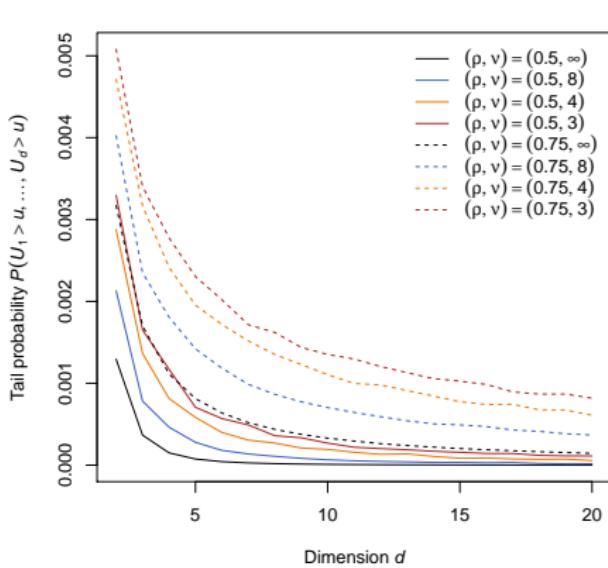
Note the different number of points in the bivariate tails (all models have the same Kendall's tau!)

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

Joint tail probabilities $\mathbb{P}(U_1 > u, \dots, U_d > u)$ for $u = 0.99$



- Homogeneous P (off-diagonal entry ρ). Note the MC randomness.
- **Left:** Clear; less mass in corners in higher dimensions.
- **Right:** $d \mapsto \frac{\mathbb{P}(U_1 > u, \dots, U_d > u)}{\mathbb{P}(V_1 > u, \dots, V_d > u)}$ $\stackrel{\text{radial}}{\equiv} \stackrel{\text{symm.}}{\equiv} \frac{C_{\nu, \rho}^t(u, \dots, u)}{C_{\rho}^{\text{Ga}}(u, \dots, u)}$ for $u = 0.99$.

Example 7.25 (Interpretation of joint tail probabilities)

- Consider 5 daily negative log-returns $\mathbf{X} = (X_1, \dots, X_5)$. Assume they follow an elliptical distribution and have pairwise correlations $\rho = 0.5$. However, we are unsure about the best joint model.
- If \mathbf{X} are multivariate normal (and thus $C_{\rho=0.5}^{\text{Ga}}$), the probability that on any day all 5 negative returns lie above their $u = 0.99$ quantiles is

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

In the long run such an event will happen once every $1/7.48 \times 10^{-5} \approx 13\,369$ trading days on average (\approx once every 51.4 years; assuming 260 trading days in a year).

- If \mathbf{X} is multivariate t_4 (and thus $C_{\nu=4, \rho=0.5}^t$), 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)

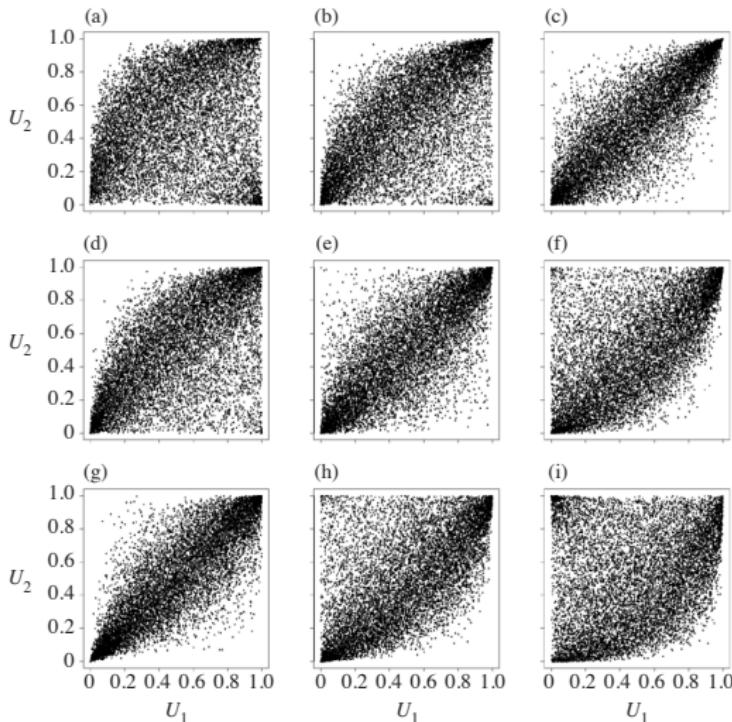
Let $\mathbf{X} \sim E_2(\mathbf{0}, P, \psi)$ with $\mathbb{P}(\mathbf{X} = \mathbf{0}) = 0$, $\rho = P_{12}$. Then $\rho_\tau = \frac{2}{\pi} \arcsin \rho$.

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 $\mathbb{P}(U_j \leq u \mid V = v) = \exp(-v\psi^{-1}(u))$.

Proof.

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

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

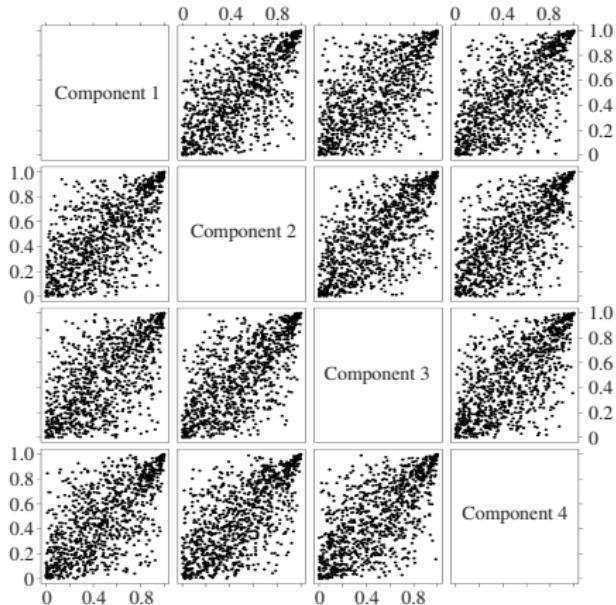
The j th marginal survival function is thus (set $x_k = 0 \ \forall k \neq j$)
 $\bar{F}_j(x_j) = \mathbb{P}(X_j > x_j) = \psi(x_j)$ (\downarrow and continuous) and therefore
 $\hat{C}(\mathbf{u}) = \bar{F}(\bar{F}_1^\leftarrow(u_1), \dots, \bar{F}_d^\leftarrow(u_d)) = \psi(\sum_{j=1}^d \psi^{-1}(u_j)).$

- 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 $V \sim G$ (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).

With copulas at hand, we can now provide a proof for subadditivity of ES.

Proposition 7.33 (Subadditivity of ES)

$$\text{ES}_\alpha(L) = \frac{\sup_{\{\tilde{Y} \sim \text{B}(1, 1-\alpha)\}} \mathbb{E}(L\tilde{Y})}{1 - \alpha}, \text{ which is subadditive; the supremum is taken over all copulas between } L \sim F_L \text{ and } \tilde{Y} \sim \text{B}(1, 1 - \alpha).$$

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. For any other (L, \tilde{Y}) with $\tilde{Y} \sim \text{B}(1, 1 - \alpha)$,
 $\mathbb{E}(L\tilde{Y}) = \text{cov}(L, \tilde{Y}) + \mathbb{E}(L)\mathbb{E}(\tilde{Y}) \stackrel{\text{Höffding}}{\leq} \text{cov}(L, Y) + \mathbb{E}(L)\mathbb{E}(Y) = \mathbb{E}(LY)$
- and thus $\text{ES}_\alpha(L) = \frac{1}{1-\alpha} \sup_{\{\tilde{Y} \sim \text{B}(1, 1-\alpha)\}} \mathbb{E}(L\tilde{Y})$. □

7.5 Fitting copulas to data

- Let $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{\text{ind.}}{\sim} F$ with cont. margins F_1, \dots, F_d and copula C .
- We assume that 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
 - ▶ $F_j = F_j(\cdot; \boldsymbol{\theta}_{0,j})$ for some $\boldsymbol{\theta}_{0,j} \in \Theta_j$, $j \in \{1, \dots, d\}$;
 $(F_j(\cdot; \boldsymbol{\theta}_j)$ is assumed to be continuous $\forall \boldsymbol{\theta}_j \in \Theta_j$, $j \in \{1, \dots, d\}$)
 - ▶ $C = C(\cdot; \boldsymbol{\theta}_{0,C})$ for some $\boldsymbol{\theta}_{0,C} \in \Theta_C$.

Thus F has the true but unknown parameter vector $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}'_{0,C}, \boldsymbol{\theta}'_{0,1}, \dots, \boldsymbol{\theta}'_{0,d})'$ to be estimated.

- Here, we focus particularly on $\boldsymbol{\theta}_{0,C}$. Whenever necessary, we assume that the margins F_1, \dots, F_d and the copula C are absolutely continuous with corresponding densities f_1, \dots, f_d and c , respectively.
- We assume the chosen copula to be appropriate (w.r.t. symmetry etc.).

7.5.1 Method-of-moments using rank correlation

- For $d = 2$ and one-parameter copulas, 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 of θ and r_n^τ is the sample version of Kendall's tau (computed via (31) 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*. If $r_{n,j_1j_2}^\tau$ denotes the sample version of Kendall's tau for data pair (j_1, j_2) , then

$$\hat{P}_{n,j_1j_2}^{\text{IKTE}} = \sin\left(\frac{\pi}{2} r_{n,j_1j_2}^\tau\right).$$

A proper correlation matrix P can be constructed as in Higham (2002).

- One can also use Spearman's rho. For Gauss copulas,

$$\rho \approx \frac{6}{\pi} \arcsin \frac{\rho}{2} \stackrel{\text{Prop.7.26}}{=} \rho_S.$$

The 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} ; see Mashal and Zeevi (2002).

7.5.2 Forming a pseudo-sample from the copula

- $\mathbf{X}_1, \dots, \mathbf{X}_n$ typically does not have $\text{U}(0, 1)$ margins. For applying the "copula approach" we thus need *pseudo-observations* from C .
- In general, we take $\hat{\mathbf{U}}_i = (\hat{U}_{i1}, \dots, \hat{U}_{id}) = (\hat{F}_1(X_{i1}), \dots, \hat{F}_d(X_{id}))$, $i \in \{1, \dots, n\}$, where \hat{F}_j denotes an estimator of F_j ; see Lemma 7.6. Note that $\hat{\mathbf{U}}_1, \dots, \hat{\mathbf{U}}_n$ are typically neither independent (even if $\mathbf{X}_1, \dots, \mathbf{X}_n$ are) nor perfectly $\text{U}(0, 1)^d$ distributed.

- Possible choices for \hat{F}_j :
 - ▶ Parametric estimators (typically if n is small). One often still uses (33) below for estimating $\theta_{0,C}$ (to keep the error due to misspecification of the margins small).
 - ▶ Semi-parametric estimators (for example EVT-based: Bodies are modelled empirically, tails semiparametrically via the GPD-based tail estimator of Smith (1987)).
 - ▶ Non-parametric estimators with scaled empirical dfs, so

$$\hat{U}_{ij} = \frac{n}{n+1} \hat{F}_{n,j}(X_{ij}) = \frac{R_{ij}}{n+1}, \quad (33)$$

where R_{ij} denotes the rank of X_{ij} among all X_{1j}, \dots, X_{nj} . The scaling is to avoid density evaluation on the boundary of $[0, 1]^d$.

If n is sufficiently large, one typically uses (33).

7.5.3 Maximum likelihood estimation

The (classical) maximum likelihood estimator

- If it exists, the density of $F(\mathbf{x}) = C(F_1(x_1), \dots, F_d(x_d))$ is

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

- 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:

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

$$\hat{\theta}_n^{\text{IFME}} \text{ solves } \left(\frac{\partial}{\partial \boldsymbol{\theta}_C} \ell, \frac{\partial}{\partial \boldsymbol{\theta}_1} \ell_1, \dots, \frac{\partial}{\partial \boldsymbol{\theta}_d} \ell_d \right) = \mathbf{0},$$

where

$$\ell = \ell(\boldsymbol{\theta}; \mathbf{X}_1, \dots, \mathbf{X}_n),$$

$$\ell_j = \ell_j(\boldsymbol{\theta}_j; X_{1j}, \dots, X_{nj}) = \sum_{i=1}^n \ell_j(\boldsymbol{\theta}_j; X_{ij}) = \sum_{i=1}^n \log f_j(X_{ij}; \boldsymbol{\theta}_j).$$

Example 7.34 (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.

- 1) MLE requires to solve a 201-dimensional optimization problem.
 - 2) 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.
 - In this case (and under more complicated marginal models), one can execute the 101 optimizations in parallel, independently of each other.

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 7.35 (Fitting the Gauss copula)

- Use pairwise inversion of Spearman's rho or Kendall's tau.
- Or the MPLE via the (copula-related) log-likelihood

$$\ell_C(P; \hat{U}_1, \dots, \hat{U}_n) = \sum_{i=1}^n \ell_C(P; \hat{U}_i) \stackrel{\text{Eq. (29)}}{=} \sum_{i=1}^n \log c_P^{\text{G}\alpha}(\hat{U}_i).$$

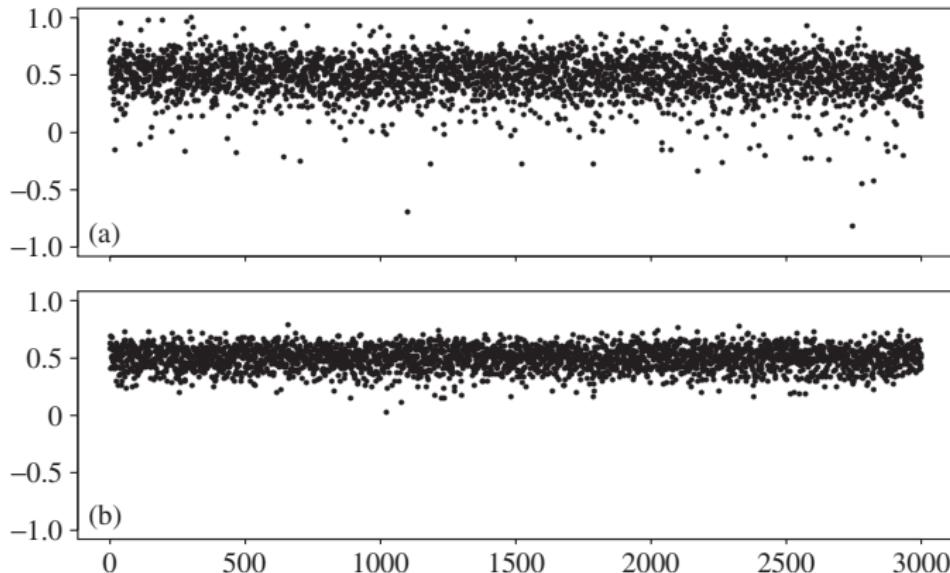
For maximization over all correlation matrices P , we can use the Cholesky factor A as reparameterization and maximize over all lower triangular matrices A with 1s on the diagonal; still this is $\mathcal{O}(d^2)$.

Example 7.36 (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 , use Mashal and Zeevi (2002):
 - 1) Estimate P via pairwise inversion of Kendall's tau (see above).
 - 2) Plug \hat{P} into the likelihood and maximize it w.r.t. ν to obtain $\hat{\nu}_n$.

Example 7.37 (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 . There are also graphical approaches not further discussed here.

8 Aggregate risk

- 8.1 Coherent and convex risk measures
- 8.2 Law-invariant coherent risk measures
- 8.3 Risk measures for linear portfolios
- 8.4 Risk aggregation
- 8.5 Capital allocation

8.1 Coherent and convex risk measures

- Consider a linear space $\mathcal{M} \subseteq \mathcal{L}^0(\Omega, \mathcal{F}, \mathbb{P})$ (a.s. finite rvs).
- Each $L \in \mathcal{M}$ (incl. constants) represents a loss over a fixed time horizon.
- A *risk measure* is a mapping $\varrho : \mathcal{M} \rightarrow \mathbb{R}$; $\varrho(L)$ gives the total amount of capital needed to back a position with loss L .
- $C \subseteq \mathcal{M}$ is *convex* if $(1 - \gamma)x + \gamma y \in C$ for all $x, y \in C$, $0 < \gamma < 1$. C is a *convex cone* if, additionally, $\lambda x \in C$ when $x \in C$, $\lambda > 0$.
- Axioms for ϱ we consider are:

Monotonicity: $L_1 \leq L_2 \Rightarrow \varrho(L_1) \leq \varrho(L_2)$.

Translation invariance: $\varrho(L + m) = \varrho(L) + m$ for all $m \in \mathbb{R}$.

Subadditivity: $\varrho(L_1 + L_2) \leq \varrho(L_1) + \varrho(L_2)$ for all $L_1, L_2 \in \mathcal{M}$.

Positive homogeneity: $\varrho(\lambda L) = \lambda \varrho(L)$ for all $\lambda \geq 0$.

Convexity: $\varrho(\gamma L_1 + (1 - \gamma)L_2) \leq \gamma \varrho(L_1) + (1 - \gamma)\varrho(L_2)$ for all $0 \leq \gamma \leq 1$, $L_1, L_2 \in \mathcal{M}$.

Definition 8.1 (Convex, coherent risk measures)

- A risk measure which satisfies monotonicity, translation invariance and convexity is called *convex*.
- A risk measure which satisfies monotonicity, translation invariance, subadditivity and positive homogeneity is called *coherent*.

A coherent risk measure is convex; the converse is not true, see below. On the other hand, for a positive-homogeneous risk measure, convexity and coherence are equivalent.

8.1.1 Risk measures and acceptance sets

Definition 8.2 (Acceptance set)

For a monotone and translation-invariant risk measure ϱ the *acceptance set of ϱ* is $A_\varrho = \{L \in \mathcal{M} : \varrho(L) \leq 0\}$ (so it contains the positions that are acceptable without any backing capital).

Proposition 8.3

Let ϱ be monotone and translation-invariant with associated A_ϱ . Then

- 1) $A_\varrho \neq \emptyset$ and A_ϱ satisfies

$$L \in A_\varrho \text{ and } \tilde{L} \leq L \Rightarrow \tilde{L} \in A_\varrho. \quad (34)$$

- 2) ϱ can be reconstructed from A_ϱ via

$$\varrho(L) = \inf\{m \in \mathbb{R} : L - m \in A_\varrho\}. \quad (35)$$

Proof. 1) is clear. For 2), note that $\inf\{m : L - m \in A_\varrho\} = \inf\{m : \varrho(L - m) \leq 0\} = \inf\{m : \varrho(L) - m \leq 0\}$ and this is equal to $\varrho(L)$. \square

Proposition 8.4

Suppose that A satisfies (34) and define

$$\varrho_A(L) = \inf\{m \in \mathbb{R} : L - m \in A\}. \quad (36)$$

Suppose $\varrho_A(L)$ is finite for all $L \in \mathcal{M}$. Then ϱ_A is monotone and translation-invariant on \mathcal{M} and A_{ϱ_A} satisfies $A_{\varrho_A} \supseteq A$.

Proof. These properties of ϱ_A are easily checked. □

Example 8.5 (Value-at-risk)

For $\alpha \in (0, 1)$, suppose we call $L \in \mathcal{M}$ *acceptable* if $\mathbb{P}(L > 0) \leq 1 - \alpha$. Then (36) is given by

$$\begin{aligned}\varrho_\alpha(L) &= \inf\{m \in \mathbb{R} : \mathbb{P}(L - m > 0) \leq 1 - \alpha\} \\ &= \inf\{m \in \mathbb{R} : \mathbb{P}(L \leq m) \geq \alpha\} = \text{VaR}_\alpha(L).\end{aligned}$$

Proposition 8.6

- 1) Let ϱ be monotone and translation-invariant. Then
 - 1.1) ϱ is convex if and only if A_ϱ is convex.
 - 1.2) ϱ is coherent if and only if A_ϱ is a convex cone.
- 2) More generally, consider a set of acceptable positions A and the associated risk measure ϱ_A (whose acceptance set may be larger than A). If A is convex, so is ϱ_A ; if A is a convex cone, then ϱ_A is coherent.

Example 8.7 (Risk measures based on loss functions)

Consider a strictly increasing and convex *loss function* $\ell : \mathbb{R} \rightarrow \mathbb{R}$ and some $c \in \mathbb{R}$. Assume that $\mathbb{E}(\ell(L))$ is finite for all $L \in \mathcal{M}$. Define an acceptance set by

$$A = \{L \in \mathcal{M} : \mathbb{E}(\ell(L)) \leq \ell(c)\},$$

and the associated risk measure by

$$\varrho_A = \inf\{m \in \mathbb{R} : \mathbb{E}(\ell(L - m)) \leq \ell(c)\}.$$

- ϱ_A is translation invariant and monotone by Proposition 8.4 since A satisfies (34).
- ϱ_A is convex by Proposition 8.6; to see this consider acceptable positions L_1 and L_2 and observe that the convexity of ℓ implies

$$\begin{aligned}\mathbb{E}(\ell(\gamma L_1 + (1 - \gamma)L_2)) &\leq \mathbb{E}(\gamma\ell(L_1) + (1 - \gamma)\ell(L_2)) \\ &\leq \gamma\ell(c) + (1 - \gamma)\ell(c) = \ell(c),\end{aligned}$$

where we have used that $\mathbb{E}(\ell(L_i)) \leq \ell(c)$ for acceptable positions. Hence $\gamma L_1 + (1 - \gamma)L_2 \in A$, so A is convex.

- Example: $\ell(x) = \exp(\alpha x)$ for some $\alpha > 0$. Then

$$\begin{aligned}\varrho_{\alpha,c}(L) &:= \inf\{m : \mathbb{E}(e^{\alpha(L-m)}) \leq e^{\alpha c}\} = \inf\{m : \mathbb{E}(e^{\alpha L}) \leq e^{\alpha c + \alpha m}\} \\ &= \frac{1}{\alpha} \log(\mathbb{E}(e^{\alpha L})) - c.\end{aligned}$$

Note that $\varrho_{\alpha,c}(0) = -c$, so $\varrho_{\alpha,c}$ cannot be coherent. For $c = 0$ and

$\lambda > 1$, the *entropic risk measure* $\varrho_{\alpha,0}$ satisfies

$$\varrho_{\alpha,0}(\lambda L) = \frac{1}{\alpha} \ln\{\mathbb{E}(e^{\alpha \lambda L})\} \geq \frac{1}{\alpha} \ln\{\mathbb{E}(e^{\alpha L})^\lambda\} = \lambda \varrho_{\alpha,0}(L),$$

where the inequality is strict if L is non-degenerate. This shows that $\varrho_{\alpha,0}$ is convex but not coherent. If L are insurance claims, $\varrho_{\alpha,0}$ is known as *exponential premium principle*.

Example 8.8 (Stress test or worst case risk measure)

Given *stress scenarios* $S \subseteq \Omega$, a *stress test risk measure* can be defined by

$$\varrho(L) = \sup\{L(\omega) : \omega \in S\},$$

that is, the worst loss on S . The associated acceptance set is

$$A_\varrho = \{L : L(\omega) \leq 0 \text{ for all } \omega \in S\}.$$

The choice of S is often guided by the underlying probability measure \mathbb{P} .

Example 8.9 (Generalized scenario risk measures)

Consider a set \mathcal{Q} of probability measures on (Ω, \mathcal{F}) and a *penalty function* $\gamma : \mathcal{Q} \rightarrow \mathbb{R}$ such that $\inf\{\gamma(Q) : Q \in \mathcal{Q}\} > -\infty$. Suppose $\sup_{Q \in \mathcal{Q}} \mathbb{E}_Q |L| < \infty$ for all $L \in \mathcal{M}$. The *generalized scenario risk measures* ϱ is defined by

$$\varrho(L) = \sup\{\mathbb{E}_Q(L) - \gamma(Q) : Q \in \mathcal{Q}\}. \quad (37)$$

The corresponding acceptance set is given by

$$A_\varrho = \{L \in \mathcal{M} : \sup\{\mathbb{E}_Q(L) - \gamma(Q) : Q \in \mathcal{Q}\} \leq 0\}.$$

- A_ϱ is convex, and thus so is ϱ .
- Every convex risk measure can be represented as (37); see Theorem 8.10.
- If $\gamma(\cdot) \equiv 0$ on \mathcal{Q} , ϱ is positive homogeneous and therefore coherent.
- The stress test risk measure of Example 8.8 is a special case of (37) in which $\gamma \equiv 0$ and \mathcal{Q} is the set of all Dirac measures $\delta_\omega(\cdot)$, $\omega \in S$, that is, $\delta_\omega(B) = I_B(\omega)$ for arbitrary measurable sets $B \subseteq \Omega$.

8.1.2 Dual representation of convex measures of risk

Theorem 8.10 (Dual representation for risk measures)

Suppose $|\Omega| = n < \infty$. Let $\mathcal{F} = \mathcal{P}(\Omega)$ (power set) and $\mathcal{M} := \{L : \Omega \rightarrow \mathbb{R}\}$. Then:

- 1) Every convex risk measure ϱ on \mathcal{M} can be written in the form

$$\varrho(L) = \max\{\mathbb{E}_{\mathbb{Q}}(L) - \alpha_{\min}(\mathbb{Q}) : \mathbb{Q} \in \mathcal{S}^1(\Omega, \mathcal{F})\}, \quad (38)$$

where $\mathcal{S}^1(\Omega, \mathcal{F})$ denotes the set of all probability measures on Ω , and where the penalty function α_{\min} is given by $\alpha_{\min}(\mathbb{Q}) = \sup\{\mathbb{E}_{\mathbb{Q}}(L) : L \in A_{\varrho}\}$.

- 2) If ϱ is coherent, it has the representation

$$\varrho(L) = \max\{\mathbb{E}_{\mathbb{Q}}(L) : \mathbb{Q} \in \mathcal{Q}\}$$

for some set $\mathcal{Q} = \mathcal{Q}(\varrho) \subseteq \mathcal{S}^1(\Omega, \mathcal{F})$.

One can show that $\alpha_{\min}(\mathbb{Q}) = \sup_{L \in \mathcal{M}} \{\mathbb{E}_{\mathbb{Q}}(L) - \varrho(L)\}$.

8.1.3 Examples of dual representations

Proposition 8.11 (ES formulas)

For $\alpha \in (0, 1)$,

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

Corollary 8.12 (ES formulas under continuous F_L)

Let F_L be continuous at $F_L^\leftarrow(\alpha)$. Then

- 1) $\text{ES}_\alpha(L) = \frac{\mathbb{E}(LI_{\{L > F_L^\leftarrow(\alpha)\}})}{1 - \alpha}$
- 2) $\text{ES}_\alpha(L) = \mathbb{E}(L | L > F_L^\leftarrow(\alpha))$ (i.e. *conditional VaR (CVaR)*)

With dual representations one can give a proof for ES_α being subadditive; see the following result.

Theorem 8.13

For $\alpha \in [0, 1]$, ES_α is coherent on $\mathcal{M} = \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. The dual representation is given by

$$\text{ES}_\alpha(L) = \max\{\mathbb{E}^{\mathbb{Q}}(L) : \mathbb{Q} \in \mathcal{Q}_\alpha\}, \quad (39)$$

where \mathcal{Q}_α is the set of all probability measures on (Ω, \mathcal{F}) that are absolutely continuous with respect to \mathbb{P} and for which the measure-theoretic density $d\mathbb{Q}/d\mathbb{P}$ is bounded by $1/(1 - \alpha)$.

8.2 Law-invariant coherent risk measures

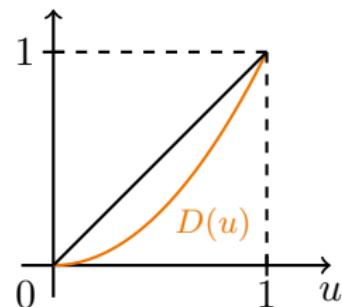
8.2.1 Distortion risk measures

Distortion risk measures are important coherent risk measures. We summarize important representations and investigate their properties.

Representations of distortion risk measures

Definition 8.14 (Distortion risk measure)

A *convex distortion function* D is a convex, increasing and absolutely continuous function on $[0, 1]$ satisfying $D(0) = 0$ and $D(1) = 1$.



The *distortion risk measure* associated with D is defined by

$$\varrho(L) = \int_0^1 F_L^\leftarrow(u) dD(u). \quad (40)$$

Note:

- A distortion risk measure is law-invariant (average of the L -quantiles).
- $D(u) = \int_0^u \phi(s) ds$ for an increasing, positive function ϕ (the right-sided derivative of D), hence

$$\varrho(L) = \int_0^1 F_L^\leftarrow(u) \phi(u) du. \quad (41)$$

A risk measure of this form is known as *spectral risk measure* and ϕ as *spectrum*.

- For $D_\alpha(u) = (1 - \alpha)^{-1}(u - \alpha)^+$ one obtains expected shortfall. The spectrum is $\phi(u) = (1 - \alpha)^{-1} I_{\{u \geq \alpha\}}$ (equal weight is placed on all quantiles beyond the α -quantile).

Lemma 8.15

The distortion risk measure ϱ associated with a convex distortion function D can be written in the form

$$\varrho(L) = \int_{\mathbb{R}} x \, dD \circ F_L(x), \quad (42)$$

where $D \circ F_L(x) = D(F_L(x))$.

Proof. $G(x) = D \circ F_L(x)$ has quantile function $G^\leftarrow = F_L^\leftarrow \circ D^\leftarrow$. Thus (42) can be written as

$$\int_{\mathbb{R}} x \, dG(x) \underset{u=G(x)}{=} \int_0^1 G^\leftarrow(u) \, du = \int_0^1 F_L^\leftarrow \circ D^\leftarrow(u) \, du = \mathbb{E}(F_L^\leftarrow \circ D^\leftarrow(U)),$$

where $U \sim U(0, 1)$. Now introduce $V = D^\leftarrow(U) \sim D$ and note that

$$\int_{\mathbb{R}} x \, dD \circ F_L(x) = \mathbb{E}(F_L^\leftarrow(V)) = \int_0^1 F_L^\leftarrow(v) \, dD(v). \quad \square$$

D distorts F_L . Since D is convex, $D(u) \leq u$, so $G = D \circ F_L$ puts more mass on high values of L than F_L .

Distortion risk measure can be represented as a weighted average of expected shortfall; see the appendix for a proof.

Proposition 8.16 (Distortion risk measures as weighted ES)

Let ϱ be a distortion risk measure associated with the convex distortion function D . Then, for a probability measure μ ,

$$\varrho(L) = \int_0^1 \text{ES}_\alpha(L) d\mu(\alpha).$$

Properties of distortion risk measures

Definition 8.17 (Comonotone additivity)

A risk measure ϱ on a space of random variables \mathcal{M} is said to be *comonotone additive* if $\varrho(L_1 + \dots + L_d) = \varrho(L_1) + \dots + \varrho(L_d)$ for comonotone L_1, \dots, L_d .

- Quantile functions (so value-at-risk) are comonotone additive. Comonotone additivity of distortion risk measures then follows from (40).

- Distortion risk measures are coherent. Monotonicity, translation invariance and positive homogeneity are obvious. Subadditivity follows from Proposition 8.16 and subadditivity of ES_α (e.g., Theorem 8.13) by observing that

$$\begin{aligned}\varrho(L_1 + L_2) &= \int_0^1 \text{ES}_\alpha(L_1 + L_2) d\mu(\alpha) \\ &\leq \int_0^1 \text{ES}_\alpha(L_1) d\mu(\alpha) + \int_0^1 \text{ES}_\alpha(L_2) d\mu(\alpha) \\ &= \varrho(L_1) + \varrho(L_2).\end{aligned}$$

- In summary, we have verified that distortion risk measures are law invariant, coherent and comonotone additive.
- It may also be shown that, on an atomless probability space (where there exists a continuous random variable), a law-invariant, coherent, comonotone-additive risk measure must be of the form (40) for some convex distortion function D .

- Parametric families of distortion risk measures can be based on convex distortion functions of the form

$$D_\alpha(u) = \Psi(\Psi^{-1}(u) + \ln(1 - \alpha)), \quad 0 \leq \alpha < 1,$$

where Ψ is a continuous df on \mathbb{R} ; for $\Psi(u) = 1 - \exp(-u)$, $u \geq 0$, one obtains the distortion function for ES.

- ▶ Such a family of convex distortion functions is strictly decreasing in α for fixed u .
- ▶ $D_0(u) = u$ (corresponding to the risk measure $\varrho(L) = \mathbb{E}(L)$) and $\lim_{\alpha \rightarrow 1} D(u) = 1_{\{u=1\}}$.
- ▶ For $\alpha_1 < \alpha_2$ and $0 < u < 1$ we have $D_{\alpha_1}(u) > D_{\alpha_2}(u)$, so that D_{α_2} distorts the original probability measure more than D_{α_1} and places more weight on outcomes in the tail.

8.2.2 The expectile risk measure

Definition 8.18 (Expectiles)

Let $L \in \mathcal{M} := L^1(\Omega, \mathcal{F}, \mathbb{P})$, so $\mathbb{E}|L| < \infty$. Then, for $\alpha \in (0, 1)$, the α -expectile $e_\alpha(L)$ is given by the unique solution y of

$$\alpha \mathbb{E}((L - y)^+) = (1 - \alpha) \mathbb{E}((L - y)^-) \quad (43)$$

where $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$.

- Since $x^+ - x^- = x$, $e_{0.5}(L) = \mathbb{E}(L)$ as $\mathbb{E}(L - y)^- = \mathbb{E}(L - y)^+$ iff $\mathbb{E}((L - y)^+ - (L - y)^-) = 0$ iff $\mathbb{E}(L - y) = 0$.
- $\mathbb{E}(L^2) < \infty$, $e_\alpha(L)$ is the minimizer of

$$\min_{y \in \mathbb{R}} \mathbb{E}(S(y, L)) \quad (44)$$

for *scoring function* $S(y, L)$. This could be relevant for the out-of-sample testing of expectile-estimates (so-called *backtesting*). The scoring func-

tion that yields the expectile is

$$S_\alpha^e(y, L) = |1_{\{L \leq y\}} - \alpha|(L - y)^2. \quad (45)$$

In fact we can compute that $\frac{d}{dy} \mathbb{E}(S_\alpha^e(y, L))$ equals

$$\begin{aligned} & \frac{d}{dy} \int_{-\infty}^{\infty} |1_{\{y \geq x\}} - \alpha|(y - x)^2 dF_L(x) \\ &= \frac{d}{dy} \int_{-\infty}^y (1 - \alpha)(y - x)^2 dF_L(x) + \frac{d}{dy} \int_y^{\infty} \alpha(y - x)^2 dF_L(x) \\ &= 2(1 - \alpha) \int_{-\infty}^y (y - x) dF_L(x) + 2\alpha \int_y^{\infty} (y - x) dF_L(x) \\ &= 2(1 - \alpha)\mathbb{E}((L - y)^-) - 2\alpha\mathbb{E}((L - y)^+) \end{aligned}$$

and setting this equal to zero yields the definition of an expectile.

- One can show that the α -quantile $F_L^\leftarrow(\alpha)$ is also a minimizer of the form (44); consider the scoring function $S_\alpha^q(y, L) = |1_{\{L \leq y\}} - \alpha||L - y|$.

The following result shows uniqueness of the α -expectile and provides a helpful formula for computing expectiles of certain distributions; see the appendix for a proof.

Proposition 8.19

Let $\alpha \in (0, 1)$ and L a rv such that $\mu := \mathbb{E}(L) < \infty$. Then $e_\alpha(L)$ may be written as $e_\alpha(L) = \tilde{F}_L^{-1}(\alpha)$ where

$$\tilde{F}_L(y) = \frac{yF_L(y) - \mu(y)}{2(yF_L(y) - \mu(y)) + \mu - y} \quad (46)$$

is a continuous df that is strictly increasing on its support and $\mu(y) := \int_{-\infty}^y x dF_L(x)$.

Example 8.20 (Bernoulli)

Let $L \sim \text{Be}(p)$ be a Bernoulli-distributed loss. Then

$$F_L(y) = \begin{cases} 0, & y < 0 \\ 1 - p, & 0 \leq y < 1, \\ 1, & y \geq 1 \end{cases} \quad \mu(y) = \begin{cases} 0, & y < 1 \\ p, & y \geq 1 \end{cases}$$

from which it follows that $\tilde{F}_L(y) = \frac{y(1-p)}{y(1-2p)+p}$, $0 \leq y \leq 1$ and

$$e_\alpha(L) = \frac{\alpha p}{(1-\alpha) + p(2\alpha-1)}.$$

Note that this can take any value in zero and one, whereas $\text{VaR}_\alpha(L) \in \{0, 1\}$, $\alpha \in (0, 1]$.

Properties of expectiles

Proposition 8.21 (Coherence of expectile risk measures)

$\varrho = e_\alpha$ is a coherent risk measure on $\mathcal{M} = L^1(\Omega, \mathcal{F}, \mathbb{P})$ for $\alpha \geq 0.5$.

- See the appendix for a proof.
- Expectiles are not comonotone additive and thus are not distortion risk measures.
- If L_1 and L_2 are comonotonic and of the same type (so that $L_2 = kL_1 + m$ for some $m \in \mathbb{R}$ and $k > 0$) then we do have comonotone additivity (by translation invariance and positive homogeneity), but for comonotonic variables that are not of the same type one can find examples where $e_\alpha(L_1 + L_2) < e_\alpha(L_1) + e_\alpha(L_2)$ for $\alpha > 0.5$.

8.3 Risk measures for linear portfolios

We now consider **linear portfolios** in

$$\mathcal{M} = \{L : L = m + \boldsymbol{\lambda}' \mathbf{X}, m \in \mathbb{R}, \boldsymbol{\lambda} \in \mathbb{R}^d\}, \quad (47)$$

for a fixed d -dimensional random vector \mathbf{X} .

- Many standard approaches to risk aggregation and capital allocation are based on the assumption that losses have a linear relationship to underlying risk factor changes.
- It is common to use linear approximations for losses due to market risks over short time horizons.

8.3.1 Coherent risk measures as stress tests

- Let $\varrho : \mathcal{M} \rightarrow \mathbb{R}$ be a positive-homogeneous risk measure. Define a *risk-measure function* $r_\varrho(\boldsymbol{\lambda}) = \varrho(\boldsymbol{\lambda}' \mathbf{X})$ (function of portfolio weights).

- If ϱ is translation-invariant, there is a one-to-one relationship between ϱ and r_ϱ given by

$$\varrho(m + \boldsymbol{\lambda}' \mathbf{X}) = m + r_\varrho(\boldsymbol{\lambda}).$$

Lemma 8.22 (Properties of r_ϱ)

Consider a translation-invariant risk measure $\varrho : \mathcal{M} \rightarrow \mathbb{R}$ with associated risk-measure function r_ϱ . Then

- 1) ϱ is a positive-homogeneous risk measure if and only if r_ϱ is a positive-homogeneous function, that is $r_\varrho(t\boldsymbol{\lambda}) = tr_\varrho(\boldsymbol{\lambda})$ for all $t > 0$, $\boldsymbol{\lambda} \in \mathbb{R}^d$.
- 2) Suppose that ϱ is positive-homogeneous. Then ϱ is subadditive if and only if r_ϱ is convex.

The main result of this section is that coherent risk measures for linear portfolios are stress tests as in Example 8.8 where the scenario set is

$$S_\varrho := \{ \mathbf{x} \in \mathbb{R}^d : \mathbf{u}' \mathbf{x} \leq r_\varrho(\mathbf{u}) \text{ for all } \mathbf{u} \in \mathbb{R}^d \}.$$

Proposition 8.23 (Coherent risk measures for linear portfolios as stress tests)

ϱ is a coherent risk measure on the set of linear portfolios \mathcal{M} in (47) if and only if for every $L = m + \boldsymbol{\lambda}' \mathbf{X} \in \mathcal{M}$ we have the representation

$$\varrho(L) = m + r_\varrho(\boldsymbol{\lambda}) = \sup\{m + \boldsymbol{\lambda}' \mathbf{x} : \mathbf{x} \in S_\varrho\}. \quad (48)$$

- S_ϱ is an intersection of the half-spaces $H_u = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{u}' \mathbf{x} \leq r_\varrho(\mathbf{u})\}$, so that S_ϱ is a closed convex set. The precise form of S_ϱ depends on the df of \mathbf{X} and on ϱ .
- If $\varrho = \text{VaR}_\alpha$, S_ϱ has an interpretation as a *depth set*. Suppose that $\mathbf{u}' \mathbf{X}$ is continuously distributed for all $\mathbf{u} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. Then for $H_u = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{u}' \mathbf{x} \leq \text{VaR}_\alpha(\mathbf{u}' \mathbf{X})\}$, $\mathbb{P}(\mathbf{u}' \mathbf{X} \in H_u) = \alpha$ so that S_{VaR_α} is the intersection of all half-spaces with probability α .

8.3.2 Elliptically distributed risk factors

Theorem 8.24 (Risk measurement for elliptical risk factors)

Let $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ and ϱ be any positive-homogeneous, translation-invariant and law-invariant risk measure on \mathcal{M} . Then:

- 1) For any $L = m + \boldsymbol{\lambda}' \mathbf{X} \in \mathcal{M}$, $\varrho(L) = m + \boldsymbol{\lambda}' \boldsymbol{\mu} + \sqrt{\boldsymbol{\lambda}' \Sigma \boldsymbol{\lambda}} \varrho(Y_1)$ for $Y_1 \sim S_1(\psi)$.
- 2) If $\varrho(Y_1) \geq 0$, then ϱ is subadditive on \mathcal{M} (e.g., VaR_α for $\alpha \geq 0.5$).
- 3) If $\mathbb{E}\mathbf{X}$ exists then, $\forall L = m + \boldsymbol{\lambda}' \mathbf{X} \in \mathcal{M}$ and $\rho_{ij} = \varrho(\Sigma)_{ij} = P_{ij}$,

$$\varrho(L - \mathbb{E}L) = \sqrt{\sum_{i=1}^d \sum_{j=1}^d \rho_{ij} \lambda_i \lambda_j \varrho(X_i - \mathbb{E}X_i) \varrho(X_j - \mathbb{E}X_j)}.$$

- 4) If $\text{cov}(\mathbf{X})$ exists and $\varrho(Y_1) > 0$ then, for every $L \in \mathcal{M}$,
 $\varrho(L) = \mathbb{E}(L) + k_\varrho \sqrt{\text{var}(L)}$ for some $k_\varrho > 0$ depending on ϱ .
- 5) If Σ^{-1} ex., $\varrho(Y_1) > 0$ then $S_\varrho = \{\mathbf{x} : (\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq \varrho(Y_1)^2\}$.

Proof.

- 1) Let $\mathbf{Y} \sim S_k(\psi)$, $AA' = \Sigma$. $L = m + \boldsymbol{\lambda}' \mathbf{X} \stackrel{d}{=} m + \boldsymbol{\lambda}' \boldsymbol{\mu} + \boldsymbol{\lambda}' A \mathbf{Y}$. By Theorem 6.15 3), $L \stackrel{d}{=} m + \boldsymbol{\lambda}' \boldsymbol{\mu} + \|A' \boldsymbol{\lambda}\| Y_1$. Thus $\varrho(L) = m + \boldsymbol{\lambda}' \boldsymbol{\mu} + \|A' \boldsymbol{\lambda}\| \varrho(Y_1) = m + \boldsymbol{\lambda}' \boldsymbol{\mu} + \sqrt{\boldsymbol{\lambda}' \Sigma \boldsymbol{\lambda}} \varrho(Y_1)$.
- 2) Set $L_1 = m_1 + \boldsymbol{\lambda}'_1 \mathbf{X}$ and $L_2 = m_2 + \boldsymbol{\lambda}'_2 \mathbf{X}$. Subadditivity follows from 1) and $\|A'(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)\| \leq \|A' \boldsymbol{\lambda}_1\| + \|A' \boldsymbol{\lambda}_2\|$ and $\varrho(Y_1) \geq 0$.
- 3) $\varrho(L - \mathbb{E}L) = \varrho(L) - \mathbb{E}(L) = \varrho(L) - (m + \boldsymbol{\lambda}' \boldsymbol{\mu}) = \sqrt{\boldsymbol{\lambda}' \Sigma \boldsymbol{\lambda}} \varrho(Y_1)$, so

$$\varrho(L - \mathbb{E}L) = \sqrt{\sum_{i=1}^d \sum_{j=1}^d \rho_{ij} \lambda_i \lambda_j \sigma_i \sigma_j \varrho(Y_1)},$$

where $\sigma_j = \sqrt{\Sigma_{jj}}$ for $j \in \{1, \dots, d\}$. For $\boldsymbol{\lambda} = \mathbf{e}_j$, $\varrho(X_j - \mathbb{E}X_j) = \varrho(\mathbf{e}'_j \mathbf{X} - \mathbb{E}(\mathbf{e}'_j \mathbf{X})) = \sigma_j \varrho(Y_1)$, from which the result follows.

- 4) $\text{cov}(\mathbf{X}) = c\Sigma$ for some $c > 0$. Since $\text{var}(L) = \text{var}(\boldsymbol{\lambda}' \mathbf{X}) = \boldsymbol{\lambda}' c \Sigma \boldsymbol{\lambda}$, 3) implies that $\varrho(L) = \mathbb{E}(L) + \sqrt{\boldsymbol{\lambda}' \Sigma \boldsymbol{\lambda}} \varrho(Y_1) = \mathbb{E}(L) + \sqrt{\text{var}(L)} \varrho(Y_1) / \sqrt{c}$.

5) 2) implies that $r_\varrho(\boldsymbol{\lambda}) = \|A'\boldsymbol{\lambda}\|\varrho(Y_1) + \boldsymbol{\lambda}'\boldsymbol{\mu}$ so that S_ϱ is

$$\begin{aligned} S_\varrho &= \left\{ \mathbf{x} \in \mathbb{R}^d : \mathbf{u}'\mathbf{x} \leq \mathbf{u}'\boldsymbol{\mu} + \|A'\mathbf{u}\| \varrho(Y_1) \quad \forall \mathbf{u} \in \mathbb{R}^d \right\} \\ &= \left\{ \mathbf{x} \in \mathbb{R}^d : \mathbf{u}'AA^{-1}(\mathbf{x} - \boldsymbol{\mu}) \leq \|A'\mathbf{u}\| \varrho(Y_1) \quad \forall \mathbf{u} \in \mathbb{R}^d \right\} \\ &= \left\{ \mathbf{x} \in \mathbb{R}^d : \mathbf{v}' \frac{A^{-1}(\mathbf{x} - \boldsymbol{\mu})}{\varrho(Y_1)} \leq \|\mathbf{v}\| \quad \forall \mathbf{v} \in \mathbb{R}^d \right\}, \end{aligned}$$

where the last line follows because $\mathbb{R}^d = \{A'\mathbf{u} : \mathbf{u} \in \mathbb{R}^d\}$. Since $\{\mathbf{y} \in \mathbb{R}^d : \mathbf{y}'\mathbf{y} \leq 1\}$ can be written as $\{\mathbf{y} \in \mathbb{R}^d : \mathbf{v}'\mathbf{y} \leq \|\mathbf{v}\| \quad \forall \mathbf{v} \in \mathbb{R}^d\}$, we conclude that, for $\mathbf{x} \in S_\varrho$, the vectors $\mathbf{y} = A^{-1}(\mathbf{x} - \boldsymbol{\mu})/\varrho(Y_1)$ describe the unit ball and therefore

$$S_\varrho = \{ \mathbf{x} \in \mathbb{R}^d : (\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) \leq \varrho(Y_1)^2 \}. \quad \square$$

- 2) gives a special case where VaR is subadditive and thus coherent. In particular, if (L_1, \dots, L_d) is jointly elliptical, VaR_α is subadditive for $\alpha \geq 0.5$.
- 3) provides a useful interpretation of risk measures on \mathcal{M} in terms of the aggregation of stress tests.
- 4) is relevant to portfolio optimization. If we consider losses $L \in \mathcal{M}$ for which $\mathbb{E}(L)$ is fixed, the weights that minimize ϱ also minimize the variance. The portfolio minimizing ϱ is thus the same as the Markowitz variance-minimizing portfolio.
- 5) shows that the scenario sets in the stress test representation of coherent risk measures are ellipsoids when X is elliptical. Different law-invariant coherent risk measures simply lead to ellipsoids of differing radius $\varrho(Y_1)$. Scenario sets of ellipsoidal form are often used in practice and this result provides a justification for this practice in the case of linear portfolios of elliptical risk factors.

8.4 Risk aggregation

- A *risk aggregation rule* is a mapping

$$f(\text{EC}_1, \dots, \text{EC}_d) = \text{EC}$$

which maps the individual capital amounts $\text{EC}_1, \dots, \text{EC}_d$ to the aggregate capital EC (economic capital). Examples are:

- ▶ *Simple summation* $\text{EC} = \text{EC}_1 + \dots + \text{EC}_d$ (a special case of and upper bound for correlation adjusted summation)
- ▶ *Correlation adjusted summation*

$$\text{EC} = \sqrt{\sum_{i=1}^d \sum_{j=1}^d \rho_{ij} \text{EC}_i \text{EC}_j}, \quad (49)$$

where $\rho_{ij} \in [0, 1]$ are parameters (referred to as *correlations*).

- Applying such rules *without* considering a multivariate model or risk measures is known as *rules-based aggregation*, otherwise, *principles-based aggregation*; we focus on the latter.
- In what follows we show that correlation adjusted summation is justified as a risk aggregation rule under various setups.

8.4.1 Aggregation based on loss distributions

- Suppose that the *overall loss* is $L = L_1 + \dots + L_d$ where L_1, \dots, L_d are the losses arising from sub-units (e.g., business units, asset classes). Consider a translation-invariant ϱ and define

$$\varrho^{\text{mean}}(\cdot) = \varrho(\cdot - \mathbb{E}(\cdot)) = \varrho(\cdot) - \mathbb{E}(\cdot),$$

that is, the capital required to cover unexpected losses.

- The *capital requirements for the sub-units* are

$$\text{EC}_j = \varrho^{\text{mean}}(L_j), \quad j \in \{1, \dots, d\},$$

and the aggregate capital should be

$$\text{EC} = \varrho^{\text{mean}}(L).$$

- We require an aggregation rule f such that $f(\text{EC}_1, \dots, \text{EC}_d) = \text{EC}$.
- If $\varrho(L) = \mathbb{E}(L) + k \text{sd}(L)$, $k > 0$, and $\mathbb{E}(L^2) < \infty$ then

$$\text{sd}(L) = \sqrt{\text{var}(\mathbf{1}' \mathbf{L})} = \sqrt{\mathbf{1}' \text{cov}(\mathbf{L}) \mathbf{1}} = \sqrt{\sum_{i=1}^d \sum_{j=1}^d \rho_{ij} \text{sd}(L_i) \text{sd}(L_j)},$$

where $(\rho_{ij})_{i,j} = \text{corr}(\mathbf{L})$, so correlation adjusted summation follows by noting that $\text{sd}(L) = \varrho^{\text{mean}}(L)/k = \text{EC}/k$ (and $\text{sd}(L_j) = \text{EC}_j/k$).

- If $L_j = m_j + \lambda'_j \mathbf{X}$ for $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ with existing $\text{cov}(\mathbf{X})$, then this formula and Theorem 8.24 4) imply that correlation adjusted summation is justified for any positive-homogeneous, translation-invariant and law-invariant risk measure ϱ .
- As the following result shows, the assumption on $\text{cov}(\mathbf{X})$ can be dropped.

Proposition 8.25 (Correlation adjusted sum. for linear portfolios)

Let $\mathbf{X} \sim E_k(\boldsymbol{\mu}, \Sigma, \psi)$ with $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$. Let $\mathcal{M} = \{L : L = m + \boldsymbol{\lambda}'\mathbf{X}, \boldsymbol{\lambda} \in \mathbb{R}^k, m \in \mathbb{R}\}$ and ϱ be a pos.-hom., translation- and law-invariant risk measure on \mathcal{M} . For $L_1, \dots, L_d \in \mathcal{M}$, let $\text{EC}_j = \varrho^{\text{mean}}(L_j)$ and $\text{EC} = \varrho^{\text{mean}}(L_1 + \dots + L_d)$. Then $\text{EC}, \text{EC}_1, \dots, \text{EC}_d$ satisfy the correlation adjusted summation for $P = \wp(\tilde{\Sigma}) = (\rho_{ij})_{ij}$ and $\tilde{\Sigma}$ is the scale matrix of the (elliptical) (L_1, \dots, L_d) .

Proof. Let $L_j = m_j + \boldsymbol{\lambda}'_j \mathbf{X}$. By Theorem 8.24 1), $\text{EC}_j = \varrho(L_j) - \mathbb{E}(L_j) = \sqrt{\boldsymbol{\lambda}'_j \Sigma \boldsymbol{\lambda}_j} \varrho(Y_1)$ where $Y_1 \sim S_1(\psi)$ and that

$$\begin{aligned} \text{EC} &= \sqrt{(\boldsymbol{\lambda}_1 + \dots + \boldsymbol{\lambda}_d)' \Sigma (\boldsymbol{\lambda}_1 + \dots + \boldsymbol{\lambda}_d)} \varrho(Y_1) \\ &= \sqrt{\sum_{i=1}^d \sum_{j=1}^d \boldsymbol{\lambda}'_i \Sigma \boldsymbol{\lambda}_j \varrho(Y_1)^2} = \sqrt{\sum_{i=1}^d \sum_{j=1}^d \frac{\boldsymbol{\lambda}'_i \Sigma \boldsymbol{\lambda}_j}{\sqrt{(\boldsymbol{\lambda}'_i \Sigma \boldsymbol{\lambda}_i)(\boldsymbol{\lambda}'_j \Sigma \boldsymbol{\lambda}_j)}} \text{EC}_i \text{EC}_j}. \end{aligned}$$

The scale matrix $\tilde{\Sigma}$ of (L_1, \dots, L_d) is $\tilde{\Sigma} = \Lambda \Sigma \Lambda'$ where $\Lambda = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_d)'$. The corresponding $P = (\rho_{ij})_{ij}$ has elements $\boldsymbol{\lambda}'_i \Sigma \boldsymbol{\lambda}_j / \sqrt{(\boldsymbol{\lambda}'_i \Sigma \boldsymbol{\lambda}_i)(\boldsymbol{\lambda}'_j \Sigma \boldsymbol{\lambda}_j)}$ and thus

$$\text{EC} = \sqrt{\sum_{i=1}^d \sum_{j=1}^d \frac{\boldsymbol{\lambda}'_i \Sigma \boldsymbol{\lambda}_j}{\sqrt{(\boldsymbol{\lambda}'_i \Sigma \boldsymbol{\lambda}_i)(\boldsymbol{\lambda}'_j \Sigma \boldsymbol{\lambda}_j)}} \text{EC}_i \text{EC}_j} = \sqrt{\sum_{i=1}^d \sum_{j=1}^d \rho_{ij} \text{EC}_i \text{EC}_j}. \quad \square$$

- Correlation adjusted summation can thus be justified under the mean-adjusted VaR or ES if L is elliptical.
- The correlations ρ_{ij} between L_1, \dots, L_d are typically difficult to estimate (data is rather available for risk factors than losses). If they are chosen by *expert judgement*, there are compatibility requirements. If (L_1, \dots, L_d) is non-elliptical, the limited range of attainable correlations for each pair (L_i, L_j) is also a relevant constraint; see Chapter 7.
- No obvious way to incorporate tail dependence between L_1, \dots, L_d .
- Simple summation only offers a conservative upper bound if ϱ is coherent.

8.4.2 Aggregation based on stressing risk factors

- Correlation adjusted summation is used in the aggregation of capital contributions $\text{EC}_1, \dots, \text{EC}_d$ computed by stressing individual risk factors (example: Standard formula approach to Solvency II).
- Let $\boldsymbol{x} = \mathbf{X}(\omega)$ be a scenario defined in terms of changes in risk factors and $L(\boldsymbol{x})$ the corresponding loss. Assume $L(\boldsymbol{x})$ is known and componentwise increasing.
- The d risk factors are stressed individually by amounts k_1, \dots, k_d . Capital contributions for each risk factor are computed by

$$\text{EC}_j = L(k_j \mathbf{e}_j) - L(\mathbb{E}(X_j) \mathbf{e}_j)$$

where $k_j > \mathbb{E}(X_j)$ so that $\text{EC}_j > 0$ (interpreted as the loss incurred by stressing risk factor j by k_j relative to the impact of stressing it by its expected change); an example is $k_j = q_\alpha(X_j)$ for large α .

- The following justifies correlation adjusted summation as a risk aggregation rule if $k_j = \varrho(X_j)$ for elliptical \mathbf{X} and $L(\mathbf{X}) = m + \boldsymbol{\lambda}' \mathbf{X}$.

Proposition 8.26 (Justification for correlation adjusted summation)

Let $\mathbf{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ with $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$. Let \mathcal{M} be the space of linear portfolios (47) and ϱ be a pos. hom., translation- and law-invariant risk measure on \mathcal{M} . Then, for any $L = L(\mathbf{X}) = m + \boldsymbol{\lambda}'\mathbf{X} \in \mathcal{M}$,

$$\text{EC} = \varrho(L - \mathbb{E}(L)) = \sqrt{\sum_{i=1}^d \sum_{j=1}^d \rho_{ij} \text{EC}_i \text{EC}_j},$$

where $\text{EC}_j = L(\varrho(X_j)\mathbf{e}_j) - L(\mathbb{E}(X_j)\mathbf{e}_j)$ and $\rho_{ij} = \varphi(\Sigma)_{i,j}$.

Proof. Note that $\text{EC}_j = m + \lambda_j \varrho(X_j) - (m + \lambda_j \mathbb{E}X_j) = \lambda_j \varrho(X_j - \mathbb{E}X_j)$ and plug this into Theorem 8.24 3) to see that the claim holds. \square

- Thus under linearity of the losses in jointly elliptical risk-factor changes, we can aggregate the effects of single-risk-factor stresses to an aggregate capital; this applies to VaR, ES or distortion risk measures. This idea underscores correlation adjusted summation in Solvency II.

- For market risk factors (returns on prices), the data may be available to estimate the ρ_{ijs} . For other risk factors (e.g. mortality and policy lapse rates in Solvency II), they are set by expert judgement (see issues mentioned earlier).

8.4.3 Modular versus fully integrated aggregation approaches

- The approaches of Sections 8.4.1 and 8.4.2 are *modular approaches*. In Sections 8.4.1 the *modules* (or *silos*) are business units or asset classes; in Section 8.4.2 they were individual risk factors; the former approach is more natural because losses are additive (and it is possible to remove risks from the enterprise by selling parts of the business).
- The aggregation approaches involved correlations and the correlation adjusted summation; however, correlations give only a partial description of dependence. It is natural to consider using copulas in aggregation.

- Consider simple summation and suppose we know/have estimated the marginal distributions F_1, \dots, F_d for each of the modules (necessary for computing $\text{EC}_j = \varrho(L_j) - \mathbb{E}(L_j)$). In the *margins-plus-copula approach*, we could attempt to choose a suitable copula C for $\mathbf{L} \sim F(\mathbf{x}) = C(F_1(x_1), \dots, F_d(x_d))$; see the converse of Sklar's Theorem. Computing the aggregate capital is then typically done by simulation and estimating the risk measures empirically.
- Problems: (Mis)specification of the copula C (*dependence uncertainty*); Data from \mathbf{L} is typically sparse.
- It is generally easier to follow a *fully integrated approach* by building a margins-plus-copula model or more dynamic models (*economic scenario generators*) for the risk-factor changes $\mathbf{X} = (X_1, \dots, X_k)$ (more data exists) and for $g_j : \mathbb{R}^k \mapsto \mathbb{R}$ which give the losses $L_j = g_j(\mathbf{X})$, $j \in \{1, \dots, d\}$, for the different portfolios/business units. Risk measures are then derived from the distribution of $L = g_1(\mathbf{X}) + \dots + g_d(\mathbf{X})$.

8.4.4 Risk aggregation and Fréchet problems

- Consider the margins-plus-copula approach where $L_j \sim F_j$, $j \in \{1, \dots, d\}$, are treated as known (estimated or postulated) and C is unknown.
- Consider $L = L_1 + \dots + L_d$. Due to the unknown C (*dependence uncertainty*), risk measures can no longer be computed explicitly.
- Our goal is to find bounds on VaR_α and ES_α under all possible C . Let

$$\mathcal{S}_d := \mathcal{S}_d(F_1, \dots, F_d) := \left\{ L = \sum_{j=1}^d L_j : L_j \sim F_j, j = 1, \dots, d \right\}$$

and consider

$$\bar{\varrho}(L) := \bar{\varrho}(\mathcal{S}_d) := \sup\{\varrho(L) : L \in \mathcal{S}_d(F_1, \dots, F_d)\} \quad (\text{worst } \varrho)$$

$$\underline{\varrho}(L) := \underline{\varrho}(\mathcal{S}_d) := \inf\{\varrho(L) : L \in \mathcal{S}_d(F_1, \dots, F_d)\} \quad (\text{best } \varrho)$$

- If $\varrho = \text{ES}_\alpha$, $\overline{\text{ES}}_\alpha(L) = \sum_{j=1}^d \text{ES}_\alpha(L_j)$ (subadditivity, com. additivity). $\underline{\text{ES}}_\alpha$, $\underline{\text{VaR}}_\alpha$, $\overline{\text{VaR}}_\alpha$ depend on whether the portfolio is *homogeneous* (that is, $F_1 = \dots = F_d$); we focus on $\overline{\text{VaR}}_\alpha$.

Summary of existing results

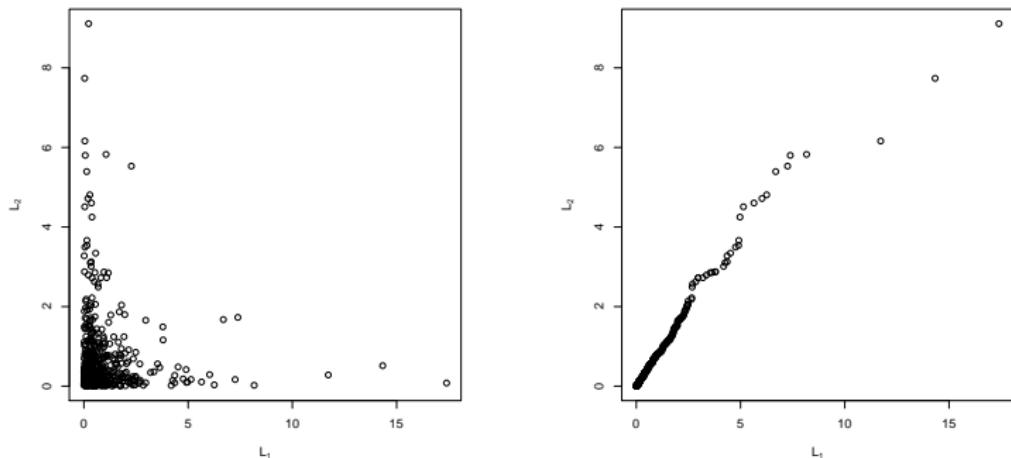
$d = 2$: Fully solved analytically

$d \geq 3$: Here we distinguish:

- ▶ **Homogeneous case ($F_1 = \dots = F_d$):**
 - $\underline{\text{ES}}_\alpha(L)$ solved analytically for decreasing densities (e.g. Pareto, Exponential)
 - $\underline{\text{VaR}}_\alpha(L)$, $\overline{\text{VaR}}_\alpha(L)$ solved analytically for tail-decreasing densities (e.g. Pareto, Log-normal, Gamma)
- ▶ **Inhomogeneous case:**
 - Few **analytical** results: current research
 - **Numerical** methods: **(Adaptive/Block) Rearrangement Algorithm**

The general problem

- We have one-period risks $L_1 \sim F_1, \dots, L_d \sim F_d$ with **given** (estimated or postulated) F_1, \dots, F_d and **unknown copula C** and want to compute $\overline{\text{VaR}}_\alpha(L)$ for $L = L_1 + \dots + L_d$.
- Iman and Conover (1982) idea for Par(2), Par(2.5) sample of size 500:



⇒ Reordering columns changes the dependence of (L_1, L_2) and F_L .

Proposition 8.27 (VaR_α in the homogeneous case)

Let $F := F_1 = \dots = F_d$ with decreasing density on $[b, \infty)$. Then, for $\alpha \in [F(b), 1]$ and $X \sim F$,

$$\overline{\text{VaR}}_\alpha(\mathcal{S}_d) = d\mathbb{E}(X \mid X \in [F^{-1}(\alpha + (d-1)c), F^{-1}(1-c)]),$$

where c is the smallest number in $[0, (1-\alpha)/d]$ such that

$$\int_{\alpha+(d-1)c}^{1-c} F^{-1}(t) dt \geq \frac{1-\alpha-dc}{d}((d-1)F^{-1}(\alpha + (d-1)c) + F^{-1}(1-c)).$$

If the density f of F is decreasing on its support, then for $\alpha \in (0, 1)$,

$$\underline{\text{VaR}}_\alpha(\mathcal{S}_d) = \max\{(d-1)F^{-1}(0) + F^{-1}(\alpha), d\mathbb{E}(X \mid X \leq F^{-1}(\alpha))\}.$$

Proof. See Wang et al. (2013) and Bernard et al. (2014). □

- The underlying numerics are non-trivial; see Hofert et al. (2015) and `qrmttools::VaR_bounds_hom()`.

Proposition 8.28 ($\underline{\text{ES}}_\alpha$ in the homogeneous case)

Let $F := F_1 = \dots = F_d$ with finite first moment and decreasing density on its support. Then, for $\alpha \in [1 - dc, 1]$, $\beta = (1 - \alpha)/d$, and $X \sim F$,

$$\begin{aligned}\underline{\text{ES}}_\alpha(\mathcal{S}_d) &= \frac{1}{\beta} \int_0^\beta ((d-1)F^{-1}((d-1)t) + F^{-1}(1-t)) dt \\ &= (d-1)^2 \text{LES}_{(d-1)\beta}(X) + \text{ES}_{1-\beta}(X),\end{aligned}$$

where c is the smallest number in $[0, 1/d]$ such that

$$\int_{(d-1)c}^{1-c} F^{-1}(t) dt \geq \frac{1-dc}{d} ((d-1)F^{-1}((d-1)c) + F^{-1}(1-c))$$

and $\text{LES}_\alpha(X) = \frac{1}{\alpha} \int_0^\alpha \text{VaR}_u(X) du = -\text{ES}_{1-\alpha}(-X)$ (*lower ES*).

Proof. See Bernard et al. (2014). □

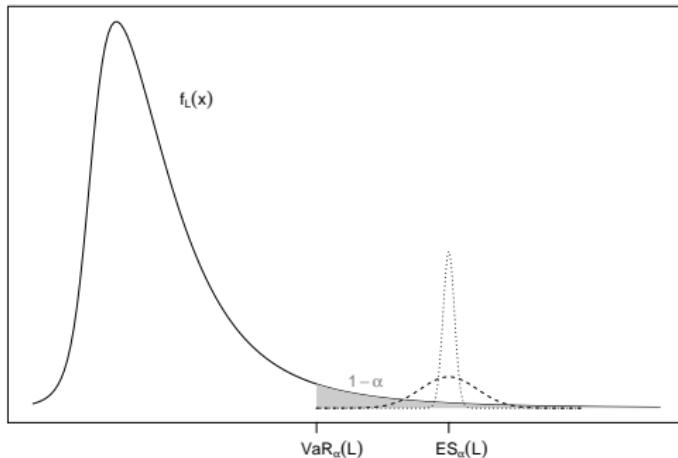
The Rearrangement Algorithm (RA)

- Two columns \mathbf{a}, \mathbf{b} are *oppositely ordered* if $(a_i - a_j)(b_i - b_j) \leq 0 \forall i, j$.
- *Minimum row-sum operator* $s(X) = \min_{1 \leq i \leq N} \sum_{1 \leq j \leq d} x_{ij}$

Algorithm 8.29 (RA for computing $\overline{\text{VaR}}_\alpha(L)$)

- 1) Fix $\alpha \in (0, 1)$, $F_1^\leftarrow, \dots, F_d^\leftarrow$, $N \in \mathbb{N}$ (# of discr. points), $\varepsilon \geq 0$ (tol.)
- 2) Compute the lower bound \underline{s}_N :
 - 2.1) Define the (N, d) -matrix $\underline{X}^\alpha = \left(F_j^\leftarrow \left(\alpha + \frac{(1-\alpha)(i-1)}{N} \right) \right)_{i,j}$.
 - 2.2) Randomly permute each column of \underline{X}^α (to avoid $\bar{s}_N - \underline{s}_N \rightarrow 0$)
 - 2.3) Iterate over all columns of \underline{X}^α and oppositely order each to the sum of all others \Rightarrow Matrix \underline{Y}^α
 - 2.4) Repeat Step 2.3) until $s(\underline{Y}^\alpha) - s(\underline{X}^\alpha) \leq \varepsilon$, then set $\underline{s}_N = s(\underline{Y}^\alpha)$.
- 3) Similarly, compute $\bar{s}_N = s(\bar{Y}^\alpha)$ based on $\bar{X}^\alpha = \left(F_j^\leftarrow \left(\alpha + \frac{(1-\alpha)i}{N} \right) \right)_{i,j}$.
- 4) Return $(\underline{s}_N, \bar{s}_N)$ (*rearrangement range*; taken as bounds on $\overline{\text{VaR}}_\alpha(L)$)

- The RA aims at maximizing the minimal row sums (solving a *maximin problem*; minimax problem for $\underline{\text{VaR}}_\alpha$).
- Intuition:** A *completely mixable matrix* (equal row sums), would minimize the variance of $L | L > F_L^-(\alpha)$ and thus concentrate more of the $1 - \alpha$ mass of F_L around the constant $\mathbb{E}[L | L > \text{VaR}_\alpha(L)]$ cont.. $\text{ES}_\alpha(L) \geq \text{VaR}_\alpha(L)$, so $\text{VaR}_\alpha(L)$ increases (F_L jumps to 1 in $\text{VaR}_\alpha(L)$ so $\text{VaR}_\alpha(L)$ is largest).



Example 8.30 (How the RA works)

1) Where it works (to compute the maximal minimal row sum):

$$\begin{array}{c}
 \begin{pmatrix} 1 & 1 & 1 \\ 2 & 3 & 2 \\ 3 & 5 & 4 \\ 4 & 7 & 8 \end{pmatrix} \xrightarrow{\sum_{-1} = \begin{pmatrix} 2 \\ 5 \\ 9 \\ 15 \end{pmatrix}} \begin{pmatrix} 4 & 1 & 1 \\ 3 & 3 & 2 \\ 2 & 5 & 4 \\ 1 & 7 & 8 \end{pmatrix} \xrightarrow{\text{here: stable}} \begin{pmatrix} 4 & 5 & 1 \\ 3 & 7 & 2 \\ 2 & 3 & 4 \\ 1 & 1 & 8 \end{pmatrix} \xrightarrow{\sum_{-3} = \begin{pmatrix} 9 \\ 10 \\ 5 \\ 2 \end{pmatrix}}
 \\
 \begin{pmatrix} 4 & 5 & 2 \\ 3 & 7 & 1 \\ 2 & 3 & 4 \\ 1 & 1 & 8 \end{pmatrix} \xrightarrow{\text{here: not}} \begin{pmatrix} 3 & 5 & 2 \\ 2 & 7 & 1 \\ 4 & 3 & 4 \\ 1 & 1 & 8 \end{pmatrix} \xrightarrow{\sum = \begin{pmatrix} 10 \\ 10 \\ 11 \\ 10 \end{pmatrix}} \widehat{\text{VaR}}_\alpha(L^+) \approx 10
 \end{array}$$

2) The RA can also fail:

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 3 & 3 \end{pmatrix} \xrightarrow{\sum_{-1} = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}} \begin{pmatrix} 3 & 1 & 1 \\ 2 & 2 & 2 \\ 1 & 3 & 3 \end{pmatrix} \xrightarrow{\sum = \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix}} \widehat{\text{VaR}}_\alpha(L^+) \approx 5 < 6$$

Example 8.31 (Par(θ) margins)

Let $L_j \sim \text{Par}(\theta)$ with $\bar{F}_j(x) = (1+x)^{-\theta}$, $j \in \{1, \dots, d\}$ (homogeneous case) and $\alpha = 0.999$. One obtains:

	$d = 8$		$d = 56$	
	$\theta = 2$	$\theta = 0.8$	$\theta = 2$	$\theta = 0.8$
$\overline{\text{VaR}}_\alpha(L)$	465	300 182	3454	4 683 172
$\text{VaR}_\alpha^+(L) = d \text{VaR}_\alpha(L_1)$	245	44 979	1715	314 855
$\text{VaR}_\alpha^\perp(L)$	96	75 877	293	862 855
$\underline{\text{VaR}}_\alpha(L)$	31	5622	53	5622
$\overline{\text{ES}}_\alpha(L) = d \text{ES}_\alpha(L_1)$	498	–	3486	–
$\text{ES}_\alpha^\perp(L)$	184	–	518	–
$\underline{\text{ES}}_\alpha(L)$	178	–	472	–

- The “+” and “ \perp ” denote the comonotonic and independent case, resp.
- $\frac{\overline{\text{ES}}_\alpha(L)}{\overline{\text{VaR}}_\alpha(L)} \underset{d \uparrow \infty}{\approx} 1$ can be explained; see McNeil et al. (2015, Prop. 8.36).
- The dependence uncertainty spread $\overline{\text{VaR}}_\alpha(L) - \underline{\text{VaR}}_\alpha(L) \geq \overline{\text{ES}}_\alpha(L) - \underline{\text{ES}}_\alpha(L)$ can be explained; see McNeil et al. (2015, Prop. 8.37).

Remark 8.32

- The RA finds approximate solutions to *maximin* (for $\overline{\text{VaR}}_\alpha(L)$) and *minimax* (for $\underline{\text{VaR}}_\alpha(L)$) *problems* and is thus of wider interest (e.g., in Operations Research).
- For $\underline{\text{ES}}_\alpha(L)$, discretize the whole support of each margin, rearrange, and approximate $\underline{\text{ES}}_\alpha(L)$ by the nonparametric ES_α estimate of the row sums.
- The *Adaptive Rearrangement Algorithm (ARA)*
 - ▶ uses relative (instead of absolute) individual tolerances;
 - ▶ uses a relative joint tolerance to guarantee that \underline{s}_N and \bar{s}_N are close;
 - ▶ chooses N adaptively to reach the joint tolerance; and
 - ▶ determines convergence after each rearranged column.
- The *Block Rearrangement Algorithm* rearranges blocks of columns.

Proposition 8.33 (Asymptotic equivalence of $\overline{\text{VaR}}_\alpha$, $\overline{\text{ES}}_\alpha$)

Suppose that $L_j \sim F_j$, $j \geq 1$ and that

- i) for some $k > 1$, $\mathbb{E}(|L_j - \mathbb{E}(L_j)|^k)$ is uniformly bounded, and
- ii) for some $\alpha \in (0, 1)$, $\liminf_{d \rightarrow \infty} \frac{1}{d} \sum_{j=1}^d \text{ES}_\alpha(L_j) > 0$.

Then, as $d \rightarrow \infty$, $\frac{\overline{\text{ES}}_\alpha(\mathcal{S}_d)}{\overline{\text{VaR}}_\alpha(\mathcal{S}_d)} = 1 + O(d^{\frac{1}{k}-1})$.

Proposition 8.34 (Dependence uncertainty spread of VaR_α vs ES_α)

Let $0 < \alpha_1 \leq \alpha_2 < 1$, assume Proposition 8.33 i) to hold and that

$\liminf_{d \rightarrow \infty} \frac{1}{d} \sum_{j=1}^d \text{LES}_{\alpha_1}(X_j) > 0$ and $\limsup_{d \rightarrow \infty} \frac{\sum_{j=1}^d \mathbb{E}(X_j)}{\sum_{j=1}^d \text{ES}_{\alpha_1}(X_j)} < 1$. Then

$$\liminf_{d \rightarrow \infty} \frac{\overline{\text{VaR}}_{\alpha_2}(\mathcal{S}_d) - \underline{\text{VaR}}_{\alpha_2}(\mathcal{S}_d)}{\overline{\text{ES}}_{\alpha_1}(\mathcal{S}_d) - \underline{\text{ES}}_{\alpha_1}(\mathcal{S}_d)} \geq 1$$

Example 8.35 (Superadditivity of VaR under 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

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

8.5 Capital allocation

How can the overall capital requirement may be disaggregated into additive contributions/units/investments? Motivation: How can we measure the risk-adjusted performance of different investments?

8.5.1 The allocation problem

- The performance of investments is usually measured using a *RORAC (return on risk-adjusted capital) approach* by considering

$$\frac{\text{expected profit of investment } j}{\text{risk capital for investment } j}.$$

- The risk capital of investment j with loss L_j can be computed as follows: Compute $\varrho(L) = \varrho(L_1 + \dots + L_d)$. Then allocate $\varrho(L)$ to the investments according to a *capital allocation principle* such that

$$\varrho(L) = \sum_{j=1}^d \text{AC}_j,$$

where the *risk contribution* AC_j is the capital allocated to investment j .

The formal set-up

- Consider an open set $\mathbf{1} \in \Lambda \subseteq \mathbb{R}^d \setminus \{\mathbf{0}\}$ of portfolio weights and define

$$L(\boldsymbol{\lambda}) = \boldsymbol{\lambda}' \mathbf{L} = \sum_{j=1}^d \lambda_j L_j, \quad \boldsymbol{\lambda} \in \Lambda.$$

- For a risk measure ϱ , define the *associated risk-measure function*

$$r_\varrho(\boldsymbol{\lambda}) = \varrho(L(\boldsymbol{\lambda})),$$

so that $r_\varrho(\mathbf{1}) = \varrho(L)$.

8.5.2 The Euler principle and examples

- If r_ϱ is positive homogeneous and differentiable at $\boldsymbol{\lambda} \in \Lambda$, Euler's rule (see the appendix) implies that

$$r_\varrho(\boldsymbol{\lambda}) = \sum_{i=1}^d \lambda_i \frac{\partial r_\varrho}{\partial \lambda_i}(\boldsymbol{\lambda}) \quad \text{so} \quad \varrho(L) = r_\varrho(\mathbf{1}) = \sum_{j=1}^d \frac{\partial r_\varrho}{\partial \lambda_j}(\mathbf{1}).$$

Note that r_ϱ is positive homogeneous if ϱ is.

Definition 8.36 (Euler capital allocation principle)

If r_ϱ is a pos.-hom. risk-measure function and differentiable at $\lambda = \mathbf{1}$, then the *Euler capital allocation principle* has risk contributions

$$\text{AC}_j = \text{AC}_j^\varrho := \frac{\partial r_\varrho}{\partial \lambda_j}(\mathbf{1}), \quad j \in \{1, \dots, d\}.$$

Examples

1) Standard deviation and the covariance principle

- Consider $r_{\text{SD}}(\boldsymbol{\lambda}) = \sqrt{\text{var}(L(\boldsymbol{\lambda}))} = \sqrt{\boldsymbol{\lambda}' \Sigma \boldsymbol{\lambda}}$ where Σ is the covariance matrix of (L_1, \dots, L_d) . Therefore

$$\text{AC}_j^\varrho = \frac{\partial r_{\text{SD}}}{\partial \lambda_j}(\mathbf{1}) = \frac{(\Sigma \mathbf{1})_j}{r_{\text{SD}}(\mathbf{1})} = \frac{\sum_{k=1}^d \text{cov}(L_j, L_k)}{r_{\text{SD}}(\mathbf{1})} = \frac{\text{cov}(L_j, L)}{\sqrt{\text{var}(L)}}.$$

This formula is known as *covariance principle*.

- If we consider the more general $\varrho(L) = \mathbb{E}(L) + \kappa \text{SD}(L)$ for some $\kappa > 0$ we get

$$r_\varrho(\boldsymbol{\lambda}) = \boldsymbol{\lambda}' \mathbb{E}(\mathbf{L}) + \kappa r_{\text{SD}}(\boldsymbol{\lambda})$$

and hence

$$\text{AC}_j^\varrho = \mathbb{E}(L_j) + \kappa \frac{\text{cov}(L_j, L)}{\sqrt{\text{var}(L)}}.$$

2) VaR and VaR contributions

Suppose that $r_{\text{VaR}}^\alpha(\boldsymbol{\lambda}) = q_\alpha(L(\boldsymbol{\lambda}))$. In this case it can be shown (non-trivial) that, subject to technical conditions,

$$\text{AC}_j^\varrho = \frac{\partial r_{\text{VaR}}^\alpha}{\partial \lambda_j}(\mathbf{1}) = \mathbb{E}(L_j \mid L = F_L^\leftarrow(\alpha)), \quad j \in \{1, \dots, d\}.$$

3) Expected shortfall and shortfall contributions

Now consider $r_{\text{ES}}^\alpha(\boldsymbol{\lambda}) = \mathbb{E}(L \mid L \geq q_\alpha(L(\boldsymbol{\lambda})))$. Then

$$r_{\text{ES}}^\alpha(\boldsymbol{\lambda}) = \frac{1}{1 - \alpha} \int_\alpha^1 r_{\text{VaR}}^u(\boldsymbol{\lambda}) \, du,$$

Assuming the differentiability of $r_{\text{VaR}}^u(\lambda)$, the Euler principle implies that

$$\frac{\partial r_{\text{ES}}^\alpha}{\partial \lambda_j}(\mathbf{1}) = \frac{1}{1-\alpha} \int_\alpha^1 \frac{\partial r_{\text{VaR}}^u}{\partial \lambda_j}(\mathbf{1}) du = \frac{1}{1-\alpha} \int_\alpha^1 \mathbb{E}(L_j | L = F_L^\leftarrow(u)) du.$$

If F_L has a differentiable inverse,

$$\frac{\partial r_{\text{ES}}^\alpha}{\partial \lambda_j}(\mathbf{1}) = \frac{1}{1-\alpha} \int_{F_L^\leftarrow(\alpha)}^\infty \mathbb{E}(L_j | L = v) f_L(v) dv = \frac{\mathbb{E}(L_j; L \geq F_L^\leftarrow(\alpha))}{1-\alpha}.$$

Hence the Euler capital allocation takes the form

$$\text{AC}_j^\varrho = \mathbb{E}(L_j | L \geq \text{VaR}_\alpha(L)), \quad L := L(\mathbf{1});$$

AC_j^ϱ is known as the *expected shortfall contribution* of investment j . This is a popular allocation principle in practice.

4) Euler allocation for elliptical loss distributions

The following result shows that allocation is very simple in the case of $L \sim E_d(\mathbf{0}, \Sigma, \psi)$: Calculate the total risk capital and then use a simple partitioning formula (regardless of the pos.-hom. risk measure).

Corollary 8.37 (Euler allocation under ellipticality)

Assume that r_ϱ is the risk-measure function of a positive-homogeneous and law invariant ϱ . Let $\mathbf{L} \sim E_d(\mathbf{0}, \Sigma, \psi)$. Then, under an Euler allocation,

$$\frac{\text{AC}_j^\varrho}{\text{AC}_k^\varrho} = \frac{\sum_{l=1}^d \Sigma_{jl}}{\sum_{l=1}^d \Sigma_{kl}}, \quad j, k \in \{1, \dots, d\}.$$

Proof. The proof of Theorem 8.24 implies that, by positive homogeneity,

$$r_\varrho(\boldsymbol{\lambda}) = \varrho(L(\boldsymbol{\lambda})) = \varrho\left(\sum_{j=1}^d \lambda_j L_j\right) = \sqrt{\boldsymbol{\lambda}' \Sigma \boldsymbol{\lambda}} \varrho(Y_1),$$

where Y_1 is the first component of $\mathbf{Y} \sim S_d(\psi)$. For the Euler allocation we get

$$\text{AC}_j^\varrho = \frac{\partial r_\varrho}{\partial \lambda_j}(\mathbf{1}) = \frac{\sum_{k=1}^d \Sigma_{jk}}{\sqrt{\mathbf{1}' \Sigma \mathbf{1}}} \varrho(Y_1)$$

from which the result follows. □

8.5.3 Economic properties of the Euler principle

- We show that the Euler principle has good economic properties.
- Assume that r_ϱ is continuously differentiable in $\mathbb{R}^d \setminus \{\mathbf{0}\}$ and by

$$\text{AC}_j^\varrho = \frac{\partial r_\varrho}{\partial \lambda_j}(\mathbf{1}), \quad j \in \{1, \dots, d\},$$

denote the associated risk contributions under the Euler principle.

Compatibility with a RORAC approach

- The *RORAC* (*return on risk adjusted capital*) is defined as

$$\text{RORAC}(L) := \frac{\mathbb{E}(-L)}{\varrho(L)}$$

and the *portfolio-related RORAC* of investment j is defined as

$$\text{RORAC}(L_j | L) := \frac{\mathbb{E}(-L_j)}{\text{AC}_j^\varrho}.$$

- The Euler principle is compatible with a RORAC approach: If investment j performs better than the overall portfolio L in the RORAC metric, then the latter is increased if one increases the weight of unit j . Hence the Euler principle gives correct signals for investment decisions.
- In mathematical terms, **RORAC compatibility** means that there is some $\varepsilon > 0$ such that for all $0 < h \leq \varepsilon$

$$\text{RORAC}(L_j | L) > \text{RORAC}(L) \Rightarrow \text{RORAC}(L + hL_j) > \text{RORAC}(L).$$

Proof. $\frac{d}{dh} \text{RORAC}(L + hL_j)|_{h=0}$

$$= \frac{d}{dh} \frac{\mathbb{E}(-(L + hL_j))}{r_\varrho(\mathbf{1} + he_j)} \Big|_{h=0} = \frac{1}{r_\varrho(\mathbf{1})^2} \left(\mathbb{E}(-L_j) \mathbf{r}_\varrho(\mathbf{1}) - \mathbb{E}(-L) \frac{\partial r_\varrho(\mathbf{1})}{\partial \lambda_j} \right),$$

$$= \frac{1}{r_\varrho(\mathbf{1})^2} (\mathbb{E}(-L_j) \varrho(\mathbf{L}) - \mathbb{E}(-L) \mathbf{AC}_j^\varrho) > 0$$

$$\text{if } \frac{\mathbb{E}(-L_j)}{\mathbf{AC}_j^\varrho} = \text{RORAC}(L_j | L) > \text{RORAC}(L) = \frac{\mathbb{E}(-L)}{\varrho(L)}. \quad \square$$

Diversification benefit

- For a subadditive ϱ , $\sum_{j=1}^d \varrho(L_j) - \varrho(L) > 0$ is known as *diversification benefit*.
- It is reasonable to require that each business unit profits from the diversification benefit in the sense that

$$\text{AC}_j^\varrho \leq \varrho(L_j), \quad j \in \{1, \dots, d\}.$$

- We now show that the Euler principle does indeed satisfy this inequality.

Proof. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex, pos.-hom. and continuously differentiable in $\mathbb{R}^d \setminus \{\mathbf{0}\}$. By convexity,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \sum_{j=1}^d (y_j - x_j) \frac{\partial f}{\partial x_j}(\mathbf{x}), \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \mathbf{x} \neq \mathbf{0}.$$

By Euler's rule, $f(\mathbf{x}) = \sum_{j=1}^d x_j \frac{\partial f}{\partial x_j}(\mathbf{x})$ and hence

$$f(\mathbf{y}) \geq \sum_{j=1}^d y_j \frac{\partial f}{\partial x_j}(\mathbf{x}).$$

For $\mathbf{y} = \boldsymbol{\lambda}$ and $\mathbf{x} = \boldsymbol{\lambda} + \tilde{\boldsymbol{\lambda}}$, we obtain

$$f(\boldsymbol{\lambda}) \geq \sum_{j=1}^d \lambda_j \frac{\partial f}{\partial \lambda_j}(\boldsymbol{\lambda} + \tilde{\boldsymbol{\lambda}}) \quad \text{for all } \boldsymbol{\lambda}, \tilde{\boldsymbol{\lambda}} : \boldsymbol{\lambda} \neq -\tilde{\boldsymbol{\lambda}}.$$

Apply this inequality with $f = r_\varrho$ (which is convex as ϱ is pos.-hom. and subadditive), $\boldsymbol{\lambda} = \mathbf{e}_j$ and $\tilde{\boldsymbol{\lambda}} = \mathbf{1} - \mathbf{e}_j$ to obtain

$$\varrho(L_j) = r_\varrho(\mathbf{e}_j) \geq \frac{\partial r_\varrho}{\partial \lambda_j}(\mathbf{1}) = \text{AC}_j^\varrho. \quad \square$$

- From a practical point of view, expected shortfall and expected shortfall contributions are typically a reasonable choice in many applications.

9 Market Risk

9.1 Risk factors and mapping

9.2 Market risk measurement

9.3 Backtesting

9.1 Risk factors and mapping

9.1.1 The loss operator

- The key idea in this section is that of a **loss operator** for expressing the change in value of a portfolio in terms of **risk-factor changes**.
- Let the current time be t and assume the current value V_t of an asset portfolio is known, or can be computed with appropriate valuation models.
- We are interested in value changes or losses over a relatively **short time period** $[t, t + 1]$, for example one day, two weeks or month.
- Scaling may be applied to derive capital requirements for longer periods.
- We assume there is **no change to the composition of the portfolio** over the time period.
- The future value V_{t+1} is modelled as a random variable.

- We want to determine the distribution of the loss distribution of $L_{t+1} = -(V_{t+1} - V_t)$.
- We map the value at time t using the formula

$$V_t = g(\tau_t, \mathbf{Z}_t)$$

where τ_t is time t expressed in units of valuation time.

The issue of time

- We will be quite precise about the modelling of time.
- The natural time unit for valuation of positions might be yearly; e.g. in Black-Scholes valuation, the volatility is expressed in annualized terms.
- On the other hand the risk modelling time horizon $[t, t + 1]$ is typically shorter.
- Let Δt be the length of the time horizon in valuation time.

- For example, suppose that valuation time is yearly. Then a monthly time horizon would be $\Delta t = 1/12$ and a trading day $\Delta t = 1/250$.
- We set $\tau_t = t(\Delta t)$ for all t so that $\tau_{t+1} - \tau_t = \Delta t$.

From the mapping to the loss operator

- The risk factor changes over the time horizon are

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

- Typically, historical risk factor data are available as a time series $\mathbf{X}_{t-n}, \dots, \mathbf{X}_{t-1}, \mathbf{X}_t$ and these are used to model the behaviour of \mathbf{X}_{t+1} .
- We have

$$\begin{aligned}
 L_{t+1} &= -(V_{t+1} - V_t) \\
 &= -(g(\tau_{t+1}, \mathbf{Z}_{t+1}) - g(\tau_t, \mathbf{Z}_t)) \\
 &= -(g(\tau_t + \Delta t, \mathbf{Z}_t + \mathbf{X}_{t+1}) - g(\tau_t, \mathbf{Z}_t)). \tag{50}
 \end{aligned}$$

- Since the risk factor values \mathbf{Z}_t are known at time t , the loss L_{t+1} is determined by the risk factor changes \mathbf{X}_{t+1} .
- Given a realization \mathbf{z}_t of \mathbf{Z}_t , the **loss operator** at time t is defined to be

$$l_{[t]}(\mathbf{x}) = -(g(\tau_t + \Delta t, \mathbf{z}_t + \mathbf{x}) - g(\tau_t, \mathbf{z}_t)), \quad (51)$$

so that

$$L_{t+1} = l_{[t]}(\mathbf{X}_{t+1}).$$

- The loss operator embodies the idea of **full revaluation**.
- From the perspective of time t the loss distribution of L_{t+1} is determined by the multivariate distribution of \mathbf{X}_{t+1} .

9.1.2 Delta and delta-gamma approximations

- If the mapping function g is differentiable and Δt is relatively small we can approximate g with a first-order Taylor series approximation

$$g(\tau_t + \Delta t, \mathbf{z}_t + \mathbf{x}) \approx g(\tau_t, \mathbf{z}_t) + g_\tau(\tau_t, \mathbf{z}_t)\Delta t + \sum_{i=1}^d g_{z_i}(\tau_t, \mathbf{z}_t)x_i, \quad (52)$$

where the τ -subscript and z_i -subscript denote partial derivatives with respect to (valuation) time and the risk factors respectively.

- This allows us to approximate the loss operator in (51) by the [linear loss operator](#) at time t given by

$$l_{[t]}^{\Delta}(\mathbf{x}) := - \left(g_{\tau}(\tau_t, \mathbf{z}_t) \Delta t + \sum_{i=1}^d g_{z_i}(\tau_t, \mathbf{z}_t) x_i \right). \quad (53)$$

- Note that, when working with a short time horizon Δt , the term $g_{\tau}(\tau_t, \mathbf{z}_t) \Delta t$ is typically small and is sometimes omitted in practice.

Example 9.1 (European call option)

- Consider portfolio consisting of one standard European call on a non-dividend paying stock S with **maturity** T and **exercise** price K .
- The Black-Scholes value of this asset at time t is $C^{BS}(t, S_t, r, \sigma)$ where

$$C^{BS}(t, S; r, \sigma) = S\Phi(d_1) - Ke^{-r(T-t)}\Phi(d_2),$$

Φ is standard normal df, r represents risk-free interest rate, σ the volatility of underlying stock, and where

$$d_1 = \frac{\log(S/K) + (r + \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}} \text{ and } d_2 = d_1 - \sigma\sqrt{T - t}.$$

- While in the BS model, it is assumed that **interest rates** and **volatilities** are constant, in reality they tend to **fluctuate** over time; they should be added to our set of risk factors.

- The risk factors: $\mathbf{Z}_t = (\log S_t, r_t, \sigma_t)'$.
- The risk factor changes: $\mathbf{X}_t = (\log(S_t/S_{t-1}), r_t - r_{t-1}, \sigma_t - \sigma_{t-1})'$.
- The mapping:

$$V_t = C^{BS}(\tau_t, S_t; r_t, \sigma_t) = g(\tau_t, \mathbf{Z}_t)$$

- For derivative positions it is quite common to use the linear loss operator

$$L_{t+1}^\Delta = l_{[t]}^\Delta(\mathbf{X}_{t+1}) = - \left(g_\tau(\tau_t, \mathbf{z}_t) \Delta t + \sum_{i=1}^3 g_{z_i}(\tau_t, \mathbf{z}_t) X_{t+1,i} \right),$$

where g_τ, g_{z_i} denote partial derivatives.

- Δt is the length of the time interval expressed in years since Black-Scholes parameters relate to units of one year.

- It is more common to write the linear loss operator as

$$l_{[t]}^{\Delta}(\boldsymbol{x}) = - \left(C_t^{BS} + C_S^{BS} S_t x_1 + C_r^{BS} x_2 + C_{\sigma}^{BS} x_3 \right),$$

in terms of the derivatives of the BS formula or [the Greeks](#).

- ▶ C_S^{BS} is known as the [delta](#) of the option.
- ▶ C_{σ}^{BS} is the [vega](#).
- ▶ C_r^{BS} is the [rho](#).
- ▶ C_t^{BS} is the [theta](#).

Note the appearance of S_t in the C_S^{BS} term. This is because the risk factor is $\ln S_t$ rather than S_t and $C_{\ln S}^{BS} = C_S^{BS} S_t$.

Quadratic loss operator

- Recall the first-order Taylor series approximation of mapping in (52).
- Let $\delta(\tau_t, \boldsymbol{z}_t) = (g_{z_1}(\tau_t, \boldsymbol{z}_t), \dots, g_{z_d}(\tau_t, \boldsymbol{z}_t))'$ be the first-order partial derivatives of the mapping with respect to the risk factors.

- Let $\omega(\tau_t, \mathbf{z}_t) = (g_{z_1\tau}(\tau_t, \mathbf{z}_t), \dots, g_{z_d\tau}(\tau_t, \mathbf{z}_t))'$ denote the mixed partial derivatives with respect to time and the risk factors.
- Let $\Gamma(\tau_t, \mathbf{z}_t)$ denote the matrix with (i, j) th element given by $g_{z_i z_j}(\tau_t, \mathbf{z}_t)$; this matrix contains **gamma sensitivities** to individual risk factors on the diagonal and **cross gamma sensitivities** to pairs of risk factors off the diagonal.
- The full second-order approximation of the mapping function is g is

$$g(\tau_t + \Delta t, \mathbf{z}_t + \mathbf{x}) \approx g(\tau_t, \mathbf{z}_t) + g_\tau(\tau_t, \mathbf{z}_t)\Delta t + \delta(\tau_t, \mathbf{z}_t)' \mathbf{x} + \\ \frac{1}{2}(g_{\tau\tau}(\tau_t, \mathbf{z}_t)(\Delta t)^2 + 2\omega(\tau_t, \mathbf{z}_t)' \mathbf{x}\Delta t + \\ \mathbf{x}'\Gamma(\tau_t, \mathbf{z}_t)\mathbf{x}).$$

- In practice, we would usually omit terms of order $o(\Delta_t)$ (terms that tend to zero faster than Δ_t). In standard continuous-time financial models like Black-Scholes the risk-factor changes \mathbf{x} are of order $\sqrt{\Delta_t}$.

- This leaves us with the quadratic loss operator

$$l_{[t]}^{\Delta\Gamma}(\boldsymbol{x}) = -(g_\tau(\tau_t, \boldsymbol{z}_t)\Delta t + \boldsymbol{\delta}(\tau_t, \boldsymbol{z}_t)' \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}'\Gamma(\tau_t, \boldsymbol{z}_t)\boldsymbol{x}) \quad (54)$$

which is more accurate than the linear loss operator (53).

Example 9.2 (European call option)

The quadratic loss operator is

$$\begin{aligned} l_{[t]}^{\Delta\Gamma}(\boldsymbol{x}) &= l_{[t]}^\Delta(\boldsymbol{x}) - 0.5 \left(C_{SS}^{BS} S_t^2 x_1^2 + C_{rr}^{BS} x_2^2 + C_{\sigma\sigma}^{BS} x_3^2 \right) \\ &\quad - \left(C_{Sr}^{BS} S_t x_1 x_2 + C_{S\sigma}^{BS} S_t x_1 x_3 + C_{r\sigma}^{BS} x_2 x_3 \right). \end{aligned}$$

The names of the second-order Greeks (with the exception of gamma) are rather obscure. Here are some of them:

- C_{SS}^{BS} is known as the [gamma](#) of the option;
- $C_{\sigma\sigma}^{BS}$ is the [vomma](#);
- $C_{S\sigma}^{BS}$ is the [vanna](#).

9.1.3 Mapping bond portfolios

Basic definitions for bond pricing

- Let $p(t, T)$ denote the price at time t of a default-free zero-coupon bond paying one at time T (also called a discount factor).
- Time is measured in years.
- Many other fixed-income instruments such as coupon bonds or standard swaps can be viewed as portfolios of zero-coupon bonds.
- The mapping $T \rightarrow p(t, T)$ for different maturities is one way of describing the so-called term structure of interest rates at time t . An alternative description is based on yields.
- The term structure $T \rightarrow p(t, T)$ is known at time t .
- However the future term structure $T \rightarrow p(t + x, T)$ for $x > 0$ is not known at time t and must be modelled stochastically.

- The continuously compounded yield of a zero-coupon bond is

$$y(t, T) = -\frac{\ln p(t, T)}{T - t}. \quad (55)$$

- We have the relation

$$p(t, T) = \exp(-(T - t)y(t, T)).$$

- The yield is the constant, annualized rate implied by the price $p(t, T)$.
Also known as spot rate.
- The mapping $T \rightarrow y(t, T)$ is referred to as the continuously compounded yield curve at time t .
- Yields are comparable across different times to maturity.

Detailed mapping of a bond portfolio

- Consider a portfolio of d default-free zero-coupon bonds with maturities T_i and prices $p(t, T_i)$ for $i = 1, \dots, d$. Assume $p(T_i, T_i) = 1$ for all i .

- By λ_i we denote the number of bonds with maturity T_i in the portfolio.
- The **portfolio value** at time t is given by

$$V(t) := \sum_{i=1}^d \lambda_i p(t, T_i) = \sum_{i=1}^d \lambda_i \exp(-(T_i - t)y(t, T_i)).$$

- In a detailed analysis of the change in value one takes all yields $y(t, T_i)$, $1 \leq i \leq d$, as risk factors.
- We want to put this in the general discrete-time framework of the mapping

$$V_t = g(\tau_t, \mathbf{Z}_t).$$

- We set

$$\tau_t = t(\Delta t), \quad V_t = V(\tau_t), \quad Z_{t,i} = y(\tau_t, T_i)$$

where Δt is risk management time horizon in years.

- We obtain a mapping of the form

$$V_t = V(\tau_t) = g(\tau_t, \mathbf{Z}_t) = \sum_{i=1}^d \lambda_i \exp(-(T_i - \tau_t) Z_{t,i}). \quad (56)$$

The loss operator and its approximations

- The portfolio loss is

$$\begin{aligned} L_{t+1} &= -(V_{t+1} - V_t) \\ &= - \sum_{i=1}^d \lambda_i e^{-(T_i - \tau_t) Z_{t,i}} \left(\exp(Z_{t,i} \Delta t - (T_i - \tau_{t+1}) X_{t+1,i}) - 1 \right). \end{aligned}$$

- Reverting to standard bond pricing notation the loss operator is

$$l_{[t]}(\mathbf{x}) = - \sum_{i=1}^d \lambda_i p(\tau_t, T_i) \left(\exp(y(\tau_t, T_i) \Delta t - (T_i - \tau_{t+1}) x_i) - 1 \right),$$

where x_i represents the change in yield of the i th bond.

- The first derivatives of the mapping function (56) are

$$g_{\tau}(\tau_t, \mathbf{z}_t) = \sum_{i=1}^d \lambda_i p(\tau_t, T_i) z_{t,i}$$

$$g_{z_i}(\tau_t, \mathbf{z}_t) = -\lambda_i (T_i - \tau_t) \exp(-(T_i - \tau_t) z_{t,i}).$$

- Inserting these in (53) and reverting to standard bond pricing notation we obtain

$$l_{[t]}^{\Delta}(\mathbf{x}) = - \sum_{i=1}^d \lambda_i p(\tau_t, T_i) \left(y(\tau_t, T_i) \Delta t - (T_i - \tau_t) x_i \right), \quad (57)$$

- For the second-order approximation we need the second derivatives with respect to yields which are

$$g_{z_i z_i}(\tau_t, \mathbf{z}_t) = \lambda_i (T_i - \tau_t)^2 \exp(-(T_i - \tau_t) z_{t,i})$$

and $g_{z_i z_j}(\tau_t, \mathbf{z}_t) = 0$ for $i \neq j$.

- The quadratic loss operator (54) is

$$l_{[t]}^{\Delta\Gamma}(\boldsymbol{x}) = - \sum_{i=1}^d \lambda_i p(\tau_t, T_i) \left(y(\tau_t, T_i) \Delta t - (T_i - \tau_t) x_i + \frac{1}{2} (T_i - \tau_t)^2 x_i^2 \right). \quad (58)$$

Relationship of linear operator to duration

- Consider a very simple model for the yield curve at time t in which

$$y(\tau_{t+1}, T_i) = y(\tau_t, T_i) + x$$

for all maturities T_i .

- In our mapping notation

$$Z_{t+1,i} = Z_{t,i} + X_{t+1}, \quad \forall i.$$

- In this model we assume that a parallel shift in level takes place along the entire yield curve.

- This is unrealistic but frequently assumed in practice.
- In this model the loss operator and its linear and quadratic approximations are functions of a scalar variable x , the change in level.
- Under the parallel shift model we can write

$$l_{[t]}^{\Delta}(x) = -V_t \left(A_t \Delta t - D_t x \right), \quad (59)$$

where

$$D_t := \sum_{i=1}^d \frac{\lambda_i p(\tau_t, T_i)}{v_t} (T_i - \tau_t), \quad A_t := \sum_{i=1}^d \frac{\lambda_i p(\tau_t, T_i)}{V_t} y(\tau_t, T_i).$$

- D_t is usually called the (Macaulay) duration of the bond portfolio.
- It is a weighted sum of the times to maturity of the different cash flows in the portfolio, the weights being proportional to the discounted values of the cash flows.

Interpreting duration

- Over short time intervals losses of value in the bond portfolio will be determined by $l_{[t]}^{\Delta}(x) \approx V_t D_t x$.
- Increases in level of yields lead to losses; decreases lead to gains.
- The duration D_t is the bond pricing analogue of the delta of an option.
- Any two bond portfolios with equal value and duration will be subject to similar losses when there is a small parallel shift of the yield curve.
- Duration is an important tool in traditional bond-portfolio or asset-liability management.
- An asset manager, who invests in various bonds to cover promised cash flows in the future, invests in such a way that the duration of the overall portfolio of bonds and liability cash flows is equal to zero.
- Portfolios are immunized against small parallel shifts in yield curve, but not changes of slope and curvature.

Relationship of quadratic operator to convexity

- It is possible to get more accurate approximations for the loss in a bond portfolio by considering second-order effects.
- The **analogue of the gamma** of an option is **convexity**. Under the parallel shift model, the quadratic loss operator (58) becomes

$$l_{[t]}^{\Delta\Gamma}(x) = -V_t \left(A_t \Delta t - D_t x + \frac{1}{2} C_t x^2 \right), \quad (60)$$

where

$$C_t := \sum_{i=1}^d \frac{\lambda_i p(\tau_t, T_i)}{V_t} (T_i - \tau_t)^2$$

is the convexity of the bond portfolio.

- The convexity is a weighted average of the squared times to maturity and is (minus) the derivative of the duration with respect to yield.

Interpreting convexity

- Consider two portfolios (1) and (2) with identical durations $D_t^{(1)} = D_t^{(2)}$ but differing convexities satisfying $C_t^{(1)} > C_t^{(2)}$.
- Ignoring terms in Δt , the difference in loss operators satisfies

$$l_{[t]}^{\Delta\Gamma,1}(x) - l_{[t]}^{\Delta\Gamma,2}(x) \approx -\frac{1}{2}V_t(C_t^{(1)} - C_t^{(2)})x^2 < 0.$$

- Since $l_{[t]}^{\Delta\Gamma,1}(x) < l_{[t]}^{\Delta\Gamma,2}(x)$ an increase in the level of yields ($x > 0$) will lead to smaller losses for portfolio (1)
- Since $-l_{[t]}^{\Delta\Gamma,1}(x) > -l_{[t]}^{\Delta\Gamma,2}(x)$ a decrease in the level of yields ($x < 0$) will lead to larger gains.
- For this reason higher convexity is considered a desirable attribute of a bond portfolio in risk management.

9.1.4 Factor models for bond portfolios

The need for factor models

- The parallel shift model is unrealistic in practice.
- For large portfolios of fixed-income instruments, such as the overall fixed-income position of a major bank, modelling changes in the yield for every cash flow maturity date becomes impractical.
- Moreover, the statistical task of estimating a distribution for X_{t+1} is difficult because the **yields are highly dependent for different times to maturity**.
- A pragmatic approach is therefore to build a factor model for yields that captures the main features of the yield curve.
- Three-factor models of the yield curve in which the factors typically represent **level**, **slope** and **curvature** are often used in practice.

The approach based on the Nelson and Siegel (1987) model

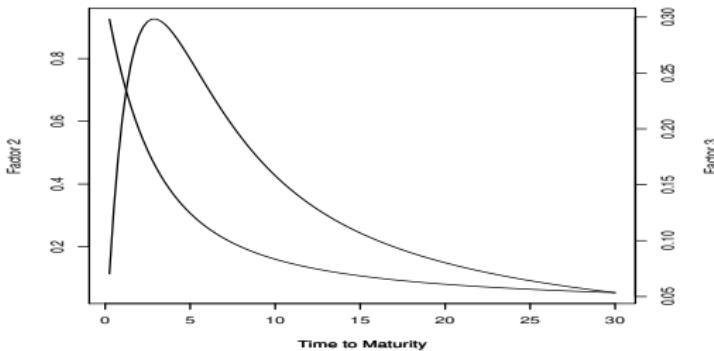
- We assume that at time t the yield curve can be modelled by

$$y(\tau_t, T) \approx Z_{t,1} + k_2(T - \tau_t, \eta_t)Z_{t,2} + k_3(T - \tau_t, \eta_t)Z_{t,3}, \quad (61)$$

where the functions k_2 and k_3 are given by

$$k_2(s, \eta) = \frac{1 - \exp(-\eta s)}{\eta s}, \quad k_3(s, \eta) = k_2(s, \eta) - \exp(-\eta s).$$

- Nelson-Siegel functions $k_2(s, \eta)$ and $k_3(s, \eta)$ for an η value of 0.623:



- η is an extra tuning parameter to improve fit.
- There are other simple factor models including the [Svensson model](#).
- Clearly $\lim_{s \rightarrow \infty} k_2(s, \eta) = \lim_{s \rightarrow \infty} k_3(s, \eta) = 0$ while $\lim_{s \rightarrow 0} k_2(s, \eta) = 1$ and $\lim_{s \rightarrow 0} k_3(s, \eta) = 0$.
- It follows that

$$\lim_{T \rightarrow \infty} y(\tau_t, T) = Z_{t,1},$$

so that the first factor is usually interpreted as a [long-term level factor](#).

- $Z_{t,2}$ is interpreted as a [slope factor](#) because the difference in short-term and long-term yields satisfies

$$\lim_{T \rightarrow \tau_t} y(\tau_t, T) - \lim_{T \rightarrow \infty} y(\tau_t, T) = Z_{t,2}.$$

- $Z_{t,3}$ has an interpretation as a [curvature factor](#).

- Using (61), the portfolio mapping (56) becomes

$$V_t = g(\tau_t, \mathbf{Z}_t) = \sum_{i=1}^d \lambda_i \exp \left(- (T_i - \tau_t) \sum_{j=1}^3 k_j (T_i - \tau_t, \eta_t) Z_{t,j} \right),$$

where $k_1(s, \eta) = 1$.

- It is then straightforward to derive the loss operator $l_{[t]}(\mathbf{x})$ or its linear version $l_{[t]}^\Delta(\mathbf{x})$ which are functions on \mathbb{R}^3 rather than \mathbb{R}^d .
- To use this method to evaluate the loss operator at time t we require realized values z_t for the risk factors \mathbf{Z}_t . We have to overcome the fact that **the Nelson-Siegel factors \mathbf{Z}_t are not directly observed** at time t . Instead **they have to be estimated** from observable yield curve data.
- Let $\mathbf{Y}_t = (y(\tau_t, \tau_t + s_1), \dots, y(\tau_t, \tau_t + s_m))'$ denote the data vector at time t , containing the yields for m different times to maturity, s_1, \dots, s_m , where m is large.

- This is assumed to follow the factor model

$$\mathbf{Y}_t = \mathbf{B}_t \mathbf{Z}_t + \boldsymbol{\varepsilon}_t,$$

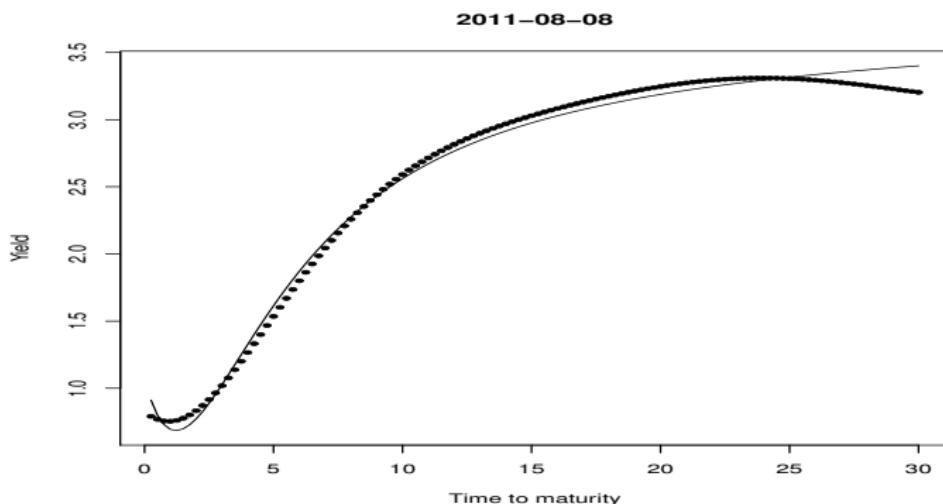
where $\mathbf{B}_t \in \mathbb{R}^{m \times 3}$ is the matrix with i th row $(1, k_2(s_i, \eta_t), k_3(s_i, \eta_t))$ and $\boldsymbol{\varepsilon}_t \in \mathbb{R}^m$ is an error vector.

- For a given value of η_t the estimation of \mathbf{Z}_t can be carried out as a cross-sectional regression using weighted least squares. It is a [fundamental](#) factor model where the loading matrix \mathbf{B}_t is known.
- To estimate η_t a more complicated optimization is carried out.

Example 9.3

- The data are daily Canadian zero-coupon bond yields for 120 different quarterly maturities ranging from 0.25 years to 30 years.
- They have been generated using pricing data for Government of Canada bonds and treasury bills.

- We model the yield curve on the 8th August 2011.
- The estimated values are $z_{t,1} = 3.82$, $z_{t,2} = -2.75$, $z_{t,3} = -5.22$ and $\hat{\eta}_t = 0.623$. Thus the curves $k_2(s, \eta)$ and $k_3(s, \eta)$ are as shown earlier.
- The fitted Nelson-Siegel curve and the data are shown below:



The approach based on PCA

- The key difference to the Nelson-Siegel approach is that here the dimension reduction via factor modelling is applied at the level of the risk factor changes \mathbf{X}_{t+1} rather than the risk factors \mathbf{Z}_t .
- We recall that PCA can be used to construct factor models of the form

$$\mathbf{X}_{t+1} = \boldsymbol{\mu} + \boldsymbol{\Gamma}_1 \mathbf{F}_{t+1} + \boldsymbol{\varepsilon}, \quad (62)$$

where \mathbf{F}_{t+1} is a p -dimensional vector of principal component factors ($p < d$), $\boldsymbol{\Gamma}_1 \in \mathbb{R}_{d \times p}$ contains the corresponding loading matrix, $\boldsymbol{\mu}$ is the mean vector of \mathbf{X}_{t+1} and $\boldsymbol{\varepsilon}$ is an error vector.

- Typically, the error term is neglected and $\boldsymbol{\mu} \approx \mathbf{0}$, so that we make the approximation $\mathbf{X}_{t+1} \approx \boldsymbol{\Gamma}_1 \mathbf{F}_{t+1}$.
- In the case of the linear loss operator for the bond portfolio in (57) we

basically replace $l_{[t]}^{\Delta}(\mathbf{X}_{t+1})$ by

$$l_{[t]}^{\Delta}(\mathbf{F}_{t+1}) = - \sum_{i=1}^d \lambda_i p(\tau_t, T_i) (y(\tau_t, T_i) \Delta t - (T_i - \tau_t)(\Gamma_1 \mathbf{F}_{t+1})_i), \quad (63)$$

so that a p -dimensional function replaces a d -dimensional function.

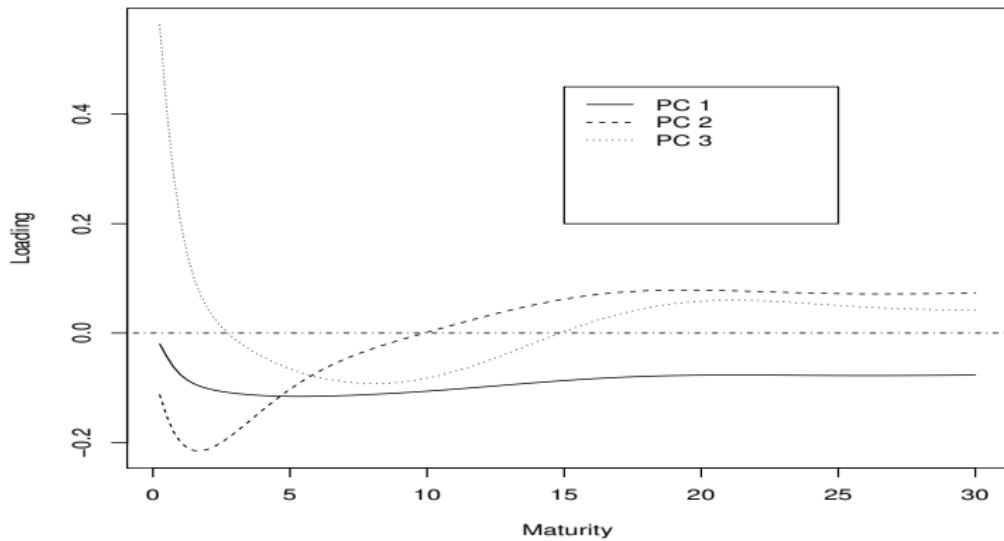
- To calibrate this function, we require an estimate for the matrix Γ_1 . This can be obtained from historical time-series data on yield changes by estimating sample principle components.

Example 9.4

- To estimate the Γ_1 matrix of principal component loadings we require longitudinal (time-series) data rather than the cross-sectional data.
- We again analyse Canadian data. Recall that we have data vectors $\mathbf{Y}_t = (y(\tau_t, \tau_t + s_1), \dots, y(\tau_t, \tau_t + s_d))$ of yields for different maturities.

- For simplicity assume that the times-to-maturity $T_1 - \tau_t, \dots, T_d - \tau_t$ of the bonds in the portfolio correspond exactly to the times to maturity s_1, \dots, s_d available in the historical dataset.
- Assume also that the risk management horizon Δt is one day.
- We analyse the first differences of the data $\mathbf{X}_t = \mathbf{Y}_t - \mathbf{Y}_{t-1}$ using PCA under an assumption of stationarity.
- In the Canadian dataset we have 2488 days of data from 2 January 2002 to 30 December 2011.
- (Note that a small error is incurred by analysing daily returns of yields with fixed times-to-maturity rather than fixed maturity date.)
- The first principle component explains 87.0% of the variance of the data, the first two components explain 95.9% and the first three components explain 97.5%.

- We choose to work with the first three principal components, meaning that we set $p = 3$ and set the columns of Γ_1 equal to the first three principal component loading vectors.
- These vectors are shown graphically below and lend themselves to a standard interpretation.
- The first principal component has negative loadings for all maturities; the second has negative loadings up to 10 years and positive loadings thereafter; the third has positive loadings for very short maturities (less than 2.5 years) and very long maturities (greater than 15 years) but negative loadings otherwise.
- This suggests that the first principal component can be thought of as **inducing a change in the level of all yields**, the second **induces a change of slope** and the third **a change in the curvature** of the yield curve.



9.2 Market risk measurement

The goal in this section is to estimate the distribution of

$$L_{t+1} = l_{[t]}(\mathbf{X}_{t+1})$$

or a linear or quadratic approximation thereof, where

- \mathbf{X}_{t+1} is the vector of risk-factor changes from time t to time $t + 1$;
- $l_{[t]}$ is the known loss operator function at time t .

The problem comprises two tasks:

- 1) on the one hand we have the statistical problem of estimating the distribution of \mathbf{X}_{t+1} ;
- 2) on the other hand we have the computational or numerical problem of evaluating the distribution of $L_{t+1} = l_{[t]}(\mathbf{X}_{t+1})$.

9.2.1 Conditional and unconditional loss distributions

- Generally, we want to compute conditional measures of risk based on the most recent information about financial markets.
- In this case, the task is to estimate $F_{\mathbf{X}_{t+1}|\mathcal{F}_t}$, the conditional distribution of risk-factor changes, given \mathcal{F}_t , the sigma field representing the available information at time t .
- The conditional loss distribution is the distribution of the loss operator $l_{[t]}(\cdot)$ under $F_{\mathbf{X}_{t+1}|\mathcal{F}_t}$, i.e. the distribution with df

$$F_{L_{t+1}|\mathcal{F}_t}(l) = \mathbb{P}(l_{[t]}(\mathbf{X}_{t+1}) \leq l | \mathcal{F}_t).$$

- In the unconditional approach we assume that $(\mathbf{X}_s)_{s \leq t}$ forms a stationary time series, at least in the recent past.
- In this case we can estimate the stationary distribution $F_{\mathbf{X}}$ and then evaluate the unconditional loss distribution of $l_{[t]}(\mathbf{X})$ where $\mathbf{X} \sim F_{\mathbf{X}}$. The unconditional loss distribution is thus $F_{L_{t+1}}(l) = \mathbb{P}(l_{[t]}(\mathbf{X}) \leq l)$.

- The unconditional approach may be appropriate for longer time intervals, or for stress testing during quieter periods.
- If the risk-factor changes form an iid series, we obviously have $F_{\mathbf{X}_{t+1}|\mathcal{F}_t} = F_{\mathbf{X}}$, so that the conditional and unconditional approaches coincide.

9.2.2 Variance-covariance method

- The variance-covariance method is an analytical method in which strong assumptions of (conditional) normality and linearity are made.
- We assume that the conditional distribution of risk-factor changes $F_{\mathbf{X}_{t+1}|\mathcal{F}_t}$ is a multivariate normal distribution.
- In other words, we assume that $\mathbf{X}_{t+1} | \mathcal{F}_t \sim N_d(\boldsymbol{\mu}_{t+1}, \Sigma_{t+1})$.
- The estimation of $F_{\mathbf{X}_{t+1}|\mathcal{F}_t}$ can be carried out in a number of ways:
 - ▶ Fit multivariate ARMA-GARCH model with multivariate normal innovations; use model to derive estimates of $\boldsymbol{\mu}_{t+1}$ and Σ_{t+1} .

- ▶ Alternatively use the exponentially weighted moving-average (EWMA) procedure; Σ_{t+1} estimated recursively by $\hat{\Sigma}_{t+1} = \theta \mathbf{X}_t \mathbf{X}'_t + (1 - \theta) \hat{\Sigma}_t$ where θ is a small positive number (typically $\theta \approx 0.04$).
- The second critical assumption in the variance–covariance method is that the linear loss operator is sufficiently accurate. The linear loss operator is a function of the form

$$l_{[t]}^{\Delta}(\mathbf{x}) = -(c_t + \mathbf{b}'_t \mathbf{x})$$

for some constant c_t and constant vector \mathbf{b}_t , known at time t .

- We infer that, conditional on \mathcal{F}_t ,

$$L_{t+1}^{\Delta} = l_{[t]}^{\Delta}(\mathbf{X}_{t+1}) \sim N(-c_t - \mathbf{b}'_t \boldsymbol{\mu}_{t+1}, \mathbf{b}'_t \Sigma_{t+1} \mathbf{b}_t).$$

- Under normality, VaR_{α} and ES_{α} may be easily calculated:

- ▶ $\widehat{\text{VaR}}_{\alpha} = -c_t - \mathbf{b}'_t \widehat{\boldsymbol{\mu}}_{t+1} + \sqrt{\mathbf{b}'_t \widehat{\Sigma}_{t+1} \mathbf{b}_t} \Phi^{-1}(\alpha)$.
- ▶ $\widehat{\text{ES}}_{\alpha} = -c_t - \mathbf{b}'_t \widehat{\boldsymbol{\mu}}_{t+1} + \sqrt{\mathbf{b}'_t \widehat{\Sigma}_{t+1} \mathbf{b}_t} \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}$.

Pros and cons, extensions

Pros: In contrast to the methods that follow, variance-covariance offers **analytical solution** with no simulation.

Cons:

- ▶ Assumption of **multivariate normality** may seriously underestimate the **tail** of the loss distribution.
- ▶ **Linearization** may be a crude approximation.

Extensions: Instead of assuming normal risk factors, the method **could be** easily **adapted to** use **multivariate Student *t*** or multivariate hyperbolic risk-factor changes without sacrificing tractability (the method **works for all elliptical distributions** but linearization is crucial here).

9.2.3 Historical simulation

- Historical simulation is by far the most popular method used by banks for the trading book.
- Instead of estimating the distribution of $l_{[t]}(\mathbf{X}_{t+1})$ under an explicit parametric model for \mathbf{X}_{t+1} , the historical simulation method can be thought of as estimating the distribution of the loss operator under the empirical distribution of historical data $\mathbf{X}_{t-n+1}, \dots, \mathbf{X}_t$.
- Construct the historically simulated losses (under the current portfolio):

$$\{\tilde{L}_s = l_{[t]}(\mathbf{X}_s) : s = t - n + 1, \dots, t\}.$$

- One may apply the linear/quadratic loss operator (if that was already used; avoids revaluation).
- \tilde{L}_s shows what would happen to the current portfolio if the risk-factor change on day s were to recur.
- Use (\tilde{L}_s) to make inferences about the loss distribution and risk measures.

■ Inference about the loss distribution

- ▶ One could use empirical quantile estimation to estimate VaR_α .
But: What about precision (sample size; confidence intervals)?
- ▶ Or fit a parametric distribution to the historical losses L_{t-n+1}, \dots, L_t and calculate risk measures from this distribution.
But: Which distribution to fit (body or tail)?
- ▶ One could use extreme value theory to estimate the tail of the loss distribution and related risk measures based on the historical losses L_{t-n+1}, \dots, L_t .

Theoretical justification

If X_{t-n+1}, \dots, X_t are iid or, more generally, stationary, convergence of the empirical distribution to the true distribution is ensured by a suitable version of the Law of Large Numbers (e.g. Glivenko–Cantelli theorem).

Pros and Cons

- Pros:**
- ▶ Easy to implement.
 - ▶ No statistical estimation of the distribution of \mathbf{X} necessary (the empirical df of \mathbf{X} is used implicitly).
- Cons:**
- ▶ It may be difficult to collect sufficient quantities of relevant, synchronized data for all risk factors.
 - ▶ Historical data may not contain examples of extreme scenarios (“driving a car by only looking in the back mirror”).
- Note:**
- ▶ The dependence here is given by the empirical df of \mathbf{X} .
 - ▶ “Historical simulation method” is a bit of a misnomer; there is no simulation in the sense of random number generation.

In its standard form HS is an unconditional method. There are a number of ways of extending historical simulation to take account of volatility dynamics (filtered HS).

9.2.4 Dynamic Historical Simulation

A univariate approach:

- Assume that $\{\tilde{L}_s = l_{[t]}(\mathbf{X}_s) : s = t - n + 1, \dots, t\}$ are realizations from a stationary process (\tilde{L}_s) of the form $\tilde{L}_s = \mu_s + \sigma_s Z_s$, where
 - ▶ μ_s and σ_s are \mathcal{F}_{s-1} -measurable;
 - ▶ (Z_s) are SWN(0, 1) innovations with distribution function F_Z .

Example: ARMA-GARCH model.

- We can easily calculate that for the next loss $L_{t+1} = l_{[t]}(\mathbf{X}_{t+1})$ ahead $F_{L_{t+1}|\mathcal{F}_t}(l) = \mathbb{P}(\mu_{t+1} + \sigma_{t+1} Z_{t+1} \leq l | \mathcal{F}_t) = F_Z((l - \mu_{t+1})/\sigma_{t+1})$.

- Writing VaR $^t_\alpha$ for $F_{L_{t+1}|\mathcal{F}_t}^\leftarrow(\alpha)$ and ES $^t_\alpha$ for ES, we obtain

$$\text{VaR}_\alpha^t = \mu_{t+1} + \sigma_{t+1} \text{VaR}_\alpha(Z),$$

$$\text{ES}_\alpha^t = \mu_{t+1} + \sigma_{t+1} \text{ES}_\alpha(Z),$$

where Z is a random variable with distribution function F_Z .

■ Estimation

- ▶ Formal parametric time series modelling to estimate μ_{t+1} , σ_{t+1} , $\text{VaR}_\alpha(Z)$ and $\text{ES}_\alpha(Z)$.
- ▶ Often $\mu_{t+1} \approx 0$ and can be neglected. We can use EWMA to estimate $\sigma_{t-n+1}, \dots, \sigma_t, \sigma_{t+1}$ and use the standardized residuals $\{\hat{Z}_s = \tilde{L}_s / \hat{\sigma}_s, s = t - n + 1, \dots, t\}$ to estimate $\text{VaR}_\alpha(Z)$ and $\text{ES}_\alpha(Z)$.

A multivariate approach:

- We (implicitly) assume risk-factor change data $\mathbf{X}_{t-n+1}, \dots, \mathbf{X}_t$ are realizations from process (\mathbf{X}_s) which satisfies

$$\mathbf{X}_s = \boldsymbol{\mu}_s + \Delta_s \mathbf{Z}_s, \quad \Delta_s = \text{diag}(\sigma_{s,1}, \dots, \sigma_{s,d}),$$

where $(\boldsymbol{\mu}_s)$ is a process of vectors and (Δ_s) a process of diagonal matrices (all assumed \mathcal{F}_{s-1} -measurable) and $(\mathbf{Z}_s) \sim \text{SWN}(\mathbf{0}, P)$ for some correlation matrix P .

- The vector $\boldsymbol{\mu}_s$ contains the conditional means and the matrix Δ_s contains the volatilities of the component series at time s .
- An example of a model that fits into this framework is the [CCC-GARCH](#) (constant conditional correlation) process.
- The key idea of the method is to apply historical simulation to the unobserved innovations (\mathbf{Z}_s) .

- The first step is to compute estimates $\{\hat{\mu}_s : s = t - n + 1, \dots, t\}$ and $\{\hat{\Delta}_s : s = t - n + 1, \dots, t\}$.
- This can be achieved by fitting univariate time series models of ARMA-GARCH type to each of the component series in turn; alternatively we can use the univariate EWMA approach for each series.
- In the second step we construct residuals

$$\{\hat{\mathbf{Z}}_s = \hat{\Delta}_s^{-1}(\mathbf{X}_s - \hat{\mu}_s) : s = t - n + 1, \dots, t\}$$

and treat these as “**observations**” of the unobserved innovations.

- We then construct the dataset

$$\{\tilde{L}_s = l_{[t]}(\hat{\mu}_{t+1} + \hat{\Delta}_{t+1}\hat{\mathbf{Z}}_s) : s = t - n + 1, \dots, t\} \quad (64)$$

and treat these as observations of $L_{t+1} = l_{[t]}(\mathbf{X}_{t+1})$.

- To estimate VaR (or expected shortfall) we can apply simple empirical estimators directly to these data.

9.2.5 Monte Carlo

- Estimate the distribution of $L = \ell_{[t]}(\mathbf{X}_{t+1})$ under some explicit parametric model for \mathbf{X}_{t+1} .
- In contrast to the variance-covariance approach we do not necessarily make the problem analytically tractable by linearizing the loss and making an assumption of normality for the risk factors.
- Instead, make inference about L using simulated risk factor data.

The method

- 1) Based on the historical risk-factor data $\mathbf{X}_{t-n+1}, \dots, \mathbf{X}_t$, estimate a suitable statistical model for the risk-factor changes.
- 2) Simulate N new risk-factor changes $\mathbf{X}_{t+1}^{(1)}, \dots, \mathbf{X}_{t+1}^{(N)}$ from this model.
- 3) Construct the simulated losses $L_k = \ell_{[t]}(\mathbf{X}_{t+1}^{(k)})$, $k \in \{1, \dots, N\}$.

- 4) Make inference about the loss distribution F_L and risk measures using L_k , $k \in \{1, \dots, N\}$ (similar possibilities as for the historical simulation method: non-parametric/parametric/EVT).

Pros and Cons

- Pros:** ▶ General. Any distribution for \mathbf{X}_{t+1} can be taken.
- Cons:** ▶ Can be time consuming if loss operator is difficult to evaluate (depends on size and complexity of the portfolio).
▶ Note that MC approach does not address the problem of determining the distribution of \mathbf{X}_{t+1} .

9.2.6 Estimating risk measures

Aim: In both the historical simulation and Monte Carlo methods we estimate risk measures using simulated loss data. Let us suppose that we have data L_1, \dots, L_n from an underlying loss distribution F_L and the aim is to estimate $\text{VaR}_\alpha = q_\alpha(F_L) = F_L^\leftarrow(\alpha)$ or $\text{ES}_\alpha = (1-\alpha)^{-1} \int_\alpha^1 q_\theta(F_L) d\theta$. In the book we consider two possibilities:

- **L-estimators.** These are linear combinations of sample order statistics. Easiest to use notation for lower order statistics $L_{(1)} \leq \dots \leq L_{(n)}$.
- **GPD-based estimators.** These are semi-parametric estimators based on GPD approximations described in EVT chapter.

L-estimators:

VaR: $\text{VaR}_\alpha(L) = \inf\{x \in \mathbb{R} : F_L(x) \geq \alpha\}$. Replacing F_L by \hat{F}_L we obtain an L-estimator.

$$\begin{aligned}
 \widehat{\text{VaR}}_\alpha(L) &= \inf\{x \in \mathbb{R} : \hat{F}_L(x) \geq \alpha\} \\
 &= \inf\left\{x \in \mathbb{R} : \sum_{i=1}^n I_{\{L_i \leq x\}} \geq \lceil n\alpha \rceil\right\} \\
 &= \inf\left\{x \in \mathbb{R} : \sum_{i=1}^n I_{\{L_{(i)} \leq x\}} \geq \lceil n\alpha \rceil\right\} = L_{(\lceil n\alpha \rceil)}.
 \end{aligned}$$

In practice, most software uses an average of two order statistics.

ES: Assume F_L is continuous so that

$$\text{ES}_\alpha(L) = \frac{\mathbb{E}(LI_{\{L > F_L^\leftarrow(\alpha)\}})}{1 - \alpha} = \frac{\mathbb{E}(LI_{\{L > F_L^\leftarrow(\alpha)\}})}{\mathbb{E}(I_{\{L > F_L^\leftarrow(\alpha)\}})}.$$

Replacing F_L by \hat{F}_L leads to the canonical estimator

$$\widehat{\text{ES}}_\alpha(L) = \frac{\sum_{i=1}^n L_i I_{\{L_i > \widehat{\text{VaR}}_\alpha(L)\}}}{\sum_{i=1}^n I_{\{L_i > \widehat{\text{VaR}}_\alpha(L)\}}}.$$

GPD-based estimators:

We set a high threshold $u = L_{(n-k)}$ at an order statistic and fit a GPD distribution to the k excess losses over u to obtain maximum likelihood estimates $\hat{\xi}$ and $\hat{\beta}$.

For $k/n > 1 - \alpha$ we can form the risk measure estimates:

$$\widehat{\text{VaR}}_\alpha = u + \frac{\hat{\beta}}{\hat{\xi}} \left(\left(\frac{1 - \alpha}{k/n} \right)^{-\hat{\xi}} - 1 \right)$$
$$\widehat{\text{ES}}_\alpha = \frac{\widehat{\text{VaR}}_\alpha}{1 - \hat{\xi}} + \frac{\hat{\beta} - \hat{\xi}u}{1 - \hat{\xi}}.$$

9.2.7 Losses over several periods and scaling

- **Goal:** Go from single-period risk measure (e.g. one day/one week VaR/ES) to multi-period risk measure using simple formula.
- **Idea:** The loss between today and h periods ahead is

$$\begin{aligned} L_{t+h}^{(h)} &= -(V_{t+h} - V_t) = -(g(\tau_{t+h}, \mathbf{Z}_{t+h}) - g(\tau_t, \mathbf{Z}_t)) \\ &= -(g(\tau_{t+h}, \mathbf{Z}_t + \mathbf{X}_{t+1} + \cdots + \mathbf{X}_{t+h}) - g(\tau_t, \mathbf{Z}_t)) \\ &= L\left(\sum_{i=1}^h \mathbf{X}_{t+i}\right). \end{aligned}$$

- **Question:** How do risk measures scale with h ?
- There is no general answer.
- If $\mathbf{X}_{t+i} \stackrel{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{Y} = \sum_{i=1}^h \mathbf{X}_{t+i} \sim \mathcal{N}(h\boldsymbol{\mu}, h\Sigma)$. Then

$$L_{t+h}^{(h)\Delta} = -g_\tau(\tau_t, \mathbf{Z}_t) - \sum_{j=1}^d g_{z_j}(\tau_t, \mathbf{Z}_t) \left(\sum_{i=1}^h X_{t+i,j} \right) = -(c_t + \mathbf{b}'_t \mathbf{Y}).$$

- We infer that $L_{t+h}^{(h)\Delta} \sim N(-c_t - h\mathbf{b}_t' \boldsymbol{\mu}, h\mathbf{b}_t' \Sigma \mathbf{b}_t)$.
- If we assume $c_t \approx 0$, $\boldsymbol{\mu} \approx \mathbf{0}$ (typical for daily data) we obtain square-root-of-time scaling formulas for VaR and ES.
- $\text{VaR}_\alpha(L_{t+h}^{(h)\Delta}) = 0 + \sqrt{h\mathbf{b}_t' \Sigma \mathbf{b}_t} \Phi^{-1}(\alpha) = \sqrt{h} \text{VaR}_\alpha(L_{t+1}^\Delta)$.
- $\text{ES}_\alpha(L_{t+h}^{(h)\Delta}) = 0 + \sqrt{h\mathbf{b}_t' \Sigma \mathbf{b}_t} \frac{\varphi(\Phi^{-1}(\alpha))}{1-\alpha} = \sqrt{h} \text{ES}_\alpha(L_{t+1}^\Delta)$.
- Note the many underlying assumptions:
 - ▶ X_{t+i} independent;
 - ▶ X_{t+i} multivariate normal;
 - ▶ The linearized loss provides a sufficiently good approximation to the true loss distribution.
- Note also that we have only considered the scaling of unconditional risk measures.

9.3 Backtesting

- Backtesting is the practice of evaluating risk measurement procedures by comparing *ex ante estimates/forecasts of risk measures* with *ex post realized losses and gains*.
- It allows us to evaluate whether a model and estimation procedure produce *credible risk measure estimates*.

9.3.1 Violation-based tests for VaR

- Let VaR_α^t denote the α -quantile of the conditional loss distribution $F_{L_{t+1} | \mathcal{F}_t}$ and consider the event indicator variable $I_{t+1} = I_{\{L_{t+1} > \text{VaR}_\alpha^t\}}$.
- The event $\{L_{t+1} > \text{VaR}_\alpha^t\}$ is a VaR *violation* or *exception*.
- Assuming a continuous loss distribution, we have, by definition of the quantile,

$$\mathbb{E}(I_{t+1} | \mathcal{F}_t) = \mathbb{P}(L_{t+1} > \text{VaR}_\alpha^t | \mathcal{F}_t) = 1 - \alpha, \quad (65)$$

- I_{t+1} is a Bernoulli variable with event probability $(1 - \alpha)$.
- Moreover, the sequence of VaR exception indicators (I_t) is an iid sequence.
- The sum of exception indicators is binomially distributed:

$$M = \sum_{t=1}^m I_t \sim B(m, 1 - \alpha).$$

- Assume exceptions occur at times $1 \leq T_1 < \dots < T_M \leq m$ and set $T_0 = 0$. The spacings $S_j = T_j - T_{j-1}$ will be independent geometrically distributed rvs with mean $1/(1 - \alpha)$, so that

$$\mathbb{P}(S_j = k) = \alpha^{k-1}(1 - \alpha), \quad k \in \mathbb{N}.$$

- Both of these properties are testable in empirical data.
- For small event probability $1 - \alpha$, the Bernoulli Trials Process may be well approximated by a Poisson process.

- Also for small $1 - \alpha$ the geometric distribution may be approximated by an exponential distribution.
- Suppose we estimate VaR_α^t at time point t by $\widehat{\text{VaR}}_\alpha^t$.
- In a backtest we consider empirical indicator variables

$$\widehat{I}_{t+1} = I_{\{L_{t+1} > \widehat{\text{VaR}}_\alpha^t\}}.$$

- The sequence $(\widehat{I}_t)_{1 \leq t \leq m}$ should behave like a realization from a Bernoulli trials process with event probability $(1 - \alpha)$.
- To test binomial behaviour for number of violations we compute a score test statistic

$$Z_m = \frac{(\sum_{t=1}^m \widehat{I}_t) - m(1 - \alpha)}{\sqrt{m\alpha(1 - \alpha)}}$$

and reject Bernoulli hypothesis at 5% level if $Z_m > \Phi^{-1}(0.95)$.

- Exponential spacings can be tested numerically or with a Q-Q plot.

9.3.2 Violation-based tests of expected shortfall

- Let ES_α^t denote the one-period expected shortfall and $\widehat{\text{ES}}_\alpha^t$ its estimate.
- Assume (L_t) follows a model of the form $L_t = \sigma_t Z_t$, where σ_t is a function of \mathcal{F}_{t-1} and the (Z_t) are $\text{SWN}(0, 1)$ innovations.
- Then we can define a process (K_t) by

$$K_{t+1} = \frac{(L_{t+1} - \text{ES}_\alpha^t)}{\text{ES}_\alpha^t} I_{\{L_{t+1} > \text{VaR}_\alpha^t\}} = \frac{Z_{t+1} - \text{ES}_\alpha(Z)}{\text{ES}_\alpha(Z)} I_{\{Z_{t+1} > q_\alpha(Z)\}},$$

and note that it is a zero-mean iid sequence.

- This suggests we form violation residuals of the form

$$\widehat{K}_{t+1} = \frac{(L_{t+1} - \widehat{\text{ES}}_\alpha^t)}{\widehat{\text{ES}}_\alpha^t} \widehat{I}_{t+1}. \quad (66)$$

- We test for mean-zero behaviour using a bootstrap test on the non-zero violation residuals (McNeil and Frey (2000)).

9.3.3 Elicitability and comparison of risk measure estimates

- The elicibility concept has been introduced into the backtesting literature by Gneiting (2011); see also important papers by Bellini and Bigozzi (2013) and Ziegel (2014).
- A key concept is that of a **scoring function** $S(y, l)$ which measures the discrepancy between a forecast y and a realized loss l .
- Forecasts are made by applying real-valued statistical functionals T (such as mean, median or other quantile) to the distribution of the loss F_L to obtain the forecast $y = T(F_L)$.
- Suppose that for some class of loss distribution functions a real-valued statistical functional T satisfies

$$T(F_L) = \arg \min_{y \in \mathbb{R}} \int_{\mathbb{R}} S(y, l) dF_L(l) = \arg \min_{y \in \mathbb{R}} \mathbb{E}(S(y, L)) \quad (67)$$

for a scoring function S and any loss distribution F_L in that class.

- Suppose moreover that $T(F_L)$ is the unique minimizing value.
- The scoring function S is said to be **strictly consistent** for T .
- The functional $T(F_L)$ is said to be **elicitable**.
- Note that (67) implies that

$$\begin{aligned}\frac{d}{dy} \mathbb{E}(S(y, L)) \Big|_{y=T(F_L)} &= \int_{\mathbb{R}} \frac{d}{dy} S(y, l) dF_L(l) \Big|_{y=T(F_L)} \\ &= \mathbb{E}(h(T(F_L), L)) = 0\end{aligned}$$

where h is the derivative of the scoring function.

- The **VaR risk measure** corresponds to $T(F_L) = F_L^\leftarrow(\alpha)$. For any $0 < \alpha < 1$ this functional **is elicitable** for strictly increasing distribution functions. The scoring function

$$S_\alpha^q(y, l) = |1_{\{l \leq y\}} - \alpha| |l - y| \quad (68)$$

is strictly consistent for T .

- The α -expectile of L is defined to be the risk measure that minimizes $\mathbb{E}(S_\alpha^e(y, L))$ where the scoring function is

$$S_\alpha^e(y, l) = |1_{\{l \leq y\}} - \alpha|(l - y)^2. \quad (69)$$

This risk measure is **elicitable by definition**.

- Bellini and Bigozzi (2013) and Ziegel (2014) show that a risk measure is coherent and elicitable if and only if it is the α -expectile risk measure for $\alpha \geq 0.5$; see also Weber (2006). **Expected shortfall is not elicitable.**
- VaR_α^t minimizes

$$\mathbb{E}\left(S_\alpha^q(\text{VaR}_\alpha^t, L_{t+1}) \mid \mathcal{F}_t\right)$$

for the scoring function in (68). We refer to $S_\alpha^q(\text{VaR}_\alpha^t, L_{t+1})$ as a (theoretical) **VaR score**.

- Assume VaR_α^t is replaced by an estimate at each time point and consider the VaR scores $\{\widehat{S}_\alpha^q(\widehat{\text{VaR}}_\alpha^t, L_{t+1}) : t = 1, \dots, m\}$
- These can be used to address questions of **relative model performance**.

- The statistic

$$Q_0 = \frac{1}{m} \sum_{t=1}^m S_\alpha^q(\widehat{\text{VaR}}_\alpha^t, L_{t+1})$$

can be used as a measure of relative model performance.

- If two models A and B deliver VaR estimates $\{\widehat{\text{VaR}}_\alpha^{tA}, t = 1, \dots, m\}$ and $\{\widehat{\text{VaR}}_\alpha^{tB}, t = 1, \dots, m\}$ with corresponding average scores Q_0^A and Q_0^B , then we expect the better model to give estimates closer to the true VaR numbers and thus a value of Q_0 that is lower.
- Of course, the power to discriminate between good models and inferior models will depend on the length of the backtest.

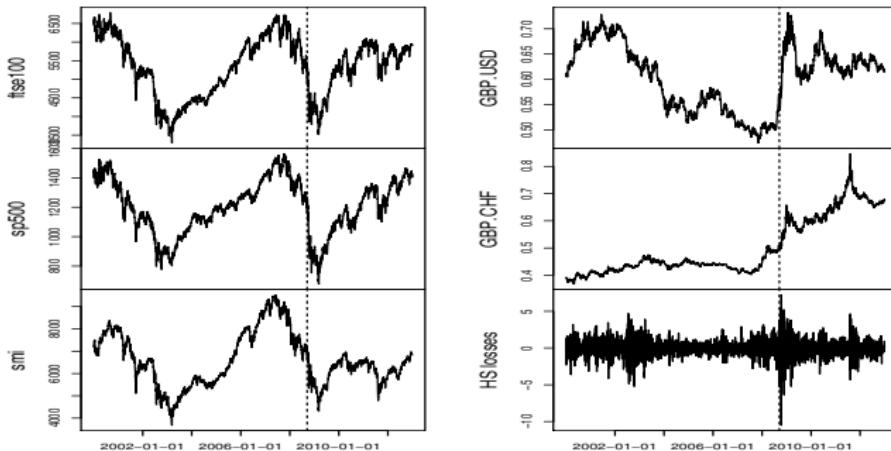
9.3.4 Empirical comparison of methods using backtesting concepts

- We apply various VaR estimation methods to the portfolio of a hypothetical investor in international equity indexes.
- The investor is assumed to have domestic currency sterling (GBP) and to invest in the Financial Times 100 Shares Index (FTSE 100), the Standard & Poor's 500 (S&P 500) and the Swiss Market Index (SMI).
- The portfolio is influenced by five risk factors.
- On any day t we standardize the total portfolio value V_t in sterling to be one and assume portfolio weights are 30%, 40% and 30%, respectively.
- The loss operator and linear loss operator are:

$$l_{[t]}(\mathbf{x}) = 1 - (0.3e^{x_1} + 0.4e^{x_2+x_4} + 0.3e^{x_3+x_5})$$

$$l_{[t]}^{\Delta}(\mathbf{x}) = -(0.3x_1 + 0.4(x_2 + x_4) + 0.3(x_3 + x_5))$$

- x_1 , x_2 and x_3 represent log-returns on the three indexes and x_4 and x_5 are log-returns on the GBP/USD and GBP/CHF exchange rates.



- The final picture shows the corresponding historical simulation data. The vertical dashed line is Lehman Brothers bankruptcy.

Estimation methods:

VC. The variance–covariance method assuming multivariate Gaussian risk-factor changes and using the multivariate EWMA method to estimate the conditional covariance matrix of risk-factor changes.

HS. The standard unconditional historical simulation method.

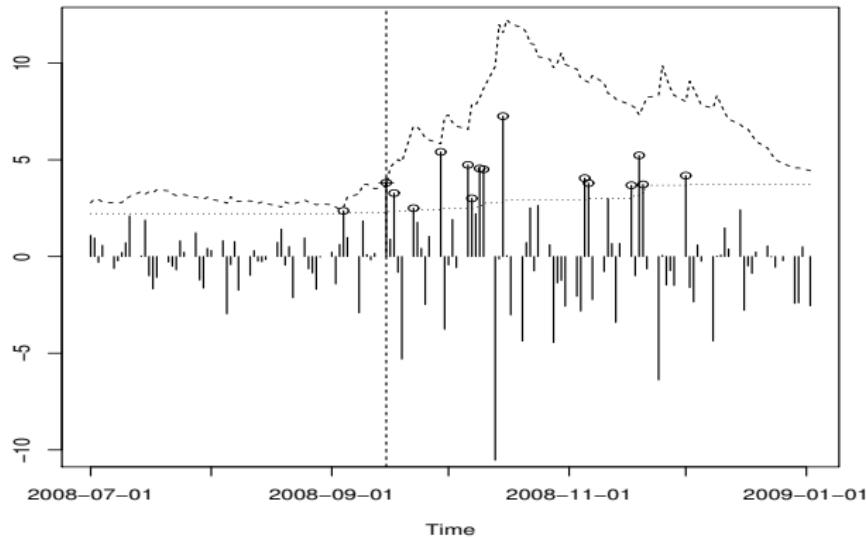
HS-GARCH. The univariate dynamic approach to historical simulation in which a GARCH(1, 1) model with a constant conditional mean term and Gaussian innovations is fitted to the historically simulated losses to estimate the volatility of the next day's loss.

HS-GARCH- t . A similar method to HS-GARCH but Student t innovations are assumed in the GARCH model.

HS-MGARCH. The multivariate dynamic approach to historical simulation in which GARCH(1, 1) models with constant conditional mean terms are fitted to each time series of risk-factor changes to estimate volatilities.

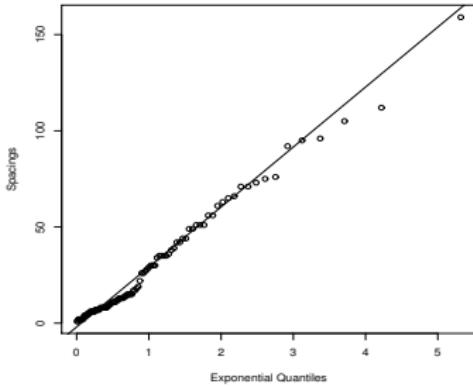
Year	2005	2006	2007	2008	2009	2010	2011	2012	All
Trading days	258	257	258	259	258	259	258	258	2065
Results for 95% VaR									
Expected no. of violations	13	13	13	13	13	13	13	13	103
VC	8	16	17	19	13	15	14	14	116
HS	0	6	28	49	19	6	10	1	119
HS-GARCH	9	13	22	22	13	14	9	15	117
HS-GARCH- <i>t</i>	9	14	23	22	14	15	10	15	122
HS-MGARCH	5	14	21	19	12	9	11	12	103
Results for 99% VaR									
Expected no. of violations	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	21
VC	2	8	8	8	2	4	5	6	43
HS	0	0	10	22	2	0	2	0	36
HS-GARCH	2	8	8	10	5	4	3	3	43
HS-GARCH- <i>t</i>	2	8	6	8	1	4	2	1	32
HS-MGARCH	0	4	4	5	0	1	2	1	17

- The HS method **does not react to changing volatility**:



- Dotted line is HS; dashed line is HS-MGARCH; vertical line is Lehmann.
- Circle is VaR violation for HS; cross is VaR violation for HS-MGARCH.

- Q-Q plot of spacings between exceptions (for HS-MGARCH):



- Violation residual test for ES (n : number of VaR violations):

	$ES_{0.95}$	n	$ES_{0.99}$	n
VC	0.00	116	0.05	43
HS	0.02	119	0.25	36
HS-GARCH	0.00	117	0.05	43
HS-GARCH- t	0.12	122	0.68	32
HS-MGARCH	0.99	103	0.55	17

9.3.5 Backtesting the predictive distribution

- As well as backtesting VaR and expected shortfall we can also devise tests that assess the **overall quality of the estimated conditional loss distribution, or its tail.**
- If L_{t+1} is a random variable with (continuous) distribution function $F_{L_{t+1}|\mathcal{F}_t}$, then $U_{t+1} = F_{L_{t+1}|\mathcal{F}_t}(L_{t+1})$ is **uniform** (probability transform).
- In actual applications we estimate $F_{L_{t+1}|\mathcal{F}_t}$ from data up to time t and we backtest our estimates by forming $\hat{U}_{t+1} = \hat{F}_{L_{t+1}|\mathcal{F}_t}(L_{t+1})$ on day $t + 1$.
- Suppose we estimate the predictive distribution on days $t = 0, \dots, n - 1$ and form backtesting data $\hat{U}_1, \dots, \hat{U}_n$; we expect these to behave like a sample of **iid uniform data**.
- The distributional assumption can be assessed by standard goodness-of-fit tests like the Kolmogorov–Smirnov test.

10 Credit risk

- 10.1 Credit risky instruments
- 10.2 Measuring credit quality
- 10.3 Structural models of default
- 10.4 Bond and CDS pricing in hazard rate models
- 10.5 Pricing with stochastic hazard rates
- 10.6 Affine models

What is credit risk?

“Credit risk is the risk of a loss arising from the failure of a counterparty to honour its contractual obligations. This subsumes both default risk (the risk of losses due to the default of a borrower or a trading partner) and downgrade risk (the risk of losses caused by a deterioration in the credit quality of a counterparty that translates into a downgrading in some rating system). ”

- *Obligor* = a counterparty who has a **financial obligation** to us; for example, a debtor who owes us money, a bond issuer who promises interest, or a counterparty in a derivatives transaction.
- *Default* = failure to fulfill that obligation, for example, failure to repay loan or pay interest/coupon on a loan/bond; generally due to **lack of liquidity** or **insolvency**; may entail **bankruptcy**.

A crucial risk category

- Credit risk is omnipresent in the portfolio of a typical financial institution.
- A **portfolio of loans or (corporate) bonds** is obviously affected by credit risk.
- Credit risk accompanies any OTC (over-the-counter) **derivative transaction** such as a swap, because the default of one of the parties involved may substantially affect the actual pay-off of the transaction.
- There is a specialized market for **credit derivatives**, such as credit default swaps.
- Credit risk relates to the core activities of most **banks** but is also highly relevant to **insurance companies**: Insurers are exposed to substantial credit risk in their investment portfolios and counterparty default risk in their reinsurance treaties.

Credit risk management: A range of tasks

- An enterprise needs to determine the **capital** it requires to absorb losses due to credit risk.
- Portfolios of credit-risky instruments should be well **diversified** and **optimized** according to risk-return considerations.
- Institutions need to manage their portfolio of traded credit derivatives, which involves **pricing**, **hedging** and **managing collateral** for such trades.
- Financial institutions need to control the **counterparty credit risk** in their trades and contracts with other institutions. This has particularly been the case since the 2007–2009 financial crisis.

10.1 Credit risky instruments

10.1.1 Loans

- May be categorized into: *retail* loans (to individuals and small or medium-sized companies), *corporate* loans (to larger companies), *interbank* loans and *sovereign* loans (to governments).
- In each of these categories there may be a number of different products. For example, retail customers may borrow money using mortgages against property, credit cards and overdrafts.
- A sum of money, known as the *principal*, is advanced to the borrower for a particular term in exchange for a series of defined *interest* payments, which may be at fixed or floating interest rates. At the end of the term the borrower is required to pay back the principal.

- A useful distinction to make is between **secured** and **unsecured** lending. If a loan is secured the borrower has pledged an asset as collateral for the loan. In a mortgage the collateral is a property. In the event of default, the lender may take possession of the asset to mitigate the loss.
- In an unsecured loan the lender has no such claim on a collateral asset.

10.1.2 Bonds

- Bonds are publicly traded securities issued by companies and governments which allow the issuer to raise funding on financial markets.
- Bonds issued by companies are **corporate bonds** and bonds issued by governments are known as **treasuries**, **sovereign bonds** or, particularly in the UK, **gilts** (gilt-edged securities).
- The security commits the bond issuer (borrower) to make a series of interest payments to the bond buyer (lender) and pay back the principal at a fixed maturity.

- The interest payments, or coupons, may be **fixed** at the issuance of the bond (so-called fixed-coupon bonds). Alternatively, there are also bonds where the interest payments vary with market rates (so-called **floating-rate notes**).
- The reference rate for the floating rates is often a LIBOR rate (London Interbank Offered Rate).
- There are also **convertible bonds** which allow the purchaser to convert them into shares of the issuing company at predetermined time points.

Risks faced by bondholders

- A bond holder is subject to a **number of risks**, particularly **interest-rate risk**, **default risk**, **downgrade risk** and **spread risk**.
- Changes in the **term structure of interest rates** affect the value of bonds.
- As for loans, default risk is the risk that promised coupon and principal payments are not made.

- Downgrade risk is the risk that the bond loses value because the issuer's credit rating is lowered.
- Historically government bonds issued by developed countries have been considered default-free; for obvious reasons, after the European debt crisis of 2010–2012, this notion was called into question.
- Spread risk is a form of [market risk](#) that refers to changes in [credit spreads](#). The credit spread of a defaultable bond measures the difference in the yield of the bond and the yield of an equivalent default-free bond.
- An increase in the spread of a bond means that the market value of the bond falls, which is generally interpreted as indicating that the financial markets perceive an increased default risk for the bond.

10.1.3 Derivative contracts subject to counterparty risk

- A substantial part of all derivative transactions is carried out over the counter and there is no central clearing counterparty such as an organized exchange to guarantee the fulfilment of the contractual obligations.
- These trades are subject to the risk that a contracting party defaults during the transaction, thus affecting the cash flows that are actually received by the other party. This risk, known as [counterparty credit risk](#), received a lot of attention during the financial crisis of 2007-2009.
- Some of the institutions heavily involved in derivative transactions experienced worsening credit quality or—in the case of Lehman Brothers—even a default event.
- Counterparty risk management is now a key issue for all financial institutions and the focus of many new regulatory developments.

Example of interest-rate swap

- Two parties A and B agree to exchange a series of interest payments on a given nominal amount of money for a given period.
- A receives payments at a fixed interest rate and makes floating payments at a rate equal to the three-month LIBOR rate.
- Suppose that A defaults at time τ_A before the maturity of the contract.
- If interest rates have risen relative to their value at inception of contract:
 - ▶ The fixed interest payments have decreased in value and the value of the contract has increased for B .
 - ▶ The default of A constitutes a loss for B .
 - ▶ The loss size depends on the term structure of interest rates at τ_A .

- If interest rates have fallen relative to their value at $t = 0$:
 - ▶ The fixed payments have increased in value so that the swap has a negative value for B .
 - ▶ B will still have to pay the value of the contract into the bankruptcy pool,
 - ▶ There is no upside for B in A 's default.
- If B defaults first the situation is reversed: falling rates lead to a counterparty-risk-related loss for A .

Management of counterparty risk

- Counterparty risk has to be taken into account in pricing and valuation. This has led to the notion of [credit value adjustments \(CVA\)](#).
- Counterparty risk needs to be controlled using risk-mitigation techniques such as [netting](#) and [collateralization](#).
- Under a netting agreement the value of all derivatives transactions between A and B is computed and only the aggregated value is subject to counterparty risk; since offsetting transactions cancel each other out, this has the potential to reduce counterparty risk substantially.
- Under a collateralization agreement the parties exchange collateral (cash and securities) that serves as a pledge for the receiver. The value of the collateral is adjusted dynamically to reflect changes in the value of the underlying transactions.

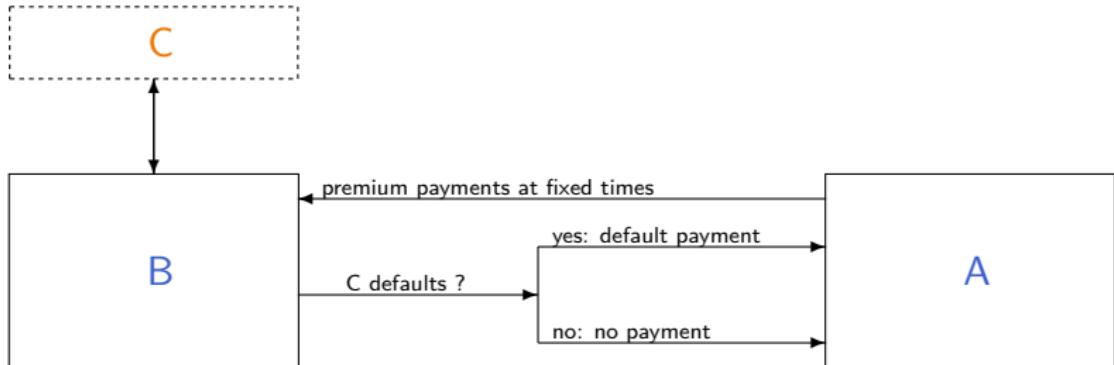
10.1.4 Credit default swaps and related credit derivatives

- Credit derivatives are securities which are primarily used for the hedging and trading of credit risk.
- The promised pay-off of a credit derivative is related to credit events affecting one or more firms.
- Major participants in the market for credit derivatives are banks, insurance companies and investment funds.
- Retail banks are typically net buyers of protection against credit events; other investors such as hedge funds and investment banks often act as both sellers and buyers of credit protection.
- Credit default swaps (CDSs) are the workhorses of the credit derivatives market and the market for CDSs written on larger corporations is fairly liquid.

Structure of CDS

Consider contract with maturity T and ignore counterparty credit risk.
Three parties are involved (only two directly):

- **C (reference entity)**; default at time $\tau_C < T$ triggers default payment.
- **A (protection buyer)**; pays premiums to B until $\min(\tau_C, T)$.
- **B (protection seller)**; makes default payment to A if $\tau_C < T$.



CDS: Payment flows

- If the reference entity experiences a default before the maturity date T of the contract, the protection seller makes a default payment to the protection buyer, which mimics the loss due to the default of a bond issued by the reference entity (the reference asset); this part of a CDS is called the **default payment leg**.
- As compensation the protection buyer makes periodic premium payments (typically quarterly or semiannually) to the protection seller (the **premium payment leg**); after the default of the reference entity, premium payments stop. There is no initial payment.
- The premium payments are quoted in the form of an annualized percentage x^* of the notional value of the reference asset; x^* is termed the (fair or market quoted) **CDS spread**.

Use of CDS

Investors enter into CDS contracts for various reasons.

- Bond investors with a large credit exposure to the reference entity may buy CDS protection to insure themselves against losses due to default of a bond (easier than reducing the original bond position as CDS contracts are more liquid).
- CDS contracts are also held for speculative reasons: so-called **naked** CDS positions, where the protection buyer does not own the bond are often assumed by investors who are speculating on the widening of the credit spread of the reference entity (similar to short-selling bonds issued by the reference entity.)
- Note that, in contrast to insurance, there is no requirement for the protection buyer to have **insurable interest**, that is, to actually own a bond issued by the reference entity.

10.1.5 PD, LGD and EAD

Exposure

- If we make a loan or buy a bond, our exposure is relatively easy to determine, since it is mainly the **principal** that is at stake. There is some additional uncertainty about the value of **lost interest payments**.
- A further source of exposure uncertainty is due to the widespread use of credit lines, essentially a ceiling up to which a corporate client can borrow money at given terms.
- For OTC derivatives the counterparty risk exposure is even more difficult to quantify, since it is a stochastic variable depending on the **unknown time** at which a counterparty defaults and the evolution of the value of the derivative up to that point.
- In practice the concept used is **exposure at default (EAD)**, which recognises that exposure often depends on the exact default time.

Probability of default (PD)

- When measuring the risk of losses over a fixed time horizon, for example one year, we are particularly concerned with estimating the probability that obligors default by the time horizon, a quantity known to practitioners as **probability of default** or PD.
- The probability of default is related to the credit quality of an obligor and we discuss models of credit quality next.
- For instruments where the loss is dependent on the exact timing of default, for example OTC derivatives with counterparty risk, the risk of default is described by the whole distribution of possible default times and not just the probability of default by a fixed horizon.
- In simple models of default time, the probability of default may be expressed in terms of a **hazard function** which measures the risk of default at any instant in time.

Loss given default (LGD)

- In the event of default, it is unlikely that the entire exposure is lost.
- When a mortgage holder defaults on a residential mortgage, and there is no realistic possibility of restructuring the debt, the lender can sell the property (the collateral asset) and the proceeds from the sale will make good some of the lost principal.
- When a bond issuer goes into administration, the bond holders join the group of creditors who will be partly recompensed for their losses by the sale of the firm's assets.
- Practitioners use the term [loss given default](#) or LGD to describe the proportion of the exposure that is actually lost in the event of default, or its converse, the [recovery](#), to describe the amount of the exposure that can be recovered through debt restructuring and asset sales.

Dependence of EAD, PD and LGD

- EAD, PD and LGD are dependent quantities. For example, in a period of financial distress, when PDs are high, asset values of firms are depressed and firms are defaulting, recoveries are likely to be correspondingly low, so that there is positive dependence between PDs and LGDs.

10.2 Measuring credit quality

Scores, ratings & measures inferred from prices

There are two philosophies for quantifying the credit quality or default probability of an obligor.

- 1) Credit quality can be described by a credit *rating or score* that is based on *empirical* data and expert judgement.
- 2) For obligors whose equity is traded on financial markets, *prices* can be used to infer the *market's view* of the credit quality of the obligor.

Credit ratings and scores fulfill a similar function—they allow us to order obligors by their credit risk and map that risk to an estimate of the PD.

Credit ratings tend to be expressed on an ordered categorical scale whereas credit scores are often expressed in points on a metric scale.

Rating and scoring

- The task of rating obligors is often outsourced to a rating agency such as Moody's or Standard & Poor's (S&P).
- In the S&P rating system there are seven pre-default rating categories labelled AAA, AA, A, BBB, BB, B, CCC, with AAA being the highest and CCC the lowest rating.
- Moody's uses nine pre-default rating categories labelled Aaa, Aa, A, Baa, Ba, B, Caa, Ca, C.
- A finer alpha-numeric system is also used by both agencies.
- Credit scores are traditionally used for retail customers and are based on so-called *scorecards*. Historical data is used to model default risk as a function of demographic, behavioural and financial covariates using techniques like logistic regression. The covariates are weighted and combined into a score.

10.2.1 Credit rating migration

- In the credit-migration approach each firm is assigned to a credit-rating category at any given time point.
- We assume that the current credit rating completely determines the default probability.
- The probability of moving from one credit rating to another over a given risk horizon (typically one year) is then specified.
- These probabilities, known as [transition probabilities](#), are typically presented in the form of a matrix. They are estimated from historical data on [empirical transition rates](#).
- The following example is taken from Ou (2013), (Exhibit 26). It gives average transition rates from one rating to another within one year. WR stands for withdrawn rating.

Initial rating	Rating at year-end (%)								WR	
	Aaa	Aa	A	Baa	Ba	B	Caa	Ca-C		
Aaa	87.20	8.20	0.63	0.00	0.03	0.00	0.00	0.00	0.00	3.93
Aa	0.91	84.57	8.43	0.49	0.06	0.02	0.01	0.00	0.02	5.48
A	0.06	2.48	86.07	5.47	0.57	0.11	0.03	0.00	0.06	5.13
Baa	0.039	0.17	4.11	84.84	4.05	7.55	1.63	0.02	0.17	5.65
Ba	0.01	0.05	0.35	5.52	75.75	7.22	0.58	0.07	1.06	9.39
B	0.01	0.03	0.11	0.32	4.58	73.53	5.81	0.59	3.85	11.16
Caa	0.01	0.02	0.02	0.12	0.38	8.70	61.71	3.72	13.34	12.00
Ca-C	0.00	0.00	0.00	0.00	0.40	2.03	9.38	35.46	37.93	14.80

- 1-year default probability for an A-rated company is estimated to be 0.06%, whereas for a Caa-rated company it is 13.3%.
- In practice a correction to the figures would probably be undertaken to account for rating withdrawals

- Rating agencies also publish cumulative default rates over longer time horizons.
- These provide estimates of cumulative default probabilities over several years. Alternative estimates of multi-year default probabilities can be inferred from one-year transition matrices as explained later.
- The data are taken from Ou (2013), (Exhibit 33).

Initial rating	Term						
	1	2	3	4	5	10	15
Aaa	0.00	0.01	0.01	0.04	0.11	0.50	0.93
Aa	0.02	0.07	0.14	0.26	0.38	0.92	1.75
A	0.06	0.20	0.41	0.63	0.87	2.48	4.26
Baa	0.18	0.50	0.89	1.37	1.88	4.70	8.62
Ba	1.11	3.08	5.42	7.93	10.18	19.70	29.17
B	4.05	9.60	15.22	20.13	24.61	41.94	52.22
Caa-C	16.45	27.87	36.91	44.13	50.37	69.48	79.18

TTC and PIT

- Default rates tend to vary with the state of the economy, being high during recessions and low during periods of economic expansion.
- Transition rates as estimated by rating agencies are historical averages over longer time horizons covering several *business cycles*.
- For instance the transition rates we show have been estimated from rating migration data over the period 1970–2012.
- Rating agencies focus on the average credit quality *through the business cycle* (TTC) when attributing a credit rating to a particular firm.
- Hence the default probabilities used in the rating migration approach are estimates of the average default probability, independent of the current economic environment.
- These can be contrasted with *point-in-time* (PIT) estimates of default probabilities which reflect the current macroeconomic environment.

10.2.2 Rating transitions as a Markov chain

- Let (R_t) denote a discrete-time stochastic process taking values in $S = \{0, 1, \dots, n\}$ at times $t = 0, 1, \dots$
- The set S defines rating states of increasing creditworthiness with 0 representing default. (R_t) models an obligor's rating over time.
- We will assume that (R_t) is a **Markov chain**. This means that it has the Markov property that

$$\mathbb{P}(R_t = k \mid R_0 = r_0, R_1 = r_1, \dots, R_{t-1} = j) = \mathbb{P}(R_t = k \mid R_{t-1} = j)$$

for all $t \geq 1$ and all $j, r_0, r_1, r_{t-2}, k \in S$.

- Conditional probabilities of rating transitions given an obligor's rating history depend only on the previous rating $R_{t-1} = j$ at the last time point and not the more distant history.
- There is evidence that rating histories show **momentum** and **stickiness** which violates the Markov assumption (Lando and Skodeberg (2002)).

Properties of Markov chains

- The Markov chain is **stationary** if, for all $t \geq 1$ and rating states j, k ,

$$\mathbb{P}(R_t = k \mid R_{t-1} = j) = \mathbb{P}(R_1 = k \mid R_0 = j).$$

- In this case we can define the transition matrix $\mathbf{P} = (p_{jk})$ with elements $p_{jk} = \mathbb{P}(R_t = k \mid R_{t-1} = j)$, for any $t \geq 1$.
- The Chapman-Kolmogorov equations say that

$$\mathbb{P}(R_t = k \mid R_{t-2} = j) = \sum_{l \in S} p_{jl} p_{lk}.$$

- An implication of this is that the matrix of transition probabilities over two time steps is given by $\mathbf{P}^2 = \mathbf{P} \times \mathbf{P}$.
- It is not clear how a matrix of transition probabilities for a fraction of a time period can be computed (one would need continuous-time chains).

Estimating default and transition probabilities

- For $t = 0, \dots, T - 1$ and $j \in S \setminus \{0\}$ let N_{tj} denote the number of companies that are rated j at time t and for which a rating is available at time $t + 1$; let N_{tjk} denote the subset of those companies that are rated k at time $t + 1$.
- Under the Markovian assumption the N_{tj} companies rated j can be thought of as being randomly allocated to the ratings $k \in S$ according to probabilities p_{jk} which satisfy $\sum_{k=0}^n p_{jk} = 1$.
- In this framework the likelihood is given by

$$L((p_{jk}); (N_{tj}), (N_{tjk})) = \prod_{t=0}^{T-1} \left(\prod_{j=1}^n \left(N_{tj}! \prod_{k=0}^n \frac{p_{jk}^{N_{tjk}}}{N_{tjk}!} \right) \right).$$

- If this is maximized subject to the constraints that $\sum_{k=0}^n p_{jk} = 1$ for $j = 1, \dots, n$ we obtain the maximum likelihood estimator

$$\hat{p}_{jk} = \frac{\sum_{t=0}^{T-1} N_{tjk}}{\sum_{t=0}^{T-1} N_{tj}}. \quad (70)$$

Continuous-time Markov transition models

- The main drawback of modelling rating transitions as a discrete-time Markov chain is that **we ignore any information about intermediate transitions** taking place between two times t and $t + 1$.
- For this reason, better to consider transitions in continuous time. Probabilities cannot be modelled directly but are instead modelled in terms of transition rates.
- Over any small time step of duration δt we assume that the probability of a transition from rating j to k is given approximately by $\lambda_{jk}\delta t$. The probability of staying at rating j is given by $1 - \sum_{k \neq j} \lambda_{jk}\delta t$.

- If we now define a matrix Λ to have off-diagonal entries λ_{jk} and diagonal entries $-\sum_{k \neq j} \lambda_{jk}$, we can summarise these transition probabilities for a small time step in the matrix

$$I_{n+1} + \Lambda \delta t.$$

- Λ is the so-called *generator matrix*.

The generator matrix

- Let $P(t)$ be the matrix of transition probabilities for the period $[0, t]$.
- Divide $[0, t]$ into N small time steps of size $\delta t = t/N$ for N large.
- The matrix of transition probabilities can be approximated by

$$P(t) \approx \left(I_{n+1} + \frac{\Lambda t}{N} \right)^N$$

- This converges, as $N \rightarrow \infty$, to the so-called matrix exponential of Λt .

$$P(t) = \exp(\Lambda t)$$

- We can compute transition probabilities for any time horizon.
- A Markov chain with generator Λ can be **constructed** in the following way. An obligor remains in rating state j for an exponentially distributed amount of time with parameter $\lambda = \sum_{k \neq j} \lambda_{jk}$. When a transition takes place the probability that it is from j to state k is given by λ_{jk}/λ .

Estimating generator in continuous time

- This construction leads to **natural estimators** for the matrix Λ .
- Since λ_{jk} is the rate of migrating from j to k we can estimate it by

$$\hat{\lambda}_{jk} = \frac{N_{jk}(T)}{\int_0^T Y_j(t) dt}, \quad (71)$$

where $N_{jk}(T)$ is the total number of observed transitions from j to k in $[0, T]$ and $Y_j(t)$ is the number of obligors with rating j at time t .

- The denominator represents the total time spent in state j by all the companies in the dataset.

- Note that this is the continuous-time analogue of the maximum likelihood estimator in (70).
- It can be shown to be the maximum likelihood estimator for the transition rates in a homogenous continuous-time Markov chain.

10.3 Structural models of default

10.3.1 The Merton model

- Merton's model (1974) is the prototype of all firm value models.
- Consider firm with stochastic asset-value (V_t), financing itself by **equity** (i.e. by issuing shares) and **debt**.
- Assume that debt consists of single zero coupon bond with face or nominal value B and maturity T .
- Denote by S_t and B_t the value at time $t \leq T$ of equity and debt so that

$$V_t = S_t + B_t, \quad 0 \leq t \leq T.$$

- Assume that default occurs if the firm misses a payment to its debt holders and hence only at T .

Equity and debt as contingent claims on assets

- At T we have two possible cases:
 - 1) $V_T > B$. In that case the debtholders receive B ; shareholders receive residual value $S_T = V_T - B$, and there is no default.
 - 2) $V_T \leq B$. In that case the firm cannot meet its financial obligations, and shareholders hand over control to the bondholders, who liquidate the firm; hence we have $B_T = V_T$, $S_T = 0$.
- In summary we obtain

$$S_T = (V_T - B)^+$$

$$B_T = \min(V_T, B) = B - (B - V_T)^+.$$

- The **value of equity** at T equals the pay-off of a European call option on V_T with exercise price equal to B .
- The **value of the debt** at T equals the nominal value of debt minus the pay-off of a European put option on V_T .

- The option interpretation explains certain **conflicts of interest** between shareholders and bondholders.
- For example, shareholders have more interest in the firm taking on risky projects/investments since the value of an option increases with the volatility of the underlying security.
- Bondholders have a short position on the firm's assets and would like to see the volatility reduced.

The asset value process

It is assumed that asset value (V_t) follows a diffusion of the form

$$dV_t = \mu_V V_t dt + \sigma_V V_t dW_t$$

for constants $\mu_V \in \mathbb{R}$, $\sigma_V > 0$, and a Brownian motion $(W_t)_{t \geq 0}$, so that

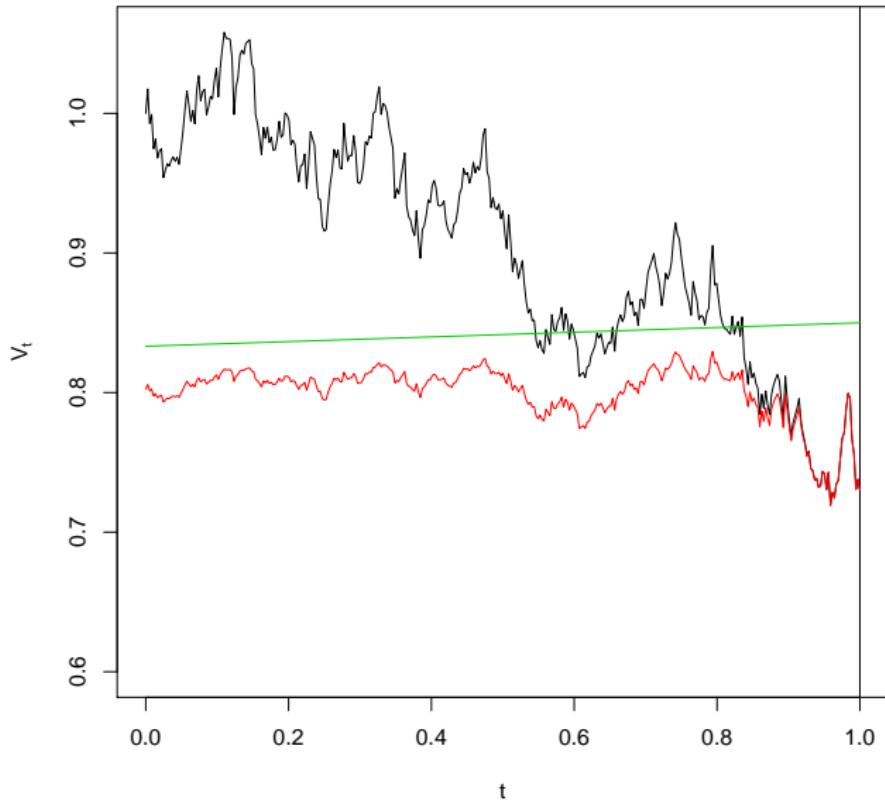
$$V_T = V_0 \exp\left((\mu_V - \frac{1}{2}\sigma_V^2)T + \sigma_V W_T\right);$$

in particular $\ln V_T \sim N(\ln V_0 + (\mu_V - \frac{1}{2}\sigma_V^2)T, \sigma_V^2 T)$. The default probability is thus

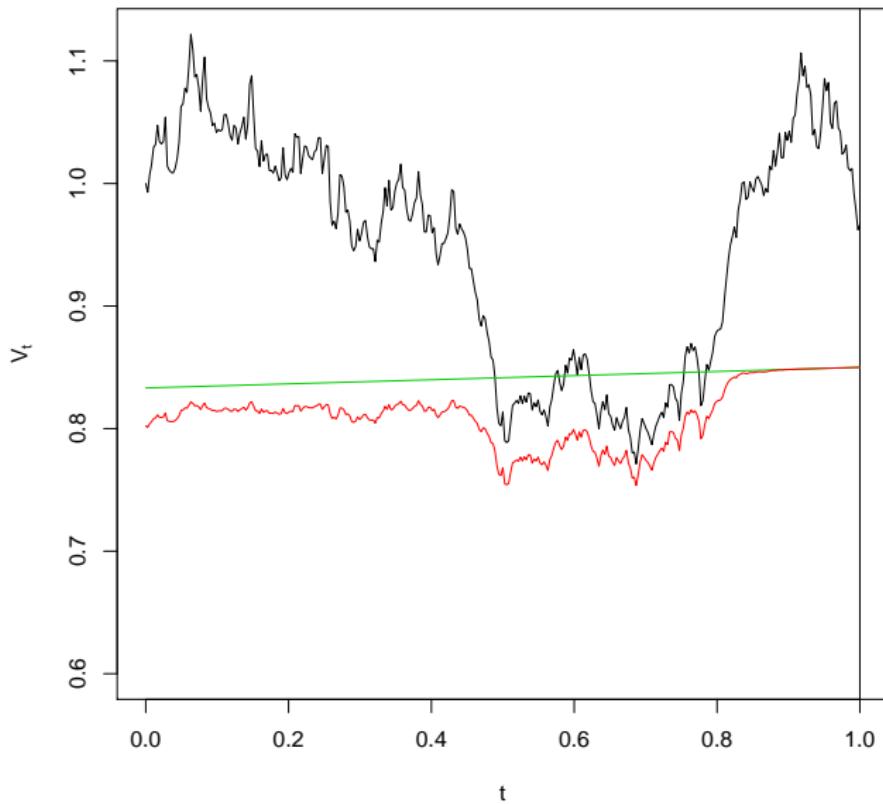
$$\mathbb{P}(V_T \leq B) = \mathbb{P}(\ln V_T \leq \ln B) = \Phi\left(\frac{\ln \frac{B}{V_0} - (\mu_V - \frac{1}{2}\sigma_V^2)T}{\sigma_V \sqrt{T}}\right); \quad (72)$$

it is increasing in B and σ_V (for $V_0 > B$) and decreasing in V_0 and μ_V .

A default path



A non-default path



10.3.2 Pricing in Merton's model

- Under some technical assumptions we can price equity and debt using the [Black–Scholes formula](#).
- The assumptions are that:
 - 1) The risk-free rate is deterministic and equal to $r \geq 0$.
 - 2) The asset-value process (V_t) is independent of the debt level B .
 - 3) The asset value (V_t) can be traded on a frictionless market.
- Recall that equity is a call option on the asset value (V_t). Hence Black–Scholes formula yields

$$S_t = C^{\text{BS}}(t, V_t; \sigma_V, r, T, B) := V_t \Phi(d_{t,1}) - B e^{-r(T-t)} \Phi(d_{t,2}),$$

where the arguments are given by

$$d_{t,1} = \frac{\ln \frac{V_t}{B} + (r + \frac{1}{2}\sigma_V^2)(T - t)}{\sigma_V \sqrt{T - t}}, \quad d_{t,2} = d_{t,1} - \sigma_V \sqrt{T - t}.$$

Pricing of debt

- The price at $t \leq T$ of a default-free zero-coupon bond with maturity T and a face value of one equals

$$p_0(t, T) = \exp(-r(T - t)).$$

- The value of the firm's debt equals the difference between the value of default-free debt and a put option on (V_t) with strike B , i.e.

$$B_t = B p_0(t, T) - P^{\text{BS}}(t, V_t; r, \sigma_V, B, T).$$

- The Black–Scholes formula for European puts now yields

$$B_t = p_0(t, T) B \Phi(d_{t,2}) + V_t \Phi(-d_{t,1}). \quad (73)$$

- The path of (B_t) is shown on the previous plots. The value of default-free debt $B p_0(t, T)$ is shown as a green curve.

Risk-neutral and physical default probabilities

- Under the risk-neutral measure \mathbb{Q} the process (V_t) satisfies the SDE $dV_t = rV_t dt + \sigma_V V_t d\tilde{W}_t$ for a standard \mathbb{Q} -Brownian motion \tilde{W} .
- The drift μ_V is replaced by the risk-free interest rate r .
- Hence the **risk-neutral default probability** is given by

$$q = \mathbb{Q}(V_T \leq B) = \Phi\left(\frac{\ln B - \ln V_0 - (r - \frac{1}{2}\sigma_V^2)T}{\sigma_V \sqrt{T}}\right).$$

- Comparison with **physical default probability** $p = \mathbb{P}(V_T \leq B)$ yields

$$q = \Phi\left(\Phi^{-1}(p) + \frac{\mu_V - r}{\sigma_V} \sqrt{T}\right). \quad (74)$$

- The correction term $(\mu_V - r)/\sigma_V$ equals the **Sharpe ratio** of the firm's assets (a popular measure of the risk premium earned by the firm).
- The formula is sometimes applied in practice to go from physical to risk-neutral default probabilities.

Credit spreads in Merton's model

- The credit spread measures the difference between the (continuously compounded) yield of a default-free zero coupon bond $p_0(t, T)$ and a defaultable zero coupon bond $p_1(t, T)$, i.e.

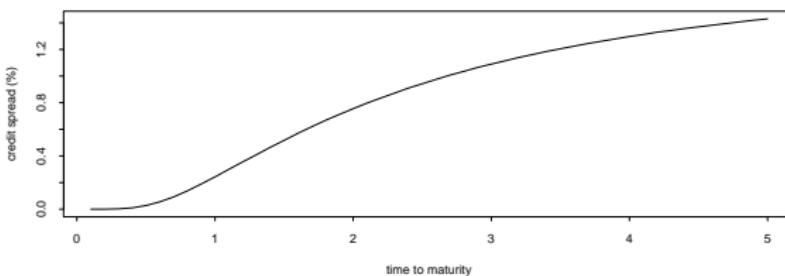
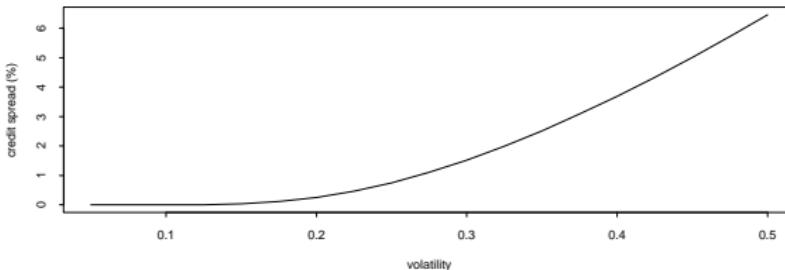
$$\begin{aligned} c(t, T) &= \frac{-1}{T-t} (\ln p_1(t, T) - \ln p_0(T-t)) \\ &= \frac{-1}{T-t} \ln \frac{p_1(t, T)}{p_0(t, T)}. \end{aligned}$$

- In Merton's model we have $p_1(t, T) = \frac{1}{B} B_t$ and hence

$$c(t, T) = \frac{-1}{(T-t)} \ln \left(\Phi(d_{t,2}) + \frac{V_t}{Bp_0(t, T)} \Phi(-d_{t,1}) \right). \quad (75)$$

- For a fixed time to maturity $c(t, T)$ depends only on σ_V and on the ratio $Bp_0(t, T)/V_t$ (a measure of indebtedness of the firm).
- In line with economic intuition it is increasing in both quantities.

Illustration of credit spreads in Merton's model



Credit spread $c(t, T)$ (%) as function of σ_V (top) and time to maturity $T - t$ (bottom) for fixed debt to firm value ratio 0.6. In upper picture $T - t = 2$; in lower picture $\sigma_V = 0.25$.

10.3.3 Structural models in practice: EDF and DD

- A number of industry models descend from the Merton model.
- An important example is the so-called [public-firm EDF model](#) that is maintained by Moody's Analytics.
- The methodology builds on earlier work by KMV (a private company named after its founders Kealhofer, McQuown and Vasicek) in the 1990s.
- Literature: Crosbie and Bohn (2002) and Sun et al. (2012).
- [Expected Default Frequency](#). The EDF is an estimate of the default probability of a given firm over a one-year horizon.
- Suppose we use Merton's model for a company issuing debt with face value B maturing at time $T = 1$. The analogous quantity would be

$$\text{EDF}_{\text{Merton}} = 1 - \Phi \left(\frac{\ln V_0 - \ln B + (\mu_V - \frac{1}{2}\sigma_V^2)}{\sigma_V} \right). \quad (76)$$

How Moody's adapt the Merton formula

- The decreasing function $1 - \Phi$ is replaced by an empirically estimated function.
- B is replaced by a new default threshold \tilde{B} representing the structure of the firm's liabilities more closely.
- The term $(\mu_V - \frac{1}{2}\sigma_V^2)$ in the numerator is usually omitted.
- The current asset value V_0 and the asset volatility σ_V are inferred or 'backed out' from information about the firm's equity value.
- Why?
 - ▶ In contrast to the assumptions underlying Merton's model, in most cases there is no market for the assets of a firm, so that the asset value is not directly observable.
 - ▶ The market value can differ widely from the value of a company as measured by accountancy rules (the so-called book value).

Inferred asset values in Merton's model

- Recall that in Merton's model we have that

$$S_t = C^{\text{BS}}(t, V_t; r, \sigma_V, B, T). \quad (77)$$

- We consider the debt structure (B and T) as well as the interest rate r to be known. Equity values (S_t) are observable.
- For fixed t , (77) is an equation with two unknowns, V_t and σ_V .
- To overcome this difficulty an iterative procedure is used.
- In step (1), an initial estimate $\sigma_V^{(0)}$ is used to infer a time series of asset values ($V_t^{(0)}$) from equity values (S_t).
- Then a new volatility estimate $\sigma_V^{(1)}$ is estimated from this time series.
- A new time series ($V_t^{(1)}$) is then constructed using (77) with $\sigma_V^{(1)}$.
- This procedure is iterated n -times, until the volatility estimates converge.

- The procedure in the public-firm EDF model is similar but a more sophisticated capital structure is assumed and the BS formula in (77) is replaced by a more complex formula.

EDF and DD

- In the public-firm EDF model a new state variable is introduced. This is the so-called **distance-to-default** (DD), given by

$$\text{DD} := (\log V_0 - \log \tilde{B})/\sigma_V . \quad (78)$$

- Here \tilde{B} represents the default threshold; in some versions of the model \tilde{B} is modelled as the sum of the liabilities payable within one year and half of the longer term debt.
- Note that (78) is in fact an approximation of the argument of (76), since μ_V and σ_V^2 are usually small.
- It is assumed that the distance-to-default **ranks** firms in the sense that firms with a higher DD exhibit a higher default probability.

- The functional relationship between DD and EDF is determined empirically; using a database of historical default events, the proportion of firms with DD in a given small range that default within a year is estimated. This proportion is the empirically estimated EDF.

Variable	J&J	RadioShack	Notes
Market value of assets V_0	\$236 bn	\$1834 m	Determined from time series of equity prices.
Asset volatility σ_V	11%	24%	
Default threshold \tilde{B}	\$39 bn	\$1042 m	Short-term liabilities and half of long-term liabilities.
DD	16.4	2.3	Given by $(\log V_0 - \log \tilde{B})/\sigma_V$.
EDF (one year)	0.01%	3.58%	Using empirical mapping between DD and EDF.

The example is taken from Sun et al. (2012); it is concerned with the situation of Johnson and Johnson (J&J) and RadioShack as of April 2012.

10.3.4 Credit migration models revisited

- In a credit migration model, consider a firm rated j at $t = 0$ with transition probabilities p_{jk} , $0 \leq k \leq n$ for the period $[0, T]$.
- Suppose that the asset-value process (V_t) of the firm follows the Merton diffusion model so that

$$V_T = V_0 \exp((\mu_V - \frac{1}{2}\sigma_V^2)T + \sigma_V W_T) \quad (79)$$

is lognormally distributed.

- We can choose thresholds $0 = \tilde{d}_0 < \tilde{d}_1 < \dots < \tilde{d}_n < \tilde{d}_{n+1} = \infty$ such that $\mathbb{P}(\tilde{d}_k < V_T \leq \tilde{d}_{k+1}) = p_{jk}$ for $k \in \{0, \dots, n\}$.
- Thus we have translated the transition probabilities into a series of thresholds for an assumed asset-value process.
- The threshold \tilde{d}_1 is the default threshold, often interpreted as the value of the firm's liabilities.

- The higher thresholds are the asset-value levels that mark the boundaries of higher rating categories.
- The firm-value model can be summarized by saying that the firm belongs to **rating class k** at the time horizon T if and only if $\tilde{d}_k < V_T \leq \tilde{d}_{k+1}$.
- The migration probabilities remain invariant under **simultaneous strictly increasing transformations** of V_T and the thresholds \tilde{d}_j .
- If we define

$$X_T := \frac{\ln V_T - \ln V_0 - (\mu_V - \frac{1}{2}\sigma_V^2)T}{\sigma_V \sqrt{T}}, \quad (80)$$

$$d_k := \frac{\ln \tilde{d}_k - \ln V_0 - (\mu_V - \frac{1}{2}\sigma_V^2)T}{\sigma_V \sqrt{T}}, \quad (81)$$

then we can also say that the firm belongs to **rating class k** at the time horizon T if and only if $d_k < X_T \leq d_{k+1}$.

- Observe that X_T is a standardized version of the **asset-value log-return** $\ln V_T - \ln V_0$ and we can easily verify that $X_T = W_T / \sqrt{T} \sim N(0, 1)$.

- In this case the formulas for the thresholds are easily obtained and are $d_k = \Phi^{-1}(\sum_{l=0}^{k-1} p_{jl})$ for $k = 1, \dots, n$.

Credit migrations and public-firm EDFs compared

Advantages of EDFs.

- 1) The EDF reacts quickly to changes in the economic prospects of a firm, whereas agencies are often slow to adjust ratings.
- 2) EDFs tend to reflect the current macroeconomic environment and tend to be better predictors of default over short time horizons.

Advantages of credit migration approach.

- 1) The EDF approach is sensitive to over- and under-reactions in equity markets. If widely followed this might have destabilizing effects.
- 2) As rating agencies focus on average credit quality “through the business cycle”, risk capital requirements based on rating transitions fluctuate less, helping to provide liquidity in credit markets.

10.4 Bond and CDS pricing in hazard rate models

10.4.1 Hazard rate models

- These are the simplest **reduced-form** credit risk models.
- A hazard rate model is a model in which the distribution of the default time of an obligor is directly specified by a hazard function without modelling the mechanism by which default occurs.
- To set up a hazard rate model we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a **random default time** τ defined on this space, i.e. an \mathcal{F} -measurable rv taking values in $[0, \infty]$.
- We denote the df of τ by $F(t) = \mathbb{P}(\tau \leq t)$ and the tail or survival function by $\bar{F}(t) = 1 - F(t)$; we assume that $\mathbb{P}(\tau = 0) = F(0) = 0$, and that $\bar{F}(t) > 0$ for all $t < \infty$.

- The jump or default indicator process (Y_t) associated with τ is

$$Y_t = I_{\{\tau \leq t\}}, \quad t \geq 0. \quad (82)$$

- (Y_t) is a right-continuous process which jumps from 0 to 1 at the default time τ .
- $1 - Y_t = I_{\{\tau > t\}}$ is the survival indicator of the firm at time t .

Definition 10.1 (cumulative hazard and hazard function)

The function $\Gamma(t) = -\ln(\bar{F}(t))$ is called the cumulative hazard function of the random time τ . If F is absolutely continuous with density f , the function

$$\gamma(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{\bar{F}(t)} = -\frac{d}{dt} \ln(\bar{F}(t))$$

is called the hazard function of τ .

- The hazard function $\gamma(t)$ gives the hazard rate at t , which is a measure of the instantaneous risk of default at t , given survival up to time t .

- We can represent the survival function of τ by

$$\bar{F}(t) = \exp\left(-\int_0^t \gamma(s) \, ds\right). \quad (83)$$

- We may show that

$$\lim_{h \rightarrow 0} \frac{1}{h} \mathbb{P}(\tau \leq t + h \mid \tau > t) = \frac{1}{\bar{F}(t)} \lim_{h \rightarrow 0} \frac{F(t+h) - F(t)}{h} = \gamma(t).$$

Example 10.2 (Weibull distribution)

For illustrative purposes we determine the hazard function for the Weibull distribution with df $F(t) = 1 - \exp(-\lambda t^\alpha)$ for parameters $\lambda, \alpha > 0$. Differentiation yields

$$f(t) = \lambda \alpha t^{\alpha-1} \exp(-\lambda t^\alpha) \text{ and } \gamma(t) = \lambda \alpha t^{\alpha-1}.$$

In particular, γ is decreasing in t if $\alpha < 1$ and increasing if $\alpha > 1$. For $\alpha = 1$ (exponential distribution) the hazard rate equals the constant λ .

Introducing filtrations

- Filtrations model information available to investors over time.
- A **filtration** (\mathcal{F}_t) on (Ω, \mathcal{F}) is an increasing family $\{\mathcal{F}_t : t \geq 0\}$ of sub- σ -algebras of $\mathcal{F} : \mathcal{F}_t \subset \mathcal{F}_s \subset \mathcal{F}$ for $0 \leq t \leq s < \infty$.
- \mathcal{F}_t represents the state of knowledge of an observer at time t . $A \in \mathcal{F}_t$ means that at time t we can determine if A has occurred.
- In this section we assume that only observable quantity is the default indicator (Y_t) associated with τ . The appropriate filtration is (\mathcal{H}_t) with

$$\mathcal{H}_t = \sigma(\{Y_u : u \leq t\}), \quad (84)$$

the **default history** up to and including time t .

- τ is a **(\mathcal{H}_t) -stopping time**, since $\{\tau \leq t\} = \{Y_t = 1\} \in \mathcal{H}_t$ for all $t \geq 0$.
- In order to study bond and CDS pricing in hazard rate models we need to compute conditional expectations with respect to the σ -algebra \mathcal{H}_t .

A useful result

Lemma 10.3

Let τ be a default time with jump indicator process $Y_t = I_{\{\tau \leq t\}}$ and natural filtration (\mathcal{H}_t) . Then, for any integrable rv X and any $t \geq 0$, we have

$$\mathbb{E}(I_{\{\tau > t\}} X \mid \mathcal{H}_t) = I_{\{\tau > t\}} \frac{\mathbb{E}(I_{\{\tau > t\}} X)}{\mathbb{P}(\tau > t)}. \quad (85)$$

This result can be used to determine conditional survival probabilities. For $t < T$, applying (85) with $X := I_{\{\tau > T\}}$ we get

$$\mathbb{P}(\tau > T \mid \mathcal{H}_t) = I_{\{\tau > t\}} \exp \left(- \int_t^T \gamma(s) \, ds \right), \quad t < T. \quad (86)$$

Martingale property of jump indicator process

Proposition 10.4

The process (M_t) defined as

$$M_t = Y_t - \int_0^t I_{\{\tau > u\}} \gamma(u) du, \quad t \geq 0$$

is an (\mathcal{H}_t) -martingale, that is $\mathbb{E}(M_s | \mathcal{H}_t) = M_t$ for all $0 \leq t \leq s < \infty$.

10.4.2 Risk-neutral pricing revisited

- According to the first fundamental theorem of asset pricing, a model for security prices is arbitrage free if and (essentially) only if it admits at least one equivalent martingale measure \mathbb{Q} .
- When building a model for pricing derivatives it is a natural shortcut to model the objects of interest—such as interest rates and default times—directly, under a martingale measure \mathbb{Q} .

Martingale modelling

- So-called **martingale modelling** is particularly convenient if the value H of the underlying assets at some maturity date T is exogenously given, as in the case of zero-coupon bonds.
- The underlying asset at time $t < T$ can be computed as the conditional expectation under \mathbb{Q} of the discounted value at maturity via the risk-neutral pricing rule

$$V_t = \mathbb{E}^{\mathbb{Q}}(e^{-\int_t^T r_s ds} H \mid \mathcal{F}_t). \quad (87)$$

- Model parameters are determined using the requirement that at time $t = 0$ the model price should coincide with the market price of the security; this is known as **calibration** to market data.

Pros and cons of Martingale modelling

- Martingale modelling ensures that the resulting model is arbitrage free, which is important for pricing many different securities simultaneously.
- The approach is frequently adopted in default-free term structure models and in reduced-form models for credit-risky securities.
- Martingale modelling is however problematic if the underlying market is incomplete (meaning that not all risk cannot be hedged away). In practice martingale modelling is best applied in situations where many liquidly traded derivatives are available.

10.4.3 Bond pricing

- It suffices to consider zero-coupon bonds.
- We use martingale modelling and work directly under some martingale measure \mathbb{Q} .

- We assume that under \mathbb{Q} the default time τ is a random time with deterministic risk-neutral hazard function $\gamma^{\mathbb{Q}}(t)$.
- The information available to investors at time t is given by the sigma algebra $\mathcal{H}_t = \sigma(\{Y_u : u \leq t\})$.
- We take **interest rates** and **recovery rates** to be **deterministic**.
- The percentage loss given default is denoted by $\delta \in (0, 1)$.
- The continuously compounded interest rate is denoted by $r(t) \geq 0$.
- The price of the default-free zero-coupon bond with maturity $T \geq t$ is $p_0(t, T) = \exp(-\int_t^T r(s) ds)$.

Analysing the payments

- The payments of a defaultable zero-coupon bond can be represented as a combination of a **survival claim** that pays one unit at the maturity date T and a **recovery payment** in case of default.
- The survival claim has pay-off $I_{\{\tau > T\}}$.
- Recall from (86) that

$$\mathbb{Q}(\tau > T \mid \mathcal{H}_t) = I_{\{\tau > t\}} \exp \left(- \int_t^T \gamma^{\mathbb{Q}}(s) ds \right)$$

and define $R(t) = r(t) + \gamma^{\mathbb{Q}}(t)$.

- Then the price of a survival claim at time t equals

$$\begin{aligned}\mathbb{E}^{\mathbb{Q}}(p_0(t, T) I_{\{\tau > T\}} \mid \mathcal{H}_t) &= \exp \left(- \int_t^T r(s) ds \right) \mathbb{Q}(\tau > T \mid \mathcal{H}_t) \\ &= I_{\{\tau > t\}} \exp \left(- \int_t^T R(s) ds \right).\end{aligned}\quad (88)$$

- Note that for $\tau > t$, this can be viewed as the price of a default-free zero-coupon bond with adjusted interest rate $R(t) > r(t)$.
- A similar relationship between defaultable and default-free bond prices can be established in many reduced-form credit risk models.

Recovery models

1) Recovery of Treasury (RT).

- The RT model was proposed by Jarrow and Turnbull (1995).
- If default occurs at some point in time $\tau \leq T$, the owner of the defaulted bond receives $(1 - \delta_\tau)$ units of the default-free zero-coupon bond $p_0(\cdot, T)$ at time τ , where $\delta_\tau \in [0, 1]$ models the percentage loss given default.
- At maturity T the holder of the defaultable bond therefore receives the payment $I_{\{\tau > T\}} + (1 - \delta_\tau)I_{\{\tau \leq T\}}$.

2) Recovery of Face Value (RF).

- Under RF, if default occurs at $\tau \leq T$, the holder of the bond receives a recovery payment of size $(1 - \delta_\tau)$ immediately at the default time τ .
- Note that even with deterministic loss given default and deterministic interest rates, the **value at maturity of the recovery payment is random** as it depends on the exact timing of default.

RF is slightly more realistic; RT is slightly easier to analyse.

Pricing recovery payment under RT

- The value of the recovery payment at the maturity date T is

$$(1 - \delta)I_{\{\tau \leq T\}} = (1 - \delta) - (1 - \delta)I_{\{\tau > T\}}.$$

- Using (88), the value of the recovery payment at time $t < T$ is hence

$$(1 - \delta)p_0(t, T) - (1 - \delta)I_{\{\tau > t\}} \exp \left(- \int_t^T R(s) \, ds \right).$$

- Hence the value of the bond is

$$p_1(t, T) = (1 - \delta)p_0(t, T) + \delta I_{\{\tau > t\}} \exp \left(- \int_t^T R(s) \, ds \right).$$

Pricing recovery payment under RF

- Under the RF-hypothesis the recovery payment takes the form $(1 - \delta)I_{\{\tau \leq T\}}$ where the payment occurs directly at time τ .
- A payments of this form is a [payment-at-default claim](#).
- The value of the recovery payment at time $t \leq T$ equals

$$\mathbb{E}^{\mathbb{Q}} \left((1 - \delta) I_{\{t < \tau \leq T\}} \exp \left(- \int_t^\tau r(s) \, ds \right) \middle| \mathcal{H}_t \right).$$

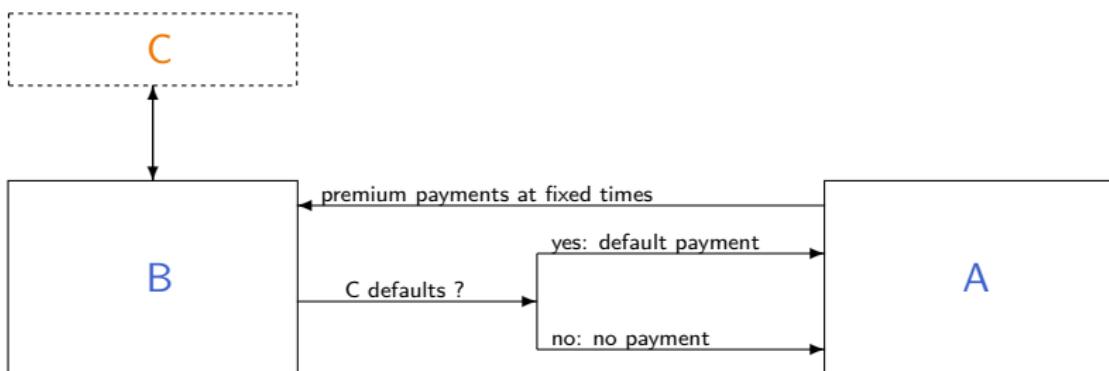
- Using (85) we may show that

$$\begin{aligned} & \mathbb{E}^{\mathbb{Q}} \left((1 - \delta) I_{\{t < \tau \leq T\}} \exp \left(- \int_t^\tau r(s) \, ds \right) \middle| \mathcal{H}_t \right) \\ &= (1 - \delta) I_{\{\tau > t\}} \int_t^T \gamma^{\mathbb{Q}}(s) \exp \left(- \int_t^s R(u) \, du \right) \, ds. \end{aligned}$$

10.4.4 CDS pricing

Recap: Structure of CDS

- C (reference entity); default at time $\tau_C < T$ triggers default payment.
- A (protection buyer); pays premiums to B until $\min(\tau_C, T)$.
- B (protection seller); makes default payment to A if $\tau_C < T$.



Payment flows

For simplicity write $\tau = \tau_C$ and consider the following contract:

- **Premium payments.**
 - ▶ These are due at times $0 < t_1 < \dots < t_N$ measured in years.
 - ▶ If $\tau > t_k$, A pays a premium of size $x^*(t_k - t_{k-1})$ at t_k , where x^* denotes the **fair swap spread**.
 - ▶ After τ premium payments stop.
 - ▶ No initial payment.
- **Default payment.**
 - ▶ If $\tau < t_N = T$, B makes a default payment δ at τ .
 - ▶ Sometimes B receives an **accrued premium payment** of size $x^*(\tau - t_k)$ for $\tau \in (t_k, t_{k-1})$. We ignore this feature for simplicity.

Valuing the premium leg

- The premium leg consists of a set of **survival claims**.
- Introduce a function of x given by

$$\begin{aligned}V_t^{\text{prem}}(x; \gamma^{\mathbb{Q}}) &= \mathbb{E}^{\mathbb{Q}} \left(\sum_{k: t_k > t} \exp \left(- \int_t^{t_k} r(u) \, du \right) x(t_k - t_{k-1}) I_{\{\tau > t_k\}} \mid \mathcal{H}_t \right) \\&= x \sum_{k: t_k > t} p_0(t, t_k) (t_k - t_{k-1}) \mathbb{Q}(\tau > t_k \mid \mathcal{H}_t),\end{aligned}$$

which is easily computed using $\mathbb{Q}(\tau > t_k \mid \mathcal{H}_t) = \exp(-\int_t^{t_k} \gamma^{\mathbb{Q}}(s) \, ds)$.

- We obtain

$$V_t^{\text{prem}}(x; \gamma^{\mathbb{Q}}) = I_{\{\tau > t\}} x \sum_{k: t_k > t} (t_k - t_{k-1}) \exp \left(- \int_t^{t_k} R(u) \, du \right).$$

Valuing the default leg

- The default payment leg is a typical payment-at-default claim.
- We obtain

$$\begin{aligned}V_t^{\text{def}}(\gamma^{\mathbb{Q}}) \\= \mathbb{E}^{\mathbb{Q}}\left(\delta I_{\{t < \tau \leq t_N\}} \exp\left(-\int_t^\tau r(s) \, ds\right) \mid \mathcal{H}_t\right) \\= I_{\{\tau > t\}} \delta \int_t^{t_N} \gamma^{\mathbb{Q}}(s) \exp\left(-\int_t^s R(u) \, du\right) \, ds.\end{aligned}$$

The fair CDS spread

- The fair CDS spread x_t^* quoted for the contract at time t is chosen such that the value of the contract is equal to zero.

- The equation $V_t^{\text{prem}}(x_t^*; \gamma^{\mathbb{Q}}) = V_t^{\text{def}}(\gamma^{\mathbb{Q}})$ yields

$$x_t^* = I_{\{\tau > t\}} \frac{\delta \int_t^{t_N} \gamma^{\mathbb{Q}}(s) \exp\left(-\int_t^s R(u) \, du\right) \, ds}{\sum_{k: t_k > t} (t_k - t_{k-1}) \exp\left(-\int_t^{t_k} R(s) \, ds\right)}. \quad (89)$$

Model calibration

- We have to calibrate our model to the available market information. Hence we have to determine the implied risk-neutral hazard function $\gamma^{\mathbb{Q}}(t)$, which ensures that the fair CDS spreads implied by the model equal the spreads quoted in the market.
- Suppose that the market information at time $t = 0$ consists of the fair spread x^* of one CDS with maturity t_N .
- In that case $\gamma^{\mathbb{Q}}(s)$ is taken constant: for all $s \geq 0$, $\gamma^{\mathbb{Q}}(s) = \bar{\gamma}^{\mathbb{Q}}$ for some $\bar{\gamma}^{\mathbb{Q}} > 0$.

- $\bar{\gamma}^{\mathbb{Q}}$ has to solve the equation

$$x^* \sum_{k=1}^N p_0(0, t_k) (t_k - t_{k-1}) e^{-\bar{\gamma}^{\mathbb{Q}} t_k} = \delta \bar{\gamma}^{\mathbb{Q}} \int_0^{t_N} p_0(0, t) e^{-\bar{\gamma}^{\mathbb{Q}} t} dt.$$

- There is a unique solution.
- If we observe spreads for several CDSs on the same reference entity but with different maturities, a constant function is not sufficient. Instead one typically uses piecewise constant or linear hazard functions.
- An exception occurs in the special case where: (1) the spread curve is flat (i.e. all CDSs on the reference entity have the same spread x^* , independent of the maturity); (2) the risk-free interest rate is constant; (3) the time points t_k are equally spaced ($t_k - t_{k-1} = \Delta t$ for all k).
- In that case the implied risk-neutral hazard rate $\bar{\gamma}^{\mathbb{Q}}$ is the solution of

$$x^* \Delta t p_0(0, \Delta t) e^{-\bar{\gamma}^{\mathbb{Q}} \Delta t} = \delta \bar{\gamma}^{\mathbb{Q}} \int_0^{\Delta t} e^{-rt} e^{-\bar{\gamma}^{\mathbb{Q}} t} dt. \quad (90)$$

- For Δt relatively small (quarterly or semi-annual spread payments) a good approximation to the solution of (90) is given by $\bar{\gamma}^Q \approx x^*/\delta$.
- This approximation is frequently used in practice and implies that the one-year default probability satisfies $Q(\tau \leq 1) = 1 - e^{-\bar{\gamma}^Q} \approx \bar{\gamma}^Q \approx x^*/\delta$.

Shortcomings of simple hazard rate models

- In the models of this section the only risk factor affecting a defaultable bond or CDS is default risk.
- In these models credit spreads evolve deterministically prior to default, which is unrealistic.
- The models are not sophisticated enough to price options on defaultable bonds or CDSs.
- To obtain more realistic models we can replace the deterministic hazard functions by stochastic hazard processes.

- This means that default times are modelled as so-called **doubly-stochastic random times**.
- We might also consider adding **stochastic interest rate models**; and **more complex assumptions on recoveries** in the event of default.

10.4.5 P versus Q : Empirical results

- There are some empirical studies of the relationship between physical and risk-neutral default probabilities.
- Risk-neutral default probabilities are generally estimated from CDS spreads. These can be compared, for example, with EDFs.
- Berndt et al. (2008) compare five-year CDSs against five-year EDFs for a large pool of firms. The five-year EDF is an annualized estimate of the physical five-year default probability.

- Let $x_{t,i}^*$ and $\text{EDF}_{t,i}$ denote the CDS spread and five-year EDF of firm i at date t . Their (most basic) model took the form

$$x_{t,i}^* = \alpha + \beta \text{EDF}_{t,i} + \varepsilon_{t,i},$$

with estimates $\alpha = 33\text{bp}$ and $\beta = 1.6$; the R^2 was 0.73.

- More crudely $x_{t,i}^*/\text{EDF}_{t,i} \approx 1.6$.
- Using $q_{t,i} = x_{t,i}^*/\delta$ as a proxy for the risk-neutral default probability yields

$$\frac{q_{t,i}}{p_{t,i}} \approx \frac{x_{t,i}^*}{\delta \text{EDF}_{t,i}} \approx 1.6 \delta^{-1}.$$

10.5 Pricing with stochastic hazard rates

Why stochastic hazard rates?

- In hazard rate models the only risk factor is default risk \Rightarrow Credit spreads evolve deterministically prior to default, which is clearly unrealistic.
- Moreover, it is not possible to price options on bonds or CDSs or to do risk management for bond portfolios in such models.
- Hence it is of interest to consider models where hazard rate is a stochastic process $(\gamma_t)_{t \geq 0}$; typically hazard rate is driven by a second stochastic process Ψ , that is $\gamma_t = \gamma(\Psi_t)$.
- Simplest such model class are doubly-stochastic random times.

10.5.1 Doubly stochastic random times

Setup. We work on $(\Omega, \mathcal{F}, \mathbb{P})$ with **background filtration** (\mathcal{F}_t) containing information about all other economic events except the default event.

Consider a random time τ , that is a measurable rv with values in $(0, \infty)$.

- $Y_t = I_{\{\tau \leq t\}}$ is the associated **default indicator** and $(\mathcal{H}_t) = \sigma\{Y_s, s \leq t\}$ is the **default history** up to t .
- Define new filtration $\mathcal{G}_t = \mathcal{F}_t \vee \mathcal{H}_t$, $t \geq 0$, i.e. \mathcal{G}_t contains background info \mathcal{F}_t and default history up to t (this is the information available to investors).

Definition 10.5

τ is called *doubly stochastic* if there is a positive (\mathcal{F}_t) -adapted process (γ_t) (the hazard rate process) such that for all $t \geq 0$

$$\mathbb{P}(\tau > t | \mathcal{F}^\infty) = \exp\left(-\int_0^t \gamma_s \, ds\right). \quad (91)$$

Comments.

- Here $\mathcal{F}_\infty = \sigma(\bigcup_{t \geq 0} \mathcal{F}_t)$. Conditioning on \mathcal{F}_∞ thus means that we know the past and future economic environment and in particular the entire trajectory $(\gamma_s(\omega))_{s \geq 0}$ of the hazard rate.
- Relation (91) implies that, given the economic environment \mathcal{F}^∞ , τ is a random time with deterministic hazard function $s \mapsto \gamma_s(\omega)$.
- In the literature doubly stochastic random times are also known as [conditional Poisson](#) or [Cox](#) random times.

Sampling doubly stochastic random times

A simple algorithm is based on the following result

Lemma 10.6

Let E be a standard exponentially distributed rv independent of \mathcal{F}^∞ , that is $\mathbb{P}(\tau > t | \mathcal{F}^\infty) = e^{-t}$. Let (γ_t) be a positive \mathcal{F}_t -adapted process with $\int_0^t \gamma_s \, ds < \infty$ for all t . Define τ by

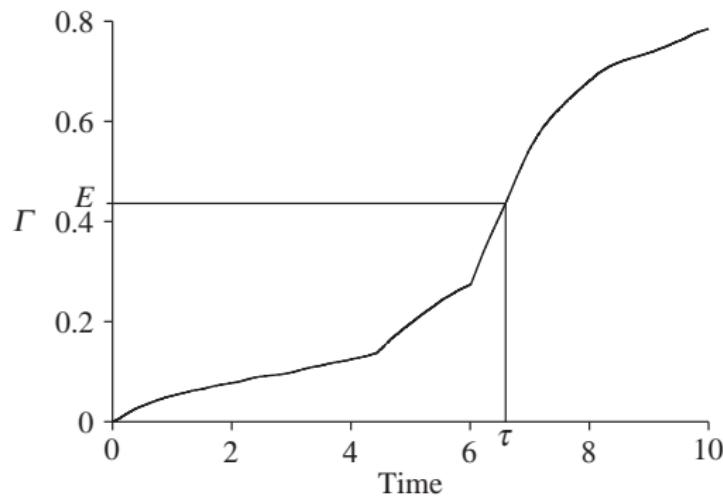
$$\tau := \inf \left\{ t \geq 0 : \int_0^t \gamma_s \, ds \geq E \right\}. \quad (92)$$

Then τ is doubly stochastic with hazard-rate process (γ_t) .

Algorithm. (threshold simulation)

- 1) Generate $E \sim \text{Exp}(1)$.
- 2) Generate a trajectory $(\gamma_s)_{s=0}^\infty$ of hazard rate process.
- 3) Return $\tau := \inf \{t \geq 0 : \int_0^t \gamma_s \, ds \geq E\}$

Graphical illustration.



A graphical illustration of threshold simulation; $E \approx 0.44$, $\tau \approx 6.59$.

Intensities

Definition 10.7

Consider a filtration (\mathcal{G}_t) and a random time τ with (\mathcal{G}_t) -adapted jump indicator process (Y_t) . A non-negative (\mathcal{G}_t) -adapted process (λ_t) is called (\mathcal{G}_t) -**(default) intensity** of the random time τ if $M_t := Y_t - \int_0^{t \wedge \tau} \lambda_s ds$ is a (\mathcal{G}_t) -martingale.

The next result extends Proposition 10.4 to doubly stochastic τ .

Proposition 10.8

Let τ be a doubly stochastic random time with (\mathcal{F}_t) -conditional hazard rate process (γ_t) . Then $M_t := Y_t - \int_0^{t \wedge \tau} \gamma_s ds$ is a (\mathcal{G}_t) -martingale, that is the hazard rate γ_t is the (\mathcal{G}_t) default intensity.

Conditional expectations

Conditional expectations wrt \mathcal{G}_t are crucial for pricing formulas.

Proposition 10.9 (Dellacherie formula)

Let τ be an arbitrary random time (not necessarily doubly stochastic) such that $\mathbb{P}(\tau > t \mid \mathcal{F}_t) > 0$ for all $t \geq 0$. Then we have for every integrable rv X that

$$\mathbb{E}(I_{\{\tau>t\}}X \mid \mathcal{G}_t) = I_{\{\tau>t\}} \frac{\mathbb{E}(I_{\{\tau>t\}}X \mid \mathcal{F}_t)}{\mathbb{P}(\tau > t \mid \mathcal{F}_t)}.$$

Corollary 10.10

Let $T > t$ and assume that τ is doubly-stochastic with hazard rate process (γ_t) . If \tilde{X} is integrable and \mathcal{F}_T -measurable, we have

$$\mathbb{E}(I_{\{\tau>T\}}\tilde{X} \mid \mathcal{G}_t) = I_{\{\tau>t\}}\mathbb{E}(e^{-\int_t^T \gamma_s ds}\tilde{X} \mid \mathcal{F}_t).$$

Application: 1-year default probabilities

γ_t gives good approximation to the one-year default probability: Let $T = t + 1$ and $\tilde{X} = 1$ to obtain

$$\mathbb{P}(\tau > t + 1 \mid \mathcal{G}_t) = I_{\{\tau > t\}} \mathbb{E} \left(\exp \left(- \int_t^{t+1} \gamma_s \, ds \right) \mid \mathcal{F}_t \right). \quad (93)$$

For $\tau > t$ and a fairly stable hazard rate over the time interval $[t, t + 1]$ the right-hand side of (93) is $\approx \exp(-\gamma_t)$ and for γ_t small,

$$\mathbb{P}(\tau \leq t + 1 \mid \mathcal{G}_t) \approx 1 - \exp(-\gamma_t) \approx \gamma_t. \quad (94)$$

10.5.2 Pricing formulas

Setup.

- Consider arbitrage-free security market model on $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{Q})$ where \mathbb{Q} is equivalent martingale measure. Prices of default-free securities (\mathcal{F}_t) -adapted; $B_t = \exp(\int_0^t r_s \, ds)$ models default-free savings account.
- Let τ be the default time of some company. As before we set $\mathcal{H}_t = \sigma(\{Y_s : s \leq t\})$ and $\mathcal{G}_t = \mathcal{F}_t \vee \mathcal{H}_t$; this is the information available to investors at time t .
- We use martingale-modelling. Hence price at t of an \mathcal{G}_T -measurable contingent claim H is given by

$$H_t = \mathbb{E}^{\mathbb{Q}} \left(\exp \left(- \int_t^T r_s \, ds \right) H \mid \mathcal{G}_t \right). \quad (95)$$

- Under \mathbb{Q} , τ is a doubly stochastic random time with background filtration (\mathcal{F}_t) and hazard rate process (γ_t) .

Key building blocks

The pricing of bonds and CDSs can be reduced to the pricing of the following building blocks:

- A **survival claim**, i.e. a promised \mathcal{F}_T -measurable payment X which is made at time T if there is no default; the actual payment of the survival claim equals $XI_{\{\tau>T\}}$.
- A **payment-at-default claim** of the form $Z_\tau I_{\{\tau \leq T\}}$, where $Z = (Z_t)_{t \geq 0}$ is an (\mathcal{F}_t) adapted stochastic process and where Z_τ is short for $Z_{\tau(\omega)}(\omega)$. Note that the payment is made directly at τ , provided that $\tau \leq T$ where T is the maturity date of the claim.

Example. Defaultable bond is a combination of a survival claim and a payment at default claim (the recovery payment).

Pricing the building blocks

Next result shows that pricing of building blocks can be reduced to pricing problem for default-free claims with adjusted interest rate.

Theorem 10.11

Define adjusted interest rate $R_t = r_t + \gamma_t$. Under the above assumptions (in particular for τ doubly stochastic) it holds that

$$\mathbb{E}^{\mathbb{Q}}\left(\exp\left(-\int_t^T r_s \, ds\right) I_{\{\tau > T\}} X \mid \mathcal{G}_t\right) = I_{\{\tau > t\}} \mathbb{E}^{\mathbb{Q}}\left(\exp\left(-\int_t^T R_s \, ds\right) X \mid \mathcal{F}_t\right),$$

$$\begin{aligned} & \mathbb{E}^{\mathbb{Q}}\left(I_{\{t < \tau \leq T\}} \exp\left(-\int_t^\tau r_s \, ds\right) Z_\tau \mid \mathcal{G}_t\right) \\ &= I_{\{\tau > t\}} \mathbb{E}^{\mathbb{Q}}\left(\int_t^T Z_s \gamma_s \exp\left(-\int_t^s R_u \, du\right) \, ds \mid \mathcal{F}_t\right). \end{aligned}$$

10.5.3 Applications

Corporate bonds and RF. The price at t of a defaultable zero-coupon bond with maturity $T \geq t$ is $p_1(t, T)$; price of corresponding default-free bond is $p_0(t, T)$.

Recall the following recovery models

- 1) **Recovery of Treasury (RT).** Under RT, if default occurs at $\tau \leq T$, the bond holder receives $(1 - \delta_\tau)$ units of $p_0(\cdot, T)$ at time τ , where $\delta \in [0, 1]$ models the percentage loss given default (LGD). Under RT the holder of the defaultable bond therefore receives the payment

$$I_{\{\tau > T\}} + (1 - \delta)I_{\{\tau \leq T\}} = 1 - \delta + \delta I_{\{\tau > T\}}.$$

- 2) **Recovery of Face Value (RF).** Under RF, if default occurs at $\tau \leq T$, the bondholder receives $(1 - \delta_\tau)$ immediately at τ . \Rightarrow Value of the recovery payment depends on the timing of default.

- 3) Recovery of market value (RM). Duffie and Singleton (1999) Under RM recovery payment equals $(1 - \delta_\tau)V_\tau I_{\{\tau \leq T\}}$, where $(\delta_t) \in (0, 1)$ gives the percentage LGDI and where the (\mathcal{F}_t) -adapted process (V_t) gives the pre-default value of the claim. This is a recursive definition (but explicit solution exists)

Application to corporate bonds.

- 1) Under RT the bond price in $t < T$ is

$$p_1(t, T) = (1 - \delta)p_0(t, T) + I_{\{\tau > t\}}\delta\mathbb{E}^{\mathbb{Q}}\left(\exp\left(-\int_t^T R_s \, ds\right) \middle| \mathcal{F}_t\right)$$

- 2) Under RF the bond is the sum of a survival claim and a payment at

default claim. One has

$$\begin{aligned} p_1(t, T) &= I_{\{\tau > t\}} \left(\mathbb{E}^{\mathbb{Q}} \left(\exp \left(- \int_t^T R_s \, ds \right) \mid \mathcal{F}_t \right) \right. \\ &\quad \left. + (1 - \delta) \mathbb{E}^{\mathbb{Q}} \left(\int_t^T \gamma_s \exp \left(- \int_t^s R_u \, du \right) \, ds \mid \mathcal{F}_t \right) \right) \end{aligned}$$

3) Under RM assumption one has

Proposition 10.12

Suppose that, under \mathbb{Q} , τ is doubly stochastic with hazard rate process (γ_t) . Then under RM the pre-default value (V_t) of a corporate bond is uniquely determined and given by

$$V_t = \mathbb{E}^{\mathbb{Q}} \left(\exp \left(- \int_t^T (r_s + \delta_s \gamma_s) \, ds \right) \mid \mathcal{F}_t \right), \quad 0 \leq t \leq T. \quad (96)$$

Special cases: $\delta = 1 \Rightarrow$ standard survival claim; $\delta = 0 \Rightarrow$ default-free.

Credit spreads

With doubly stochastic default times hazard rate process (γ_t) and credit spread $c(t, T) = -\frac{1}{T-t}(\ln p_1(t, T) - \ln p_0(t, T))$ of defaultable bonds are closely related. Analytic results for the instantaneous credit spread

$$c(t, t) = \lim_{T \rightarrow t} c(t, T) = -\frac{\partial}{\partial T} \Big|_{T=t} (\ln p_1(t, T) - \ln p_0(t, T)). \quad (97)$$

Proposition 10.13

In all recovery models $c(t, t) = \delta \gamma_t^{\mathbb{Q}}$.

- Instantaneous credit spreads are product of LGD and instantaneous default probability.
- In hazard-rate models short term spreads strictly positive.
- For $T > t$ spreads in the three models differ.

CDS contracts

Here premium payments constitute a sequence of survival claims; default payment is a payment-at-default claim. This gives the following formula for the fair CDS spread x^* at t :

$$x^* = \frac{\delta \mathbb{E}^{\mathbb{Q}}\left(\int_t^T \gamma_s e^{-\int_t^s R_u du} ds \mid \mathcal{F}_t\right)}{\sum_{T_k > t} (t_k - t_{k-1}) \mathbb{E}^{\mathbb{Q}}\left(\exp\left(-\int_t^{t_k} R_s ds\right) \mid \mathcal{F}_t\right)}. \quad (98)$$

- For $T \rightarrow t$ we get that x^* converges to $\delta \gamma_t$.
- The formula is a generalization of (89).

10.6 Affine models

In most models with doubly stochastic default time used in practice it is assumed that (r_t) and (γ_t) are functions of some **Markov process** (Ψ_t) on $D \subset \mathbb{R}^p$.

- Natural background filtration is $(\mathcal{F}_t) = \sigma(\{\Psi_s : s \leq t\})$.
- $R_t := r_t + \gamma_t$ is of the form $R_t = R(\Psi_t)$ for some $R : D \subseteq \mathbb{R}^p \rightarrow \mathbb{R}_+$.

To evaluate general pricing formulas we hence have to compute conditional expectations of the form

$$\mathbb{E}\left(e^{-\int_t^T R(\Psi_s) \, ds} g(\Psi_T) + \int_t^T h(\Psi_s) e^{-\int_t^s R(\Psi_u) \, du} \, ds \mid \mathcal{F}_t\right) \quad (99)$$

for generic $g, h : D \rightarrow \mathbb{R}_+$. Since (Ψ_t) is Markov, (99) is a function $f(t, \Psi_t)$ of time and of Ψ_t . f can sometimes be computed by solving a PDE; this is the well-known **Feynman-Kac formula**.

Theorem 10.14 (Feynman–Kac)

Consider generic $R, g: D \rightarrow \mathbb{R}_+$. Suppose that $f : [0, T] \times D \rightarrow \mathbb{R}$ is bounded, continuous and solves the terminal-value problem

$$f_t + \mu(\psi)f_\psi + \frac{1}{2}\sigma^2(\psi)f_{\psi\psi} = R(\psi)f, \quad (t, \psi) \in [0, T] \times D, \quad (100)$$

with $f(T, \psi) = g(\psi)$, $\psi \in D$. Suppose that (Ψ_t) is the unique solution of the SDE

$$d\Psi_t = \mu(\Psi_t) dt + \sigma(\Psi_t) dW_t, \quad \Psi_0 = \psi \in D, \quad (101)$$

with state space $D \subseteq \mathbb{R}$, (W_t) a standard, Brownian motion and μ and σ continuous functions from D to \mathbb{R} resp. \mathbb{R}_+ . Then

$$E\left(e^{-\int_t^T R(\Psi_s) ds} g(\Psi_T) \mid \mathcal{F}_t\right) = f(t, \Psi_t). \quad (102)$$

Comments

The Feynman Kac formula can be used in two ways:

- We can use probabilistic techniques or Monte-Carlo simulation to compute the (conditional) expectation (102) in order to solve numerically the PDE (100).
- We can solve the PDE (100) perhaps numerically, in order to compute the expectation (102).

For an extension to the d -dimensional case (and weaker regularity conditions on f) we refer to the literature such as Karatzas and Shreve (1988)

Affine term structure

Consider a model where r and γ are functions of a diffusion Ψ . Define a function f by

$$f(t, \Psi_t) = \mathbb{E}\left(e^{-\int_t^T R(\Psi_s)ds} e^{u\Psi_T} \mid \mathcal{F}_t\right)$$

where u, D are such $u\psi \leq 0$ for all $\psi \in D$. Note that for $u = 0$ we have a bond price with zero recovery.

Definition 10.15

The model has an *affine* (defaultable) term structure if

$$f(t, \psi) = \exp(\alpha(t, T) + \beta(t, T)\psi) \tag{103}$$

for deterministic functions $\alpha(\cdot, T)$ and $\beta(\cdot, T)$.

The following assumption guarantees an affine term structure.

Assumption 10.16 (affine term structure)

R , μ and σ^2 are *affine functions* of ψ , i.e. there are constants ρ^0 , ρ^1 , k^0 , k^1 , h^0 and h^1 such that

$$R(\psi) = \rho^0 + \rho^1\psi, \quad \mu(\psi) = k^0 + k^1\psi, \quad \sigma^2(\psi) = h^0 + h^1\psi.$$

Moreover, for all $\psi \in D$ we have $h^0 + h^1\psi \geq 0$ and $\rho_0 + \rho_1\psi \geq 0$.

An ODE system for α and β

The educated guess $f(t, \psi) = \exp(\alpha(t, T) + \beta(t, T)\psi)$ gives

$$f_t = (\alpha' + \beta'\psi)f, \quad f_\psi = \beta f \text{ and } f_{\psi\psi} = \beta^2 f.$$

Substituting this in the PDE $f_t + \mu(\psi)f_\psi + \frac{1}{2}\sigma^2(\psi)f_{\psi\psi} = R(\psi)f$, using the special form of μ, σ^2 and rearranging terms gives the following ODE system

$$\beta'(t, T) = \rho^1 - k^1\beta(t, T) - \frac{1}{2}h^1\beta^2(t, T), \quad \beta(T, T) = u, \quad (104)$$

$$\alpha'(t, T) = \rho^0 - k^0\beta(t, T) - \frac{1}{2}h^0\beta^2(t, T), \quad \alpha(T, T) = 0. \quad (105)$$

Comments.

- The ODE (104) for $\beta(\cdot, T)$ is a so-called [Riccati equation](#).
- The ODE (105) for $\alpha(\cdot, T)$ can be solved by (numerical) integration once β has been determined.

Summary. Suppose that the affine-term-structure assumption holds, that the ODE system (104), (105) has a unique solution (α, β) on $[0, T]$ and that there is some C such that $\beta(t, T)\psi \leq C$ for all $t \in [0, T]$, $\psi \in D$. Then the model has an affine term structure.

10.6.1 The CIR square-root diffusion

The **CIR** or **square-root** diffusion model due to Cox et al. (1985) is a popular affine model.

CIR dynamics.

$$d\Psi_t = \kappa(\bar{\theta} - \Psi_t) dt + \sigma\sqrt{\Psi_t} dW_t, \quad \Psi_0 = \psi > 0, \quad (106)$$

for parameters $\kappa, \bar{\theta}, \sigma > 0$ and state space $D = [0, \infty)$.

Properties.

- (106) is an affine model; the parameters are given by $k^0 = \kappa\bar{\theta}$, $k^1 = -\kappa$, $h^0 = 0$ and $h^1 = \sigma^2$.
- The SDE (106) admits a global solution (non-trivial)
- (106) implies that (Ψ_t) is *mean reverting*.
- Mean reversion sufficiently strong \Rightarrow trajectories never reach zero: Let $\tau_0(\Psi) := \inf\{t \geq 0 : \Psi_t = 0\}$. For $\kappa\bar{\theta} \geq \frac{1}{2}\sigma^2$, $\mathbb{P}(\tau_0(\Psi) < \infty) = 0$; for $\kappa\bar{\theta} < \frac{1}{2}\sigma^2$, $\mathbb{P}(\tau_0(\Psi) < \infty) = 1$.

CIR term structure

Theorem 10.17

Suppose that the factor Ψ follows the CIR model and that adjusted interest rate is an affine function of the state, $R(\psi) = \rho^0 + \rho^1\psi$. Then it holds that

$$\mathbb{E}\left(\exp\left(-\int_t^T (\rho^0 + \rho^1\Psi_s) ds\right) \middle| \Psi_t\right) = \exp(\alpha(T-t) + \beta(T-t)\Psi_t),$$

where

$$\beta(\tau) = \frac{-2\rho^1(e^{\gamma\tau} - 1)}{\gamma - \kappa + e^{\gamma\tau}(\gamma + \kappa)},$$

$$\alpha(\tau) = -\rho^0\tau + 2\frac{\kappa\bar{\theta}}{\sigma^2} \ln\left(\frac{2\gamma e^{\tau(\gamma+\kappa)/2}}{\gamma - \kappa + e^{\gamma\tau}(\gamma + \kappa)}\right),$$

and $\tau := T - t$, $\gamma := \sqrt{\kappa^2 + 2\sigma^2\rho^1}$

10.6.2 Extensions

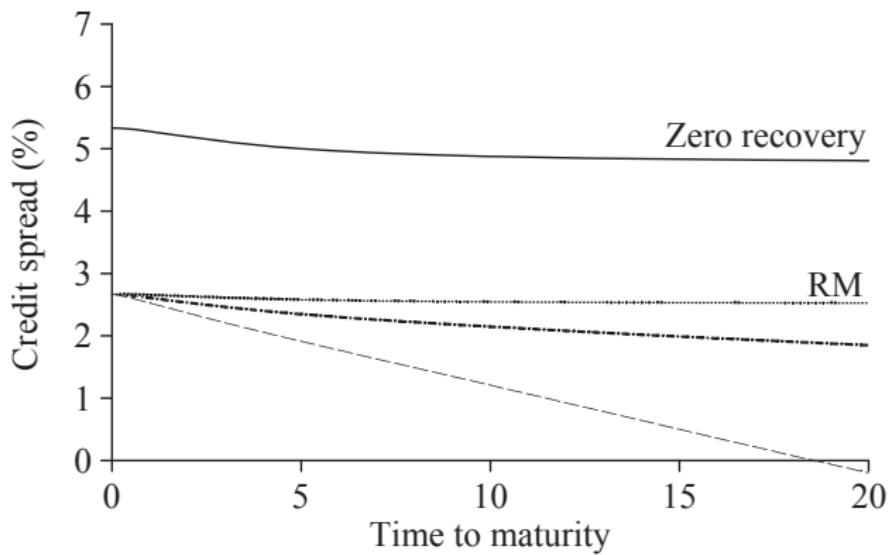
- It is possible to extend the above to CIR models with **jumps**,

$$d\Psi_t = \kappa(\bar{\theta} - \Psi_t) dt + \sigma\sqrt{\Psi_t} dW_t + dZ_t$$

where $Z_t = \sum_{T_n \leq t} Z_n$ is a compound Poisson process with jump intensity $\lambda_t = \lambda_0 + \lambda_1 \Psi_t$, λ_0 and $\lambda_1 > 0$ and the Z_n are iid positive rvs, for instance exponentially distributed.

- The computation of **payment-at-default** claims is also possible with “affine model technology”.

Numerical example



Spreads of defaultable zero-coupon bonds in an affine model for various recovery assumptions. It holds $\Psi_0 \approx 0.0533$, $r = 6\%$ and $\delta = 0.5$. Note that under the RF recovery model (dashed line) the spread becomes negative for large times to maturity; this is not true under other recovery assumptions.

11 Portfolio credit risk management

11.1 Threshold models

11.2 Mixture models

11.3 Asymptotics for large portfolios

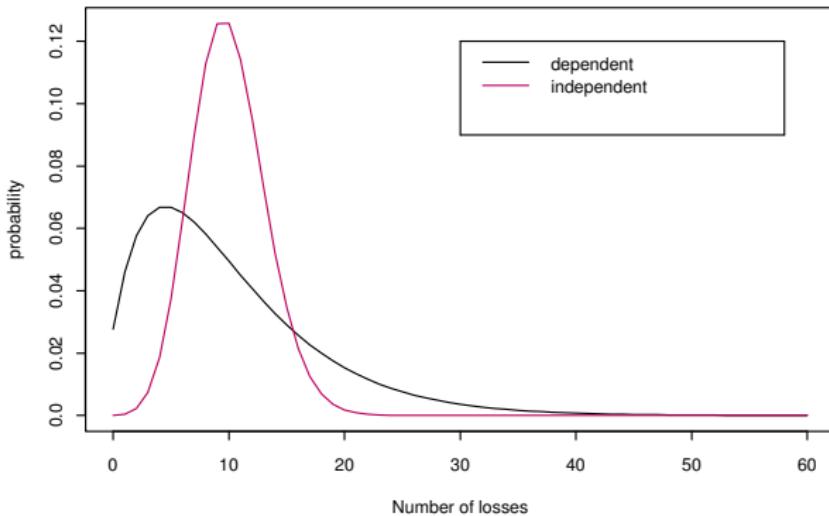
11.4 Monte Carlo methods

11.5 Statistical inference for portfolio credit models

Importance of default dependence

Dependence between defaults (and downgrades) is a key issue in credit risk management. There are two main sources of dependence between defaults:

- Dependence caused by **common factors** (for example, interest rates and changes in economic growth) affecting all obligors
- Default of company A may have direct impact on default probability of company B and vice versa because of **direct business relations**, a phenomenon known as **contagion**



Comparison of the loss distribution of a homogeneous portfolio of 1000 loans with a default probability of $p_1 = \dots = p_{1000} = 1\%$ assuming (i) independent defaults and (ii) a default correlation of $\rho(Y_i, Y_j) = 0.5\%$. Case (ii) can be considered as roughly representative for BB-rated loans.

11.1 Threshold models

11.1.1 Notation for one-period portfolio models

- Consider portfolio of m firms and time horizon $T = 1$ (say one year).
- For $1 \leq i \leq m$, let R_i be a state indicator for obligor i at time T taking values in the set $\{0, 1, \dots, n\}$; we interpret the value 0 as default and non-zero values as states of increasing credit quality. At time $t = 0$ obligors are assumed to be in some non-default state.
- Mostly we will concentrate on the binary outcomes of default and non-default. We write Y_i for the default indicator variables so that $Y_i = 1 \iff R_i = 0$ and $Y_i = 0 \iff R_i > 0$.
- The random vector $\mathbf{Y} = (Y_1, \dots, Y_m)'$ is a vector of default indicators for the portfolio and $p(\mathbf{y}) = \mathbb{P}(Y_1 = y_1, \dots, Y_m = y_m)$, $\mathbf{y} \in \{0, 1\}^m$, is its joint probability function; the marginal default probabilities are denoted by $p_i = \mathbb{P}(Y_i = 1)$, $i = 1, \dots, m$.

- Default or event correlation. Noting that

$$\text{var}(Y_i) = \mathbb{E}(Y_i^2) - p_i^2 = \mathbb{E}(Y_i) - p_i^2 = p_i - p_i^2,$$

we obtain, for firms i and j with $i \neq j$, the formula

$$\rho(Y_i, Y_j) = \frac{\mathbb{E}(Y_i Y_j) - p_i p_j}{\sqrt{(p_i - p_i^2)(p_j - p_j^2)}}. \quad (107)$$

- Let the rv $M := \sum_{i=1}^m Y_i$ denote the **number of defaulted obligors** at T .
- The actual loss if company i defaults is modelled by the random quantity $\delta_i e_i$, where e_i represents the overall exposure to company i and $0 \leq \delta_i \leq 1$ represents the LGD.
- We denote the **overall portfolio loss** by $L := \sum_{i=1}^m \delta_i e_i Y_i$.
- It is possible to set up different credit risk models leading to the same multivariate distribution for \mathbf{R} or \mathbf{Y} . We call two models with state vectors \mathbf{R} and $\tilde{\mathbf{R}}$ (or \mathbf{Y} and $\tilde{\mathbf{Y}}$) **equivalent** if $\mathbf{R} \stackrel{d}{=} \tilde{\mathbf{R}}$ (or $\mathbf{Y} \stackrel{d}{=} \tilde{\mathbf{Y}}$).

The exchangeable special case

- It is common to group obligors together to form homogeneous groups.
This corresponds to the mathematical concept of [exchangeability](#).
- A random vector \mathbf{R} is exchangeable if

$$(R_1, \dots, R_m) \stackrel{\text{d}}{=} (R_{\Pi(1)}, \dots, R_{\Pi(m)}) ,$$

for any permutation $(\Pi(1), \dots, \Pi(m))$ of $(1, \dots, m)$.

- We talk of an [exchangeable default model](#) if the default indicator vector $\mathbf{Y} = (Y_1, \dots, Y_m)'$ is exchangeable.
- Note that this permits a simple notation for default probabilities:

$$\pi_k := \mathbb{P}(Y_{i_1} = 1, \dots, Y_{i_k} = 1), \quad \{i_1, \dots, i_k\} \subset \{1, \dots, m\},$$
$$\pi := \pi_1 = P(Y_i = 1), \quad i \in \{1, \dots, m\}.$$

- In the exchangeable case the default correlation is given by

$$\rho_Y := \rho(Y_i, Y_j) = \frac{\pi_2 - \pi^2}{\pi - \pi^2}, \quad i \neq j. \tag{108}$$

11.1.2 Threshold models and copulas

Definition 11.1

Let $\mathbf{X} = (X_1, \dots, X_m)'$ be an m -dimensional random vector and let $D \in \mathbb{R}^{m \times n}$ be a deterministic matrix with elements d_{ij} such that, for every i , the elements of the i th row form a set of increasing thresholds satisfying $d_{i1} < \dots < d_{in}$. Augment these thresholds by setting $d_{i0} = -\infty$ and $d_{i(n+1)} = \infty$ for all obligors and then set

$$R_i = j \iff d_{ij} < X_i \leq d_{i(j+1)}, \quad j \in \{0, \dots, n\}, \quad i \in \{1, \dots, m\}.$$

Then (\mathbf{X}, D) is said to define a threshold model for $\mathbf{R} = (R_1, \dots, R_m)'$.

- \mathbf{X} are the **critical variables** and the i th row of D contains the **critical thresholds** for firm i .
- Default occurs if $X_i \leq d_{i1}$ so that the default probability of company i is given by $p_i = F_{X_i}(d_{i1})$.

- When working with a default-only model we simply write $d_i = d_{i1}$ and denote the threshold model by (\mathbf{X}, \mathbf{d}) .
- **Default correlation and asset correlation.** It is important to distinguish the default correlation $\rho(Y_i, Y_j)$ of two firms $i \neq j$ from the correlation of the critical variables X_i and X_j .
- Since the critical variables are often interpreted in terms of asset values, the latter correlation is often referred to as **asset correlation**.
- For given default probabilities, $\rho(Y_i, Y_j)$ is determined by $\mathbb{E}(Y_i Y_j)$ according to (107), and in a threshold model $\mathbb{E}(Y_i Y_j) = \mathbb{P}(X_i \leq d_{i1}, X_j \leq d_{j1})$, so default correlation depends on the joint df of X_i and X_j .
- If \mathbf{X} is multivariate normal, as in many models used in practice, the correlation of X_i and X_j determines the **copula** of their joint distribution and hence the default correlation.

Copulas: Key facts for this chapter

- A copula is a multivariate distribution function (df) with standard uniform marginal distributions (margins).
- If the df of $\mathbf{U} = (U_1, \dots, U_m)'$ is the copula C , we have that $\mathbb{P}(U_i \leq u_i) = C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$.

Sklar's Theorem

Let F be a joint distribution function with margins F_1, \dots, F_m . There exists a copula C such that for all x_1, \dots, x_m in $[-\infty, \infty]$

$$F(x_1, \dots, x_m) = C(F_1(x_1), \dots, F_m(x_m)).$$

If the margins are continuous then C is unique. **Conversely**, if C is a copula and F_1, \dots, F_m are univariate distribution functions, then F defined above is a multivariate df with margins F_1, \dots, F_m .

- Sklar's Theorem shows that every multivariate df can be written as a copula function of its marginal dfs.
- The converse shows that copulas can be used to create multivariate distributions with arbitrary margins.
- Let $\mathbf{X} = (X_1, \dots, X_m)'$ be a random vector with df F and continuous margins.
 - ▶ We refer to the copula C contained in F as the copula of F (or \mathbf{X}).
 - ▶ Invariance property: C is also the copula of $(T_1(X_1), \dots, T_m(X_m))$ for strictly increasing transformations T_1, \dots, T_m .
 - ▶ The copula C can be viewed as a representation of the dependence structure of F (or \mathbf{X}).
- The copula of a multivariate normal vector \mathbf{X} is known as the Gauss copula. In view of the invariance property it only depends on the correlation matrix \mathbb{P} of \mathbf{X} .

- The copula of a multivariate Student t random vector \mathbf{X} is known as the **t copula**. It depends on the correlation matrix of \mathbf{X} and the degree of freedom ν .
- The Gauss and t copula do not have simple analytical forms. They are, however, **flexible dependence models** because they have at least one parameter for every pairs of marginal distributions.
- There are other copulas with simpler forms like the Archimedean copulas. These take the form

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

where the **(Archimedean) generator** $\psi : [0, \infty) \rightarrow [0, 1]$ is strictly decreasing on $[0, \inf\{t : \psi(t) = 0\}]$ and satisfies $\psi(0) = 1$ and $\lim_{t \rightarrow \infty} \psi(t) = 0$. Examples include **Clayton** ($\psi(t) = (1+t)^{-1/\theta}$, $\theta \in (0, \infty)$) and **Gumbel** ($\psi(t) = \exp(-t^{1/\theta})$, $\theta \in [1, \infty)$). These are distributions for exchangeable random vectors \mathbf{U} .

Copulas in threshold models

Lemma 11.2

Let (\mathbf{X}, D) and $(\tilde{\mathbf{X}}, \tilde{D})$ be a pair of threshold models with state vectors $\mathbf{R} = (R_1, \dots, R_m)'$ and $\tilde{\mathbf{R}} = (\tilde{R}_1, \dots, \tilde{R}_m)'$, respectively. The models are equivalent if the following conditions hold.

- (i) The marginal distributions of the random vectors \mathbf{R} and $\tilde{\mathbf{R}}$ coincide, i.e.

$$\mathbb{P}(R_i = j) = \mathbb{P}(\tilde{R}_i = j), \quad j \in \{1, \dots, n\}, \quad i \in \{1, \dots, m\}.$$

- (ii) \mathbf{X} and $\tilde{\mathbf{X}}$ admit the same copula C .

- The copula is **critical for joint default probabilities**. Consider a subgroup of k companies $\{i_1, \dots, i_k\} \subset \{1, \dots, m\}$.

- We have

$$\begin{aligned}\mathbb{P}(Y_{i_1} = 1, \dots, Y_{i_k} = 1) &= \mathbb{P}(X_{i_1} \leq d_{i_1}, \dots, X_{i_k} \leq d_{i_k}) \\ &= C_{i_1, \dots, i_k}(F_{i_1}(d_{i_1}), \dots, F_{i_k}(d_{i_k})) \\ &= C_{i_1, \dots, i_k}(p_{i_1}, \dots, p_{i_k}),\end{aligned}$$

where C_{i_1, \dots, i_k} is a k -dimensional marginal df of C .

- For $S \subset \{1, \dots, d\}$, C_S is obtained from $C(u_1, \dots, u_d)$ by setting $u_i = 1$ for $i \notin S$.
- The copula C crucially determines higher order joint default probabilities and thus **extreme risk** that many companies default.
- In an **exchangeable default model** we have

$$\pi_k = C_{1, \dots, k}(\pi, \dots, \pi), \quad 2 \leq k \leq m.$$

11.1.3 Gaussian threshold models

Multivariate Merton model:

- Assume that the multivariate asset-value process $\mathbf{V}_t = (V_{t,1}, \dots, V_{t,m})'$ follows an m -dimensional GBM with drift vector $\boldsymbol{\mu}_V = (\mu_1, \dots, \mu_m)'$, vector of volatilities $\boldsymbol{\sigma}_V = (\sigma_1, \dots, \sigma_m)'$ and correlation matrix P .
- This means that (\mathbf{V}_t) solves the stochastic differential equations

$$dV_{t,i} = \mu_i V_{t,i} dt + \sigma_i V_{t,i} dW_{t,i}, \quad i = 1, \dots, m,$$

for correlated BMs with correlation $\rho(W_{t,i}, W_{t,j}) = \rho_{ij}$, $t \geq 0$.

- For all i the asset value $V_{T,i}$ is of the form

$$V_{T,i} = V_{0,i} \exp((\mu_i - \frac{1}{2}\sigma_i^2)T + \sigma_i W_{T,i}),$$

where $\mathbf{W}_T \sim \mathcal{N}_m(\mathbf{0}, TP)$.

- In its basic form the Merton model is a default-only model where the firm defaults if $V_{T,i} \leq B_i$ and B_i is the liability of firm i .

- Writing $\mathbf{B} = (B_1, \dots, B_m)'$ the threshold model representation is $(\mathbf{V}_T, \mathbf{B})$.
- The multivariate Merton model is equivalent to the model (\mathbf{X}, \mathbf{d}) with

$$X_i := \frac{\ln V_{T,i} - \ln V_{0,i} - (\mu_i - \frac{1}{2}\sigma_i^2)T}{\sigma_i\sqrt{T}},$$

$$d_i := \frac{\ln B_i - \ln V_{0,i} - (\mu_i - \frac{1}{2}\sigma_i^2)T}{\sigma_i\sqrt{T}}.$$

- The transformed variables satisfy $\mathbf{X} \sim N_m(\mathbf{0}, P)$ and their copula is the Gauss copula C_P^{Ga} .

Gaussian threshold models in practice

- In practice it is usual to start directly with threshold models of the form (\mathbf{X}, \mathbf{d}) with $\mathbf{X} \sim N_m(\mathbf{0}, P)$.

- There are two practical challenges:
 - 1) calibration of \boldsymbol{d} (or, in the case of a multi-state model, the threshold matrix D) in line with exogenously given default and transition probabilities;
 - 2) calibration of the correlation matrix P in a **parsimonious** way.
- The problem of embedding state transition probabilities in a threshold matrix D has already been discussed. In a default-only model we set $d_i = \Phi^{-1}(p_i)$ for $i = 1, \dots, m$.

Factor models

- In its most general form P has $m(m - 1)/2$ distinct parameters.
- m is typically large and it is important to use a more parsimonious parametrization of this matrix based on a factor model.

- Factor models also lend themselves to **economic interpretation** and the factors are commonly interpreted as **country and industry effects**.
- We assume that

$$X_i = \sqrt{\beta_i} \tilde{F}_i + \sqrt{1 - \beta_i} \varepsilon_i, \quad (109)$$

where \tilde{F}_i and $\varepsilon_1, \dots, \varepsilon_m$ are independent standard normal variables, and where $0 \leq \beta_i \leq 1$ for all i .

- In this formulation \tilde{F}_i are the **systematic** variables, which are correlated, and ε_i are **idiosyncratic** variables.
- It follows that β_i can be viewed as a measure of the **systematic risk** of X_i : that is, the part of the variance of X_i which is explained by the systematic variable.
- The systematic variables are assumed to be of the form $\tilde{F}_i = \mathbf{a}'_i \mathbf{F}$ where \mathbf{F} is a vector of common factors satisfying $\mathbf{F} \sim N_p(\mathbf{0}, \Omega)$ with $p < m$, and where Ω is a correlation matrix.

- These factors typically represent country and industry effects.
- The assumption that $\text{var}(\tilde{F}_i) = 1$ means that $\mathbf{a}'_i \Omega \mathbf{a}_i = 1$ for all i .
- Since $\text{var}(X_i) = 1$ and since \tilde{F}_i and $\varepsilon_1, \dots, \varepsilon_m$ are independent and standard normal, the asset correlations in this model are given by

$$\rho(X_i, X_j) = \text{cov}(X_i, X_j) = \sqrt{\beta_i \beta_j} \text{cov}(\tilde{F}_i, \tilde{F}_j) = \sqrt{\beta_i \beta_j} \mathbf{a}'_i \Omega \mathbf{a}_j .$$

- In order to set up the model we have to determine \mathbf{a}_i and β_i for each obligor and Ω , with the additional constraint that $\mathbf{a}'_i \Omega \mathbf{a}_i = 1$ for all i .
- Since Ω has $p(p - 1)/2$ parameters, the loading vectors \mathbf{a}_i and coefficients β_i have collectively $mp + m$ parameters, and we are applying m constraints, this gives $mp + p(p - 1)/2$ parameters.

The one-factor model

- We often consider the special case of a one-factor model.

- This corresponds to a model where $\tilde{F}_i = F$ for a single common standard normal factor so that the equation in (109) takes the form

$$X_i = \sqrt{\beta_i} F + \sqrt{1 - \beta_i} \varepsilon_i. \quad (110)$$

- If, moreover, every obligor has the same systematic variance $\beta_i = \rho$ we get that $\rho(X_i, X_j) = \rho$ for all $i \neq j$, which is often referred to as an **equicorrelation model**.

11.1.4 Models based on alternative copulas

t copula model

- Suppose Z_i follows the Gaussian factor model

$$Z_i = \sqrt{\beta_i} \tilde{F}_i + \sqrt{1 - \beta_i} \varepsilon_i,$$

with $\tilde{F}_i = \mathbf{a}'_i \mathbf{F}$ and all assumptions as before.

- Let $X_i = \sqrt{W} Z_i$ for $i = 1, \dots, m$, where W has an inverse gamma distribution, $W \sim \text{IG}(\frac{1}{2}\nu, \frac{1}{2}\nu)$, or equivalently, $\nu/W \sim \chi^2_\nu$.

- The vector (X_1, \dots, X_m) has a multivariate t distribution with ν degrees of freedom, location vector $\mathbf{0}$ and correlation matrix P identical to that of (Z_1, \dots, Z_m) .
- The critical variables have the t copula $C_{\nu, P}^t$.
- The class of threshold models based on the t copula can be thought of as containing the Gaussian threshold models as limiting cases when $\nu \rightarrow \infty$. However, the additional parameter ν adds a great deal of flexibility.

Archimedean copulas

- We recall that these take the form

$$C(u_1, \dots, u_m) = \psi(\psi^{-1}(u_1) + \dots + \psi^{-1}(u_m)),$$

where the **generator** $\psi : [0, \infty) \rightarrow [0, 1]$ is a continuous, decreasing function satisfying $\psi(0) = 1$ and $\lim_{t \rightarrow \infty} \psi(t) = 0$, and ψ^{-1} is its inverse.

- The Clayton copula has generator $\psi_\theta(t) = (1 + \theta t)^{-1/\theta}$, where $\theta > 0$, leading to the expression

$$C_\theta^{\text{Cl}}(u_1, \dots, u_m) = (u_1^{-\theta} + \dots + u_m^{-\theta} + 1 - m)^{-1/\theta}.$$

- Suppose that \mathbf{X} is a random vector with an Archimedean copula and marginal distributions F_{X_i} , $1 \leq i \leq m$, so that (\mathbf{X}, \mathbf{d}) specifies a threshold model with individual default probabilities $F_{X_i}(d_i)$.
- Consider the Clayton copula and assume a homogeneous situation where all individual default probabilities are identical to π .
- We can calculate that

$$\pi_k = (k\pi^{-\theta} - k + 1)^{-1/\theta}.$$

- Essentially, the dependent default mechanism of the homogeneous group is now determined by this equation and the parameters π and θ .

11.2 Mixture models

11.2.1 Bernoulli mixture models

- In a mixture model the default risk of an obligor is assumed to depend on a set of **common factors**, usually interpreted as macroeconomic variables, which are also modelled stochastically.
- Given a realization of the factors, defaults of individual firms are assumed to be independent.
- Dependence between defaults stems from the dependence of individual default probabilities on the set of common factors.
- Bernoulli mixture models provide a way of capturing the dependence between Bernoulli events (i.e. defaults/non-defaults).
- They can be extended to multinomial mixture models to capture dependent migrations in a rating system.

Definition 11.3 (Bernoulli mixture model)

Given some $p < m$ and a p -dimensional random vector $\Psi = (\Psi_1, \dots, \Psi_p)'$, the default indicator vector \mathbf{Y} follows a Bernoulli mixture model with factor vector Ψ if there are functions $p_i : \mathbb{R}^p \rightarrow (0, 1)$, such that conditional on Ψ the components of \mathbf{Y} are independent Bernoulli rvs with $\mathbb{P}(Y_i = 1 | \Psi = \psi) = p_i(\psi)$.

The **conditional independence given factors** makes these models relatively easy to analyse. For $\mathbf{y} = (y_1, \dots, y_m)'$ in $\{0, 1\}^m$ we get

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} | \Psi = \psi) = \prod_{i=1}^m p_i(\psi)^{y_i} (1 - p_i(\psi))^{1-y_i}$$

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \int_{\mathbb{R}^p} \prod_{i=1}^m p_i(\psi)^{y_i} (1 - p_i(\psi))^{1-y_i} g(\psi) d\psi,$$

where $g(\psi)$ is the probability density of the factors. The default probabilities are given by $p_i = \mathbb{E}(Y_i = 1) = \mathbb{E}(p_i(\Psi))$.

- Consider the portfolio loss $L = \sum_{i=1}^m e_i \delta_i Y_i$ in the case where the exposures e_i and LGDs δ_i are deterministic.
- It is difficult to compute the df F_L of L .
- However, it is easy to use the conditional independence of the defaults to show that the *Laplace–Stieltjes transform* of F_L is for $t \in \mathbb{R}$ given by

$$\begin{aligned}\hat{F}_L(t) &= \mathbb{E}(e^{-tL}) = \mathbb{E}\left(\mathbb{E}(e^{-t\sum_{i=1}^m e_i \delta_i Y_i} \mid \Psi)\right) \\ &= \mathbb{E}\left(\prod_{i=1}^m \mathbb{E}(e^{-te_i \delta_i Y_i} \mid \Psi)\right) \\ &= \mathbb{E}\left(\prod_{i=1}^m (p_i(\Psi)e^{-te_i \delta_i} + 1 - p_i(\Psi))\right)\end{aligned}$$

which can be obtained by integrating over distribution of factors Ψ .

- This is useful for: sampling losses from model with **importance sampling**; approximating probability mass function using **Fourier inversion**.

11.2.2 One-factor Bernoulli mixture models

Often it is useful to work with a one factor model ($p = 1$):

- Fitting to default data is relatively easy because evaluation of joint distribution/likelihood involves a one-dimensional integral at worst.
- The behaviour of large portfolios is easy to understand in terms of the distribution of the common factor.
- A one-factor model underlies the Basel II formula for computing capital under the IRB approach.

Thus we consider a rv Ψ and functions $p_i(\Psi)$ such that, conditional on Ψ , the default indicator vector \mathbf{Y} is a vector of independent Bernoulli random variables with

$$\mathbb{P}(Y_i = 1 \mid \Psi = \psi) = p_i(\psi).$$

Exchangeable special case

- In the exchangeable special case of a one-factor model the conditional default probabilities $p_i(\Psi)$ are identical, $\forall i$, making \mathbf{Y} exchangeable.
- In an exchangeable model we will introduce a new rv $Q = p_1(\Psi)$ with df $G(q)$ for the **conditional default probability**.
- Obviously Q has a distribution on $[0, 1]$.
- Recalling the π_k notation for exchangeable models we can calculate that

$$\pi := \mathbb{P}(Y_i = 1) = \mathbb{E}(Y_i) = \mathbb{E}(\mathbb{E}(Y_i | Q)) = \mathbb{E}(Q)$$

$$\pi_k := \mathbb{P}(Y_{i_1} = 1, \dots, Y_{i_k} = 1) = \mathbb{E}(Q^k) = \int_0^1 q^k dG(q). \quad (111)$$

- Conditional on $Q = q$, the number of defaults M is the sum of m independent Bernoulli variables and thus has a binomial distribution. The unconditional distribution of M is

$$\mathbb{P}(M = k) = \binom{m}{k} \int_0^1 q^k (1 - q)^{m-k} dG(q). \quad (112)$$

- In an exchangeable model the default probability and the higher order joint default probabilities are moments of the distribution of Q .
- Recall that the default correlation between two firms $i \neq j$ is defined to be the correlation between the default indicators Y_i and Y_j .
- In exchangeable Bernoulli mixtures we have

$$\text{cov}(Y_i, Y_j) = \pi_2 - \pi^2 = \text{var}(Q) \geq 0.$$

- The **default correlation** is given by

$$\rho_Y := \text{corr}(Y_i, Y_j) = \frac{\pi_2 - \pi^2}{\pi - \pi^2} = \frac{\text{var}(Q)}{\mathbb{E}(Q) - \mathbb{E}(Q)^2}.$$

Common mixing distributions

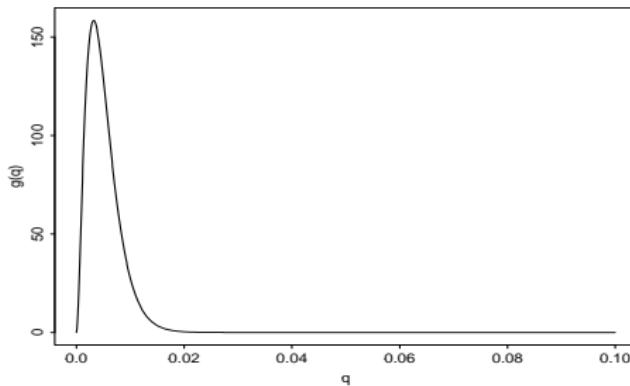
- **Beta.** $Q \sim \text{Beta}(a, b)$ where $a, b > 0$. This model corresponds quite closely to an exchangeable version of an industry model called CreditRisk⁺ which we study later.

- **Probit-Normal.** $Q = \Phi(\mu + \sigma\Psi)$, $\Psi \sim N(0, 1)$

This corresponds to an exchangeable version of the threshold model with a Gaussian copula.

- **Logit-Normal.** $Q = (1 + \exp(-\mu - \sigma\Psi))^{-1}$, $\Psi \sim N(0, 1)$

In all these 2-parameter examples, if we fix default probability π and default correlation ρ_Y (or π_2) we fully calibrate the model. Picture shows beta density $g(q)$ of mixing variable Q in exchangeable Bernoulli mixture model with $\pi = 0.005$ and $\rho_Y = 0.0018$.



Example 11.4 (Beta mixing distribution)

- The density of a beta distribution is given by

$$g(q) = \frac{1}{\beta(a, b)} q^{a-1} (1-q)^{b-1}, \quad a, b > 0, \quad 0 < q < 1,$$

where $\beta(a, b)$ denotes the **beta function**.

- The beta function satisfies the following recursion formula:
$$\beta(a+1, b) = (a/(a+b))\beta(a, b).$$
- Using (111) we obtain for the higher-order default probabilities

$$\pi_k = \frac{1}{\beta(a, b)} \int_0^1 q^k q^{a-1} (1-q)^{b-1} dq = \frac{\beta(a+k, b)}{\beta(a, b)}, \quad k = 1, 2, \dots$$

- The recursion formula for the beta function yields:

$$\pi_k = \prod_{j=0}^{k-1} (a+j)/(a+b+j).$$

- In particular, $\pi = a/(a+b)$, $\pi_2 = \pi(a+1)/(a+b+1)$ and
$$\rho_Y = (a+b+1)^{-1}.$$

- The rv M in (112) has a so-called *beta-binomial distribution*:

$$\begin{aligned}\mathbb{P}(M = k) &= \binom{m}{k} \frac{1}{\beta(a, b)} \int_0^1 q^{k+a-1} (1-q)^{m-k+b-1} dq \\ &= \binom{m}{k} \frac{\beta(a+k, b+m-k)}{\beta(a, b)}.\end{aligned}\tag{113}$$

One-factor models with covariates

- It is straightforward to extend the one-factor probit-normal and logit-normal mixture models to include *covariates*.
- These might be indicators for group membership, such as *rating class* or *industry sector*, or key *ratios taken from a company's balance sheet*.

- Writing $\boldsymbol{x}_i \in \mathbb{R}^k$ for a vector of covariates, we assume that

$$\begin{aligned} p_i(\Psi) &= h(\mu_i + \sigma_i \Psi) \\ \mu_i &= \mu + \boldsymbol{\beta}' \boldsymbol{x}_i \\ \sigma_i &= \exp(\delta + \boldsymbol{\gamma}' \boldsymbol{x}_i) \end{aligned} \tag{114}$$

where $\Psi \sim N(0, 1)$, $h(x) = \Phi(x)$ or $h(x) = (1 + \exp(-x))^{-1}$.

- The vectors $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$ and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_k)'$ contain regression parameters, and $\mu \in \mathbb{R}$ and $\delta \in \mathbb{R}$ are intercept parameters.
- Similar specifications are commonly used in the class of [generalized linear models](#) in statistics.
- The regression structure in (114) includes partially exchangeable models where we define a number of groups within which risks are exchangeable. These groups might represent rating classes.

- If the covariates \mathbf{x}_i are of the form $\mathbf{x}_i = \mathbf{e}_{r(i)}$, where $r(i) \in \{1, \dots, k\}$ indicates the rating class of firm i , then the model (114) can be written in the form

$$p_i(\Psi) = h(\mu_{r(i)} + \sigma_{r(i)}\Psi) \quad (115)$$

for parameters $\mu_r := \mu + \beta_r$ and $\sigma_r := \exp(\delta + \gamma_r)$ for $r = 1, \dots, k$.

- Suppose there are m_r obligors in rating category r for $r = 1, \dots, k$, and write M_r for the number of defaults.
- The conditional distribution of the vector $\mathbf{M} = (M_1, \dots, M_k)'$ is

$$\mathbb{P}(\mathbf{M} = \mathbf{l} \mid \Psi = \psi) = \prod_{r=1}^k \binom{m_r}{l_r} (h(\mu_r + \sigma_r \psi))^{l_r} (1 - h(\mu_r + \sigma_r \psi))^{m_r - l_r},$$

where $\mathbf{l} = (l_1, \dots, l_k)'$.

11.2.3 Recovery risk in mixture

- In standard portfolio risk models it is assumed that the loss given default is independent of the default event.

- However, economic intuition suggests that recovery rates depend on similar risk factors to default probabilities.
- For example, during a property crisis many mortgages default. At the same time property prices are low, so that real estate can be sold only for very low prices in a foreclosure, leading to low recovery rates.
- The presence of systematic recovery risk is confirmed in a number of empirical studies, including Frye (2000) and Hamilton et al. (2005).
- The latter estimated that the relationship between the one-year default rate q and recovery rate R for corporate bonds was $R(q) \approx (0.52 - 6.9q)^+$.
- Systematics recoveries (LGDs) can be incorporated in the mixture-model framework by we replacing the constant δ_i with some function $\delta_i(\psi)$.
- The challenge lies in calibrating the function $\delta_i(\cdot)$.

11.2.4 Threshold models as mixture models

- Although the mixture models of this section seem, at first glance, to be different in structure to the threshold models, it is important to realize that the majority of useful threshold models, including all the examples we have given, [can be represented as Bernoulli mixture models](#).
- Recall the threshold models based on a vector of multivariate Gaussian critical variables. These models can be motivated by a multivariate firm value model. Industry models such as Moody's public-firm EDF model or CreditMetrics belong in this category.
- Default occurs for counterparty i if a critical variable X_i lies below a critical threshold d_i . $\mathbf{X} = (X_1, \dots, X_m)'$ is a random vector with standard normal margins (since we can standardize X_i and d_i without altering the default probability.)

- Moreover X_i follows a linear factor model

$$X_i = \sqrt{\beta_i} \mathbf{a}'_i \mathbf{F} + \sqrt{1 - \beta_i} \varepsilon_i$$

where

- ▶ $\mathbf{F} \sim N_p(\mathbf{0}, \Omega)$ is a random vector of normally distributed common economic factors;
- ▶ $0 \leq \beta_i \leq 1$ and $\text{var}(\mathbf{a}'_i \mathbf{F}) = 1$;
- ▶ $\varepsilon_1, \dots, \varepsilon_m$ are iid standard normal and are also independent of \mathbf{F} .
- We will write the Gaussian threshold model as a Bernoulli mixture model with factor vector $\Psi = -\mathbf{F}$. (This makes the conditional default probabilities increasing in the factors for positive \mathbf{a}_i .)
- Conditioning on $\Psi = -\mathbf{F}$, the vector \mathbf{X} is multivariate normally distributed with a diagonal covariance matrix and therefore the components of \mathbf{X} are conditionally independent.

- The conditional default probabilities are

$$\begin{aligned}
 p_i(\psi) &= \mathbb{P}(Y_i = 1 \mid \Psi = \psi) = \mathbb{P}(X_i \leq d_i \mid \Psi = \psi) \\
 &= \mathbb{P}(X_i \leq d_i \mid \mathbf{F} = -\psi) \\
 &= \mathbb{P}(\sqrt{1 - \beta_i} \varepsilon_i \leq d_i + \sqrt{\beta_i} \mathbf{a}'_i \psi) \\
 &= \Phi\left(\frac{d_i + \sqrt{\beta_i} \mathbf{a}'_i \psi}{\sqrt{1 - \beta_i}}\right) \\
 &= \Phi\left(\frac{\Phi^{-1}(p_i) + \sqrt{\beta_i} \mathbf{a}'_i \psi}{\sqrt{1 - \beta_i}}\right).
 \end{aligned}$$

- $p_i(\Psi)$ has a probit-normal distribution. The parameters μ_i and σ_i are

$$\mu_i = \Phi^{-1}(p_i)/\sqrt{1 - \beta_i} \quad \text{and} \quad \sigma_i^2 = \beta_i/(1 - \beta_i).$$

Special cases

1) One-Factor Model

$$p_i(\psi) = \Phi \left(\frac{\Phi^{-1}(p_i) + \sqrt{\beta_i} \psi}{\sqrt{1 - \beta_i}} \right),$$

where Ψ is a standard normally distributed factor.

2) One Factor Model with Equicorrelation Structure

$$p_i(\psi) = \Phi \left(\frac{\Phi^{-1}(p_i) + \sqrt{\rho} \psi}{\sqrt{1 - \rho}} \right), \quad (116)$$

where ρ is the asset correlation between any two critical variables $X_i \neq X_j$.

3) Fully exchangeable model. We set $p_1 = \dots = p_m$ in equicorrelation model.

11.2.5 Poisson mixture models and CreditRisk+

- Since default is typically a rare event, it is possible to approximate Bernoulli indicator rvs for default with Poisson rvs and Bernoulli mixture models with Poisson mixture models.
- By choosing independent gamma distributions for the economic factors Ψ , we obtain a tractable model known as CreditRisk+, proposed by Credit Suisse Financial Products in 1997.
- Assume that, given the factors Ψ , the default indicators Y_1, \dots, Y_m for a particular time horizon are conditionally independent Bernoulli variables satisfying $\mathbb{P}(Y_i = 1 | \Psi = \psi) = p_i(\psi)$.
- Moreover assume that the distribution of Ψ is such that the conditional default probabilities $p_i(\psi)$ tend to be very small.
- The Y_i variables can be approximated by conditionally independent Poisson variables \tilde{Y}_i satisfying $\tilde{Y}_i | \Psi = \psi \sim \text{Poi}(p_i(\psi))$.

- This follows because

$$\mathbb{P}(\tilde{Y}_i = 0 \mid \Psi = \psi) = e^{-p_i(\psi)} \approx 1 - p_i(\psi),$$

$$\mathbb{P}(\tilde{Y}_i = 1 \mid \Psi = \psi) = p_i(\psi)e^{-p_i(\psi)} \approx p_i(\psi).$$

- The portfolio loss $L = \sum_{i=1}^m e_i \delta_i Y_i$ can be approximated by $\tilde{L} = \sum_{i=1}^m e_i \delta_i \tilde{Y}_i$.
- It is possible for a company to “**default more than once**” in the approximating Poisson model, albeit with a very low probability.
- In [CreditRisk+](#) the parameter $\lambda_i(\Psi)$ of the conditional Poisson distribution for firm i is assumed to take the form

$$\lambda_i(\Psi) = k_i \mathbf{w}'_i \Psi \tag{117}$$

for $k_i > 0$, non-negative weights $\mathbf{w}_i = (w_{i1}, \dots, w_{ip})'$ satisfying $\sum_j w_{ij} = 1$, and p independent $\text{Ga}(\alpha_j, \beta_j)$ -distributed factors Ψ_1, \dots, Ψ_p .

- The parameters are set to be $\alpha_j = \beta_j = \sigma_j^{-2}$ for $\sigma_j > 0$ and $j = 1, \dots, p$.

- This parametrization of the gamma variables ensures that we have $\mathbb{E}(\Psi_j) = 1$ and $\text{var}(\Psi_j) = \sigma_j^2$.
- It is easy to verify that

$$\mathbb{E}(\tilde{Y}_i) = \mathbb{E}(\mathbb{E}(\tilde{Y}_i | \boldsymbol{\Psi})) = \mathbb{E}(\lambda_i(\boldsymbol{\Psi})) = k_i \mathbb{E}(\mathbf{w}'_i \boldsymbol{\Psi}) = k_i,$$

so k_i is the expected number of defaults for obligor i in the time period.

- The assumptions in CreditRisk+ make it possible to compute the distribution of the number of defaults and the aggregate portfolio loss fairly explicitly using techniques for compound distributions and mixture distributions that are well known in actuarial mathematics.
- The exchangeable version of CreditRisk+ with $k_i = k, \forall i$, and a single gamma-distributed factor can be shown to be very close to an exchangeable Bernoulli mixture model with beta mixing distribution.

Distribution of the number of defaults

In CreditRisk+ we have that given $\Psi = \psi$, $\tilde{Y}_i \sim \text{Poi}(k_i \mathbf{w}'_i \psi)$, which implies that the distribution of the number of defaults $\tilde{M} := \sum_{i=1}^m \tilde{Y}_i$ satisfies

$$\tilde{M} \mid \Psi = \psi \sim \text{Poi}\left(\sum_{i=1}^m k_i \mathbf{w}'_i \psi\right). \quad (118)$$

- This uses the fact that the **sum of independent Poisson variables is also Poisson** with a rate parameter given by the sum of the rate parameters
- To compute the unconditional distribution of \tilde{M} we require a well-known result on **mixed Poisson distributions**.

Proposition 11.5

If the rv N is conditionally Poisson with a gamma-distributed rate parameter $\Lambda \sim \text{Ga}(\alpha, \beta)$, then N has a **negative binomial distribution**, $N \sim \text{NB}(\alpha, \beta/(\beta + 1))$.

In the case when $p = 1$ we may apply this result directly to (118) to deduce that \tilde{M} has a negative binomial distribution. The general result is:

Proposition 11.6

\tilde{M} is distributed as a sum of p independent negative binomial rvs.

This follows by observing that

$$\sum_{i=1}^m k_i \mathbf{w}_i' \boldsymbol{\Psi} = \sum_{i=1}^m k_i \sum_{j=1}^p w_{ij} \Psi_j = \sum_{j=1}^p \Psi_j \left(\sum_{i=1}^m k_i w_{ij} \right).$$

Now consider rvs $\tilde{M}_1, \dots, \tilde{M}_p$ such that \tilde{M}_j is conditionally Poisson with mean $(\sum_{i=1}^m k_i w_{ij})\psi_j$ conditional on $\Psi_j = \psi_j$. The independence of the components Ψ_1, \dots, Ψ_p implies that the \tilde{M}_j are independent, and by construction we have $\tilde{M} \stackrel{d}{=} \sum_{j=1}^p \tilde{M}_j$. Moreover, the rvs $(\sum_{i=1}^m k_i w_{ij})\Psi_j$ are gamma distributed, so that each of the \tilde{M}_j has a negative binomial distribution by Proposition 11.5.

Distribution of the aggregate loss

- To obtain a tractable model, exposures are **discretized** in CreditRisk+ using the concept of **exposure bands**.
- The LGD is subsumed in the exposure by multiplying the actual exposure by a typical value for the LGD for an obligor with the same credit rating.
- The losses arising from the individual obligors are of the form $\tilde{L}_i = e_i \tilde{Y}_i$ where the e_i are known (LGD-adjusted) exposures.
- For all i , the **exposure e_i is discretized in units of an amount ϵ** so that e_i is replaced by a value $\ell_i \epsilon \geq e_i$ where ℓ_i is a positive integer multiplier.
- Exposure bands $b = 1, \dots, n$ are defined corresponding to the distinct values $\ell^{(1)}, \dots, \ell^{(n)}$ for the multipliers so that obligors are grouped in exposure bands according to the values of their discretized exposures.
- It is then possible to derive the distribution of the aggregate loss
$$\tilde{L} = \sum_{i=1}^m \ell_i \epsilon \tilde{Y}_i.$$

Theorem 11.7

Let \tilde{L} represent the aggregate loss in the general p -factor CreditRisk+ model with exposures discretized into exposure bands as described above. Then the following hold.

- i) The Laplace–Stieltjes transform of the df of \tilde{L} is given by

$$\hat{F}_{\tilde{L}}(s) = \prod_{j=1}^p \left(1 + \sigma_j^2 \sum_{i=1}^m k_i w_{ij} \left(1 - \sum_{b=1}^n e^{-s\epsilon\ell^{(b)}} q_{jb} \right) \right)^{-\sigma_j^{-2}}, \quad (119)$$

where $q_{jb} = \sum_{i \in s_b} k_i w_{ij} / \sum_{i=1}^m k_i w_{ij}$ for $b = 1, \dots, n$.

- ii) The distribution of \tilde{L} has the structure $\tilde{L} \stackrel{d}{=} \sum_{j=1}^p Z_j$ where the Z_j are independent variables that follow a compound negative binomial distribution. More precisely, it holds that $Z_j \sim \text{CNB}(\sigma_j^{-2}, \theta_j, G_{X_j})$ with $\theta_j = (1 + \sigma_j^2 \sum_{i=1}^m k_i w_{ij})^{-1}$ and G_{X_j} the df of a multinomial random variable X_j taking the value $\epsilon\ell^{(b)}$ with probability q_{jb} .

11.3 Asymptotics for large portfolios

11.3.1 Exchangeable one-factor models

- We begin the study of asymptotics with the special case of an exchangeable Bernoulli mixture model.
- We consider an infinite sequence of obligors indexed by $i \in \mathbb{N}$ with identical exposures $e_i = e$ and LGD equal to 100%.
- We assume that, given a mixing variable $Q \in [0, 1]$ the default indicators Y_i are independent Bernoulli random variables with conditional default probability $\mathbb{P}(Y_i = 1 | Q = q) = q$.
- We are interested in the asymptotic behaviour of the relative loss (the loss expressed as a proportion of total exposure).

- Writing $L^{(m)} = \sum_{i=1}^m eY_i$ for the total loss of the first m companies, the corresponding relative loss is given by

$$\frac{L^{(m)}}{me} = \frac{1}{m} \sum_{i=1}^m Y_i.$$

- Conditioning on $Q = q$ the Y_i are independent with mean q and the **strong law of large numbers (SLLN)** implies that, given $Q = q$,

$$\lim_{m \rightarrow \infty} \frac{L^{(m)}}{me} = q$$

almost surely.

- This shows that, for large m , the behaviour of the relative loss is essentially governed by the mixing distribution $G(q)$ of Q . In particular, it can be shown that, for G strictly increasing,

$$\lim_{m \rightarrow \infty} \text{var}_\alpha \left(\frac{L^{(m)}}{me} \right) = q_\alpha(Q).$$

Example: Vasicek distribution

- Consider the exchangeable model that is equivalent to a Gaussian threshold model with default probability π and default correlation ρ .
- In this model Q has a probit-normal distribution given by

$$Q = \Phi \left(\frac{\Phi^{-1}(\pi) + \sqrt{\rho}\Psi}{\sqrt{1-\rho}} \right).$$

- The idea of using this distribution as an approximation to the large portfolio loss is attributed to *Vasicek*.
- Similar asymptotic ideas can be applied to more complicated models. We find in general that the distribution of the loss is driven by the distribution of the *systematic factor* and idiosyncratic risks become negligible.
- The asymptotic analysis is also described as analysing *infinitely fine-grained portfolios*.

The large portfolio results can be used to give approximations when m is large. Consider again the simple example.

- For tail probabilities we have

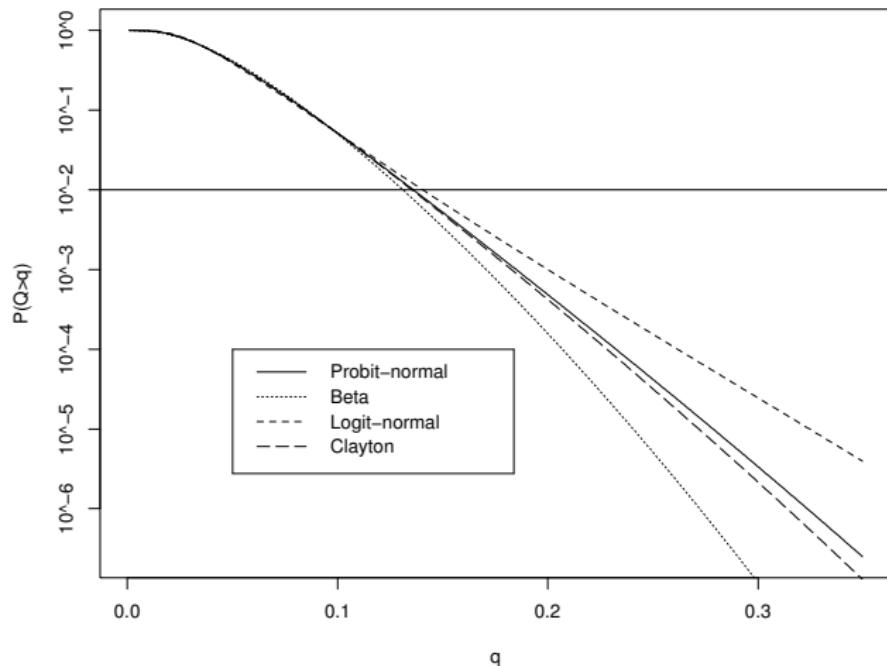
$$\mathbb{P}\left(L^{(m)} > l\right) \approx \mathbb{P}\left(Q > \frac{l}{me}\right) = \Phi\left(\frac{\Phi^{-1}(\pi) - \sqrt{1-\rho}\Phi^{-1}(l/(me))}{\sqrt{\rho}}\right).$$

- For value-at-risk we have

$$\text{VaR}_\alpha(L^{(m)}) \approx me \cdot q_\alpha(Q) = me \cdot \Phi\left(\frac{\Phi^{-1}(\pi) + \sqrt{\rho}\Phi^{-1}(\alpha)}{\sqrt{1-\rho}}\right).$$

An important insight is that the tail of distribution of Q determines the tail of portfolio loss distribution.

Comparison of 4 models with same values for π and π_2 .



Horizontal line at 0.01 shows models diverge at 99th percentile of Q .

11.3.2 General results

- Let $(e_i)_{i \in \mathbb{N}}$ be an infinite sequence of positive deterministic exposures, $(Y_i)_{i \in \mathbb{N}}$ be the corresponding sequence of default indicators and $(\delta_i)_{i \in \mathbb{N}}$ a sequence of random variables with values in $(0, 1]$ representing percentage losses given that default occurs.
- In this setting the loss for a portfolio of size m is given by $L^{(m)} = \sum_{i=1}^m L_i$ where $L_i = e_i \delta_i Y_i$ are the individual losses. We now make some technical assumptions for our model.
- We introduce the notation $a_m = \sum_{i=1}^m e_i$ for the aggregate exposure to the first m obligors.
- We now make some technical assumptions for our model.

Assumptions

- A1. There is a p -dimensional random vector Ψ such that, conditional on Ψ , the $(L_i)_{i \in \mathbb{N}}$ form a sequence of independent random variables.
We extend conditional independence assumption to losses.
- A2. There is a function $\bar{\ell} : \mathbb{R}^p \rightarrow [0, 1]$ such that

$$\lim_{m \rightarrow \infty} \frac{1}{a_m} \mathbb{E}(L^{(m)} \mid \Psi = \psi) = \bar{\ell}(\psi)$$

for all $\psi \in \mathbb{R}^p$. We call $\bar{\ell}(\psi)$ the asymptotic relative loss function.

We preserve composition of portfolio as it grows.

- A3. The sequence of exposures satisfies

$$\lim_{m \rightarrow \infty} a_m = \infty \quad \text{and} \quad \sum_{i=1}^{\infty} (e_i/a_i)^2 < \infty.$$

The portfolio may not be dominated by a few large exposures.

First result

- We can show that in large portfolios the portfolio loss is essentially determined by the asymptotic relative loss function $\bar{\ell}$ and the realisation of the factor random vector Ψ .
- Consider a sequence $L^{(m)} = \sum_{i=1}^m L_i$ satisfying the assumptions.
- Denote by $\mathbb{P}(\cdot | \Psi = \psi)$ the conditional distribution of the sequence $(L_i)_{i \in \mathbb{N}}$ given $\Psi = \psi$. Then

$$\lim_{m \rightarrow \infty} \frac{1}{a_m} L^{(m)} = \bar{\ell}(\psi), \quad \mathbb{P}(\cdot | \Psi = \psi) \text{ a.s.}$$

- The proof uses a version of the law of large numbers for non-identically-distributed random variable by Petrov (1995).

Second result

- For one-factor Bernoulli mixture models we can obtain a stronger result which links the quantiles of $L^{(m)}$ to quantiles of the mixing distribution.

- Consider a sequence $L^{(m)} = \sum_{i=1}^m L_i$ satisfying the assumptions with a one-dimensional mixing variable Ψ with distribution function $G(\psi)$.
- Technical: assume that the conditional asymptotic loss function $\bar{\ell}(\psi)$ is *strictly increasing* and continuous and that G is strictly increasing at $q_\alpha(\Psi)$, i.e. that $G(q_\alpha(\Psi) + \delta) > \alpha$ for every $\delta > 0$.
- Then

$$\lim_{m \rightarrow \infty} \text{var}_\alpha \left(\frac{1}{a_m} L^{(m)} \right) = \bar{\ell}(q_\alpha(\Psi)). \quad (120)$$

- The proof is based on the following simple intuition. Since $L^{(m)}/a_m$ converges to $\bar{\ell}(\Psi)$ and since $\bar{\ell}$ is strictly increasing by assumption we have for large m

$$q_\alpha \left(\frac{L^{(m)}}{a_m} \right) \approx q_\alpha(\bar{\ell}(\Psi)) = \bar{\ell}(q_\alpha(\Psi)).$$

Application to exchangeable groups

- Consider the one-factor Bernoulli mixture model for k exchangeable groups defined by (115).
- Denote by $r(i)$ the group of obligor i and assume that, within each group r , the exposures, LGDs and conditional default probabilities are identical and given by e_r , δ_r and $p_r(\psi)$ respectively.
- Suppose that we allow the portfolio to grow and write $m_r^{(m)}$ for the number of obligors in group r when the portfolio size is m .
- The relative exposure to group r is given by

$$\lambda_r^{(m)} = e_r m_r^{(m)} / \sum_{r=1}^k e_r m_r^{(m)}$$

and we assume that $\lambda_r^{(m)} \rightarrow \lambda_r$ as $m \rightarrow \infty$.

- The asymptotic relative loss function is

$$\begin{aligned}\bar{\ell}(\psi) &= \lim_{m \rightarrow \infty} \sum_{i=1}^m \frac{e_{r(i)}}{\sum_{i=1}^m e_{r(i)}} \delta_{r(i)} p_{r(i)}(\psi) \\ &= \lim_{m \rightarrow \infty} \sum_{r=1}^k \frac{e_r m_r^{(m)}}{\sum_{r=1}^k e_r m_r^{(m)}} \delta_r h(\mu_r + \sigma\psi) = \sum_{r=1}^k \lambda_r \delta_r h(\mu_r + \sigma\psi).\end{aligned}$$

- Since Ψ is assumed to have a standard normal distribution, (120) implies that

$$\lim_{m \rightarrow \infty} q_\alpha \left(\frac{L^{(m)}}{\sum_{i=1}^m e_i} \right) = \sum_{r=1}^k \lambda_r \delta_r h(\mu_r + \sigma \Phi^{-1}(\alpha)). \quad (121)$$

- For large m , since $\lambda_r \sum_{i=1}^m e_i \approx m_r^{(m)} e_r$, we get that

$$\text{var}_\alpha(L^{(m)}) \approx \sum_{r=1}^k m_r^{(m)} e_r \delta_r h(\mu_r + \sigma_r \Phi^{-1}(\alpha)). \quad (122)$$

- This formula is the **basis of the Basel IRB formula**.

11.3.3 The Basel IRB formula

- Under the Basel framework a bank is required to hold 8% of the so-called *risk-weighted assets* (RWA) of its credit portfolio as risk capital.
- The RWA of a portfolio is given by the sum of the RWA of the individual risks in the portfolio, i.e. $\text{RWA}^{\text{portfolio}} = \sum_{i=1}^m \text{RWA}_i$.
- The quantity RWA_i reflects the exposure size and riskiness of obligor i and takes the form $\text{RWA}_i = w_i e_i$, where w_i is a risk weight and e_i denotes exposure size.
- Banks may choose between two options for determining the risk weight w_i , which must then be implemented for the entire portfolio. Under the simpler *standardized approach*, the risk weight w_i is determined by the type (sovereign, bank or corporate) and the credit rating of counterparty i .
- For instance, $w_i = 50\%$ for a corporation with a Moody's rating in the range of A+ to A-.

- Under the *internal-ratings-based* (IRB) approach, the risk weight is determined by a formula where inputs may be determined by the bank.
- The IRB risk weights takes the form

$$w_i = (0.08)^{-1} c \delta_i \Phi \left(\frac{\Phi^{-1}(p_i) + \sqrt{\beta_i} \Phi^{-1}(0.999)}{\sqrt{1 - \beta_i}} \right).$$

- Here c is a technical adjustment factor, p_i represents the default probability, and δ_i is the percentage loss given default of obligor i .
- $\beta_i \in (0.12, 0.24)$ measures the systematic risk of obligor i .
- Estimates for p_i and (under the so-called advanced IRB approach) for δ_i and e_i are provided by the individual bank.
- The adjustment factor c and, **most importantly**, the value of β_i are determined by fixed rules within the Basel II Accord.
- The risk capital to be held for counterparty i is thus given by

$$RC_i = 0.08 RWA_i = c \delta_i e_i \Phi \left(\frac{\Phi^{-1}(p_i) + \sqrt{\beta_i} \Phi^{-1}(0.999)}{\sqrt{1 - \beta_i}} \right). \quad (123)$$

11.4 Monte Carlo methods

11.4.1 Basics of importance sampling

In a generic Monte Carlo problem we have an rv X with density f and we wish to compute an [expected value](#) of the form

$$\theta = \mathbb{E}(h(X)) = \int_{-\infty}^{\infty} h(x)f(x) dx, \quad (124)$$

for some known function h . For an event probability $h(x) = I_{\{x \in A\}}$ for some set $A \subset \mathbb{R}$; for expected shortfall computation $h(x) = xI_{\{x \geq c\}}$ where $c = \text{VaR}_{\alpha}$.

Where analytical evaluation of θ is difficult we can use an [MC approach](#):

- 1) Simulate X_1, \dots, X_n independently from density f .
- 2) Compute the standard MC estimate $\hat{\theta}_n^{\text{MC}} = \frac{1}{n} \sum_{i=1}^n h(X_i)$.

- The MC estimator converges to θ by the strong law of large numbers, (SLLN) but the speed of convergence may not be particularly fast, particularly when we are dealing with rare event simulation.
- Suppose we want to estimate $ES_{0.99}(L)$ in credit risk context. Only 1% of our standard Monte Carlo draws will lead to a portfolio loss higher than $VaR_{0.99}(L)$. The standard MC estimator of, which consists of averaging the simulated values of L over all draws leading to a simulated portfolio loss $L \geq VaR_\alpha(L)$, will be unstable and subject to high variability, unless the number of simulations is very large.
- The technique of *importance sampling* is a way of reducing this variability and is well suited to problems of the kind we consider.
- Importance sampling is based on an alternative representation of θ in (124).

Importance sampling algorithm

- We consider an *importance sampling density* g (whose support should contain that of f) and define the *likelihood ratio* $r(x)$ by $r(x) := f(x)/g(x)$ whenever $g(x) > 0$ and $r(x) = 0$ otherwise.
- The integral may be written as

$$\theta = \int_{-\infty}^{\infty} h(x)r(x)g(x) dx = \mathbb{E}_g(h(X)r(X)), \quad (125)$$

where \mathbb{E}_g denotes expectation with respect to the density g .

We can approximate the integral with the following algorithm;

- 1) Simulate X_1, \dots, X_n independently from density g .
- 2) Compute the IS estimate $\hat{\theta}_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^n h(X_i) r(X_i)$.

Reducing the variance

- The art of importance sampling is in choosing g such that for fixed n the variance of the IS estimator is considerably smaller than that of the standard Monte Carlo estimator.

$$\begin{aligned}\text{var}_g(\hat{\theta}_n^{\text{IS}}) &= \frac{\mathbb{E}_g(h(X)^2 r(X)^2) - \theta^2}{n} = \frac{\mathbb{E}(h(X)^2 r(X)) - \theta^2}{n}, \\ \text{var}(\hat{\theta}_n^{\text{MC}}) &= \frac{\mathbb{E}(h(X)^2) - \theta^2}{n}.\end{aligned}$$

- The aim is to make $\mathbb{E}(h(X)^2 r(X))$ small compared to $\mathbb{E}(h(X)^2)$.
- Consider the case of estimating a tail probability where $h(x) = I_{\{x \geq c\}}$ for c significantly larger than the mean of X . We try to choose g so that the likelihood ratio $r(x) = f(x)/g(x)$ is small for $x \geq c$; in other words we make the event $\{X \geq c\}$ more likely under the IS density g than it is under the original density f .

- A technique for constructing importance sampling densities is known as *exponential tilting*.
- For $t \in \mathbb{R}$ write $M_X(t) = \mathbb{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$ for the moment generating function of X , which we assume is finite for $t \in \mathbb{R}$.
- We can define a density $g_t(x) = e^{tx} f(x)/M_X(t)$ which can be used for importance sampling when X is **light tailed**.
- The likelihood ratio is $r_t(x) = f(x)/g_t(x) = M_X(t)e^{-tx}$.
- Define μ_t to be the mean of X with respect to the density g_t i.e.

$$\mu_t := \mathbb{E}_{g_t}(X) = \mathbb{E}(X \exp(tX)/M_X(t)).$$

- How can we choose t optimally for a particular importance sampling problem?
- In the case of tail probability estimation theory suggests we should choose t as the solution of $\mu_t = c$.

Example 11.8 (Exponential tilting for normal distribution)

- We illustrate the concept of exponential tilting in the simple case of a standard normal random variable. Suppose $X \sim N(0, 1)$ with density $\phi(x)$.
- Using exponential tilting we obtain the new density

$$g_t(x) = \exp(tx)\phi(x)/M_X(t).$$

The moment generating function of X is known to be $M_X(t) = \exp(t^2/2)$.

- Hence

$$g_t(x) = \frac{1}{\sqrt{2\pi}} \exp\left(tx - \frac{1}{2}(t^2 + x^2)\right) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{(x-t)^2}{2}\right),$$

so that under the tilted distribution, $X \sim N(t, 1)$.

- Exponential tilting is a convenient way of shifting the mean of X .

An abstract view of importance sampling

To handle the more complex application to portfolio credit risk we consider importance sampling from a slightly more general viewpoint.

- Given densities f and g we define probability measures \mathbb{P} and \mathbb{Q} by

$$\mathbb{P}(A) = \int_A f(x) dx \quad \text{and} \quad \mathbb{Q}(A) = \int_A g(x) dx, \quad A \subset \mathbb{R}.$$

- With this notation (125) becomes

$$\theta = \mathbb{E}_{\mathbb{P}}(h(X)) = \mathbb{E}_{\mathbb{Q}}(h(X)r(X)),$$

so that $r(X)$ equals $d\mathbb{P}/d\mathbb{Q}$, the (measure-theoretic) density of \mathbb{P} with respect to \mathbb{Q} .

- Using this more abstract view, exponential tilting can be applied in more general situations.

- Given a rv X on (Ω, \mathcal{F}, P) such that $M_X(t) = \mathbb{E}_{\mathbb{P}}(\exp(tX)) < \infty$, define the measure \mathbb{Q}_t on (Ω, \mathcal{F}) by

$$\frac{d\mathbb{Q}_t}{d\mathbb{P}} = \frac{\exp(tX)}{M_X(t)}$$

and note that $(d\mathbb{Q}_t/d\mathbb{P})^{-1} = r_t(X)$.

- Event probabilities are now calculated according to

$$\mathbb{Q}_t(A) = \mathbb{E}_{\mathbb{Q}}(I_A) = \mathbb{E}_{\mathbb{P}}(r_t(X)^{-1} I_A) = \mathbb{E}_{\mathbb{P}}\left(\frac{\exp(tX)}{M_X(t)} I_A\right),$$

- The IS algorithm remains essentially unchanged: simulate independent realizations X_i under the measure \mathbb{Q}_t and set $\hat{\theta}^{\text{IS}} = \frac{1}{n} \sum_{i=1}^n X_i r_t(X_i)$ as before.

11.4.2 Application to Bernoulli-mixture models

- Consider a portfolio loss of the form $L = \sum_{i=1}^m e_i Y_i$, where the e_i are deterministic, positive exposures and the Y_i are default indicators with default probabilities p_i . \mathbf{Y} follows a Bernoulli mixture model with factor vector Ψ and conditional default probabilities $p_i(\Psi)$.
- We study the problem of estimating $\theta = \mathbb{P}(L \geq c)$ for c substantially larger than $\mathbb{E}(L)$ using importance sampling.
- We consider first the situation where the default indicators Y_1, \dots, Y_m are **independent** and discuss subsequently the extension to the case of **conditionally independent default indicators**.
- Here we use the more general IS approach and set $\Omega = \{0, 1\}^m$, the state space of \mathbf{Y} . The probability measure \mathbb{P} is given by

$$\mathbb{P}(Y_1 = y_1, \dots, Y_m = y_m) = \prod_{i=1}^m p_i^{y_i} (1 - p_i)^{1-y_i}, \quad (y_1, \dots, y_m) \in \Omega.$$

- The moment generating function of L is

$$M_L(t) = \mathbb{E} \left(\exp \left(t \sum_{i=1}^m e_i Y_i \right) \right) = \prod_{i=1}^m \mathbb{E} \left(e^{t e_i Y_i} \right) = \prod_{i=1}^m \left(e^{t e_i p_i} + 1 - p_i \right).$$

- Under the new measure \mathbb{Q}_t we have

$$\begin{aligned} \mathbb{Q}_t(Y_1 = y_1, \dots, Y_m = y_m) &= \mathbb{E}_{\mathbb{P}} \left(\frac{e^{tL}}{M_L(t)} I_{\{Y_1=y_1, \dots, Y_m=y_m\}} \right) \\ &= \frac{e^{(t \sum_{i=1}^m e_i y_i)}}{M_L(t)} \mathbb{P}(Y_1 = y_1, \dots, Y_m = y_m). \end{aligned}$$

- We obtain

$$= \prod_{i=1}^m \frac{\exp(te_i y_i)}{\exp(te_i)p_i + 1 - p_i} p_i^{y_i} (1 - p_i)^{1 - y_i}.$$

- Define $q_{t,i} := \exp(te_i)p_i / (\exp(te_i)p_i + 1 - p_i)$.

- It follows that

$$\mathbb{Q}_t(Y_1 = y_1, \dots, Y_m = y_m) = \prod_{i=1}^m q_{t,i}^{y_i} (1 - q_{t,i})^{1-y_i}$$

- We see that the default indicators remain independent but with new default probability $q_{t,i}$.
- The optimal value of t is chosen such that $\mathbb{E}_{\mathbb{Q}_t}(L) = c$, leading to the equation $\sum_{i=1}^m e_i q_{t,i} = c$.

Conditionally independent defaults

- The first step is obvious: given a realization ψ of the economic factors, the conditional exceedance probability $\theta(\psi) := \mathbb{P}(L \geq c \mid \Psi = \psi)$ is estimated using the approach for independent default indicators.
- This gives an estimate $\hat{\theta}_{n_1}^{\text{IS},1}(\psi)$ where n_1 is the number of random draws of (Y_1, \dots, Y_m) .

- Our aim is to estimate

$$\theta = \mathbb{P}(L \geq c) = \mathbb{E}(\mathbb{P}(L \geq c | \Psi)) = \mathbb{E}(\theta(\Psi)).$$

- In a naive approach we could generate n realizations of Ψ and estimate θ by calculating the average $\frac{1}{n} \sum_{i=1}^n \hat{\theta}_{n_1}^{\text{IS},1}(\Psi_i)$.
- However a dramatic improvement can be obtained by also [applying importance sampling to the distribution of \$\Psi\$](#) .
- We will illustrate this idea with a one-factor example.
- Consider the one-factor Gaussian threshold model with conditional default probabilities $p_i(\Psi)$ where $\Psi \sim N(0, 1)$.
- Instead of generating Ψ_1, \dots, Ψ_n from a standard normal $N(0, 1)$ distribution we should use exponential tilting to generate them from a $N(\mu, 1)$ distribution for some sensibly chosen value of μ .
- Note that we will not discuss the optimal choice of μ ; see Glasserman and Li (2005) for more details.

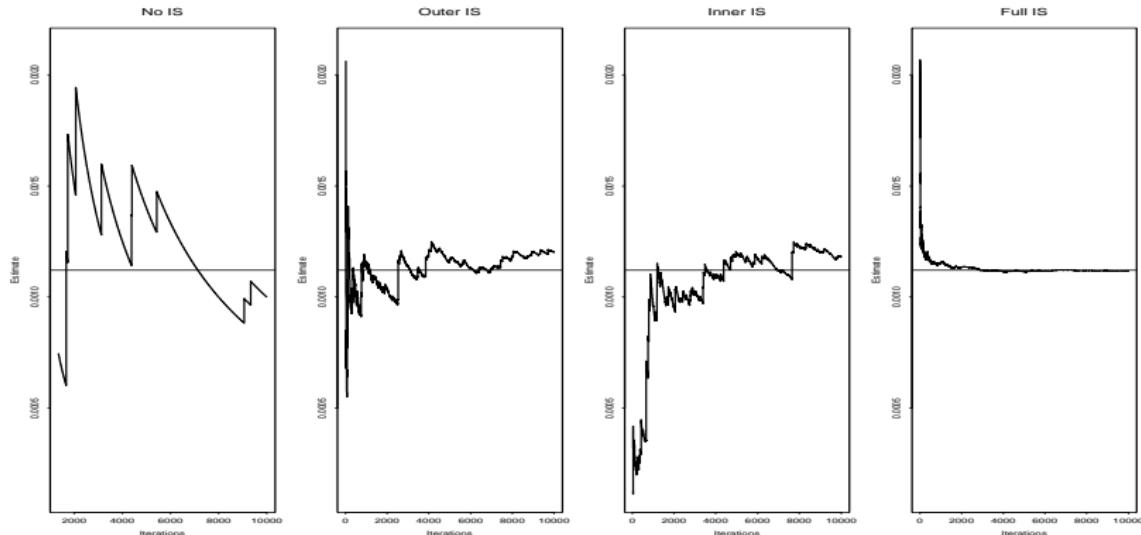
We obtain the algorithm:

- 1) Generate $\Psi_1, \dots, \Psi_n \sim N(\mu, 1)$ independently.
- 2) For each Ψ_i calculate $\hat{\theta}_{n_1}^{\text{IS},1}(\Psi_i)$ by importance sampling.
- 3) Determine the full IS estimator: $\hat{\theta}_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^n r_\mu(\Psi_i) \hat{\theta}_n^{\text{IS},1}(\Psi_i)$ where $r_\mu(\psi) = \exp(-\mu\psi + \frac{1}{2}\mu^2)$.

Example 11.9 (Exchangeable portfolio)

- We consider an exchangeable portfolio of 100 firms with identical unit exposures, default probabilities 0.05 and asset correlations (i.e. values of ρ) 0.05.
- Aim is to calculate the tail probability $\mathbb{P}(L \geq 20)$ by IS.
- In such a simple model it can in fact be calculated analytically to be 0.00112.
- We compare the following MC/IS methods:
 - 1) Naive Monte Carlo ($n = 10000$) (No IS)

- 2) IS for factor distribution ($n = 10000$) ([outer IS](#))
- 3) Naive Monte Carlo for factor ($n = 10000$) and IS for conditional default distribution ($n1 = 50$) ([inner IS](#))
- 4) IS for factor distribution ($n = 10000$) and conditional default distribution ($n1 = 50$) ([Full IS](#))



11.5 Statistical inference for portfolio credit models

11.5.1 Industry factor models

- Recall that portfolio models in industry often take the form of a Gaussian threshold model (\mathbf{X}, \mathbf{d}) with $\mathbf{X} \sim N_m(\mathbf{0}, P)$, where the random vector \mathbf{X} contains the critical variables, the deterministic vector \mathbf{d} contains the critical default thresholds and P is the so-called asset correlation matrix, which is estimated with the help of a factor model for \mathbf{X} .
- Industry models generally separate the calibration of the vector \mathbf{d} (or the threshold matrix D in a multi-state model) and the calibration of the factor model for \mathbf{X} .
- In a default-only model the threshold d_i is usually set at $d_i = \Phi^{-1}(p_i)$ where p_i is an estimate of the default probability for obligor i for the time period in question (generally one year).

- The default probability may be estimated in different ways: for larger corporates it may be estimated using credit ratings or using a firm-value approach, such as the Moody's public-firm EDF model; for retail obligors it may be estimated on the basis of credit scores.
- Recall that the **factor model** for \mathbf{X} takes the form

$$X_i = \sqrt{\beta_i} \tilde{F}_i + \sqrt{1 - \beta_i} \varepsilon_i, \quad i = 1, \dots, m, \quad (126)$$

where \tilde{F}_i and $\varepsilon_1, \dots, \varepsilon_m$ are independent standard normal variables, and where $0 \leq \beta_i \leq 1$ for all i .

- The systematic variables \tilde{F}_i are assumed to be of the form $\tilde{F}_i = \mathbf{a}'_i \mathbf{F}$ where \mathbf{F} is a vector of common factors satisfying $\mathbf{F} \sim N_p(\mathbf{0}, \Omega)$ with $p < m$, and where Ω is a correlation matrix.
- The factors typically represent country and industry effects.**
- The assumption that $\text{var}(\tilde{F}_i) = 1$ implies that $\mathbf{a}'_i \Omega \mathbf{a}_i = 1$ for all i .

- Different industry models use different data for \mathbf{X} to calibrate the factor model (126).
- The Moody's Analytics Global Correlation or GCorr model has sub-models for many different kinds of obligor including public corporate firms, private firms, small and medium enterprises (SMEs), retail customers and sovereigns. Huang et al. (2012)
- The sub-model for public firms (GCorr Corporate) is calibrated using data on weekly asset value returns, where asset values are determined as part of the public-firm EDF methodology.
- In the CreditMetrics framework weekly equity returns are viewed as a proxy for asset returns and used to estimate the factor model.
- We sketch a generic procedure for estimating a factor model for corporates where the factors have country and industry-sector interpretations.

Estimating a credit risk factor model

- We assume that we have a **high-dimensional multivariate time series** $(\mathbf{X}_t)_{1 \leq t \leq n}$ of asset returns (or other proxy data for changing credit quality) over a period of time in which **stationarity** can be assumed.
 - We also assume that each component time series has been **scaled to have mean zero and variance one**.
- 1) We first fix the structure of the factor vector \mathbf{F} so that, for example, the first block of components might represent country factors and the second block of components might represent industry factors. We then assign vectors of factor weights a_i to each obligor based on our knowledge of the companies. The elements of a_i may simply consist of ones and zeros if the company can be clearly identified with a single country and industry, but may also consist of weights if the company has significant activity in more than one country or sector.

2) We then use cross-sectional estimation techniques to estimate the factor values \mathbf{F}_t at each time point t . Effectively the factor estimates $\hat{\mathbf{F}}_t$ are constructed as weighted sums of the $X_{t,i}$ data for obligors i that are exposed to each factor. One way of achieving this is to construct a matrix A with rows a_i and then to estimate a fundamental factor model of the form $\mathbf{X}_t = A\mathbf{F}_t + \varepsilon_t$ at each time point t .

- We have a regression model

$$\mathbf{X}_t = A\mathbf{F}_t + \varepsilon_t, \quad (127)$$

where $\mathbf{X}_t \in \mathbb{R}^m$ are the return data, $A \in \mathbb{R}^{m \times p}$ is a known matrix of factor loadings, $\mathbf{F}_t \in \mathbb{R}^p$ are the factors to be estimated and ε_t are errors with diagonal covariance matrix Υ .

- Note that the components of the error vector ε_t can not generally be assumed to have equal variance, so that (127) is a regression problem with so-called heteroskedastic errors.

- Unbiased estimators of the factors \mathbf{F}_t may be obtained by forming the ordinary least squares (OLS) estimates

$$\hat{\mathbf{F}}_t^{\text{OLS}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}_t.$$

- Since the errors are heteroskedastic, slightly more efficient estimators can be obtained by using the method of generalized least squares (GLS).

- 3) The raw factor estimates form a multivariate time series of dimension p . We standardize each component series to have mean zero and variance one to obtain $(\hat{\mathbf{F}}_t)_{1 \leq t \leq n}$ and calculate the sample covariance matrix of the standardized factor estimates, which serves as our estimate of Ω .
- 4) We then scale the vectors of factor weights \mathbf{a}_i so that the conditions $\mathbf{a}_i'\hat{\Omega}\mathbf{a}_i = 1$ are met for each obligor.
- 5) Time series of estimated systematic variables for each obligor are then constructed by calculating $\hat{\tilde{\mathbf{F}}}_{t,i} = \mathbf{a}_i'\hat{\mathbf{F}}_t$ for $t = 1, \dots, n$.

6) Finally we estimate the β_i parameters by performing a time series regression of $X_{t,i}$ on $\hat{F}_{t,i}$ for each obligor.

Note that the accurate estimation of the β_i in the last step is particularly important (as it effects tail behaviour). The estimate of β_i is the so-called R-squared of the time series regression model in Step 6 and will be largest for the firms whose credit-quality changes are best explained by systematic factors.

11.5.2 Exchangeable Bernoulli-mixture models

- We discuss the estimation of default probabilities and default correlations for homogeneous groups, e.g. groups with the same credit rating.
- Suppose that we observe historical default numbers over n periods of time for a homogeneous group; typically these might be yearly data.
- For $t = 1, \dots, n$, let m_t denote the number of observed companies at the start of period t .

- Let M_t denote the number that defaulted during the period; the former will be treated as fixed at start of each period and the latter as an rv.
- Suppose further that within a time period these defaults are generated by an [exchangeable Bernoulli mixture model](#).
- In other words, assume that, given some mixing variable Q_t taking values in $(0, 1)$ and the cohort size m_t , the number of defaults M_t is conditionally binomially distributed and satisfies $M_t | Q_t = q \sim B(m_t, q)$.
- Further assume that the mixing variables Q_1, \dots, Q_n are identically distributed.
- We consider using a simple [method of moments](#) to estimate the fundamental parameters of the mixing distribution $\pi = \pi_1, \pi_2$ and ρ_Y (default correlation).
- For $1 \leq t \leq n$, let $Y_{t,1}, \dots, Y_{t,m_t}$ be default indicators for the m_t

companies in the cohort. Suppose we define the rv

$$Z_{t,k} := \sum_{\{i_1, \dots, i_k\} \subset \{1, \dots, m_t\}} Y_{t,i_1} \cdots Y_{t,i_k} \quad (128)$$

which represents the number of possible subgroups of k obligors among the defaulting obligors in period t .

- By taking expectations in (128) we get

$$\mathbb{E}(Z_{t,k}) = \binom{m_t}{k} \pi_k.$$

- Note that we can write

$$Z_{t,k} = \binom{M_t}{k} I_{\{M_t \geq k\}},$$

the number of subgroups of size k among the defaulting obligors.

- We can estimate the unknown theoretical moment π_k by taking a natural

empirical average (129) constructed from the n years of data:

$$\hat{\pi}_k = \frac{1}{n} \sum_{t=1}^n \frac{Z_{t,k}}{\binom{m_t}{k}} = \frac{1}{n} \sum_{t=1}^n \frac{M_t(M_t - 1) \cdots (M_t - k + 1)}{m_t(m_t - 1) \cdots (m_t - k + 1)}. \quad (129)$$

- For $k = 1$ we get the standard estimator of default probability

$$\hat{\pi} = \frac{1}{n} \sum_{t=1}^n \frac{M_t}{m_t},$$

and ρ_Y can obviously be estimated by taking

$$\hat{\rho}_Y = \frac{\hat{\pi}_2 - \hat{\pi}^2}{\hat{\pi} - \hat{\pi}^2}.$$

- The estimator is unbiased for π_k and consistent as $n \rightarrow \infty$.
- For more details see Frey and McNeil (2001).

14 Multivariate time series

14.1 Fundamentals of multivariate time series

14.2 Multivariate GARCH Processes

14.1 Fundamentals of multivariate time series

14.1.1 Basic definitions

Definition 14.1

The mean function $\mu(t)$ and the covariance matrix function $\Gamma(t + h, t)$ of $(\mathbf{X}_t)_{t \in \mathbb{Z}}$ are given by

$$\mu(t) = \mathbb{E}(\mathbf{X}_t), \quad t \in \mathbb{Z},$$

$$\Gamma(t + h, t) = \mathbb{E}((\mathbf{X}_{t+h} - \mu(t + h))(\mathbf{X}_t - \mu(t))'), \quad t, h \in \mathbb{Z}.$$

- Analogously to the univariate case, we have $\Gamma(t, t) = \text{cov}(\mathbf{X}_t)$. By observing that the elements $\gamma_{ij}(t + h, t)$ of $\Gamma(t + h, t)$ satisfy

$$\gamma_{ij}(t + h, t) = \text{cov}(X_{t+h,i}, X_{t,j}) = \text{cov}(X_{t,j}, X_{t+h,i}) = \gamma_{ji}(t, t + h),$$

it is clear that $\Gamma(t + h, t) = \Gamma(t, t + h)'$ for all t, h .

- However, the matrix Γ **need not be symmetric**, so in general $\Gamma(t + h, t) \neq \Gamma(t, t + h)$. **One series can lead other series.**

Definition 14.2 (strict stationarity)

The multivariate time series $(\mathbf{X}_t)_{t \in \mathbb{Z}}$ is **strictly stationary** if

$$(\mathbf{X}'_{t_1}, \dots, \mathbf{X}'_{t_n}) \stackrel{\text{d}}{=} (\mathbf{X}'_{t_1+k}, \dots, \mathbf{X}'_{t_n+k}),$$

for all $t_1, \dots, t_n, k \in \mathbb{Z}$ and for all $n \in \mathbb{N}$.

Definition 14.3 (covariance (weak, second-order) stationarity)

The multivariate time series $(\mathbf{X}_t)_{t \in \mathbb{Z}}$ is **covariance stationary** if the first two moments exist and satisfy

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

$$\Gamma(t+h, t) = \Gamma(h, 0), \quad t, h \in \mathbb{Z}.$$

- For a covariance-stationary process we write $\Gamma(h) := \Gamma(h, 0)$.
- Note that $\Gamma(0) = \text{cov}(\mathbf{X}_t)$, for all t .

- Write Δ for the diagonal matrix whose entries are the square roots of the diagonal entries of $\Gamma(0)$ (standard deviations of component series).

Definition 14.4 (correlation matrix function)

The correlation matrix function $P(h)$ of a covariance-stationary multivariate time series is

$$P(h) = \Delta^{-1} \Gamma(h) \Delta^{-1}, \quad \forall h \in \mathbb{Z}. \quad (130)$$

- The diagonal entries $\rho_{ii}(h)$ of this matrix-valued function give the **autocorrelation function of the i th component series** ($X_{t,i}$) $_{t \in \mathbb{Z}}$.
- The off-diagonal entries give so-called **cross-correlations between different component series** at different times.

Definition 14.5 (multivariate white noise)

$(\mathbf{X}_t)_{t \in \mathbb{Z}}$ is multivariate white noise if it is covariance stationary with correlation matrix function given by

$$P(h) = \begin{cases} P, & h = 0, \\ 0, & h \neq 0, \end{cases}$$

for some positive-definite correlation matrix P .

Such a process has [no cross-correlation between component series](#), except for contemporaneous cross-correlation at lag zero.

Definition 14.6 (multivariate strict white noise)

$(\mathbf{X}_t)_{t \in \mathbb{Z}}$ is multivariate strict white noise if it is a series of [iid random vectors](#).

A strict white noise process with mean zero and covariance matrix Σ will be denoted $\text{SWN}(\mathbf{0}, \Sigma)$.

14.1.2 Analysis in the time domain

- Assume we have a random sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ from a covariance-stationary multivariate time series model $(\mathbf{X}_t)_{t \in \mathbb{Z}}$.
- In the time domain we construct empirical estimators of the covariance matrix function and the correlation matrix function.
- The *sample covariance matrix function* is calculated according to

$$\hat{\Gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (\mathbf{X}_{t+h} - \bar{\mathbf{X}})(\mathbf{X}_t - \bar{\mathbf{X}})', \quad 0 \leq h < n,$$

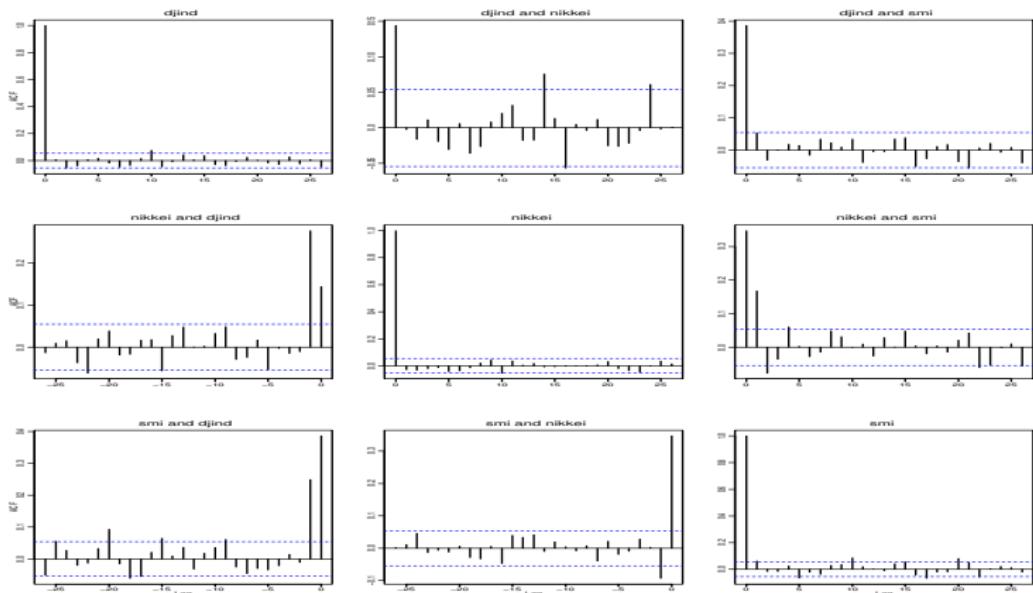
where $\bar{\mathbf{X}} = \sum_{t=1}^n \mathbf{X}_t / n$ is the sample mean.

- Writing $\hat{\Delta}$, for the diagonal matrix of sample standard deviations (square root of the diagonal of $\hat{\Gamma}(0)$) the *sample correlation matrix function* is

$$\hat{P}(h) = \hat{\Delta}^{-1} \hat{\Gamma}(h) \hat{\Delta}^{-1}, \quad 0 \leq h < n.$$

- The information contained in the elements $\hat{\rho}_{ij}(h)$ of the sample correlation matrix function is generally displayed in the *cross-correlogram*.

Cross-correlogram of index returns



The US market leads Europe and Japan.

14.1.3 Multivariate ARMA processes

- ARMA models extend to higher dimensions where they are called VARMA. They provide **models for the conditional mean vector**.
- The VAR class is most widely used in practice.
- The first-order VAR process satisfies the set of equations

$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \forall t. \tag{131}$$

where $\Phi \in \mathbb{R}^{d \times d}$ is a matrix and $(\boldsymbol{\varepsilon}_t)$ is a white noise process.

- The process is covariance stationary if and only if all eigenvalues of the matrix Φ are less than one in absolute value.
- The covariance matrix function of this process is

$$\Gamma(h) = \Phi^h \Gamma(0), \quad h = 0, 1, 2, \dots$$

14.2 Multivariate GARCH Processes

Recall that the Cholesky factor A of a positive-definite matrix Σ is the lower-triangular matrix satisfying $AA' = \Sigma$.

Definition 14.7

Let $(\mathbf{Z}_t)_{t \in \mathbb{Z}}$ be SWN($\mathbf{0}, I_d$). The process $(\mathbf{X}_t)_{t \in \mathbb{Z}}$ is said to be a multivariate GARCH process if it is strictly stationary and satisfies equations of the form

$$\mathbf{X}_t = A_t \mathbf{Z}_t, \quad t \in \mathbb{Z}, \tag{132}$$

where $A_t \in \mathbb{R}^{d \times d}$ is the Cholesky factor of a positive-definite matrix Σ_t which is measurable with respect to $\mathcal{F}_{t-1} = \sigma(\{\mathbf{X}_s : s \leq t-1\})$, the history of the process up to time $t-1$.

Conditional moments:

- $\mathbb{E}(\mathbf{X}_t | \mathcal{F}_{t-1}) = \mathbf{0}$

- $\text{cov}(\mathbf{X}_t \mid \mathcal{F}_{t-1}) = A_t A_t' = \Sigma_t$ is the **conditional covariance matrix**.
- Could add a non-zero conditional mean term $\boldsymbol{\mu}_t$ so that $\mathbf{X}_t = \boldsymbol{\mu}_t + A_t \mathbf{Z}_t$ where, for example,
 - ▶ $\boldsymbol{\mu}_t = \boldsymbol{\mu}$ for a constant conditional mean;
 - ▶ or $\boldsymbol{\mu}_t$ could follow a VARMA specification, such as $\boldsymbol{\mu}_t = \Phi \mathbf{X}_{t-1}$.
- We can write $\Sigma_t = \Delta_t P_t \Delta_t$, where Δ_t is the diagonal **volatility matrix** and P_t is the **conditional correlation matrix**.
- The art of building multivariate GARCH models is to specify the dependence of Σ_t (or of Δ_t and P_t) on the past in such a way that Σ_t always remains **symmetric and positive definite**.
- The innovations are generally taken to be from either a multivariate Gaussian distribution ($\mathbf{Z}_t \sim N_d(\mathbf{0}, I_d)$) or an appropriately scaled **spherical** multivariate t distribution ($\mathbf{Z}_t \sim t_d(\nu, \mathbf{0}, (\nu - 2)I_d/\nu)$). Any distribution with mean zero and covariance matrix I_d is permissible.

14.2.1 Models for conditional correlation

Definition 14.8

The process $(\mathbf{X}_t)_{t \in \mathbb{Z}}$ is a CCC-GARCH process if it is a multivariate GARCH process with conditional covariance matrix of the form $\Sigma_t = \Delta_t P_c \Delta_t$, where

- P_c is a constant, positive-definite correlation matrix; and
- Δ_t is a diagonal volatility matrix with elements $\sigma_{t,k}$ satisfying

$$\sigma_{t,k}^2 = \alpha_{k0} + \sum_{i=1}^{p_k} \alpha_{ki} X_{t-i,k}^2 + \sum_{j=1}^{q_k} \beta_{kj} \sigma_{t-j,k}^2, \quad k = 1, \dots, d, \quad (133)$$

where $\alpha_{k0} > 0$, $\alpha_{ki} \geq 0$, $i = 1, \dots, p_k$, $\beta_{kj} \geq 0$, $j = 1, \dots, q_k$.

- Alternatives to ordinary GARCH(p_k, q_k) model may of course be used.
- In the CCC GARCH model the process $\mathbf{Y}_t = \Delta_t^{-1} \mathbf{X}_t$ (known as the **de-volatilized process**) satisfies $(\mathbf{Y}_t)_{t \in \mathbb{Z}} \sim \text{SWN}(\mathbf{0}, P_c)$.

- Estimation can be accomplished in **two stages**:
 - 1) Fit univariate GARCH models to each component series;
 - 2) Form residuals $\hat{\mathbf{Y}}_t = \hat{\Delta}_t^{-1} \mathbf{X}_t$, for $t = 1, \dots, n$ and estimate P_c (either by using the standard correlation estimator or by fitting an appropriate distribution).
- Alternatively all parameters can be maximized in **one step**.
- The CCC model is often a **useful starting point** from which to proceed to more complex models.
- In some empirical settings it gives an adequate performance, but it is generally considered that the constancy of conditional correlation in this model is **an unrealistic feature** and that the impact of news on financial markets requires models that allow a **dynamic evolution of conditional correlation as well as a dynamic evolution of volatilities**.

Definition 14.9

The process $(\mathbf{X}_t)_{t \in \mathbb{Z}}$ is a DCC-GARCH process if it is a multivariate GARCH process where the volatilities comprising Δ_t follow univariate GARCH specifications as in (133) and the conditional correlation matrices P_t satisfy, for $t \in \mathbb{Z}$, the equations

$$P_t = \wp \left(\left(1 - \sum_{i=1}^p \alpha_i - \sum_{j=1}^q \beta_j \right) P_c + \sum_{i=1}^p \alpha_i \mathbf{Y}_{t-i} \mathbf{Y}'_{t-i} + \sum_{j=1}^q \beta_j P_{t-j} \right), \quad (134)$$

where

- P_c is a positive-definite correlation matrix,
- \wp is the operator that extracts correlation matrices from covariance matrices,
- $\mathbf{Y}_t = \Delta_t^{-1} \mathbf{X}_t$ denotes the devolatilized process,
- and the coefficients satisfy $\alpha_i \geq 0$, $\beta_j \geq 0$ and $\sum_{i=1}^p \alpha_i + \sum_{j=1}^q \beta_j < 1$.

- If all the α_i and β_j coefficients (134) are zero, model reduces to CCC.
- In a covariance-stationary **univariate GARCH model** with unconditional variance σ^2 , the volatility equation can be written

$$\sigma_t^2 = \left(1 - \sum_{i=1}^p \alpha_i - \sum_{j=1}^q \beta_j\right) \sigma^2 + \sum_{i=1}^p \alpha_i X_{t-i} + \sum_{j=1}^q \beta_j \sigma_{t-j}^2.$$

- Thus, in DCC, the correlation matrix P_c in (134) can be thought of as representing the **long-run correlation structure**.
- The usual estimation method for the DCC model is as follows.
 - 1) Fit **univariate GARCH-type models** to the component series to estimate the volatility matrix Δ_t . Form an estimated realization of the devolatilized process by taking $\hat{Y}_t = \hat{\Delta}_t^{-1} \mathbf{X}_t$.
 - 2) Estimate P_c by estimating correlation matrix of the devolatilized data.
 - 3) Estimate the remaining parameters α_i and β_j in equation (134) by fitting the implied dynamic model to the devolatilized data (\hat{Y}_t).

Consider a first-order model ($p = q = 1$):

- Given \mathcal{F}_{t-1} (comprising $\mathbf{Y}_{t-k}, P_{t-k}, k = 1, 2, \dots$) and using an estimate of P_c (known as **variance targeting**), we have

$$\mathbf{Y}_t = B_t \mathbf{Z}_t, \quad \text{where}$$

$$B_t B_t' = P_t, \quad (\text{Cholesky decomposition})$$

$$P_t = \wp(Q_t), \quad (\text{correlation from covariance})$$

$$Q_t = (1 - \alpha_1 - \beta_1)P_c + \alpha_1 \mathbf{Y}_{t-1} \mathbf{Y}_{t-1}' + \beta_1 P_{t-1}$$

- Usually estimated by **conditional maximum likelihood**.
- The likelihood is built up recursively from **starting values** (for example $P_0 = \mathbf{Y}_0 \mathbf{Y}_0' = P_c$).
- There are two parameters to estimate for dynamics, in addition to parameters of innovation distribution (if non-Gaussian).

Relationship to dynamic copula models:

- In terms of **copulas**, using a Gaussian innovation distribution means estimating a 2-parameter model where

$$\mathbf{Y}_t \mid \mathcal{F}_{t-1} \sim C_{P_t}^{\text{G}\alpha}(\Phi, \dots, \Phi).$$

- Using a Student innovation distribution means estimating a 3-parameter model where

$$\mathbf{Y}_t \mid \mathcal{F}_{t-1} \sim C_{\nu, P_t}^{\text{t}}(F_\nu, \dots, F_\nu)$$

where F_ν is a scaled Student t distribution.

Copula-MGARCH models

- Note that models of the form

$$\mathbf{Y}_t \mid \mathcal{F}_{t-1} \sim C_{\nu, P_t^*}^{\text{t}}(F_{\nu_1}, \dots, F_{\nu_d})$$

with P_t^* updating as in (134) have also been considered.

This has $d + 3$ parameters.

- Previous model **doesn't quite fit into the DCC class** as we have defined it because $\text{cov}(\mathbf{Y}_t \mid \mathcal{F}_{t-1}) = P_t \neq P_t^*$.
- It is **not the parameters of the conditional correlation matrix** but rather the **parameters of the copula** that update according to (134).
- However it fits into a bigger class of **copula-MGARCH models** where

$$\mathbf{X}_t = \boldsymbol{\mu}_t + \Delta_t \mathbf{Y}_t, \quad \mathbf{Y}_t \mid \mathcal{F}_{t-1} \sim C_t(F_1, \dots, F_d)$$

and

- ▶ the volatility components of Δ_t follow GARCH schemes;
- ▶ the conditional mean terms $\boldsymbol{\mu}_t$ follow VARMA schemes;
- ▶ the conditional copula C_t evolves as function of information in \mathcal{F}_{t-1} ;
- ▶ F_1, \dots, F_d are zero-mean, unit-variance distributions.
- See Patton (2006), Patton (2012), and Fan and Patton (2014)

14.2.2 Dimension reduction in MGARCH

- While the multi-stage estimation procedure for DCC makes it possible to estimate in quite high dimensions, it is usual to first apply dimension reduction through factor modelling and then fit MGARCH models to the most important factors.
- Can easily fit MGARCH models to factors derived from macroeconomic and fundamental factor models. The factors are typically correlated.
- The use of so-called PC-GARCH (principal components GARCH) is quite popular and avoids need for multivariate models.
 - ▶ Here we assume that the principal components of X_t follow a CCC model with $P = I_d$.
 - ▶ To estimate such a model, we estimate the principal components of the data and fit univariate GARCH models to each principal components series.

17 Introduction to counterparty risk

17.1 Introduction

17.2 Credit value adjustments

17.1 Introduction

- A substantial part of all derivative transactions is carried out over the counter and there is no central clearing counterparty to guarantee fulfilment of the contractual obligations.
- These trades are subject to the risk that a contracting party defaults during the transaction, thus affecting the cash flows that are actually received by the other party. This is known as **counterparty credit risk**.
- Counterparty risk received a lot of attention during the financial crisis of 2007-2009 as some of the institutions heavily involved in derivative transactions experienced worsening credit quality or—in the case of Lehman Brothers—even a default event.
- **Counterparty risk management** is now a key issue for all financial institutions and the focus of many new regulatory developments.

Example of Interest-Rate Swap

- Two parties A and B agree to exchange a series of interest payments on a given nominal amount of money for a given period.
- A receives payments at a fixed interest rate and makes floating payments at a rate equal to the three-month LIBOR rate.
- Suppose that A defaults at time τ_A before the maturity of the contract.
- If interest rates have risen relative to their value at inception of contract:
 - ▶ The fixed interest payments have decreased in value and the value of the contract has increased for B .
 - ▶ The default of A constitutes a loss for B ; the loss size depends on the term structure of interest rates at τ_A .
- If interest rates have fallen relative to their value at $t = 0$:
 - ▶ The fixed payments have increased in value so that the swap has a negative value for B .

- ▶ B will still have to pay the value of the contract into the bankruptcy pool, and there is **no upside for B** in A 's default.
- If B defaults first the situation is **reversed**: falling rates lead to a counterparty-risk-related loss for A .

Management of counterparty risk

- Counterparty risk has to be taken into account in pricing and valuation. This has led to the notion of **credit value adjustments (CVA)**.
- Counterparty risk needs to be controlled using risk-mitigation techniques such as **netting** and **collateralization**.
- Under a **netting agreement** the value of all derivatives transactions between A and B is computed and only the aggregated value is subject to counterparty risk; since offsetting transactions cancel each other out, this has the potential to reduce counterparty risk substantially.

- Under a **collateralization agreement** the parties exchange collateral (cash and securities) that serves as a pledge for the receiver. The value of the collateral is adjusted dynamically to reflect changes in the value of the underlying transactions.

17.2 Credit value adjustments

General definition. The price (for the protection buyer) satisfies

$$\text{True price} = \text{(counterparty) risk-free price}$$
$$- \text{adjustment for default of seller (CVA)}$$
$$+ \text{adjustment for default of buyer (DVA)} ,$$

where CVA and DVA stand for Credit Value Adjustment and Debt Value Adjustment respectively.

General adjustment formulas

Denote by V_t the market value of the CDS (assuming that B and S are default-free), by $\tau = \min\{\tau_R, \tau_S, \tau_B\}$ the first default time and by $\xi \in \{R, S, B\}$ the identity of first defaulting firm. Recall that $x^+ = \max(x, 0)$ and $x^- = -\min(x, 0)$ and denote by $D(0, t)$ the discount factor over the period $[0, t]$ (with constant interest rate, $D(0, t) = e^{-rt}$).

It can be shown that

$$\text{CVA} = \mathbb{E}^{\mathbb{Q}}(I_{\{\tau < T\}} I_{\{\xi=S\}} D(0, \tau) \delta^S V_{\tau}^+)$$

$$\text{DVA} = \mathbb{E}^{\mathbb{Q}}(I_{\{\tau < T\}} I_{\{\xi=B\}} D(0, \tau) \delta^B V_{\tau}^-)$$

Comments.

- CVA gives loss of B due to premature default of S; DVA gives loss of S due to premature default of B.
- The value adjustments involve an option on the market value $V = (V_t)_{t \leq T}$ of the swap with strike $K = 0$ (a call for the CVA and a put for the DVA).
- Similar formula holds if V is the market value of another derivative such as an interest swap or even a reinsurance contract.
- DVA is a bit problematic: a worsening credit quality of B leads to an accounting profit for B.

A simplified formula

In order to evaluate the CVA and DVA formulas one needs a model with stochastic credit spreads that takes **dependence** between the default of S, B and the **market value V** of the CDS into account (a **dynamic portfolio credit risk model**). Markets often work with a simpler formula that assumes that the default of S and B and V are **independent**:

$$\text{CVA}^{\text{indep}} = \delta^S \int_0^T \bar{F}_B(t) D(0, t) \mathbb{E}^{\mathbb{Q}}(V_t^+) f_S(t) dt,$$

$$\text{DVA}^{\text{indep}} = \delta^B \int_0^T \bar{F}_S(t) D(0, t) \mathbb{E}^{\mathbb{Q}}(V_t^-) f_B(t) dt.$$

Here f_S is the density of τ_S and \bar{F}_B resp \bar{F}_S is the survival function of τ_B resp τ_S .

Comments.

- In order to evaluate the simplified formula one only needs to determine the marginal distribution of τ_S and τ_B and the so-called **expected exposures** $\mathbb{E}^{\mathbb{Q}}(V_t^+)$ and $\mathbb{E}^{\mathbb{Q}}(V_t^-)$.
- The independence assumption underlying the simplified value adjustment formula between the price of the CDS on R , that is V_t , and the default event of S and B is often unrealistic; in practice this is known as **wrong way risk**.

Examples:

- ▶ CDS on a financial institution: given that S defaults it is quite likely that credit quality of R is low.
- ▶ Reinsurance.

For further reading on counterparty risk see Gregory (2012).

References

- Acharya, B V., Cooley, V. V., Richardson, M., and Walter, I. (2009), Manufacturing tail risk: A perspective on the financial crisis of 2007–2009, *Foubdations and Trends in Finance*, 4(4), 247–325.
- Artzner, P., Delbaen, F., Eber, J. M., and Heath, D. (1999), Coherent measures of risk, *Mathematical Finance*, 9, 203–228.
- Balkema, A. A. and de Haan, L. (1974), Residual life time at great age, *The Annals of Probability*, 2, 792–804.
- Basel Committee on Banking Supervision (2004), Basel II: International convergence of capital measurement and capital standards: A revised framework, Publication No. (107), Bank of International Settlements.
- Bellini, F. and Bignozzi, V. (2013), Elicitable risk measures, Working paper, available at SSRN: <http://ssrn.com/abstract=2334746>.

- Bernard, Carole, Jiang, Xiao, and Wang, Ruodu (2014), Risk aggregation with dependence uncertainty, *Insurance: Mathematics and Economics*, 54, 93–108.
- 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
Chicago Mercantile Exchange.

Cox, J. C., Ingersoll, J. E., and Ross, S. A. (1985), A Theory of the Term Structure of Interest Rates, *Econometrica. Journal of the Econometric Society*, 53(2), 385–407.

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.

Duffie, D. and Singleton, K. (1999), Modeling term structures of defaultable bonds, *Review of Financial Studies*, 12, 687–720.

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, Paul and Hofert, Marius (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, Paul and Hofert, Marius (2013b), Statistical inference for copulas in high dimensions: A simulation study, *ASTIN Bulletin*, 43(2), 81–95, doi:10.1017/asb.2013.6.

Fan, Y. and Patton, A.J. (2014), Copulas in econometrics, *Annual Review of Economics*, 6, 179–200.

- 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, Christian 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.
- Gregory, J. (2012), Counterparty Credit Risk and Credit Value Adjustment: A Continuing Challenge for Global Financial Markets, Chichester: John Wiley and Sons Ltd.

- 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, Nick (2002), Computing the nearest correlation matrix – A problem from finance, *IMA Journal of Numerical Analysis*, 22, 329–343.
- Hofert, Marius (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, Marius and Mächler, Martin (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, Marius and Scherer, Matthias (2011), CDO pricing with nested Archimedean copulas, *Quantitative Finance*, 11(5), 775–787, doi:<http://dx.doi.org/10.1080/14697680903508479>.

Hofert, Marius, Memartoluie, Amir, Saunders, David, and Wirjanto, Tony (2015), Improved Algorithms for Computing Worst Value-at-Risk: Numerical Challenges and the Adaptive Rearrangement Algorithm.

Huang, Jing-Zhi, 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.

Iman, R. L. and Conover, W. J. (1982), A distribution-free approach to inducing rank correlation among input variables, *Communications in Statistics – Simulation and Computation*, 11(3), 311–334.

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.
- Karatzas, I. and Shreve, S. E. (1988), Brownian Motion and Stochastic Calculus, Berlin: Springer.
- 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.
- Mashal, R. and Zeevi, A. (2002), Beyond correlation: Extreme co-movements between financial assets, Preprint, Columbia Business School.
- McNeil, A. J. and Frey, R. (2000), Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach, *Journal of Empirical Finance*, 7, 271–300.
- 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.
- Nelson, C. R. and Siegel, A. F. (1987), Parsimonious modeling of yield curves, *Journal of Business*, 60, 473–489.
- Ou, S. (2013), Annual Default Study: Corporate Default and recovery Rates, 1920-2012, Technical document, Moody's Investor Service.
- Patton, A. J. (2006), Modelling asymmetric exchange rate dependence, *International Economic Review*, 47, 527–556.
- Patton, A.J. (2012), A review of copula models for economic time series, *Journal of Multivariate Analysis*, 110, 4–18.
- 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/2008shreve.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.

- Wang, Ruodu, Peng, Liang, and Yang, Jingping (2013), Bounds for the sum of dependent risks and worst Value-at-Risk with monotone marginal densities, *Finance and Stochastics*, 17(2), 395–417.
- Weber, S. (2006), Distributions-invariant risk measures, information, and dynamic consistency, *Mathematical Finance*, 16(2), 419–441.
- Ziegel, J. F. (2014), Coherence and elicability, *Mathematical Finance*, to appear.

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

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 at least sample from F , we would 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 . How do we get a $(1 - \beta)$ -confidence interval for θ ? How $\text{var}(\hat{\theta}_n)$? How the distribution of $\hat{\theta}_n$?
- The (nonparametric) bootstrap treats \hat{F}_n as the true df F (first approximation) and samples from that (i.e. resamples X_1, \dots, X_n) to approximate the distribution of $\hat{\theta}_n$ (second approximation).
- 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.

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):

- $\hat{\mu}_{B,n} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_{b,n} \approx \mathbb{E}(\hat{\theta}_n)$; $\hat{\sigma}_{B,n}^2 = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_{b,n} - \hat{\mu}_{B,n})^2 \approx \text{var}(\hat{\theta}_n)$; the empirical df based on the $\hat{\theta}_{b,n}$'s approximates the df of $\hat{\theta}_n$.
- An *bootstrapped $(1 - \beta)$ -confidence interval* for θ based on $\hat{\theta}_{(1),n} \leq \dots \leq \hat{\theta}_{(B),n}$ is

$$\left[\hat{\theta}_{\left(\max\left\{\left\lfloor \frac{\beta}{2}B \right\rfloor, 1\right\}\right),n}, \hat{\theta}_{\left(\left\lceil (1-\frac{\beta}{2})B \right\rceil\right),n} \right].$$

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) $\hat{F}_n(x) \xrightarrow[n \rightarrow \infty]{} F(x) \quad \forall x \in C(F) \Leftrightarrow \hat{F}_n^\leftarrow(u) \xrightarrow[n \rightarrow \infty]{} F^\leftarrow(u) \quad \forall u \in C(F^\leftarrow)$;
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 , $x_{(i)} = \hat{F}_n^\leftarrow(i/n) \approx \hat{F}_n^\leftarrow(p_i) \stackrel{2)}{\approx} F^\leftarrow(p_i)$ (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.

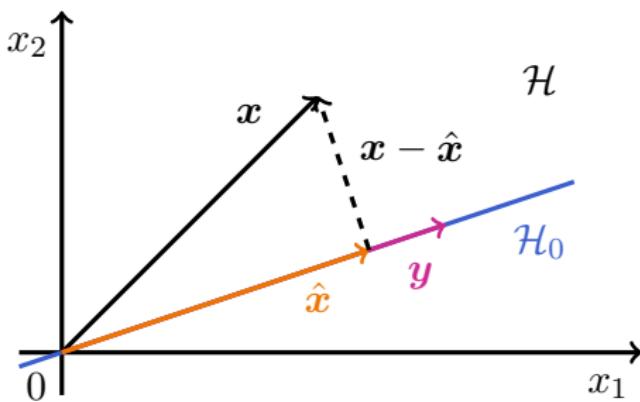
- *Hilbert's Projection Theorem* (see Brockwell and Davis (1991, p. 51)):
If \mathcal{H}_0 is a closed subspace of the Hilbert space \mathcal{H} and $x \in \mathcal{H}$, then:
 - i) There exists a unique $\hat{x} \in \mathcal{H}_0 : \|x - \hat{x}\| = \inf_{y \in \mathcal{H}_0} \|x - y\|$;
 - ii) $\hat{x} \in \mathcal{H}_0, \|x - \hat{x}\| = \inf_{y \in \mathcal{H}_0} \|x - y\|$ 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 $n \in \mathbb{N}$,

$$\begin{aligned}\phi_{n,n} &\stackrel{(*)}{=} \frac{\gamma(n) - \sum_{k=1}^{n-1} \gamma(n-k)\phi_{n-1,k}}{\gamma(0) - \sum_{k=1}^{n-1} \gamma(n-k)\phi_{n-1,n-k}} \\ &= \frac{\rho(n) - \sum_{k=1}^{n-1} \rho(n-k)\phi_{n-1,k}}{1 - \sum_{k=1}^{n-1} \rho(n-k)\phi_{n-1,n-k}},\end{aligned}$$

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

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

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

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

and $\sum_{k=1}^{n-1} \gamma(n-k) \phi_{n,k} + \phi_{n,n} \gamma(0) = \gamma(n)$. Multiplying with Γ_{n-1}^{-1} leads (**). For (*), use the k th row $\phi_{n,k} = \phi_{n-1,k} - \phi_{n,n} \phi_{n-1,n-k}$ in (***) and solve w.r.t. $\phi_{n,n}$. □

Definition 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

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

- 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 $|z| \leq 1$. \square

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 GARCH(p, q) model with Gaussian innovations,

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

where

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

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

$$\underbrace{N_u!}_{\substack{\text{ordered} \\ \text{sample prob. of } N_u \text{ samples}}} \underbrace{e^{-\Lambda(A)} \frac{\Lambda(A)^{N_u}}{N_u!}}_{\substack{\text{density of } \tilde{X}_i}} \prod_{i=1}^{N_u} \underbrace{\lambda(\tilde{X}_i)}_{\substack{\text{density of } \tilde{X}_i}} = e^{-\Lambda(A)} \prod_{i=1}^{N_u} \lambda(\tilde{X}_i) = e^{-\tau_u} \prod_{i=1}^{N_u} \lambda(\tilde{X}_i).$$

- 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

$$\begin{aligned}&= -\tau_u + \sum_{i=1}^{N_u} \log \lambda(\tilde{X}_i) = -\tau_u + N_u \log \tau_u + \underbrace{\sum_{i=1}^{N_u} (\log \lambda(\tilde{X}_i) - \log \tau_u)}_{=\log g_{\xi, \beta}(\tilde{X}_i - u)} \\ &= \ell_{\text{Poi}}(\tau_u; N_u) - N_u \log(T) + \ell_{\text{GPD}}(\xi, \beta; \tilde{X}_1 - u, \dots, \tilde{X}_{N_u} - u),\end{aligned}\tag{135}$$

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

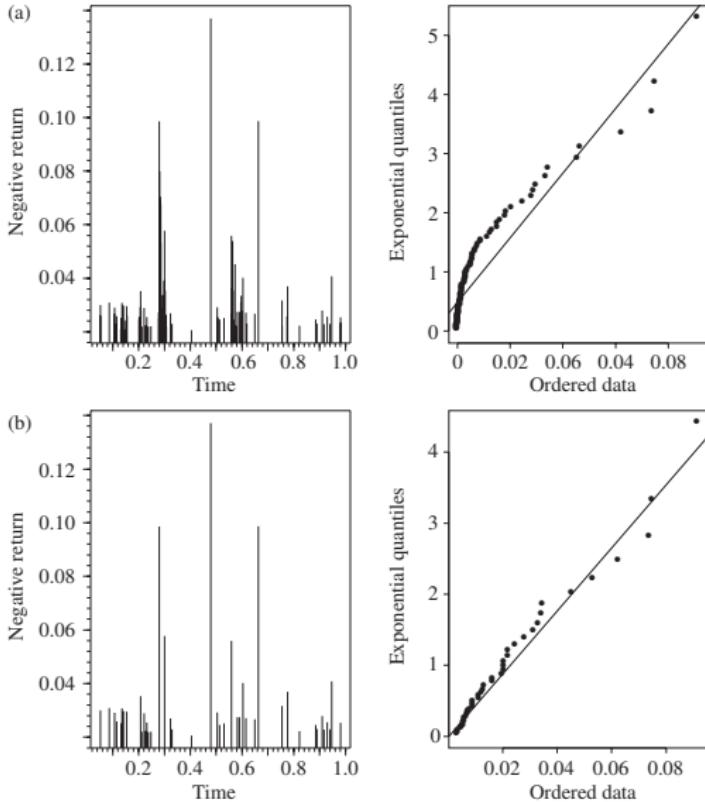
- One advantage of the two-dimensional Poisson point process model is that ξ , μ and σ do not depend on u (unlike β in the GPD model).
 - ⇒ In practice, we would expect the estimated parameters of the Poisson model to be roughly stable over a range of high thresholds.
- The intensity λ is thus often 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.

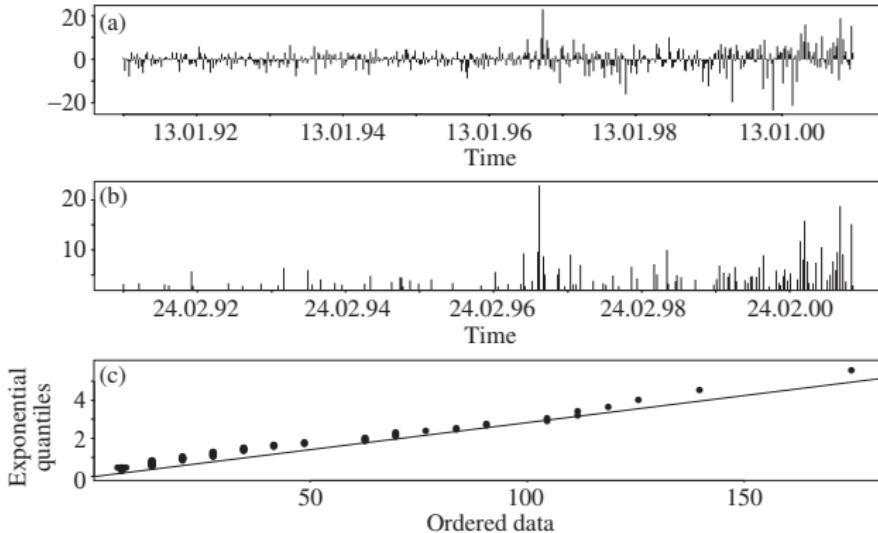
- Declustering may circumvent the problem. One identifies clusters (ad hoc; not easy) of exceedances and then applies the POT model to cluster maxima only.
- A possible declustering algorithm is the *runs method*. A run size r is fixed and two successive exceedances are said to belong to two different 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 (135), 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 (16). We have

$$\begin{aligned} & \int_{(-\infty, x_1]} F_{X_2|X_1}(x_2 | z) dF_{X_1}(z) \\ &= \int_{\mathbb{R}^d} I_{\{z \leq x_1\}} \mathbb{E}(I_{\{X_2 \leq x_2\}} | X_1 = z) dF_{X_1}(z) \\ &= \mathbb{E}(I_{\{X_1 \leq x_1\}} \mathbb{E}(I_{\{X_2 \leq x_2\}} | X_1)) = \mathbb{E}(\mathbb{E}(I_{\{X_1 \leq x_1, X_2 \leq x_2\}} | X_1)) \\ &\stackrel{\substack{\text{tower} \\ \text{property}}}{=} \mathbb{E}(I_{\{X_1 \leq x_1, X_2 \leq x_2\}}) = F(\mathbf{x}), \end{aligned}$$

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,

$$\phi'_Z(t) = \int_{-\infty}^{\infty} \sin(tx)(-x) \varphi(x) dx = \int_{-\infty}^{\infty} \sin(tx) \varphi'(x) dx \stackrel{\text{by parts}}{=} -t\phi_Z(t).$$

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

$$\mathbf{X}_n \xrightarrow[n \uparrow \infty]{\text{d}} \mathbf{X} \iff \mathbf{a}' \mathbf{X}_n \xrightarrow[n \uparrow \infty]{\text{d}} \mathbf{a}' \mathbf{X} \quad \forall \mathbf{a} \in \mathbb{R}^d$$

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

$$\mathbf{X} \stackrel{\text{d}}{=} \mathbf{Y} \iff \mathbf{a}' \mathbf{X} \stackrel{\text{d}}{=} \mathbf{a}' \mathbf{Y} \quad \forall \mathbf{a} \in \mathbb{R}^d.$$

Properties of multivariate normal variance mixtures

Proof of Lemma 6.10. W.l.o.g. assume $\mu = \mathbf{0}$.

$$\begin{aligned} \Rightarrow \mathbb{E}|X_i| \mathbb{E}|X_j| &\stackrel{\text{ind.}}{=} \mathbb{E}(|X_i||X_j|) = \mathbb{E}(W|Z_i||Z_j|) \stackrel{\text{ind.}}{=} \mathbb{E}(W) \mathbb{E}|Z_i| \mathbb{E}|Z_j| \\ &\stackrel{\text{Jensen}}{\geq} \mathbb{E}(\sqrt{W})^2 \mathbb{E}|Z_i| \mathbb{E}|Z_j| \stackrel{\text{ind.}}{=} \mathbb{E}|\sqrt{W}Z_i| \mathbb{E}|\sqrt{W}Z_j| = \mathbb{E}|X_i| \mathbb{E}|X_j| \end{aligned}$$

\Rightarrow We must have “=” in Jensen’s inequality. This holds if and only if W is constant a.s.; so $\mathbf{X} \sim N_d(\mathbf{0}, WI_d)$ in this case.

\Leftarrow W a.s. constant $\Rightarrow \mathbf{X} \sim N_d(\mathbf{0}, WI_d) \Rightarrow X_i, X_j$ independent. \square

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} \Rightarrow we can define $\psi(\|\mathbf{t}\|^2) = \phi_{\mathbf{Y}}(\mathbf{t})$.

$2) \Rightarrow 3)$: $\phi_{Y_1}(t) = \phi_{\mathbf{Y}}(te_1) \stackrel{2)}{=} \psi(t^2)$ (*). Now $\phi_{\mathbf{a}'\mathbf{Y}}(t) = \phi_{\mathbf{Y}}(t\mathbf{a}) \stackrel{2)}{=}$

$$\psi(t^2\|\mathbf{a}\|^2) = \psi((t\|\mathbf{a}\|)^2) \stackrel{(*)}{=} \phi_{Y_1}(t\|\mathbf{a}\|) = \phi_{\|\mathbf{a}\|Y_1}(t)$$

$3) \Rightarrow 1)$: $\phi_{U\mathbf{Y}}(\mathbf{t}) = \mathbb{E}(\exp(i(U'\mathbf{t})'\mathbf{Y})) \stackrel{U'\mathbf{t}:=\mathbf{a}}{=} \mathbb{E}(\exp(i\mathbf{a}'\mathbf{Y})) \stackrel{3)}{=} \mathbb{E}(\exp(i\|\mathbf{a}\|Y_1))$
 $= \mathbb{E}(\exp(i\|\mathbf{t}\|Y_1)) \stackrel{3)}{=} \mathbb{E}(\exp(i\mathbf{t}'\mathbf{Y})) = \phi_{\mathbf{Y}}(\mathbf{t})$ \square

Proof of Theorem 6.16. Let Ω_d be the characteristic generator of \mathbf{S} .

" \Rightarrow " $\mathbf{Y} \sim S_d(\psi) \Rightarrow \phi_{\mathbf{Y}}(\|\mathbf{t}\|\mathbf{u}) \stackrel{2)}{=} \psi(\|\mathbf{t}\|^2\mathbf{u}'\mathbf{u}) = \psi(\|\mathbf{t}\|^2)$ for all $\mathbf{u} \in \mathbb{R}^d$:
 $\|\mathbf{u}\| = 1$. Replacing \mathbf{u} by \mathbf{S} and integrating leads to $\psi(\|\mathbf{t}\|^2) = \mathbb{E}_{\mathbf{S}}(\phi_{\mathbf{Y}}(\|\mathbf{t}\|\mathbf{S})) = \mathbb{E}_{\mathbf{S}}(\mathbb{E}_{\mathbf{Y}}(e^{i\|\mathbf{t}\|\mathbf{S}'\mathbf{Y}})) \stackrel{\text{Fubini}}{=} \mathbb{E}_{\mathbf{Y}}(\mathbb{E}_{\mathbf{S}}(e^{i\|\mathbf{t}\|\mathbf{S}'\mathbf{Y}})) = \mathbb{E}_{\mathbf{Y}}(\phi_{\mathbf{S}}(\|\mathbf{t}\|\mathbf{Y})) \stackrel{2)}{=} \mathbb{E}_{\mathbf{Y}}(\Omega_d(\|\mathbf{t}\|^2\mathbf{Y}'\mathbf{Y}))$. 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 $\|\mathbf{Z}/\|\mathbf{Z}\|\| = \|\mathbf{Z}\|/\|\mathbf{Z}\| = 1$, $\mathbf{S} \stackrel{d}{=} \mathbf{Z}/\|\mathbf{Z}\|$. As such, \mathbf{S} itself is spherical, since $U\mathbf{S} \stackrel{d}{=} U\mathbf{Z}/\|\mathbf{Z}\| \stackrel{d}{=} \mathbf{Z}/\|\mathbf{Z}\| \stackrel{d}{=} \mathbf{S}$ for any orthogonal $U \in \mathbb{R}^{d \times d}$. Theorem 6.15 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 of $\|\mathbf{t}\|^2$ 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 ,

$$\begin{aligned} f_{\mathbf{Y}}(U\mathbf{y}) &\stackrel{\text{inv.}}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i(U'\mathbf{t})' \mathbf{y}} \phi_{\mathbf{Y}}(\mathbf{t}) d\mathbf{t} \\ &\stackrel{\text{subs.}}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-is' \mathbf{y}} \phi_{\mathbf{Y}}(Us) ds \end{aligned}$$

$$\begin{aligned}
&\stackrel{2)}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-is' \mathbf{y}} \psi((Us)' Us) ds \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-is' \mathbf{y}} \psi(s' s) ds \underset{\text{backwards}}{=} \dots = f_{\mathbf{Y}}(\mathbf{y}).
\end{aligned}$$

This implies that $f_{\mathbf{Y}}(\mathbf{y}) = g(\|\mathbf{y}\|^2)$ for a function $g : [0, \infty) \rightarrow [0, \infty)$, 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. $\mathbf{X} | ((\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \geq c) \stackrel{(24)}{\stackrel{d}{=}} \boldsymbol{\mu} + RAS | (R^2 \geq c) \stackrel{\text{ind.}}{=} \boldsymbol{\mu} + \tilde{R}AS$

where $\tilde{R} \stackrel{d}{=} (R | R^2 \geq c)$. 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 $\hat{B} = (F'F)^{-1}F'X$.

- 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 $X \sim \text{B}(1, \lambda)$, independent of each other. Furthermore, let

$$\mathbf{U} = \begin{cases} \mathbf{U}_1, & \text{if } X = 1, \\ \mathbf{U}_2, & \text{if } X = 0. \end{cases}$$

The Law of Total Probability implies 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.

$$\begin{aligned} \text{“}\Rightarrow\text{” } \mathbb{P}(F_j(X_j) \leq u_j \forall j) &\stackrel{\text{cont.}}{=} \mathbb{P}(F_j(X_j) < u_j \forall j) \stackrel{\text{(GI5)}}{=} \mathbb{P}(X_j < F_j^\leftarrow(u_j) \forall j) \\ &\stackrel{\text{cont.}}{=} \mathbb{P}(X_j \leq F_j^\leftarrow(u_j) \forall j) = F(F_1^\leftarrow(u_1), \dots, F_d^\leftarrow(u_d)) \stackrel{\text{Sklar}}{=} C(\mathbf{u}). \end{aligned}$$

“ \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

$$\begin{aligned}
 & \Delta_{(\frac{1}{2}, 1]} W \\
 &= \sum_{\mathbf{i} \in \{0,1\}^d} (-1)^{\sum_{j=1}^d i_j} W\left(\frac{1}{2}^{i_1}, \dots, \frac{1}{2}^{i_d}\right) \\
 &= \max\{1 + 1 + 1 + \dots + 1 - d + 1, 0\} \quad (i_j = 0 \ \forall j) \\
 &\quad - d \max\{\frac{1}{2} + 1 + 1 + \dots + 1 - d + 1, 0\} \quad (\exists! j : i_j = 1)
 \end{aligned}$$

$$\begin{aligned}
& + \binom{d}{2} \max\left\{\frac{1}{2} + \frac{1}{2} + 1 + \cdots + 1 - d + 1, 0\right\} \quad (\exists! \text{ two } j : i_j = 1) \\
& - \cdots + (-1)^d \max\left\{\frac{1}{2} + \cdots + \frac{1}{2} - d + 1, 0\right\} \quad (i_j = 1 \ \forall j) \\
& = 1 - \frac{d}{2} < 0 \quad \text{for } d \geq 3.
\end{aligned}$$

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

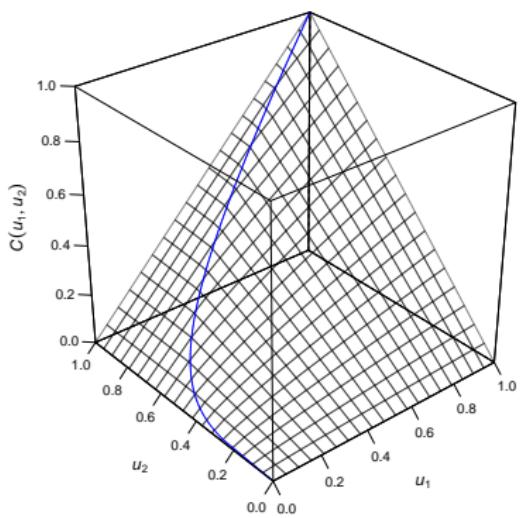
$$C(\mathbf{u}) = \left(\prod_{j=1}^d u_j \right)^{A\left(\frac{\log u_1}{\log \Pi(\mathbf{u})}, \dots, \frac{\log u_d}{\log \Pi(\mathbf{u})}\right)}$$

for a *Pickands dependence function* A ; see Ressel (2013) for a characterization of A .

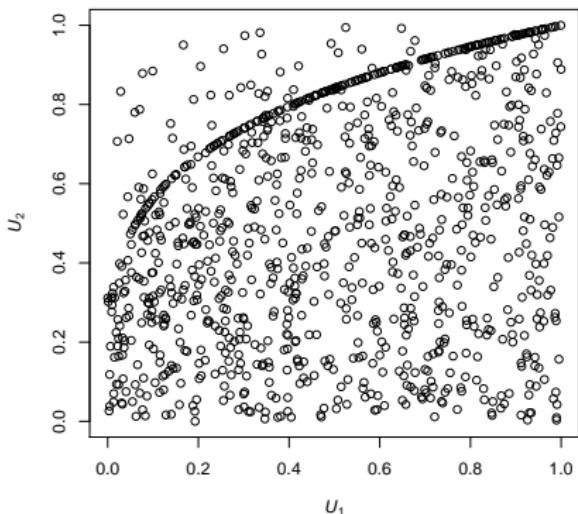
- Examples: Gumbel copula, Marshall-Olkin copulas.
- For more details, see Jaworski et al. (2010, Chapter 6).

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

$$\begin{aligned} C(\mathbf{u}) &\stackrel{\text{L.7.6}}{=} \mathbb{P}(F_1(X_1) \leq u_1, F_2(F_2^\leftarrow(1 - F_1(X_1))) \leq u_2) \\ &\stackrel{(GI4)}{=} \mathbb{P}(F_1(X_1) \leq u_1, 1 - F_1(X_1) \leq u_2) \\ &= \mathbb{P}(1 - u_2 < U \leq u_1) = W(u_1, u_2) \quad \text{for } U \sim \text{U}(0, 1). \end{aligned}$$

“ \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 $u \in [0, 1]$.

$$1) \quad T \text{ left-continuous} \Rightarrow T(\textcolor{brown}{u}) \leq x \Leftrightarrow \textcolor{brown}{u} \leq \textcolor{blue}{u}_x := \sup\{u : T(u) \leq x\}$$

$$2) \quad 1) \Rightarrow \{T(\textcolor{brown}{U}) \leq x\} = \{\textcolor{brown}{U} \leq \textcolor{blue}{u}_x\} \Rightarrow F_{T(U)}(x) = F_U(\textcolor{blue}{u}_x) = \textcolor{blue}{u}_x.$$

$$\Rightarrow F_{T(U)}^\leftarrow(u) \leq \textcolor{brown}{x} \stackrel{(\text{GI5})}{\Leftrightarrow} F_{T(U)}(x) \geq u \stackrel{2)}{\Leftrightarrow} \textcolor{blue}{u}_x \geq u \stackrel{1)}{\Leftrightarrow} T(u) \leq \textcolor{brown}{x}$$

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

$$F_{\sum_{j=1}^d X_j}(x) = \mathbb{P}\left(\sum_{j=1}^d X_j \leq x\right) = \mathbb{P}\left(\sum_{j=1}^d F_j^\leftarrow(U) \leq x\right) = \mathbb{P}(T(U) \leq x)$$

and thus $F_{\sum_{j=1}^d X_j}^\leftarrow(\alpha) = F_{T(U)}^\leftarrow(\alpha) = T(\alpha) = \sum_{j=1}^d F_j^\leftarrow(\alpha)$. □

Linear correlation

Proof of Proposition 7.16. Let (X'_1, X'_2) be an iid-copy of (X_1, X_2) . Consider

$$\begin{aligned} & 2 \operatorname{cov}(X_1, X_2) \\ &= \mathbb{E}((X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)) + \mathbb{E}((X'_1 - \mathbb{E}X'_1)(X'_2 - \mathbb{E}X'_2)) \\ &= \mathbb{E}(((X_1 - \mathbb{E}X_1) - (X'_1 - \mathbb{E}X'_1)) \cdot ((X_2 - \mathbb{E}X_2) - (X'_2 - \mathbb{E}X'_2))) \\ &\stackrel{\text{check}}{=} \mathbb{E}((X_1 - X'_1)(X_2 - X'_2)). \end{aligned}$$

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 $\mathbf{Y} \stackrel{d}{=} RA\mathbf{U} \stackrel{d}{=} RA \begin{pmatrix} \cos \Theta \\ \sin \Theta \end{pmatrix}$, $\Theta \sim U(-\pi, \pi)$ independent of R .
- With $\varphi = \arcsin \rho$, we have $\mathbf{Y} \stackrel{d}{=} R \begin{pmatrix} \cos \Theta \\ \sin \varphi \cos \Theta + \cos \varphi \sin \Theta \end{pmatrix} = \begin{pmatrix} \cos \Theta \\ \sin(\varphi + \Theta) \end{pmatrix}$.
- Thus $\mathbb{P}(X_1 > 0, X_2 > 0) = \mathbb{P}(Y_1 > 0, Y_2 > 0) = \mathbb{P}(\cos \Theta > 0, \sin(\varphi + \Theta) > 0) = \mathbb{P}(\Theta \in (-\frac{\pi}{2}, \frac{\pi}{2}), \varphi + \Theta \in (0, \pi)) = \mathbb{P}(\Theta \in (-\frac{\pi}{2}, \frac{\pi}{2}), \Theta \in (-\varphi, \pi - \varphi)) = \mathbb{P}(\Theta \in (-\varphi, \frac{\pi}{2})) = (\frac{\pi}{2} - (-\varphi))/(2\pi)$. \square

Lemma A.19 (Representation of Spearman's rho)

Let $(U_1, U_2) \sim C$ and $\tilde{U}_1, \bar{U}_2 \stackrel{\text{ind.}}{\sim} U(0, 1)$ 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. $\mathbf{X} \stackrel{d}{=} \sqrt{W}\mathbf{Z}$ for $\mathbf{Z} \sim N_2(\mathbf{0}, P)$. Let $\tilde{Z}, \bar{Z} \sim N(0, 1)$ and assume $\mathbf{Z}, \tilde{Z}, \bar{Z}, W, \tilde{W}$ and \bar{W} are all independent. Let

$$\tilde{X} = \sqrt{\tilde{W}}\tilde{Z}, \quad \bar{X} = \sqrt{\bar{W}}\bar{Z},$$

$$Y_1 = X_1 - \tilde{X} = \sqrt{W}Z_1 - \sqrt{\tilde{W}}\tilde{Z},$$

$$Y_2 = X_2 - \bar{X} = \sqrt{W}Z_2 - \sqrt{\bar{W}}\bar{Z}.$$

$$\rho_S(X_1, X_2) \stackrel{\substack{\text{L.A.19} \\ \Phi^{-1}}}{=} 12\mathbb{P}(X_1 \leq \tilde{X}_1, X_2 \leq \bar{X}_2) - 3$$

$$\begin{aligned}
&= 6\mathbb{P}((X_1 - \tilde{X}_1)(X_2 - \bar{X}_2) > 0) - 3 \\
&= 3(2\mathbb{E}(\mathbb{P}(Y_1 Y_2 > 0 | W, \tilde{W}, \bar{W})) - 1) \\
&= 3(4\mathbb{E}(\mathbb{P}(Y_1 > 0, Y_2 > 0 | W, \tilde{W}, \bar{W})) - 1).
\end{aligned}$$

Now note that $\mathbf{Y} | W, \tilde{W}, \bar{W} \sim N_2(\mathbf{0}, (\frac{W+\tilde{W}}{W\rho} \frac{W\rho}{W+\bar{W}}))$ 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 $\mathbf{X} \stackrel{d}{=} RAU$ and $\mathbf{X}' \stackrel{d}{=} R'AU' (\stackrel{d}{=} -\mathbf{X}')$ we have $\mathbf{Y} = \mathbf{X} - \mathbf{X}' \stackrel{d}{=} \mathbf{0} + A(RU - R'U')$. 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

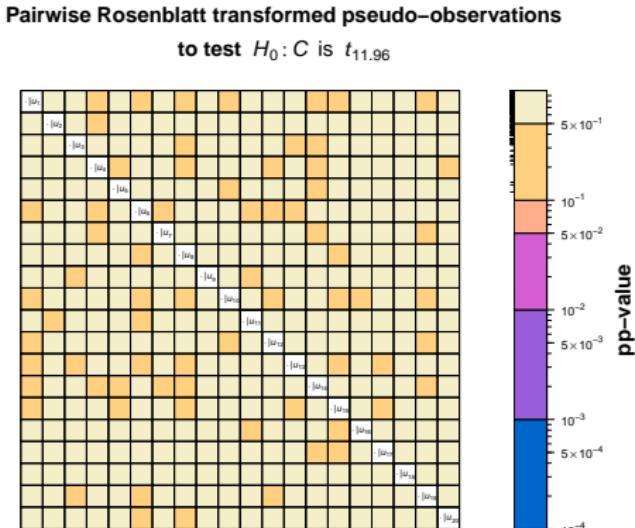
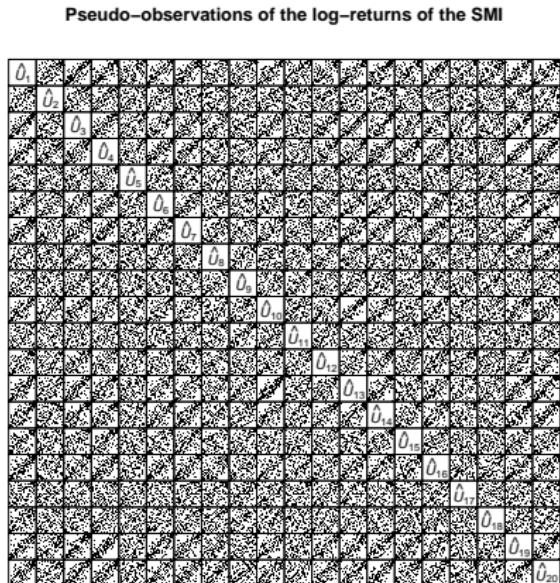
$$\begin{aligned}\rho_{\tau} &= 2\mathbb{P}(Y_1 Y_2 > 0) - 1 = 2(\mathbb{P}(Y_1 > 0, Y_2 > 0) + \mathbb{P}(Y_1 < 0, Y_2 < 0)) - 1 \\ &= 4\mathbb{P}(\mathbf{Y} > \mathbf{0}) - 1 \stackrel{\substack{\text{cont.} \\ \text{L.A.18}}}{=} \frac{2}{\pi} \arcsin \rho.\end{aligned}$$

□

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.

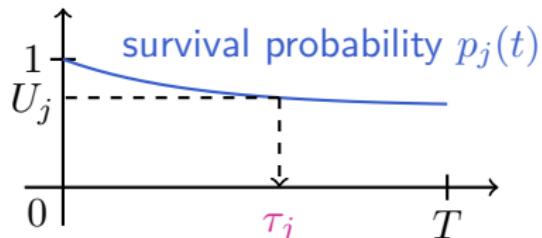


How intensity-/copula-based default models work

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

Proof of Proposition 8.11.

1) Recall that $L \stackrel{d}{=} F_L^\leftarrow(U)$, $U \sim U(0, 1)$. Therefore,

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

2) First note that

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

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

□

Proof of Corollary 8.12.

- 1) Since $\bar{F}_L(F_L^\leftarrow(\alpha)) = 1 - F_L(F_L^\leftarrow(\alpha)) = 1 - \alpha$ for all $\alpha \in \text{ran } F_L \cup \{\inf F_L, \sup F_L\} \supseteq (0, 1)$, the claim follows from Proposition 8.11 2).
- 2) First note that

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

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

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

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 $I_{\{L > F_L^\leftarrow(\alpha)\}}^{(\alpha)} =$

$\frac{1 - \bar{F}_L(F_L^\leftarrow(\alpha))}{\mathbb{P}(L = F_L^\leftarrow(\alpha))}$. 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 8.11 2) 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 8.11 2), 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

$$\begin{aligned}
 & \left(\sum_{j=1}^2 \mathbb{E}(L_j I_{\{L_j > F_{L_j}^{\leftarrow}(\alpha)\}}^{(\alpha)}) \right) - \mathbb{E}((L_1 + L_2) I_{\{L_1 + L_2 > F_{L_1 + L_2}^{\leftarrow}(\alpha)\}}^{(\alpha)}) \\
 & \stackrel{\text{Linearity}}{=} \sum_{j=1}^2 \mathbb{E}(L_j (I_{\{L_j > F_{L_j}^{\leftarrow}(\alpha)\}}^{(\alpha)} - I_{\{L_1 + L_2 > F_{L_1 + L_2}^{\leftarrow}(\alpha)\}}^{(\alpha)})). \tag{136}
 \end{aligned}$$

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

In both cases, we make the expectations in (136) smaller by replacing L_j by $F_{L_j}^{\leftarrow}(\alpha)$. Hence

$$(136) \geq \sum_{j=1}^2 F_{L_j}^{\leftarrow}(\alpha) \underbrace{\mathbb{E}(I_{\{L_j > F_{L_j}^{\leftarrow}(\alpha)\}}^{(\alpha)} - I_{\{L_1 + L_2 > F_{L_1 + L_2}^{\leftarrow}(\alpha)\}}^{(\alpha)})}_{\substack{= (1-\alpha)-(1-\alpha)=0 \\ \text{Lem. A.21 2)}} \geq 0. \quad \square$$

Proof of Proposition 8.16. $\nu([0, t]) := \phi(t)$ defines a measure on $[0, 1]$. For every function $f : [0, 1] \rightarrow \mathbb{R}$ we have

$$\int_0^1 f(\alpha) d\nu(\alpha) = f(0)\phi(0) + \int_0^1 f(\alpha) d\phi(\alpha).$$

Define a further measure μ on $[0, 1]$ by $\frac{d\mu}{d\nu}(\alpha) = 1 - \alpha$, that is, $\int_0^1 f(\alpha) d\mu(\alpha) = f(0)\phi(0) + \int_0^1 f(\alpha)(1 - \alpha) d\phi(\alpha)$. Using this for $f \equiv 1$,

$$\begin{aligned}\mu([0, 1]) &= 1 \cdot \phi(0) + \int_0^1 1 \cdot (1 - \alpha) d\phi(\alpha) \stackrel{\text{by parts}}{=} \phi(0) + \int_0^1 \phi(\alpha) d\alpha - \phi(0) \\ &= \int_0^1 \phi(\alpha) d\alpha = D(1) - D(0) = 1,\end{aligned}$$

so μ is a probability measure. By Fubini's Theorem we have that

$$\begin{aligned}\varrho(L) &= \int_0^1 q_u(L)\phi(u) du = \int_0^1 q_u(L) \int_0^u 1 d\nu(\alpha) du \\ &= \int_0^1 \int_0^1 q_u(L) 1_{\{\alpha \leq u\}} d\nu(\alpha) du \stackrel{\text{Fubini}}{=} \int_0^1 \int_\alpha^1 q_u(L) du d\nu(\alpha) \\ &= \int_0^1 \text{ES}_\alpha(L)(1 - \alpha) d\nu(\alpha) = \int_0^1 \text{ES}_\alpha(L) d\mu(\alpha).\end{aligned}$$

□

Proof of Proposition 8.19. The expectile y solves $\alpha\mathbb{E}(|L - y|) = \alpha\mathbb{E}((L - y)^+) + \alpha\mathbb{E}((L - y)^-) \stackrel{(43)}{=} \mathbb{E}((L - y)^-) = \int_{-\infty}^y (y - x) dF_L(x) = yF_L(y) - \mu(y)$, where

$$\begin{aligned}\mathbb{E}(|L - y|) &= \int_{-\infty}^y (y - x) dF_L(x) + \int_y^{\infty} (x - y) dF_L(x) \\ &= 2 \int_{-\infty}^y (y - x) dF_L(x) + \int_{-\infty}^{\infty} (x - y) dF_L(x) \\ &= 2(yF_L(y) - \mu(y)) + \mu - y\end{aligned}$$

and hence $\alpha = \tilde{F}_L(y)$ with \tilde{F}_L as defined in (46). We have left to show that \tilde{F}_L is a df. The derivative of \tilde{F}_L can be easily computed to be

$$\tilde{f}_L(y) = \frac{\mu F_L(y) - \mu(y)}{(2(yF_L(y) - \mu(y)) + \mu - y)^2} = \frac{F_L(y)(\mu - \mathbb{E}(L | L \leq y))}{(2(yF_L(y) - \mu(y)) + \mu - y)^2},$$

which is nonnegative for all y and strictly positive on $D = \{y : 0 < F_L(y) < 1\}$; hence \tilde{F}_L is increasing for all y and strictly increasing on D . It is easy to check that for $y_0 = \inf D$ and $y_1 = \sup D$, $[y_0, y_1]$ is the support of \tilde{F}_L and $\lim_{y \rightarrow y_0} \tilde{F}_L(y) = 0$ and $\lim_{y \rightarrow y_1} \tilde{F}_L(y) = 1$. \square

Proof of Proposition 8.21. For $L \in \mathcal{M}$ and $y \in \mathbb{R}$ define

$$\begin{aligned}g(L, y, \alpha) &= \alpha \mathbb{E}((L - y)^+) - (1 - \alpha) \mathbb{E}((L - y)^-) \\&= (2\alpha - 1) \mathbb{E}((L - y)^+) + (1 - \alpha) \mathbb{E}(L - y).\end{aligned}$$

For fixed L , g is a decreasing function of y and, for fixed y , g is monotonic in L so that $L_1 \leq L_2 \Rightarrow g(L_1, y, \alpha) \leq g(L_2, y, \alpha)$.

Translation invariance and positive homogeneity follow easily from the fact that if $g(L, y, \alpha) = 0$ (i.e. if $e_\alpha(L) = y$) then $g(L + m, y + m, \alpha) = 0$ for $m \in \mathbb{R}$ and $g(\lambda L, \lambda y, \alpha) = 0$ for $\lambda > 0$.

For monotonicity, fix α and let $y_1 = e_\alpha(L_1)$, $y_2 = e_\alpha(L_2)$. If $L_2 \geq L_1$ then $g(L_2, y_1, \alpha) \geq g(L_1, y_1, \alpha) = 0 = g(L_2, y_2, \alpha)$. Since g is decreasing in y , it must be the case that $y_2 \geq y_1$.

For subadditivity, again let $y_1 = e_\alpha(L_1)$, $y_2 = e_\alpha(L_2)$. We have that

$$\begin{aligned} & g(L_1 + L_2, y_1 + y_2, \alpha) \\ &= (2\alpha - 1)\mathbb{E}((L_1 + L_2 - y_1 - y_2)^+) + (1 - \alpha)\mathbb{E}(L_1 + L_2 - y_1 - y_2) \\ &= (2\alpha - 1)\mathbb{E}((L_1 + L_2 - y_1 - y_2)^+) + (1 - \alpha)\mathbb{E}(L_1 - y_1) + (1 - \alpha)\mathbb{E}(L_2 - y_2) \end{aligned}$$

and, since $(2\alpha - 1)\mathbb{E}((L_i - y_i)^+) + (1 - \alpha)\mathbb{E}(L_i - y_i) = 0$ for $i = 1, 2$, we get

$$\begin{aligned} & g(L_1 + L_2, y_1 + y_2, \alpha) \\ &= (2\alpha - 1) \left(\mathbb{E}((L_1 + L_2 - y_1 - y_2)^+) - \mathbb{E}((L_1 - y_1)^+) - \mathbb{E}((L_2 - y_2)^+) \right) \end{aligned}$$

where we have used the fact that $(x_1 + x_2)^+ \leq x_1^+ + x_2^+$. So $g(L_1 + L_2, y_1 + y_2, \alpha) \leq 0$ and thus, as g is decreasing in y , $g(L_1 + L_2, y, \alpha) = 0$ for some $y \leq y_1 + y_2 = e_\alpha(L_1) + e_\alpha(L_2)$ with $y = e_\alpha(L_1 + L_2)$. \square

Proof of Euler's rule. Let $f : \mathbb{R}^d \setminus \{\mathbf{0}\} \rightarrow \mathbb{R}$ be differentiable and positive homogeneous of order $r > 0$, so that

$$f(tx) = t^r f(x), \quad \text{for all } t > 0, \quad x \in \mathbb{R}^d \setminus \{\mathbf{0}\}.$$

Differentiating w.r.t. t (use the chain rule on the left-hand side) leads to $rt^{r-1}f(x) = \nabla f(tx)x = \sum_{j=1}^d \frac{\partial f}{\partial x_j}(tx) \cdot x_j$ and thus and thus (for $t = 1$)

$$rf(x) = \nabla f(x)x = \sum_{j=1}^d \frac{\partial f}{\partial x_j}(x) \cdot x_j. \quad (137)$$

On the other hand, if $f : \mathbb{R}^d \setminus \{\mathbf{0}\} \rightarrow \mathbb{R}$ fulfills (137) and is differentiable, consider $g(t) = f(tx)$, $t > 0$. Then g satisfies

$$g'(t) = \nabla f(tx)x = \frac{1}{t} \sum_{j=1}^d \frac{\partial f}{\partial x_j}(tx) \cdot tx_j \stackrel{(137)}{=} \frac{1}{t} rf(tx) = \frac{r}{t} g(t)$$

with the initial value $g(1) = f(x)$. The only solution is $g(t) = t^r f(x)$, so $f(tx) = t^r f(x)$ and thus f is homogeneous of order $r > 0$. \square