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.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 payed 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 (Q) measure: An example

- Consider a defaultable bond with principal 1 and maturity $T = 1$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, 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.

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 $\rho(L) =: \rho(F_L)$ 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 $\rho(F_L)$ via the point forecast $\rho(\hat{F}_L)$.
- 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

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

  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 $\rho(F_L)$ for any S. Hence, for two competing point forecasts of $\rho(F_L)$, one cannot tell which performs the best by comparing their forecasting error, no matter what S is used.

A.3 Empirical properties of financial data

Non-normality and heavy tails

Justification for P-P and Q-Q plots:

1) Glivenko–Cantelli: $\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \overset{\text{a.s.}}{\underset{n\uparrow\infty}{\to}} 0$

2) $\hat{F}_n(x) \underset{n\to\infty}{\to} F(x)\ \forall\, x \in C(F) \Leftrightarrow \hat{F}_n^{\leftarrow}(u) \underset{n\to\infty}{\to} F^{\leftarrow}(u)\ \forall\, u \in C(F^{\leftarrow})$; 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, $x_{(i)} = \hat{F}_n^{\leftarrow}(i/n) \approx \hat{F}_n^{\leftarrow}(p_i) \underset{2)}{\approx} F^{\leftarrow}(p_i)$ (a justification for both P-P and Q-Q plots).

A.4 Financial time series

Conditional expectations

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

1) $\boldsymbol{Y} \in \mathcal{G}$ ($\boldsymbol{Y}$ is $\mathcal{G}$-measurable);

2) $\mathbb{E}|\boldsymbol{Y}| < \infty$; and

3) $\mathbb{E}(\boldsymbol{Y} I_G) = \int_G \boldsymbol{Y}\, d\mathbb{P} = \int_G \boldsymbol{X}\, d\mathbb{P} = \mathbb{E}(\boldsymbol{X} I_G)$ for all $G \in \mathcal{G}$

is called *conditional expectation of $\boldsymbol{X}$ given $\mathcal{G}$* and denoted by $\mathbb{E}(\boldsymbol{X} \,|\, \mathcal{G})$. $\mathbb{P}(A \,|\, \mathcal{G}) = \mathbb{E}(I_A \,|\, \mathcal{G})$ 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.2 (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.1 Part 3) to $\mathbb{E}(\mathbb{E}(X \,|\, \mathcal{G}) \,|\, \mathcal{F})$ and then to $\mathbb{E}(X \,|\, \mathcal{G})$ implies that $\mathbb{E}(\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{G}) \,|\, \mathcal{F}]I_G] = \mathbb{E}(\mathbb{E}[X \,|\, \mathcal{G})I_G) = \mathbb{E}(XI_G)$. □

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 $\hat{x} \in \mathcal{H}_0 : \|x - \hat{x}\| = \inf_{y \in \mathcal{H}_0} \|x - y\|$;

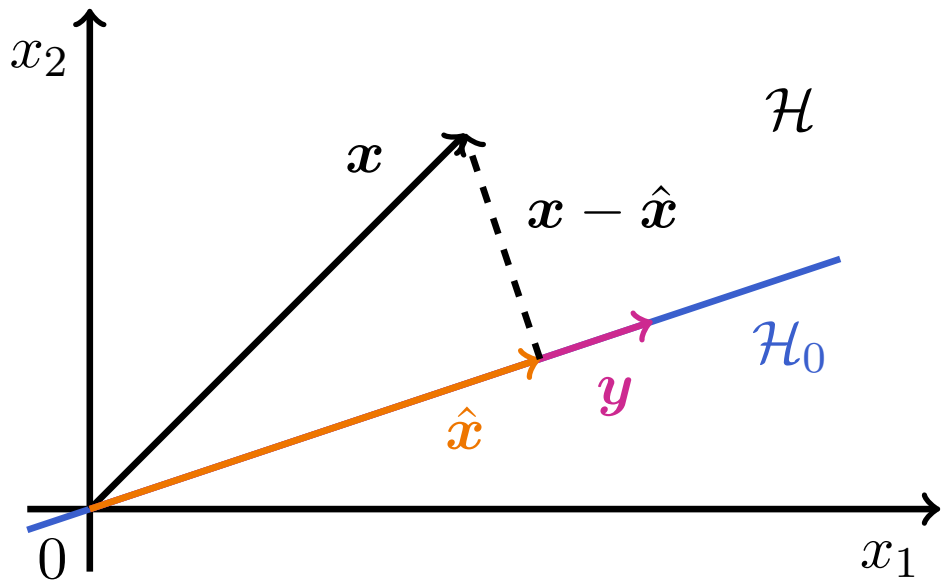
  ii) $\hat{x} \in \mathcal{H}_0$, $\|x - \hat{x}\| = \inf_{y \in \mathcal{H}_0} \|x - y\|$ if and only if $\hat{x} \in \mathcal{H}_0$, $x - \hat{x} \in \mathcal{H}_0^{\perp} = \{x \in \mathcal{H} : \langle x, y \rangle = 0 \text{ for all } y \in \mathcal{H}_0\}$.

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.3

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



- 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

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

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

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

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

$$\Leftrightarrow \quad \gamma(n-l) = \sum_{k=1}^{n-1} \gamma(n-k-l)\phi_{n-1,k}$$

$$\Leftrightarrow \quad \gamma(h) = \sum_{k=1}^{n-1} \gamma(h-k)\phi_{n-1,k}, \quad \forall\, h \in \{1, \dots, n-1\}$$

stationary

$$\Leftrightarrow \quad \Gamma_{n-1}\boldsymbol{\phi}_{n-1} = \boldsymbol{\gamma}_{n-1}, \quad (\textit{Yule–Walker equations})$$

where

$$\boldsymbol{\phi}_{n-1} = (\phi_{n-1,1}, \dots, \phi_{n-1,n-1}),$$
$$\boldsymbol{\gamma}_{n-1} = (\gamma(1), \dots, \gamma(n-1)),$$
$$\Gamma_{n-1} = (\gamma(|i-j|))_{i,j=1}^{n-1}.$$

Hilbert's Projection Theorem ii) $\Rightarrow$ there exists at least one solution $\boldsymbol{\phi}_{n-1}$ and all of them lead to the same $\hat{X}_n$ (unique by i)). If Γ_{n-1}

is regular (invertible), $\boldsymbol{\phi}_{n-1}$ is unique. This holds, e.g. if $\gamma(0) > 0$, $\gamma(h) \to 0$ $(h \to \infty)$; see Brockwell and Davis (1991, p. 167).

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

$$\begin{aligned}\phi_{n,n} &\underset{(*)}{=} \frac{\gamma(n) - \sum_{k=1}^{n-1} \gamma(n-k)\phi_{n-1,k}}{\gamma(0) - \sum_{k=1}^{n-1} \gamma(n-k)\phi_{n-1,n-k}} \\ &= \frac{\rho(n) - \sum_{k=1}^{n-1} \rho(n-k)\phi_{n-1,k}}{1 - \sum_{k=1}^{n-1} \rho(n-k)\phi_{n-1,n-k}},\end{aligned}$$

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

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

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

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

and $\sum_{k=1}^{n-1} \gamma(n-k)\phi_{n,k} + \phi_{n,n}\gamma(0) \underset{(***)}{=} \gamma(n)$. Multiplying with Γ_{n-1}^{-1} leads $(**)$. For $(*)$, use the kth 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.4 (PACF)

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

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

- PACF for $\mathrm{MA}(1)$: 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 $\Leftrightarrow P_h \boldsymbol{\phi}_h = \boldsymbol{\rho}_h$. One can show by induction (or the Durbin–Levinson algorithm) that

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

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

In particular, $\phi(h) = \phi_{h,h} \searrow 0$ exponentially.

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

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

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 $X_1, \ldots, X_{h-1}$, so $\sum_{k=1}^{p} \phi_k X_{h-k} = P_{\mathcal{H}_{h-1}} X_h$. Hence,

$$\phi(h) = \operatorname{corr}\Big(\underbrace{X_0 - P_{\mathcal{H}_{h-1}} X_0}_{\in \operatorname{span}\{X_0,\ldots,X_{h-1}\}} , \underbrace{X_h - P_{\mathcal{H}_{h-1}} X_h}_{=\varepsilon_h} \Big) \underset{\text{causality}}{=} 0.$$

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 $\varepsilon > 0 \Rightarrow 1/\phi(z) = \sum_{k=0}^{\infty} a_k z^k$, $a_k(1+\varepsilon/2)^k \to 0$ $(k \to \infty)$ $\Rightarrow \exists\, c > 0 : |a_k| < c(1+\varepsilon/2)^{-k}$, $k \in \mathbb{N}_0 \Rightarrow \sum_{k=0}^{\infty} |a_k| < \infty$. 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 $(\varepsilon_t) \sim \mathrm{WN}(0, \sigma^2)$, we obtain $\eta_k = \theta_k$, $k \in \{0, \ldots, q\}$, 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 dyanmics but the 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 $\mathrm{GARCH}(p, q)$ model with Gaussian innovations,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \underset{(n\to\infty)}{\xrightarrow{\mathrm{d}}} \mathrm{N}_{p+q+1}(\mathbf{0}, I(\boldsymbol{\theta})^{-1}),$$

where

$$I(\boldsymbol{\theta}) := \mathbb{E}\Big(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big)'\Big) = -\mathbb{E}\Big(\frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\Big) =: J(\boldsymbol{\theta})$$

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

$$\bar{I}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\Big(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big(\frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big)'\Big) \quad \text{and} \quad \bar{J}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2},$$

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

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \underset{(n\to\infty)}{\overset{\mathrm{d}}{\to}} \mathrm{N}_{p+q+1}(\mathbf{0}, J(\boldsymbol{\theta})^{-1} I(\boldsymbol{\theta}) J(\boldsymbol{\theta})^{-1}),$$

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 SARIMA$(p,d,q)\times(\tilde{p},\tilde{d},\tilde{q})_s$ (Seasonal; Integrated) process if

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

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

A.5 Extreme value theory

The convergence to types theorem

Theorem A.5 (Convergence to types)
Suppose $(M_n)_n$ is a sequence of rvs such that $\frac{M_n - d_n}{c_n} \xrightarrow{\mathsf{d}} Y$ for a rv Y and $d_n \in \mathbb{R}$, $c_n > 0$. Then

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

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

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

in which case $Z \overset{\mathsf{d}}{=} cY + d$ (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.6 (Gumbel MDA)

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

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

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

If $F \in \mathrm{MDA}(H_0)$, 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

$$
\begin{aligned}
e(\log u) &= \mathbb{E}(\log X - \log u \,|\, \log X > \log u) \\
&= \frac{1}{\bar{F}(u)} \int_u^\infty (\log x - \log u)\, dF(x) = -\frac{1}{\bar{F}(u)} \int_u^\infty \log\Big(\frac{x}{u}\Big)\, d\bar{F}(x) \\
&= -\frac{1}{\bar{F}(u)} \Big(\underbrace{\Big[\log\Big(\frac{x}{u}\Big)\bar{F}(x)\Big]_u^\infty}_{=0} - \int_u^\infty \bar{F}(x) \frac{1}{x}\, dx \Big) \\
&= \frac{1}{\bar{F}(u)} \int_u^\infty \frac{\bar{F}(x)}{x}\, dx = \frac{1}{\bar{F}(u)} \int_u^\infty x^{-\alpha-1} L(x)\, dx.
\end{aligned}
$$

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

$$\frac{1}{\alpha} \approx e_n(\log X_{k,n}) = \frac{\sum_{i=1}^{n} (\log X_i - \log X_{k,n}) I_{\{\log X_i > \log X_{k,n}\}}}{\sum_{i=1}^{n} I_{\{\log X_i > \log X_{k,n}\}}}$$

$$= \frac{\sum_{i=1}^{k-1} (\log X_{i,n} - \log X_{k,n})}{k-1} = \frac{1}{k-1} \sum_{i=1}^{k-1} \log X_{i,n} - \log X_{k,n}$$

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 $X_1, \ldots, X_n$ 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 $F \in \mathrm{MDA}(H_\xi)$; can be extended to dependent processes with extremal index $\theta = 1$).
- Let $u_n(x) = c_n x + d_n$ (x fixed). We know $F^n(u_n(x)) \underset{n\uparrow\infty}{\to} H_\xi(x)$. Taking $-\log(\cdot)$ and using $-\log y \approx 1 - y$ for $y \to 1$, we obtain $n\bar{F}(u_n(x)) \approx -n \log F(u_n(x)) = -\log(F^n(u_n(x))) \underset{n\uparrow\infty}{\to} -\log H_\xi(x)$.
- $N_{u_n(x)}$ (exceedances among $X_1, \ldots, X_n$) fulfills $N_{u_n(x)} \sim \mathrm{B}(n, \bar{F}(u_n(x)))$
- The Poisson Limit Theorem ($n \to \infty$, $p = \bar{F}(u_n(x)) \to 0$, $np = n\bar{F}(u_n(x)) \to \lambda = -\log H_\xi(x)$) implies $N_{u_n(x)} \underset{n\uparrow\infty}{\to} \mathrm{Poi}(-\log H_\xi(x))$.
- 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 $Y_1, \ldots, Y_n$ 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:

  1) For $A \subseteq \mathcal{X}$ and $k \geq 0$,

$$\mathbb{P}(N(A) = k) = \begin{cases} e^{-\Lambda(A)} \frac{\Lambda(A)^k}{k!}, & \text{if } \Lambda(A) < \infty, \\ 0, & \text{if } \Lambda(A) = \infty. \end{cases}$$

  2) $N(A_1), \ldots, N(A_m)$ are independent for any mutually disjoint subsets $A_1, \ldots, A_m$ of $\mathcal{X}$.

- Note that $\mathbb{E}N(A) = \Lambda(A)$. Also, the *intensity (function)* is the function $\lambda(\boldsymbol{x})$ which satisfies $\Lambda(A) = \int_A \lambda(\boldsymbol{x})\, d\boldsymbol{x}$.

2) Asymptotic behaviour of the point process of exceedances

- For $n \in \mathbb{N}$ and $i \in \{1, \ldots, n\}$ 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 $n \to \infty$.
- Embrechts et al. (1997, Theorem 5.3.2) show that $N_n(\cdot)$ converges in distribution on $\mathcal{X}$ to a Poisson process $N(\cdot)$ with intensity $\Lambda(\cdot)$ satisfying $\Lambda(A) = (t_2 - t_1)\lambda(x)$ for $A = (t_1, t_2) \subseteq \mathcal{X}$, $\lambda(x) = -\log H_\xi(x)$.

- In particular, $\mathbb{E}N_n(A) \underset{n\uparrow\infty}{\to} \mathbb{E}N(A) = \Lambda(A) = (t_2 - t_1)\lambda(x)$. 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 $\lambda = -\log H_\xi(x) = -\log H_\xi((u - d_n)/c_n) = -\log H_{\xi,\mu=d_n,\sigma=c_n}(u)$.

  $\Rightarrow$ 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 $\tilde{X}_1 < \cdots < \tilde{X}_{N_u}$, $A = (0,1] \times (u, \infty)$ and $\Lambda(A) = \tau(u) =: \tau_u$, the likelihood $L(\xi, \sigma, \mu; \tilde{X}_1, \ldots, \tilde{X}_{N_u})$ is

$$\underbrace{N_u!}_{\text{ordered sample}} \underbrace{e^{-\Lambda(A)} \frac{\Lambda(A)^{N_u}}{N_u!}}_{\text{prob. of } N_u \text{ samples}} \prod_{i=1}^{N_u} \underbrace{\frac{\lambda(\tilde{X}_i)}{\Lambda(A)}}_{\text{density of } \tilde{X}_i} = e^{-\Lambda(A)} \prod_{i=1}^{N_u} \lambda(\tilde{X}_i) = e^{-\tau_u} \prod_{i=1}^{N_u} \lambda(\tilde{X}_i).$$

- Reparametrizing λ by $\tau_u = -\log H_{\xi,\mu,\sigma}(u) = (1 + \xi \frac{u-\mu}{\sigma})^{-1/\xi}$ and

$\beta = \sigma + \xi(u-\mu)$, we obtain

$$\lambda(x) = \frac{1}{\sigma}\Big(1+\xi\frac{x-\mu}{\sigma}\Big)^{-\frac{1}{\xi}-1} = \frac{1}{\sigma}\Big(\Big(1+\xi\frac{u-\mu}{\sigma}\Big)\Big(1+\frac{\xi\frac{x-u}{\sigma}}{1+\xi\frac{u-\mu}{\sigma}}\Big)\Big)^{-\frac{1}{\xi}-1}$$

$$= \frac{\tau_u}{\sigma(1+\xi\frac{u-\mu}{\sigma})}\Big(1+\frac{\xi\frac{x-u}{\sigma}}{1+\xi\frac{u-\mu}{\sigma}}\Big)^{-\frac{1}{\xi}-1} = \frac{\tau_u}{\beta}\Big(1+\frac{\xi(x-u)}{\sigma+\xi(u-\mu)}\Big)^{-\frac{1}{\xi}-1}$$

$$= \frac{\tau_u}{\beta}\Big(1+\frac{\xi(x-u)}{\beta}\Big)^{-\frac{1}{\xi}-1} = \tau_u g_{\xi,\beta}(x-u),$$

where $\xi \in \mathbb{R}$ and $\tau_u, \beta > 0$. Therefore, $\ell(\xi, \sigma, \mu; \tilde{X}_1, \ldots, \tilde{X}_{N_u})$ equals

$$= -\tau_u + \sum_{i=1}^{N_u} \log \lambda(\tilde{X}_i) = -\tau_u + N_u \log \tau_u + \sum_{i=1}^{N_u} \overbrace{(\log \lambda(\tilde{X}_i) - \log \tau_u)}^{=\log g_{\xi,\beta}(\tilde{X}_i - u)}$$

$$= \ell_{\text{Poi}}(\tau_u; N_u) - N_u \log(T) + \ell_{\text{GPD}}(\xi, \beta; \tilde{X}_1 - u, \ldots, \tilde{X}_{N_u} - u), \tag{32}$$

where ℓ_{Poi} is the log-likelihood for a one-dimensional homogeneous Poisson process with rate τ_u and ℓ_{GPD} is the log-likelihood for fitting a GPD to the excesses $\tilde{X}_i - u$, $i \in \{1, \ldots, N_u\}$.

- 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).

  $\Rightarrow$ In practice, we would expect the estimated parameters of the Poisson model to be roughly stable over a range of high thresholds.

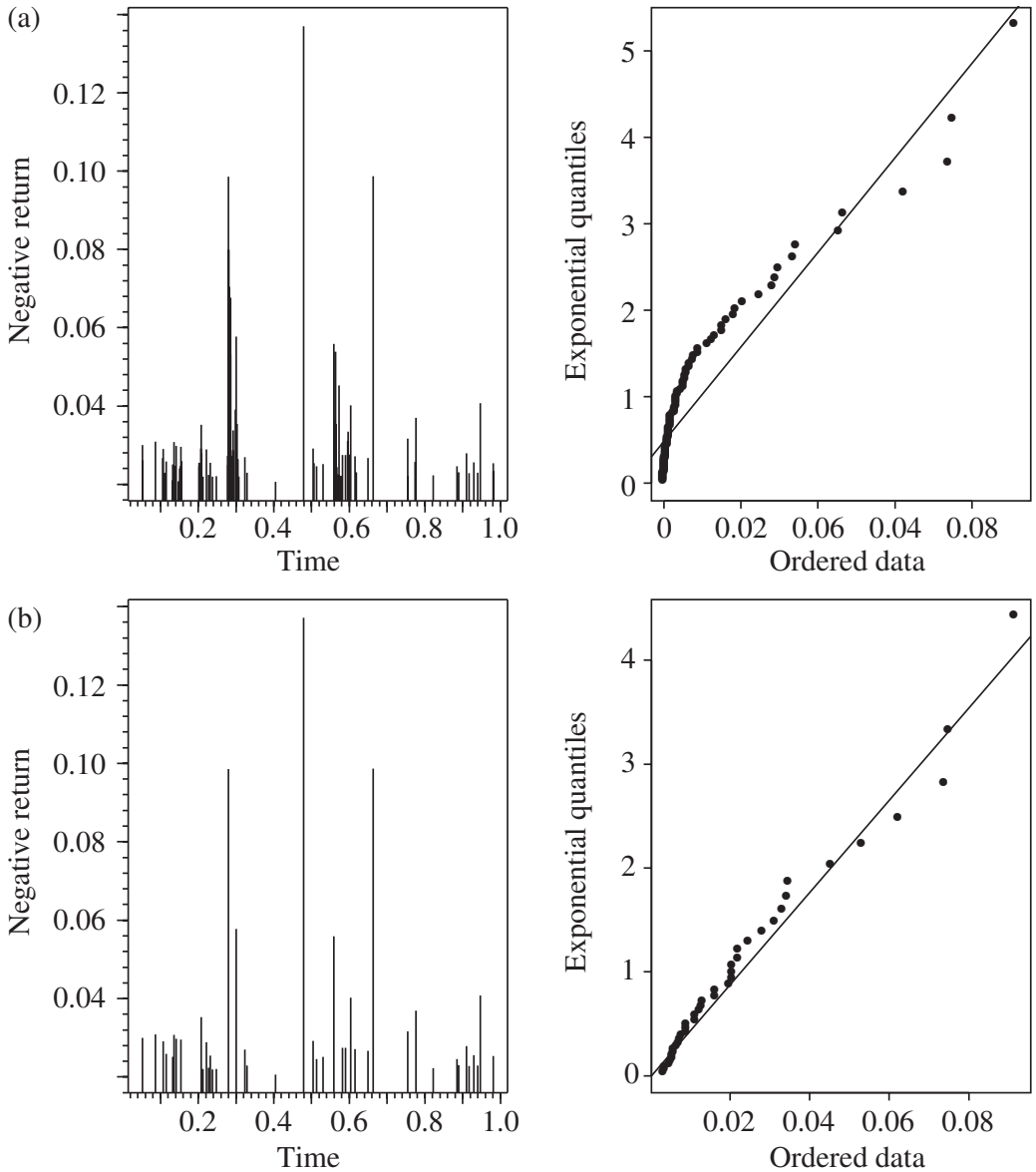
- The intensity λ is thus often used to introduce covariates to obtain Poisson processes which are non-homogeneous in time, e.g. by replacing μ and σ by parameters that vary over time as functions of covariates; see, e.g. Chavez-Demoulin et al. (2014).

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.

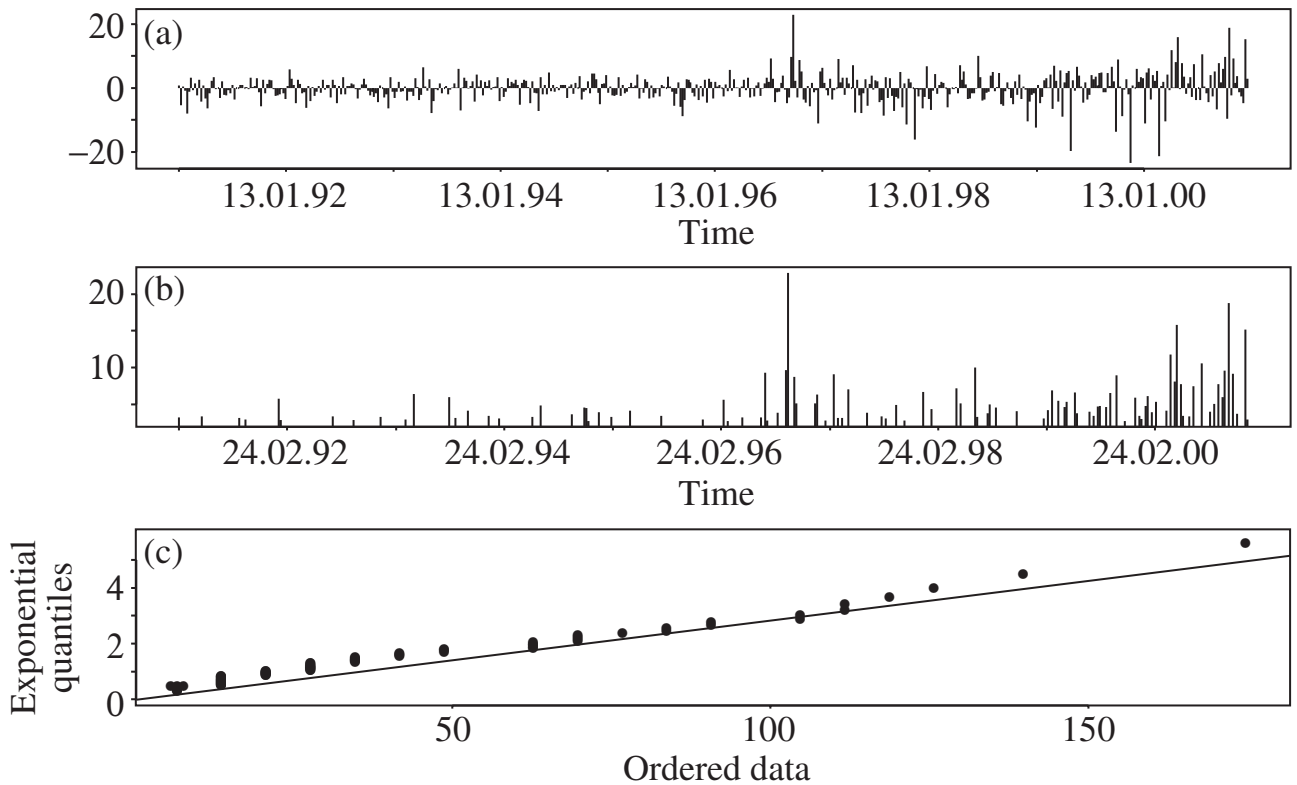
- *Declustering* may circumvent the problem. One identifies clusters (ad hoc; not easy) of exceedances and then applis the POT model to cluster maxima only.
- A possible declustering algorithm is the *runs method*. A run size r is fixed and two successive exceedances are said to belong to two different 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.7 (POT analysis of AT&T weekly losses (continued))

Consider the 102 weekly percentage losses exceeding $u = 2.75\%$:



- Inter-exceedance times seem to follow an exponential distribution.
- But exceedances become more frequent over time (which contradicts a homogeneous Poisson process)

- Using the log-likelihood (32), we fit a two-dimensional Poisson model to the 102 exceedances of $u = 2.75\%$. The parameter estimates are $\hat{\xi} = 0.22$, $\hat{\mu} = 19.9$ and $\hat{\sigma} = 5.95$.
- The implied GPD scale parameter is $\hat{\beta} = \hat{\sigma} + \hat{\xi}(u - \hat{\mu}) = 2.1 \Rightarrow$ The same $\hat{\xi}$ and $\hat{\beta}$ as in Example 5.21.
- 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

$u = H^{-1}_{\hat{\xi},\hat{\mu},\hat{\sigma}}(\exp(-1)) = 19.9$ 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 (15). We have

$$\begin{aligned}
&\int_{(-\infty, \boldsymbol{x}_1]} F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \,|\, \boldsymbol{z})\, dF_{\boldsymbol{X}_1}(\boldsymbol{z}) \\
&= \int_{\mathbb{R}^d} I_{\{\boldsymbol{z} \le \boldsymbol{x}_1\}} \mathbb{E}(I_{\{\boldsymbol{X}_2 \le \boldsymbol{x}_2\}} \,|\, \boldsymbol{X}_1 = \boldsymbol{z})\, dF_{\boldsymbol{X}_1}(\boldsymbol{z}) \\
&= \mathbb{E}(I_{\{\boldsymbol{X}_1 \le \boldsymbol{x}_1\}} \mathbb{E}(I_{\{\boldsymbol{X}_2 \le \boldsymbol{x}_2\}} \,|\, \boldsymbol{X}_1)) = \mathbb{E}(\mathbb{E}(I_{\{\boldsymbol{X}_1 \le \boldsymbol{x}_1, \boldsymbol{X}_2 \le \boldsymbol{x}_2\}} \,|\, \boldsymbol{X}_1)) \\
&\underset{\text{property}}{\overset{\text{tower}}{=}} \mathbb{E}(I_{\{\boldsymbol{X}_1 \le \boldsymbol{x}_1, \boldsymbol{X}_2 \le \boldsymbol{x}_2\}}) = F(\boldsymbol{x}),
\end{aligned}$$

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 $\mathrm{N}(0,1)$*; see Proposition 6.3.* The rv $Z \sim \mathrm{N}(0,1)$ 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

$$\phi_Z(t) = \int_{-\infty}^{\infty} (\cos(tx) + i\sin(tx))\varphi(x)\,dx \underset{\text{i)}}{=} \int_{-\infty}^{\infty} \cos(tx)\varphi(x)\,dx.$$

Hence,

$$\phi_Z'(t) = \int_{-\infty}^{\infty} \sin(tx)(-x)\varphi(x)\,dx \underset{\text{ii)}}{=} \int_{-\infty}^{\infty} \sin(tx)\varphi'(x)\,dx \underset{\text{by parts}}{=} -t\phi_Z(t).$$

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.8 (Cramér–Wold)

Let $\boldsymbol{X}, \boldsymbol{X}_n$, $n \in \mathbb{N}$, be random vectors. Then

$$\boldsymbol{X}_n \xrightarrow[n\uparrow\infty]{\mathrm{d}} \boldsymbol{X} \quad \Longleftrightarrow \quad \boldsymbol{a}'\boldsymbol{X}_n \xrightarrow[n\uparrow\infty]{\mathrm{d}} \boldsymbol{a}'\boldsymbol{X} \quad \forall\, \boldsymbol{a} \in \mathbb{R}^d$$

Proof.

"$\Rightarrow$" This follows from the Continuous Mapping Theorem with the map $g(\boldsymbol{x}) = \boldsymbol{a}'\boldsymbol{x}$.

"$\Leftarrow$" Note that $\phi_{\boldsymbol{X}_n}(\boldsymbol{t}) = \mathbb{E}(\exp(i \cdot 1 \cdot \boldsymbol{t}'\boldsymbol{X}_n)) = \phi_{\boldsymbol{t}'\boldsymbol{X}_n}(1) \underset{n\uparrow\infty}{\to} \phi_{\boldsymbol{t}'\boldsymbol{X}}(1) = \phi_{\boldsymbol{X}}(\boldsymbol{t})$ for all $\boldsymbol{t}$ and apply the Lévy Continuity Theorem. □

Corollary A.9

Let $\boldsymbol{X}, \boldsymbol{Y}$ be two random vectors. Then

$$\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{Y} \quad \Longleftrightarrow \quad \boldsymbol{a}'\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{a}'\boldsymbol{Y} \quad \forall\, \boldsymbol{a} \in \mathbb{R}^d.$$

Properties of multivariate normal variance mixtures

Proof of Lemma 6.10. W.l.o.g. assume $\boldsymbol{\mu} = \mathbf{0}$.

"$\Rightarrow$" $\mathbb{E}|X_i|\,\mathbb{E}|X_j| \overset{\text{ind.}}{=} \mathbb{E}(|X_i||X_j|) = \mathbb{E}(W|Z_i||Z_j|) \overset{\text{ind.}}{=} \mathbb{E}(W)\,\mathbb{E}|Z_i|\,\mathbb{E}|Z_j|$

$\underset{\text{Jensen}}{\geq} \mathbb{E}(\sqrt{W})^2\,\mathbb{E}|Z_i|\,\mathbb{E}|Z_j| \overset{\text{ind.}}{=} \mathbb{E}|\sqrt{W}Z_i|\,\mathbb{E}|\sqrt{W}Z_j| = \mathbb{E}|X_i|\mathbb{E}|X_j|$

$\Rightarrow$ We must have "=" in Jensen's inequality. This hols if and only if W is constant a.s.; so $\boldsymbol{X} \sim \mathrm{N}_d(\mathbf{0}, WI_d)$ in this case.

"$\Leftarrow$" W a.s. constant $\Rightarrow \boldsymbol{X} \sim \mathrm{N}_d(\mathbf{0}, WI_d) \Rightarrow X_i, X_j$ independent. $\square$

Spherical distributions

Proof of Theorem 6.15.

1) $\Rightarrow$ 2): $\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = \phi_{U\boldsymbol{Y}}(\boldsymbol{t}) = \phi_{\boldsymbol{Y}}(U'\boldsymbol{t})$ for all $U \in \mathbb{R}^{d\times d}$ orthogonal. Since U can only change the direction of $\boldsymbol{t}$ but not its length, $\phi_{\boldsymbol{Y}}(\boldsymbol{t})$ only depends on $\|\boldsymbol{t}\|$, i.e. the length of $\boldsymbol{t}$ $\Rightarrow$ we can define $\psi(\|\boldsymbol{t}\|^2) = \phi_{\boldsymbol{Y}}(\boldsymbol{t})$.

2) $\Rightarrow$ 3): $\phi_{Y_1}(t) = \phi_{\boldsymbol{Y}}(te_1) \underset{2)}{=} \psi(t^2)$ $(*)$. Now $\phi_{\boldsymbol{a}'\boldsymbol{Y}}(t) = \phi_{\boldsymbol{Y}}(t\boldsymbol{a}) \underset{2)}{=} \psi(t^2\|\boldsymbol{a}\|^2) = \psi((t\|\boldsymbol{a}\|)^2) \underset{(*)}{=} \phi_{Y_1}(t\|\boldsymbol{a}\|) = \phi_{\|\boldsymbol{a}\|Y_1}(t)$

3) $\Rightarrow$ 1): $\phi_{U\boldsymbol{Y}}(\boldsymbol{t}) = \mathbb{E}(\exp(i(U'\boldsymbol{t})'\boldsymbol{Y})) \underset{U'\boldsymbol{t}=:\boldsymbol{a}}{=} \mathbb{E}(\exp(i\boldsymbol{a}'\boldsymbol{Y})) \underset{3)}{=} \mathbb{E}(\exp(i\|\boldsymbol{a}\|Y_1))$
$= \mathbb{E}(\exp(i\|\boldsymbol{t}\|Y_1)) \underset{3)}{=} \mathbb{E}(\exp(i\boldsymbol{t}'\boldsymbol{Y})) = \phi_{\boldsymbol{Y}}(\boldsymbol{t})$ $\square$

Proof of Theorem 6.16. Let Ω_d be the characteristic generator of $\boldsymbol{S}$.

"$\Rightarrow$" $\boldsymbol{Y} \sim S_d(\psi) \Rightarrow \phi_{\boldsymbol{Y}}(\|\boldsymbol{t}\|\boldsymbol{u}) \underset{2)}{=} \psi(\|\boldsymbol{t}\|^2\boldsymbol{u}'\boldsymbol{u}) = \psi(\|\boldsymbol{t}\|^2)$ for all $\boldsymbol{u} \in \mathbb{R}^d$: $\|\boldsymbol{u}\| = 1$. Replacing $\boldsymbol{u}$ by $\boldsymbol{S}$ and integrating leads to $\psi(\|\boldsymbol{t}\|^2) = \mathbb{E}_{\boldsymbol{S}}(\phi_{\boldsymbol{Y}}(\|\boldsymbol{t}\|\boldsymbol{S})) = \mathbb{E}_{\boldsymbol{S}}(\mathbb{E}_{\boldsymbol{Y}}(e^{i\|\boldsymbol{t}\|\boldsymbol{S}'\boldsymbol{Y}})) \underset{\text{Fubini}}{=} \mathbb{E}_{\boldsymbol{Y}}(\mathbb{E}_{\boldsymbol{S}}(e^{i\|\boldsymbol{t}\|\boldsymbol{S}'\boldsymbol{Y}})) = \mathbb{E}_{\boldsymbol{Y}}(\phi_{\boldsymbol{S}}(\|\boldsymbol{t}\|\boldsymbol{Y})) \underset{2)}{=} \mathbb{E}_{\boldsymbol{Y}}(\Omega_d(\|\boldsymbol{t}\|^2\boldsymbol{Y}'\boldsymbol{Y}))$. We thus obtain that $\phi_{\boldsymbol{Y}}(\boldsymbol{t}) \underset{2)}{=} \psi(\|\boldsymbol{t}\|^2) \underset{R:=\|\boldsymbol{Y}\|}{=} \mathbb{E}_R(\Omega_d(\|\boldsymbol{t}\|^2R^2)) = \int_0^\infty \Omega_d(\|\boldsymbol{t}\|^2r^2)\,dF_R(r)$
$\underset{2)}{=} \int_0^\infty \phi_{\boldsymbol{S}}(r\boldsymbol{t})\,dF_R(r) = \phi_{R\boldsymbol{S}}(\boldsymbol{t})$ for all $\boldsymbol{t} \in \mathbb{R}^d$.

"⇐" Let $\boldsymbol{Z} \sim \mathrm{N}_d(\boldsymbol{0}, I_d)$. Since $\boldsymbol{Z}$ is spherical and $\|\boldsymbol{Z}/\|\boldsymbol{Z}\|\| = \|\boldsymbol{Z}\|/\|\boldsymbol{Z}\| = 1$, $\boldsymbol{S} \overset{\mathrm{d}}{=} \boldsymbol{Z}/\|\boldsymbol{Z}\|$. As such, $\boldsymbol{S}$ itself is spherical, since $U\boldsymbol{S} \overset{\mathrm{d}}{=} U\boldsymbol{Z}/\|\boldsymbol{Z}\| \overset{\mathrm{d}}{=} \boldsymbol{Z}/\|\boldsymbol{Z}\| \overset{\mathrm{d}}{=} \boldsymbol{S}$ for any orthogonal $U \in \mathbb{R}^{d\times d}$. Theorem 6.15 Part 2) implies that $\phi_{\boldsymbol{S}}(\boldsymbol{t}) = \Omega_d(\|\boldsymbol{t}\|^2)$, so $\phi_{R\boldsymbol{S}}(\boldsymbol{t}) = \mathbb{E}(\exp(i\boldsymbol{t}'R\boldsymbol{S})) = \mathbb{E}_R(\,\mathbb{E}(\exp(i\boldsymbol{t}'R\boldsymbol{S})\,|\,R)\,) = \mathbb{E}_R(\phi_{\boldsymbol{S}}(R\boldsymbol{t})) = \mathbb{E}_R(\Omega_d(R^2\|\boldsymbol{t}\|^2))$, which is a function in $\|\boldsymbol{t}\|^2$ and thus, by 2), $R\boldsymbol{S}$ is spherical. □

Density of $\boldsymbol{Y} \sim S_d(\psi)$ constant on spheres

If $\boldsymbol{Y}$ admits a density $f_{\boldsymbol{Y}}$, then $f_{\boldsymbol{Y}}(\boldsymbol{y})$ is constant on hyperspheres in $\mathbb{R}^d$.

The *inversion formula* $f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\boldsymbol{t}'\boldsymbol{y}}\phi_{\boldsymbol{Y}}(\boldsymbol{t})\,d\boldsymbol{t}$ and Theorem 6.15 Part 2) show that for any orthogonal U,

$$
\begin{aligned}
f_{\boldsymbol{Y}}(U\boldsymbol{y}) &\underset{\text{inv.}}{=} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i(U'\boldsymbol{t})'\boldsymbol{y}}\phi_{\boldsymbol{Y}}(\boldsymbol{t})\,d\boldsymbol{t} \\
&\underset{\text{subs.}}{=} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\boldsymbol{s}'\boldsymbol{y}}\phi_{\boldsymbol{Y}}(U\boldsymbol{s})\,d\boldsymbol{s}
\end{aligned}
$$

$$\underset{2)}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-is'\boldsymbol{y}} \psi((U\boldsymbol{s})'U\boldsymbol{s})\, d\boldsymbol{s}$$
$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-is'\boldsymbol{y}} \psi(\boldsymbol{s}'\boldsymbol{s})\, d\boldsymbol{s} \underset{\text{backwards}}{=} \cdots = f_{\boldsymbol{Y}}(\boldsymbol{y}).$$

This implies that $f_{\boldsymbol{Y}}(\boldsymbol{y}) = g(\|\boldsymbol{y}\|^2)$ for a function $g : [0,\infty) \to [0,\infty)$, the density generator. For $\boldsymbol{Y} \sim t_d(\nu, \boldsymbol{0}, I_d)$, 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.10

Let $\boldsymbol{X} \sim \mathrm{E}_d(\boldsymbol{\mu}, \Sigma, \psi)$ for positive definite Σ and $\mathbb{E}(R^2) < \infty$ (i.e. $\mathrm{cov}(\boldsymbol{X})$ finite). For any $c \geq 0$ such that $\mathbb{P}((\boldsymbol{X}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \geq c) > 0$,

$$\mathrm{corr}(\boldsymbol{X} \,|\, (\boldsymbol{X}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \geq c) = \mathrm{corr}(\boldsymbol{X}).$$

Proof. $\boldsymbol{X} \,|\, ((\boldsymbol{X}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \geq c) \overset{\mathrm{d}}{=} \underset{(23)}{\boldsymbol{\mu} + RA\boldsymbol{S}} \,|\, (R^2 \geq c) \overset{\text{ind.}}{=} \boldsymbol{\mu} + \tilde{R}A\boldsymbol{S}$ where $\tilde{R} \overset{\mathrm{d}}{=} (R \,|\, R^2 \geq c)$. A (and thus Σ) remains the same. □

Estimating scale and correlation

- Suppose $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim \mathrm{E}_d(\boldsymbol{\mu}, \Sigma, \psi)$. How can we estimate $\boldsymbol{\mu}, \Sigma$ and P? (P is the correlation matrix corresponding to Σ; this always exists)
- $\bar{\boldsymbol{X}}$, 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 $D_i = \sqrt{(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})' \hat{\Sigma}^{-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})}$ (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.11 (M-estimators of location and scale)

1) Set $k = 1$, $\hat{\boldsymbol{\mu}}^{[1]} = \bar{\boldsymbol{X}}$ and $\hat{\Sigma}^{[1]} = S$.

2) Repeat until convergence:

   2.1) For $i \in \{1, \ldots, n\}$ set $D_i = \sqrt{(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})' \hat{\Sigma}^{[k]-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})}$.

   2.2) Update:

$$\hat{\boldsymbol{\mu}}^{[k+1]} = \frac{\sum_{i=1}^n w_1(D_i) \boldsymbol{X}_i}{\sum_{i=1}^n w_1(D_i)},$$

   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:

$$\hat{\Sigma}^{[k+1]} = \frac{1}{n} \sum_{i=1}^n w_2(D_i^2)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})',$$

   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.12 (One-factor/equicorrelation model)

Let $\mathbb{E}(\boldsymbol{X}) = \mathbf{0}$, $\Sigma = \operatorname{cov}(\boldsymbol{X}) = \rho J_d + (1-\rho) I_d$ ($J_d = (1) \in \mathbb{R}^{d \times d}$).

- Then $\Sigma = BB' + \Upsilon$ for $B = \sqrt{\rho}\mathbf{1}$ and $\Upsilon = (1-\rho) I_d$.
- Any Y with $\mathbb{E}Y = 0$, $\operatorname{var} Y = 1$ independent of $\boldsymbol{X}$ leads to the *factor decomposition* of $\boldsymbol{X}$

$$F = \frac{\sqrt{\rho}}{1 + \rho(d-1)} \sum_{j=1}^{d} X_j + \sqrt{\frac{1-\rho}{1+\rho(d-1)}} Y, \quad \varepsilon_j = X_j - \sqrt{\rho} F.$$

  We have $\mathbb{E}(F) = 0$, $\operatorname{var}(F) = 1$, so $\boldsymbol{X} = \mathbf{0} + BF + \boldsymbol{\varepsilon} = \sqrt{\rho}\mathbf{1}F + \boldsymbol{\varepsilon}$.
- The requirements of Definition 6.24 are fulfilled since $\operatorname{cov}(F, \varepsilon_j) = 0$, $\operatorname{cov}(\varepsilon_j, \varepsilon_k) = 0$ for all $j \neq k$.
- $\operatorname{var}(\bar{X}_n) = \operatorname{var}(\sqrt{\rho} F + \bar{\varepsilon}_d) = \rho + \frac{1-\rho}{d} \underset{(d \to \infty)}{\to} \rho$ (systematic factor matters!)

- If $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, take $Y \sim \mathrm{N}(0,1)$ (then F is also normal). One typically writes this (one-factor) equicorrelation model as $\boldsymbol{X} = \sqrt{\rho} F + \sqrt{1-\rho}\boldsymbol{Z}$, where $F, Z_1, \ldots, Z_d \overset{\text{ind.}}{\sim} \mathrm{N}(0,1)$.

Multivariate regression

- Here, construct large matrices:

$$X = \underbrace{\begin{pmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix}}_{n \times d}, \quad F = \underbrace{\begin{pmatrix} 1 & \boldsymbol{F}_1' \\ \vdots & \vdots \\ 1 & \boldsymbol{F}_n' \end{pmatrix}}_{n \times (p+1)}, \quad \tilde{B} = \underbrace{\begin{pmatrix} \boldsymbol{a}' \\ B' \end{pmatrix}}_{(p+1) \times d}, \quad E = \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_1' \\ \vdots \\ \boldsymbol{\varepsilon}_n' \end{pmatrix}}_{n \times d}.$$

  This model can be expressed by $X = F\tilde{B} + E$ (estimate $\tilde{B}$).

- Assume the unobserved $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ form a white noise process. Then, conditional on $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_n$, we have a multivariate linear regression, see, e.g. Mardia et al. (1979), with estimator $\hat{\tilde{B}} = (F'F)^{-1}F'X$.

- Now examine the condititions of Definition 6.24: 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 $\hat{E} = X - F\hat{\hat{B}}$ (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 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with identical distribution, unknown mean vector $\boldsymbol{\mu}$ and covariance matrix Σ with the spectral decomposition $\Sigma = \Gamma\Lambda\Gamma'$ as before.
- Estimate $\boldsymbol{\mu}$ by $\bar{\boldsymbol{X}}$ and Σ by $S_x = \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})'$.
- Apply the spectral decomposition to S_x to get $S_x = GLG'$, where G is the eigenvector matrix and $L = \mathrm{diag}(l_1, \ldots, l_d)$ is the diagonal matrix consisting of ordered eigenvalues.

- Define the "sample principle component transforms" $\boldsymbol{Y}_t = G'(\boldsymbol{X}_t - \bar{\boldsymbol{X}})$, $t \in \{1, \ldots, n\}$. The jth component $Y_{t,j} = \boldsymbol{g}_j'(\boldsymbol{X}_t - \bar{\boldsymbol{X}})$ is the *jth sample principal component at time t* ($\boldsymbol{g}_j$ is the jth column of G).
- The rotated vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ have sample covariance matrix L:

$$S_y = \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})' = \frac{1}{n}\sum_{t=1}^{n}\boldsymbol{Y}_t\boldsymbol{Y}_t'$$
$$= \frac{1}{n}\sum_{t=1}^{n}G'(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})'G = G'S_xG = L.$$

  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 $\boldsymbol{Y}_t$ to calibrate an approximate factor model. We assume our data are realizations from the model

$$\boldsymbol{X}_t = \bar{\boldsymbol{X}} + G_1\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \quad t \in \{1, \ldots, n\},$$

where G_1 consists of the first k columns of G and $\boldsymbol{F}_t = (Y_{t,1}, \ldots, Y_{t,k})$, $t \in \{1, \ldots, n\}$.

- In practice, the errors $\boldsymbol{\varepsilon}_t$ do not have a diagonal covariance matrix and are not uncorrelated with $\boldsymbol{F}_t$. 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

Basic properties

Lemma A.13
For any copula C, $|C(\boldsymbol{b}) - C(\boldsymbol{a})| \leq \sum_{j=1}^{d} |b_j - a_j|$ for all $\boldsymbol{a}, \boldsymbol{b} \in [0,1]^d$.

Proof. Using a telescoping sum expansion and the triangle inequality, we obtain

$$|C(\boldsymbol{b}) - C(\boldsymbol{a})| \leq \sum_{j=1}^{d} |C(b_1, \ldots, b_{j-1}, b_j, a_{j+1}, \ldots, a_d) - C(b_1, \ldots, b_{j-1}, a_j, a_{j+1}, \ldots, a_d)|.$$

W.l.o.g. let $\boldsymbol{a} \leq \boldsymbol{b}$. By d-increasingness, $C \nearrow$ component-wise, so omit $|\cdot|$. Since, by d-increasingness, the jth summand is $\nearrow$ in each component $\neq j$, let $b_1, \ldots, b_{j-1}, a_{j+1}, \ldots, a_d \nearrow 1$ to obtain the upper bound $\sum_{j=1}^{d} C(1, \ldots, 1, b_j, 1, \ldots, 1) - C(1, \ldots, 1, a_j, 1, \ldots, 1) = b_j - a_j$ for summand j. □

Generalized inverses

$T \nearrow$ means that T is *increasing*; $T \uparrow$ means that T is *strictly increasing*; $\operatorname{ran} T = \{T(x) : x \in \mathbb{R}\}$ denotes the *range* of T.

Proposition A.14 (Working with generalized inverses)
Let $T : \mathbb{R} \to \mathbb{R} \nearrow$ 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 $\operatorname{ran} T$ denote the *range of* T, i.e. $\operatorname{ran} T = \{T(x) : x \in \mathbb{R}\}$. $T^{\leftarrow}(y) < \infty$ implies $T(T^{\leftarrow}(y)) \geq y$. Furthermore, $y \in \operatorname{ran} T \cup \{\inf T, \sup T\}$ 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 $\operatorname{ran} T$.

(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

$$\operatorname{ran} X = \{x \in \mathbb{R} : \mathbb{P}(X \in (x-h, x]) > 0 \text{ for all } h > 0\}.$$

Since F is continuous on $\mathbb{R}$, (GI7) implies that $F^{\leftarrow} \uparrow$ on $[\inf F, \sup F] = [0,1]$. Thus

$$\begin{aligned}\mathbb{P}(F(X) \leq u) &\underset{\text{(GI7)}}{=} \mathbb{P}(F^{\leftarrow}(F(X)) \leq F^{\leftarrow}(u)) \underset{\text{(GI3)}}{=} \mathbb{P}(X \leq F^{\leftarrow}(u)) \\ &= F(F^{\leftarrow}(u)) \underset{\text{(GI4)}}{=} u, \quad u \in [0,1],\end{aligned}$$

where (GI3) applies since $F \uparrow$ on $\operatorname{ran} X$. □

Proof of Lemma 7.6.

"⇒" $\mathbb{P}(F_j(X_j) \leq u_j \,\forall\, j) \underset{\text{cont.}}{=} \mathbb{P}(F_j(X_j) < u_j \,\forall\, j) \underset{\text{(GI5)}}{=} \mathbb{P}(X_j < F_j^{\leftarrow}(u_j) \,\forall\, j)$

$$\underset{\text{cont.}}{=} \mathbb{P}(X_j \leq F_j^{\leftarrow}(u_j) \,\forall\, j) = F(F_1^{\leftarrow}(u_1), \ldots, F_d^{\leftarrow}(u_d)) \underset{\text{Sklar}}{=} C(\boldsymbol{u}).$$

"⇐" Since $F_j \uparrow$ on $\operatorname{ran} X_j$, $j \in \{1, \ldots, d\}$,

$$F(\boldsymbol{x}) \underset{\text{(GI3)}}{=} \mathbb{P}(F_j^{\leftarrow}(F_j(X_j)) \leq x_j \,\forall\, j) \underset{\text{(GI5)}}{=} \mathbb{P}(F_j(X_j) \leq F_j(x_j) \,\forall\, j)$$

$$\underset{\text{ass.}}{=} C(F_1(x_1), \ldots, F_d(x_d)) \quad \underset{\text{Sklar}}{\Rightarrow} \quad \boldsymbol{X} \text{ has copula } C \qquad \square$$

Proof of Theorem 7.8.

1) ■ By Lemma A.13, $1 - C(\boldsymbol{u}) = C(\mathbf{1}) - C(\boldsymbol{u}) \leq \sum_{j=1}^d (1 - u_j) = d - \sum_{j=1}^d u_j$, so $C(\boldsymbol{u}) \geq \sum_{j=1}^d u_j - d + 1$. Also, $C(\boldsymbol{u}) \geq 0$. So $C(\boldsymbol{u}) \geq W(\boldsymbol{u})$.

■ Since copulas are componentwise increasing, $C(\boldsymbol{u}) \leq C(1, \dots, 1, u_j, 1, \dots, 1) = u_j$ for all j. Hence, $C(\boldsymbol{u}) \leq \min_{1 \leq j \leq d}\{u_j\} = M(\boldsymbol{u})$.

2) W is a copula for $d = 2$ since $(U, 1-U) \sim W$ for $U \sim \mathrm{U}(0,1)$. W is not a copula for $d \geq 3$ since

$$
\begin{aligned}
&\Delta_{(\frac{\mathbf{1}}{\mathbf{2}}, \mathbf{1}]} W \\
&= \sum_{\boldsymbol{i} \in \{0,1\}^d} (-1)^{\sum_{j=1}^d i_j} W(\tfrac{1}{2}^{i_1}, \dots, \tfrac{1}{2}^{i_d}) \\
&= \max\{1 + 1 + 1 + \cdots + 1 - d + 1, 0\} \quad (i_j = 0 \ \forall\, j) \\
&\quad - d \max\{\tfrac{1}{2} + 1 + 1 + \cdots + 1 - d + 1, 0\} \quad (\exists!\, j : i_j = 1)
\end{aligned}
$$

$$+\binom{d}{2}\max\{\tfrac{1}{2}+\tfrac{1}{2}+1+\cdots+1-d+1,0\} \quad (\exists!\text{ two } j: i_j=1)$$
$$-\cdots+(-1)^d\max\{\tfrac{1}{2}+\cdots+\tfrac{1}{2}-d+1,0\} \quad (i_j=1\ \forall\, j)$$
$$=1-\tfrac{d}{2}<0 \quad \text{for } d\geq 3.$$

3) M is a copula for all $d\geq 2$ since $(U,\ldots,U)\sim M$ for $U\sim \mathrm{U}(0,1)$. $\square$

Extreme value and Marshall–Olkin copulas

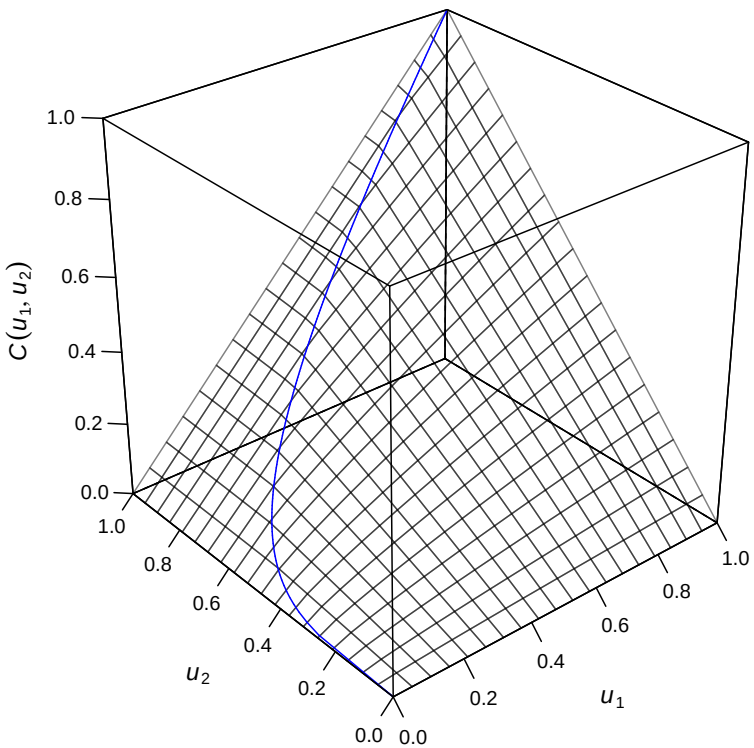
- *Extreme value copulas* are the copulas C of limiting distributions of properly location-scale transformed componentwise maxima of a sequence of random vectors.
- They are given by

$$C(\boldsymbol{u}) = \left(\prod_{j=1}^{d} u_j \right)^{A\left(\frac{\log u_1}{\log \Pi(\boldsymbol{u})}, \ldots, \frac{\log u_d}{\log \Pi(\boldsymbol{u})} \right)}$$

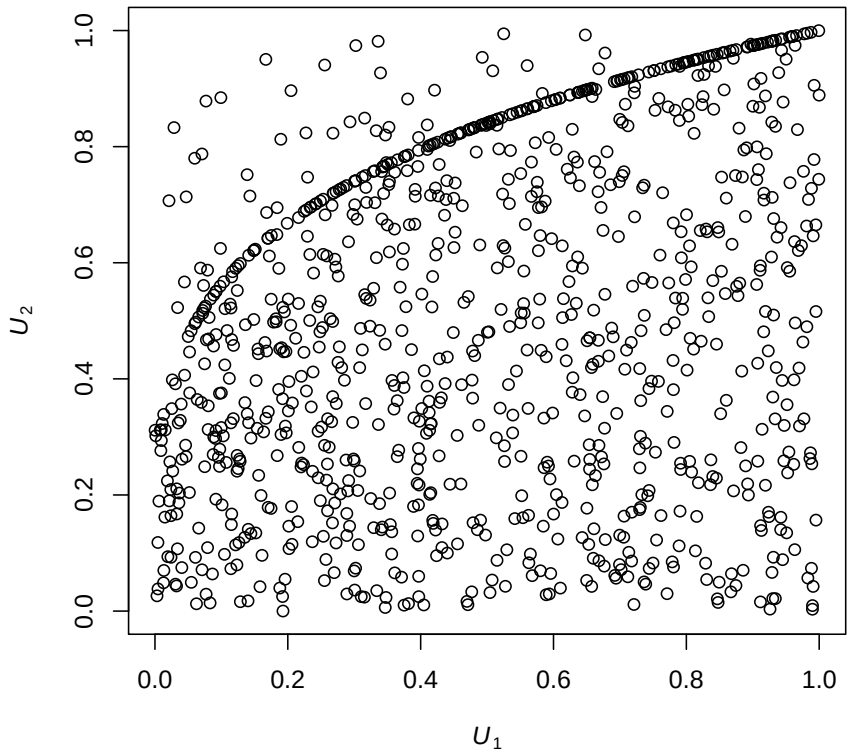
  for a *Pickands dependence function* A; see Ressel (2013) for a characterization of A.
- Examples: Gumbel copula, Marshall-Olkin copulas.
- For more details, see Jaworski et al. (2010, Chapter 6).

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 $\mathbb{P}(F_2^{\leftarrow}(1 - F_1(X_1)) \leq x) \underset{\text{(GI5)}}{=} \mathbb{P}(1 - F_1(X_1) \leq F_2(x)) = F_2(x)$. If (X_1, X_2) has copula C, then

$$\begin{aligned} C(\boldsymbol{u}) &\underset{\text{"only if"}}{\overset{\text{L.7.6}}{=}} \mathbb{P}(F_1(X_1) \leq u_1, F_2(F_2^{\leftarrow}(1 - F_1(X_1))) \leq u_2) \\ &\underset{\text{(GI4)}}{=} \mathbb{P}(F_1(X_1) \leq u_1, 1 - F_1(X_1) \leq u_2) \\ &= \mathbb{P}(1 - u_2 < U \leq u_1) = W(u_1, u_2) \quad \text{for } U \sim \mathrm{U}(0,1). \end{aligned}$$

"$\Leftarrow$" $W(u_1, u_2) = 0$ 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 $F_2 \uparrow \operatorname{ran} X_2$, we thus obtain $\mathbb{P}(X_2 = F_2^{\leftarrow}(1 - F_1(X_1))) = \mathbb{P}(F_2(X_2) = F_2(F_2^{\leftarrow}(1 - F_1(X_1)))) \underset{\text{(GI4)}}{=} \mathbb{P}(F_2(X_2) = 1 - F_1(X_1)) = \mathbb{P}(U_2 = 1 - U_1) = 1$. $\square$

Proof of Proposition 7.15. Consider $T(u) = F_1^{\leftarrow}(u) + \cdots + F_d^{\leftarrow}(u)$ $\nearrow$, left-continuous and let $U \sim \mathrm{U}(0,1)$. We first show that $F^{\leftarrow}_{T(U)}(u) = T(u)$, for all $u \in [0,1]$.

1) T left-continuous $\Rightarrow T(u) \le x \Leftrightarrow \; u \le u_x := \sup\{u : T(u) \le x\}$

2) 1) $\Rightarrow \{T(U) \le x\} = \{U \le u_x\} \; \Rightarrow F_{T(U)}(x) = F_U(u_x) = u_x$.

$\Rightarrow F^{\leftarrow}_{T(U)}(u) \le x \underset{\text{(GI5)}}{\Leftrightarrow} F_{T(U)}(x) \ge u \underset{2)}{\Leftrightarrow} u_x \ge u \underset{1)}{\Leftrightarrow} T(u) \le x$

Choosing $x = T(u)$ and $x = F^{\leftarrow}_{T(U)}(u)$ in the last line, we see that $F^{\leftarrow}_{T(U)}(u) = T(u)$. Now Proposition 7.14 2) implies that

$$(X_1, \dots, X_d) \overset{\mathrm{d}}{=} (F_1^{\leftarrow}(U), \dots, F_d^{\leftarrow}(U)),$$

so that

$$F_{\sum_{j=1}^d X_j}(x) = \mathbb{P}\left(\sum_{j=1}^d X_j \le x\right) = \mathbb{P}\left(\sum_{j=1}^d F_j^{\leftarrow}(U) \le x\right) = \mathbb{P}(T(U) \le x)$$

and thus $F^{\leftarrow}_{\sum_{j=1}^d X_j}(\alpha) = F^{\leftarrow}_{T(U)}(\alpha) = T(\alpha) = \sum_{j=1}^d F_j^{\leftarrow}(\alpha)$. $\square$

Linear correlation

Proof of Proposition 7.16. Let (X_1', X_2') be an iid-copy of (X_1, X_2). Consider

$$\begin{aligned} & 2\,\mathrm{cov}(X_1, X_2) \\ &= \mathbb{E}((X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)) + \mathbb{E}((X_1' - \mathbb{E}X_1')(X_2' - \mathbb{E}X_2')) \\ &\overset{\text{check}}{=} \mathbb{E}(((X_1 - \mathbb{E}X_1) - (X_1' - \mathbb{E}X_1')) \cdot ((X_2 - \mathbb{E}X_2) - (X_2' - \mathbb{E}X_2'))) \\ &= \mathbb{E}((X_1 - X_1')(X_2 - X_2')). \end{aligned}$$

With $b - a = \int_{-\infty}^{\infty} (I_{\{a \le x\}} - I_{\{b \le x\}})\, dx$ for all $a, b \in \mathbb{R}$, we obtain that

$$\begin{aligned} & 2\,\mathrm{cov}(X_1, X_2) \\ &= \mathbb{E}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (I_{\{X_1' \le x_1\}} - I_{\{X_1 \le x_1\}})(I_{\{X_2' \le x_2\}} - I_{\{X_2 \le x_2\}})\, dx_1 dx_2\right] \\ &\overset{\text{Fubini}}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{E}(\dots)\, dx_1 dx_2 \overset{\text{ind.}}{\overset{\text{multiply}}{=}} 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (F(x_1, x_2) - F_1(x_1)F_2(x_2))\, dx_1 dx_2. \end{aligned}$$

□

Rank correlation

To overcome (some) of the deficiencies of ρ, Scarsini (1984) introduced:

Definition A.15 (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) $C_1(\boldsymbol{u}) \leq C_2(\boldsymbol{u})$ for all $\boldsymbol{u} \in [0,1]^2$ implies $\kappa(C_1) \leq \kappa(C_2)$;
7) $C_n \to C$ $(n \to \infty)$ pointwise implies $\lim_{n\to\infty} \kappa(C_n) = \kappa(C)$.

Proposition A.16 (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 $\operatorname{ran} X_j$, $j \in \{1, 2\}$, 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

$$\begin{aligned}\rho_\tau &= \mathbb{P}((X_1 - X_1')(X_2 - X_2') > 0) - \mathbb{P}((X_1 - X_1')(X_2 - X_2') < 0) \\ &= 2\underbrace{\mathbb{P}((X_1 - X_1')(X_2 - X_2') > 0)}_{=2\mathbb{P}(X_1<X_1', X_2<X_2')} - 1 = 4\mathbb{P}(U_1 \le U_1', U_2 \le U_2') - 1 \\ &= 4\int_0^1 \int_0^1 \mathbb{P}(U_1 \le u_1, U_2 \le u_2)\, dC(u_1, u_2) - 1 \qquad \square\end{aligned}$$

For computing ρ_τ, $\int_{[0,1]^2} C(\boldsymbol{u})\, d\tilde{C}(\boldsymbol{u}) = \frac{1}{2} - \int_{[0,1]^2} D_1 C(\boldsymbol{u}) D_2 \tilde{C}(\boldsymbol{u})\, d\boldsymbol{u}$ is often helpful; see Li et al. (2002). One can also show that for any bivariate copulas C, $\tilde{C}$, $\int_{[0,1]^2} C(\boldsymbol{u})\, d\tilde{C}(\boldsymbol{u}) = \int_{[0,1]^2} \tilde{C}(\boldsymbol{u})\, dC(\boldsymbol{u})$.

Rank correlation for elliptical copulas

Lemma A.17

Let $\boldsymbol{X} \sim \mathrm{E}_2(\mathbf{0}, \Sigma, \psi)$ with $\mathbb{P}(\boldsymbol{X} = \mathbf{0}) = 0$ and $\rho = P_{12} = \mathrm{corr}(\Sigma)_{12}$. Then

$$\mathbb{P}(X_1 > 0, X_2 > 0) = \frac{1}{4} + \frac{\arcsin \rho}{2\pi}.$$

Proof.

- Note that $\boldsymbol{Y} = \left(\begin{smallmatrix} 1/\sqrt{\sigma_{11}} & 0 \\ 0 & 1/\sqrt{\sigma_{22}} \end{smallmatrix}\right) \boldsymbol{X} \sim \mathrm{E}_2(\mathbf{0}, P, \psi)$ with $P = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$.
- Let $A = \left(\begin{smallmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{smallmatrix}\right)$ so that $AA' = P$. Then $\boldsymbol{Y} \overset{\mathrm{d}}{=} RA\boldsymbol{U} \overset{\mathrm{d}}{=} RA\left(\begin{smallmatrix} \cos\Theta \\ \sin\Theta \end{smallmatrix}\right)$, $\Theta \sim \mathrm{U}(-\pi, \pi)$ independent of R.
- With $\varphi = \arcsin \rho$, we have $\boldsymbol{Y} \overset{\mathrm{d}}{=} R\left(\begin{smallmatrix} \cos\Theta \\ \sin\varphi\cos\Theta + \cos\varphi\sin\Theta \end{smallmatrix}\right) = \left(\begin{smallmatrix} \cos\Theta \\ \sin(\varphi+\Theta) \end{smallmatrix}\right)$.
- Thus $\mathbb{P}(X_1 > 0, X_2 > 0) = \mathbb{P}(Y_1 > 0, Y_2 > 0) = \mathbb{P}(\cos\Theta > 0, \sin(\varphi + \Theta) > 0) = \mathbb{P}(\Theta \in (-\frac{\pi}{2}, \frac{\pi}{2}), \varphi + \Theta \in (0, \pi) = \mathbb{P}(\Theta \in (-\frac{\pi}{2}, \frac{\pi}{2}), \Theta \in (-\varphi, \pi - \varphi) = \mathbb{P}(\Theta \in (-\varphi, \frac{\pi}{2}]) = (\frac{\pi}{2} - (-\varphi))/(2\pi)$. □

Lemma A.18 (Representation of Spearman's rho)
Let $(U_1, U_2) \sim C$ and $\tilde{U}_1, \bar{U}_2 \overset{\text{ind.}}{\sim} \mathrm{U}(0,1)$ be independent. Then $\rho_{\mathsf{S}} = \rho_{\mathsf{S}}(U_1, U_2) = 12\mathbb{P}(U_1 \le \tilde{U}_1, U_2 \le \bar{U}_2) - 3$.

Proof. $12\mathbb{P}(U_1 \le \tilde{U}_1, U_2 \le \bar{U}_2) - 3 = 12\mathbb{E}(\mathbb{P}(\tilde{U}_1 > U_1, \bar{U}_2 > U_2 \,|\, U_1, U_2)) - 3 \underset{\text{ind.}}{=} 12\mathbb{E}((1-U_1)(1-U_2)) - 3 = 12\mathbb{E}(U_1 U_2) - 3 = \rho_{\mathsf{S}}(U_1, U_2)$. □

Proof of Proposition 7.26. $\boldsymbol{X} \overset{\mathsf{d}}{=} \sqrt{W}\boldsymbol{Z}$ for $\boldsymbol{Z} \sim \mathrm{N}_2(\mathbf{0}, P)$. Let $\tilde{Z}, \bar{Z} \sim \mathrm{N}(0,1)$ and assume $\boldsymbol{Z}$, $\tilde{Z}$, $\bar{Z}$, W, $\tilde{W}$ and $\bar{W}$ are all independent. Let

$$\tilde{X} = \sqrt{\tilde{W}}\tilde{Z}, \quad \bar{X} = \sqrt{\bar{W}}\bar{Z},$$
$$Y_1 = X_1 - \tilde{X} = \sqrt{W}Z_1 - \sqrt{\tilde{W}}\tilde{Z},$$
$$Y_2 = X_2 - \bar{X} = \sqrt{W}Z_2 - \sqrt{\bar{W}}\bar{Z}.$$

$$\rho_{\mathsf{S}}(X_1, X_2) \underset{\Phi^{-1}}{\overset{\text{L. A.18}}{=}} 12\mathbb{P}(X_1 \le \tilde{X}_1, X_2 \le \bar{X}_2) - 3$$

$$= 6\mathbb{P}((X_1 - \tilde{X}_1)(X_2 - \bar{X}_2) > 0) - 3$$
$$= 3(2\mathbb{E}(\mathbb{P}(Y_1 Y_2 > 0 \,|\, W, \tilde{W}, \bar{W})) - 1)$$
$$= 3(4\mathbb{E}(\mathbb{P}(Y_1 > 0, Y_2 > 0 \,|\, W, \tilde{W}, \bar{W})) - 1).$$

Now note that $\boldsymbol{Y} \,|\, W, \tilde{W}, \bar{W} \sim \mathrm{N}_2(\mathbf{0}, \left(\begin{smallmatrix} W+\tilde{W} & W\rho \\ W\rho & W+\bar{W} \end{smallmatrix}\right))$ with $\rho(Y_1, Y_2)$ $= \frac{W\rho}{\sqrt{(W+\tilde{W})(W+\bar{W})}}$. Apply Lemma A.17 to see that

$$\rho_S(X_1, X_2) = 3\Big(4\mathbb{E}\Big(\frac{1}{4} + \frac{\arcsin\rho}{2\pi}\Big) - 1\Big) = \frac{12}{2\pi}\mathbb{E}(\arcsin\rho(Y_1, Y_2)).$$

For Gauss copulas, $F_W(x) = I_{\{x \geq 1\}}$, thus $W = \tilde{W} = \bar{W} = 1$ a.s. and the result follows. □

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 $\boldsymbol{X} \overset{\mathrm{d}}{=} RA\boldsymbol{U}$ and $\boldsymbol{X}' \overset{\mathrm{d}}{=} R'A\boldsymbol{U}'(\overset{\mathrm{d}}{=} -\boldsymbol{X}')$ we have $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{X}' \overset{\mathrm{d}}{=} \mathbf{0} + A(R\boldsymbol{U} - R'\boldsymbol{U}')$. Note that the characteristic function of $-\boldsymbol{X}'$ is

$\phi_{-\boldsymbol{X}'}(\boldsymbol{t}) = \phi_{\boldsymbol{X}'}(\boldsymbol{t}) = \phi_{\boldsymbol{X}}(\boldsymbol{t})$ so that $\phi_{\boldsymbol{Y}}(\boldsymbol{t}) \underset{\text{ind.}}{=} \phi_{\boldsymbol{X}}(\boldsymbol{t})\phi_{-\boldsymbol{X}'}(\boldsymbol{t}) = \phi_{\boldsymbol{X}}(\boldsymbol{t})^2$, hence $\boldsymbol{Y} \sim \mathrm{E}_2(\boldsymbol{0}, P, \psi^2)$.

- We thus obtain that

$$\begin{aligned}\rho_\tau &= 2\mathbb{P}(Y_1 Y_2 > 0) - 1 = 2(\mathbb{P}(Y_1 > 0, Y_2 > 0) + \mathbb{P}(Y_1 < 0, Y_2 < 0)) - 1 \\ &= 4\mathbb{P}(\boldsymbol{Y} > \boldsymbol{0}) - 1 \underset{\text{L.A.17}}{\overset{\text{cont.}}{=}} \frac{2}{\pi} \arcsin \rho. \end{aligned}$$ □

For a generalization to componentwise n.d. $\boldsymbol{X}$, 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.