Fat-Tailed Models: Building Blocks

Prof. Dr. Svetlozar (Zari) Rachev

Frey Family Foundation Chair-Professor, Applied Mathematics and Statistics, Stony Brook University
Chief Scientist, FinAnalytica

STONY BROOK
STATE UNIVERSITY OF NEW YORK

FinAnalytica
Risk Solutions for the Real World
Univariate Fat-Tailed Models
   Student’s $t$ Distribution
   Stable Paretian Distributions
   Extreme Value Theory

Copula Functions
   Dependence and Correlation
   Modeling Multivariate Data
   Copula Functions
   Scenario Generation from $t$-copula

Estimation Methodologies
   Maximum Likelihood Estimation
   Hypothesis Tests and Confidence Intervals
   Fitting Stable Distributions
   Comparing Probability Distributions
Outline: Fat-Tailed Models’ Building Blocks

Univariate Fat-Tailed Models
- Student’s $t$ Distribution
- Stable Paretian Distributions
- Extreme Value Theory

Copula Functions
- Dependence and Correlation
- Modeling Multivariate Data
- Copula Functions
- Scenario Generation from $t$-copula

Estimation Methodologies
- Maximum Likelihood Estimation
- Hypothesis Tests and Confidence Intervals
- Fitting Stable Distributions
- Comparing Probability Distributions
Risk models are employed in financial modeling to provide a measure of risk that could be employed in portfolio selection, risk management, and derivatives pricing.

A risk model is typically a combination of a probability distribution model and a risk measure.

A successful univariate probability distribution model
- Accounts for returns’ temporal dynamics, such as autocorrelations, volatility clustering, and long memory.
- Employs a distributional assumption flexible enough to accommodate various degrees of skewness and heavy-tailedness.
- Is scalable and practical—can be extended to a multivariate model covering a large number of assets.

Next, we discuss alternatives for building the probability distribution model. Time series models are covered in the 3pm-6pm session, while risk measures are our focus in tomorrow’s morning session.
The Basics: Normal Distribution

The main reasons for the Gaussian distribution’s traditional popularity are several.

1. Its analytical tractability means that deriving theoretical results and employing it in applications is (relatively) straightforward. Numerical methods are widely available and implementable.

2. Central results in statistics—Central Limit Theorem and Law of Large Numbers—underlie its importance. (However, recall that it is the stable distributions that have a domain of attraction.)

3. It has an intuitive appeal—random variables distributed with the Gaussian distribution tend to assume values around the average, with the odds of deviation from the average decreasing exponentially as one moves away from it.
Prominent financial frameworks are built around the normal distribution:

- Markowitz’s modern portfolio theory.
- The Capital Asset Pricing Model.
- The Black-Scholes option pricing model.

- All of these three frameworks assume (or imply) that asset returns follow a normal distribution and reflect a long-standing paradigm that rational investors’ preferences can be described exclusively in terms of expected returns and risk as measured by the variance of the return distribution.

- They are inherently static frameworks. The underlying dynamic of returns is either given exogenously or is based on the assumption that returns have independent and identical distributions.

- Such characteristics do not fit adequately with the empirically-observed features of financial returns and investor choice.\(^1\)

\(^1\)For a survey of theoretical and practical aspects of heavy-tailed distributions applied to finance, see Rachev (1994).
We already defined the classical Student’s $t$ distribution through its density function and saw that its degrees-of-freedom (DOF) parameter controls the degree of tail-fatness.

The $t$-distribution can be expressed in an alternative way which is very useful in obtaining simulations from that distribution. This equivalent representation is a \textit{scale mixture of normal distributions}.

The mixing variable is distributed with the inverse-gamma distribution.

The random variable $X \sim t_\nu(\mu, \sigma)$ can be expressed as

$$X = \mu + \sqrt{W} Z,$$

where $W \sim \text{inv-gamma} \left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ and $Z \sim N(0, \sigma^2)$, independent of $W$. 
A variable $W$ with the **inverse-gamma distribution** has a positive support ($W > 0$) and a PDF given by

$$f(w|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} w^{-\alpha-1} e^{-\frac{\beta}{w}},$$

where $\alpha$ and $\beta$ are shape and scale parameters, respectively.

In the subordinated representation above, the mixing variable $W$ modifies the Gaussian scale (standard deviation), so that the resulting distribution is heavier-tailed than the Gaussian one.

In the terminology of mathematical finance, this scale mixing process is referred to as “subordinated” process. We will have more to say about subordination later this session.
The major limitation of the classical $t$-distribution for financial modeling is its symmetry. If there is significant asymmetry (skewness) in the data, it will not be reflected in the probability model and, therefore, in the risk estimate.

The *skewed Student’s $t$ distribution* is obtained as a mean-scale mixture of normal distributions. A variable $X$ with this distribution can be expressed as follows:

$$X = \mu + \gamma W + \sqrt{W} Z,$$

where $W \sim \text{inv-gamma} \left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ and $Z \sim N(0, \sigma^2)$, independent of $W$.

The parameter $\gamma$ takes real values. Its sign and magnitude control the asymmetry in $X$.

Notice that here the mixing variable modifies the Gaussian mean, in addition to its scale, resulting in distributional asymmetry.
Skewed Student’s $t$ Distribution

- Conditional on the value of $W$, $X$’s distribution is Gaussian:

$$X \mid W = w \sim N(\mu + \gamma w, w\sigma^2).$$

- The unconditional distribution is what is defined as the skewed Student’s $t$ and its density is given by the expression

$$f(x \mid \mu, \sigma, \gamma, \nu) = A \times \frac{\exp \left( \frac{(x-\mu)\gamma}{\sigma^2} \right)}{\left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{(\nu+1)/2}} \times \frac{K_{(\nu+1)/2}(B)}{B^{-(\nu+1)/2}},$$

where

$$A = \frac{2^{1-(\nu+1)/2}}{\Gamma(\frac{\nu}{2}) (\pi \nu)^{1/2} \sigma}, \quad B = \sqrt{\left(\nu + \frac{(x-\mu)^2}{\sigma^2}\right) \frac{\gamma^2}{\sigma^2}},$$

and $K_\lambda(\cdot)$ is the so-called "modified Bessel function with index $\lambda".\textsuperscript{\dagger}

\textsuperscript{\dagger}The modified Bessel function is given by $K_\lambda(x) = \sum_{m=0}^{\infty} \frac{1}{m! \Gamma(m+\lambda+1)} \left(\frac{x}{2}\right)^{2m+\lambda}$. 
The typical approach to risk modeling based on the Student’s $t$ distribution includes building an autoregressive and volatility clustering components, as well as assuming that the DOF parameter is the same for all assets’ returns. The latter assumption is essential for extending the classical $t$-distribution to a multivariate one.

However, it is an empirical fact that assets are not homogeneous with respect to the degree of non-normality of their returns. Moreover, tail thickness and shape are not constant through time either.

To illustrate, on the next slide we provide two plots:

1. We fit the $t$-distribution to the returns on the constituents of the S&P 500 index for the period 1/2/91—6/30/11 and we plot the fitted DOF. (Returns are filtered for GARCH effects.)

2. We fit the $t$-distribution to the returns on the DJIA returns for the period 1/1/97—6/30/11 and observe the temporal dynamics of DOF.
**Across Assets**

- 44% of returns have $\text{DOF} < 5$: very heavy-tailed.
- 54% of stocks have $\text{DOF}$ between 5 and 10.
- Only 3 stocks are close to Gaussian ($\text{DOF} > 15$).
- **Attention**: Portfolio risk contributors & diversifiers.

**Through Time**

- When markets are "normal", $\text{DOF} > 30$.
- When markets are "rough", $\text{DOF}$ small.
- **Attention**: Risk budgeting & management.
Simulations from both the ”classical” and the skewed $t$-distributions make use of their normal mixture representations:

- Generate an observation $w$ from inv-gamma\( \left( \frac{\nu}{2}, \frac{\nu}{2} \right) \).
- Generate an observation $z$ from $N\left( 0, \sigma^2 \right)$.
- Compute the corresponding observation of the $t$- or skewed $t$-distribution, respectively, as

\[
x = \mu + \sqrt{w}z \quad \text{and} \quad y = \mu + w\gamma + \sqrt{w}z.
\]
Stable Paretian Distributions

- Research on stable distributions in the field of finance has a long history.\(^1\)

- The practical implementation of stable distributions to risk modeling, however, has only recently been developed. Reasons for the late penetration are the complexity of the associated algorithms for fitting and simulating stable models, as well as the multivariate extensions.

- To distinguish between Gaussian and non-Gaussian stable distributions, the latter are commonly referred to as stable Paretian, Lévy stable or \(\alpha\)-stable distributions.

- Stable Paretian tails decay more slowly than the tails of the normal distribution and therefore better describe the extreme events present in the data. Like the Student’s \(t\) distribution, stable Paretian distributions have a parameter responsible for the tail behavior, called tail index or index of stability.

---

\(^1\)See Rachev and Mittnik (2000) for extensive discussion of stable distributions in finance. Also Rachev, Menn, Fabozzi (2005) and Samorodnitsky and Taqqu (1994).
The first definition establishes the stable distribution as having a domain of attraction. That is, (properly normalized) sums of $i.i.d.$ random variables are distributed with the $\alpha$-stable distribution as the number of summands $n$ goes to infinity.

Define:

\[ Y_1, Y_2, \ldots, Y_n = i.i.d. \text{ random variables.} \]
\[ \{a_n\} \text{ and } \{b_n\} = \text{sequences of real and positive numbers, respectively.} \]

A variable $X$ is said to have the stable Paretian distribution if

\[
\frac{\sum_{i=1}^{n} Y_i - a_n}{b_n} \xrightarrow{d} X,
\]

where "$\xrightarrow{d}$" denotes convergence in distribution.
A characteristic function provides a third possibility (besides the PDF and the CDF) to uniquely define a probability distribution.

It is a mapping from the set of real numbers into the set of complex numbers and is denoted by

$$\varphi_X(t) = E(e^{itX}),$$

which represents the so-called "Fourier transform" of the distribution of the random variable $X$.

Knowing the characteristic function is mathematically equivalent to knowing the CDF or PDF. For example, the relationship between the PDF and the characteristic function is provided by the expression:

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t) \, dt$$
The density function of the stable Paretian distribution is not available in a closed-form expression in the general case. Therefore, the distribution of a stable random variable $X$ is alternatively defined through its characteristic function.

The characteristic function of the $\alpha$-stable distribution is given by

$$
\varphi_X(t) = \begin{cases} 
\exp\{i\mu t - \sigma^\alpha |t|^\alpha (1 - i\beta \text{sign}(t) \tan \frac{\pi \alpha}{2})\}, & \alpha \neq 1 \\
\exp\{i\mu t - \sigma |t| (1 - i\beta \text{sign}(t) \log (t))\}, & \alpha = 1,
\end{cases}
$$

where $\text{sign}(t)$ is 1 if $t > 0$, 0 if $t = 0$, and -1 if $t < 0$.

The PDF of the $\alpha$-stable distribution can be obtained through a numerical method, as we explain further below.
The four parameters uniquely determining the $\alpha$-stable distribution are:

- $\alpha$: index of stability or tail index, $0 < \alpha \leq 2$.
- $\beta$: skewness parameter, $-1 \leq \beta \leq 1$.
- $\sigma$: scale parameter, $\sigma > 0$.
- $\mu$: location parameter, $\mu \in \mathbb{R}$.

Denote the distribution by $S_{\alpha}(\sigma, \beta, \mu)$.

The roles of $\alpha$ and $\beta$ are illustrated in the plots on the next slide.
Stable Paretian Distributions: Effects of Alpha and Beta

Alpha ($\alpha$)

- Controls tail-fatness.
- The lower $\alpha$, the heavier the tails.
- The closer to 2, the more Gaussian-like the distribution.
- $\alpha = 2 \Rightarrow N(\mu, 2\sigma^2)$.
- $\alpha < 2 \Rightarrow$ variance is infinite.

Beta ($\beta$)

- Controls asymmetry.
- $\beta > 0 \Rightarrow$ Skewness to right.
- $\beta < 0 \Rightarrow$ Skewness to left.
- $\alpha = 2 \Rightarrow \beta = 0$. 

Prof. Dr. Zari Rachev
Fat-Tailed Models: Building Blocks
Next Generation Risk Management
Stable Paretian Distribution: Basic Properties

- **Power-tail decay.** Tail decays like a power function (slower than exponential decay). For a constant $C$, the property can be expressed as

$$P (|X| > x) \propto Cx^{-\alpha}, \quad \text{as} \ x \to \infty. \quad (1)$$

- **Existence of raw moments.** The magnitude of $\alpha$ determines the order up to which raw moments exist. For non-Gaussian $\alpha$-stable distributions ($\alpha < 2$), the variance (as well as higher moments, such as skewness and kurtosis) does not exist.

- **Stability.** The distributional form is preserved under linear transformations. This property is governed by $\alpha$ in the following way. Suppose $X_1, \ldots, X_n$ are $i.i.d.$ random variables, independent copies of a r.v. $X$. Then, for $C_n > 0$ and a real number $D_n$, $X$ follows the stable distribution if

$$X_1 + X_2 + \ldots + X_n \overset{d}{=} C_n X + D_n.$$

The constant $C_n = n^{1/\alpha}$ determines the stability property.
• $\alpha$ varies through time: bottom plot.
• Stable distribution is fitted to GARCH-filtered DJIA returns.
• $\alpha$ close to 2 in calm markets.
• $\alpha$ decreases in build-up to a crash.
Similar to the Student’s $t$ distribution, stable distributions can be represented as mixtures of other distributions.

More generally, mixture representations are analyzed within the framework of intrinsic time change and subordination.

The price and return dynamics can be considered under two different time scales—the physical time and an intrinsic (also called market) time.

The intrinsic time is best thought of as the cumulative trading volume process which measures the cumulative trading volume of the transactions up to a point on the calendar-time scale. It is a measure of market activity.\(^1\)

\(^1\)See Rachev and Mittnik (2000) and Janicki and Weron (1994). Classical texts are Clark (1973) and Feller (1966).
Stable Paretian Distributions

An Aside: Subordination

- Denote the intrinsic time process by \( T(t) \) and the time-evolving random variable such as price or return by \( X(t) \). \( X(t) \) is assumed independent of \( T(t) \). The compound process

\[
X(T(t))
\]

is said to be subordinated to \( X \) by the intrinsic time \( T(t) \) and \( T(t) \) is referred to as a subordinator.

- Since the increments of the intrinsic time \( \Delta T(t) = T(t) - T(t - \Delta t) \) are non-decreasing and positive, distributions such as gamma, Poisson, and inverse-Gaussian can be used to describe them in probabilistic terms.

- Another distributional alternative is the completely skewed to the right \( \alpha \)-stable distribution, \( S_\alpha(\sigma, 1, 0) \), for \( 0 < \alpha < 1 \), whose support is the positive real line.

- Therefore, when \( 0 < \alpha < 2 \), the subordinator is a stable distribution given by \( S_{\frac{\alpha}{2}}(\sigma, 1, 0) \).
Stable Paretian Distributions
An Aside: Subordination

- Subordinated models with random intrinsic time are leptokurtic. They have heavier tails and higher peaks around the mode than the normal distribution. As such, they provide a natural way to model the tail effects observed in prices and returns.

- The general form of a subordinated model is given by the expression

\[ X = \mu + \gamma W + g(W)Z, \]

where \( W \) is responsible for the market timing.

- Subordinated representations are useful for two main reasons:
  - Provide practical ways of simulating random numbers from the corresponding models.
  - In multi-asset settings, each marginal distribution in general has a different tail heaviness. This across-asset heterogeneity can be modeled by having subordinators with different parameters for each asset.
The subordinated representation of the $\alpha$-stable distribution can be expressed in the following way.

Let $Z$ be a standard normal random variable, $Z \sim N(0, 1)$, and $Y$ be a positive $\alpha/2$-stable random variable independent of $Z$, $Y \sim S_{\alpha/2}(s, 1, 0)$, where

$$s = \frac{\sigma^2}{2} \cos \left(\frac{\pi \alpha}{4}\right)^{2/\alpha}.$$

Then, the variable

$$X = Y^{1/2} Z$$

is symmetric $\alpha$-stable: $X \sim S_{\alpha}(\sigma, 0, 0)$.

Every symmetric stable variable is conditionally Gaussian (conditional on the value of the stable subordinator). Unconditionally, the symmetric $\alpha$-stable distribution is expressed as a scale mixture of normal distributions.
Extreme Value Theory

- Extreme Value Theory (EVT) has been applied in modeling severe weather, earthquakes, and other extreme natural phenomena.

- Extreme value distributions are the asymptotic distributions for the normalized largest observations of i.i.d. random variables. There are two main categories of models for extreme values: **block maxima** models and **threshold exceedances** models.

- In financial applications, block maxima could refer to the maximal observations within certain pre-defined periods of time. For e.g., daily return data could be divided into quarterly blocks and the largest daily observations within these blocks collected and analyzed.

- When the block size is large, so as block maxima are independent, the limit distribution is given by EVT.

- In block maxima models, the number of blocks determines the size of the data sample available for analysis and fitting. In contrast, in threshold exceedances models, the sample size is not pre-determined but depends on the a priori selected threshold level.
The two plots illustrate the two types of data categories considered in EVT.

Plot on the left: block maxima data.
Plot on the right: threshold exceedances data.
The block maxima model is represented by the **generalized extreme value** (GEV) distribution. Its distribution function has the form

\[ F_X(x \mid \xi, \mu, \sigma) = \exp \left( - \left( 1 + \xi \frac{x - \mu}{\sigma} \right)^{-1/\xi} \right), \]

where \( 1 + \xi(x - \mu)/\sigma > 0 \) and \( \xi \in \mathbb{R} \) is the shape, \( \mu \in \mathbb{R} \)—the location, and \( \sigma > 0 \)—the scale parameter.

The parametric form above encompasses three distributions, depending on the value of \( \xi \): Weibull \( (\xi < 0) \), Gumbel \( (\xi = 0) \), and Fréchet \( (\xi > 0) \).

One of the block maxima method’s major drawback is its “wastefulness” of data: all but the largest observation within each block are discarded. For this reason, a more common approach to EVT modeling is the threshold exceedance method.
In the **threshold exceedance models**, the extreme events exceeding a predetermined high level (that is, events in the tail) are modeled with the *generalized Pareto distribution* (GPD).

Its distribution function is given by

\[ F_X(x | \xi, \sigma) = 1 - \left(1 + \frac{\sigma}{\xi} x\right)^{-1/\xi}, \]  

(2)

where \( \sigma > 0 \) and \( x \geq 0 \) when \( \xi \geq 0 \) and \( 0 < x < -\sigma/\xi \) when \( \xi < 0 \). The parameters \( \xi \) and \( \sigma \) are the shape and scale parameters, respectively.

Again, three distributions are subsumed in the parametric form above: Pareto (\( \xi > 0 \)) with \( \alpha = 1/\xi \) and \( k = \sigma/\xi \), exponential (\( \xi = 0 \)), and a short-tailed distribution, called Pareto type II (\( \xi < 0 \)).

---

1For more details on EVT, see Embrechts, Klüppelberg, and Mikosch (1997).
Denote the available data sample by $X_1, \ldots, X_n$. Define the upper and lower threshold level by $u_U$ and $u_L$, resp. The data points beyond the thresholds constitute the tails of the data distribution, which are modeled with EVT.

Define the exceedances of $u_U$ and $u_L$ as the differences between the data points and the thresholds.

A statistical result implies that the distribution of the exceedances from a large class of underlying distributions converge to a GPD as the threshold level increases. That is, GPD is the limiting distribution as $u_U$ ($u_L$) increases to infinity.

But where does the tail start? How does one determine $u_U$ and $u_L$?

Two of the most popular tools for selection of the threshold level are the mean excess plot and the Hill plot. Both of them rely on visual inspection to determine the threshold. We discuss them next.
Mean Excess Plot

- The *mean excess function* describes the average exceedance above a threshold $u$, as a function of $u$. Formally, it is defined as

$$m(u) = E(X - u | X > u).$$

- In the case of the GPD, $m(u)$ can be shown to equal

$$m(u) = \frac{\sigma}{1 - \xi} + \frac{\xi}{1 - \xi} u.$$

- The mean excess function is linear in the threshold level. This linearity is used to motivate a graphical check that the data conform to a GPD model.

- If the mean excess function plot is approximately linear for high threshold values, the GPD may be employed to describe the distribution of the exceedances. The level above which linearity is evident may be taken as the threshold level.
Hill Estimator Plot

- The *Hill estimator* approach offers a way to estimate the tail index $\alpha = 1/\xi$.

- Denote the $i$th order statistics of the data sample by $X_{(i)}$. The Hill estimator of $\alpha$ is defined with the expression

\[
H_{i,k} = \left( \frac{1}{k} \sum_{i=1}^{k} \ln X_{(i)} - \ln X_{(k)} \right)^{-1},
\]

where $1 \leq k \leq n$ and $k$ is a sufficiently high number.

- For $\xi > 0$, the Hill estimator is equal to $\alpha$ asymptotically, as the sample size $n$ and the number of extremes $k$ increase without bound.

- In practical applications, the Hill estimator is computed for different values of $k$ and plotted against these values. The plot is expected to stabilize above a certain value of $k$ and the threshold level is then estimated by $X_{(k)}$. 
Applying EVT has two major challenges:

1. In order to obtain a sufficiently large number of observations in the tail, an extremely large sample of historical data is needed.
   - Academic studies indicate that the minimum requirement is for samples of at least 1,000 data points.
   - Using a sample this large obscures any current fat-tailed market behavior and can only be suitable for long-term projections.

2. We need to know where the body of the distribution ends and where the tail begins.
   - Even though the plots of the Hill estimator and the mean excess function provide a method for threshold identification, such identification is intrinsically subjective, as it is based on visual inspection.
   - It is difficult to automate it for large-scale applications.
Outline: Fat-Tailed Models’ Building Blocks

Univariate Fat-Tailed Models
- Student’s $t$ Distribution
- Stable Paretian Distributions
- Extreme Value Theory

Copula Functions
- Dependence and Correlation
- Modeling Multivariate Data
- Copula Functions
- Scenario Generation from $t$-copula

Estimation Methodologies
- Maximum Likelihood Estimation
- Hypothesis Tests and Confidence Intervals
- Fitting Stable Distributions
- Comparing Probability Distributions
Dependence and Correlation

- Correlation is a widespread concept in modern finance and risk management and is used to describe the dependence between random variables.

- It is often incorrectly used to denote and mean any kind of dependence.

- Correlation is one particular measure of dependence among many. It measures linear dependence only and does not capture asymmetric “tail dependence” (the higher chance of extreme co-movements in a market shock).

- In a bivariate context, lower (upper) tail dependence measures the conditional probability of one variable taking an extremely low (high) value, given an extremely low (high) realization of the other variable.

- Since financial theory and risk management analysis rely crucially on the dependence structure of assets, we discuss the limitations of correlation as a measure of dependence between two random variables and introduce an alternative measure to overcome these limitations—copulas.
There are at least three major drawbacks of the correlation as a measure of dependence between two random variables, $X$ and $Y$.

1. The variances of $X$ and $Y$ must be finite or the correlation is not defined. This causes problems when we have heavy-tailed data modeled, for example, with sub-Gaussian stable distributions ($\alpha < 2$) because the variance is infinite.

2. Independence between two random variables implies zero correlation. However, generally the opposite is not true—zero correlation does not imply independence. Only in the case of multivariate normal distribution are uncorrelatedness and independence interchangeable notions.

3. The correlation is not invariant under nonlinear strictly increasing transformations:

$$\text{corr}(T(X), T(Y)) \neq \text{corr}(X, Y),$$

for some nonlinear increasing function, $T$. 
Correlation: Deficiencies

Example

- Assume that $X$ and $Y$ represent the continuous returns (log-returns) of two financial assets over some holding period $[0, T]$.

- If one knows the correlation between the returns, $X$ and $Y$, this does not imply that the dependence structure between the asset prices themselves is known. Denote the two prices by $P$ and $Q$, respectively. Then, $P_T = P_0 \exp(X)$ and $Q_T = Q_0 \exp(Y)$.

- The asset prices are strictly increasing functions of the returns but the correlation structure is not maintained by this transformation. This implies that the returns could be uncorrelated whereas the prices are strongly correlated and vice versa.
There are two paths one can take to model jointly multiple assets’ returns (i.e., two ways in which univariate models can be extended to multivariate ones):

1. Consider the realized returns in each time period as the realization from some multivariate distribution.

2. Model the return on each asset separately, as following some univariate process, and then impose a dependence structure on returns using a copula function.

- The multivariate distributions that are easy enough to work with describe well symmetric, unimodal data.
- These distributions are elliptical, the most well-known of which are the multivariate normal and the multivariate Student’s $t$. 
Under a multivariate normal model, the covariance matrix defining the dependence structure only determines linear dependencies, and is symmetric.

Moreover, tail events are asymptotically independent: if a large bivariate sample is generated from a bivariate normal distribution with a correlation very close (but not equal) to +1, the extreme values (losses and gains) will be approximately uncorrelated.

To overcome the covariance matrix deficiency, the multivariate Student’s $t$ distribution is often used to account for tail dependence. However, this leads to two implicit assumptions that significantly limit the flexibility of the model:

- Each variable has one and the same tail behavior.
- Dependence among variables is elliptical.
In general, each multivariate distribution implicitly contains information about both the marginal behavior of individual returns and their dependence structure.

The copula approach allows us to separate out the description of the dependence structure.

The copula can be applied to any marginal distribution model. It offers several distinct advantages:

- General (non-linear) dependencies can be modeled.
- Dependence of extreme events can be modeled.
- Copulas are indifferent to continuously increasing transformations (not only linear transformations, as is the case for the correlation).

---

Suppose that returns on $N$ assets are observed: $Y_1, \ldots, Y_N$. The probability distribution of the random vector $Y = (Y_1, \ldots, Y_n)$ is given by

$$F_Y(y_1, \ldots, y_N) = P(Y_1 \leq y_1, \ldots, Y_N \leq y_N).$$

Suppose that the risk factors’ marginal distributions are given by the $N$ cumulative distribution functions (CDFs) $F_1, \ldots, F_N$.

Since a distribution function can only take values between 0 and 1, computing $F_i$ for $i = 1, \ldots, N$ essentially transforms each return into a uniform random variable. We use this fact to help us define what copula is.
Copula Functions: Definition

- For the CDF, we can write:\(^1\)

\[
F_Y(y_1, \ldots, y_N) = P(F_1(Y_1) \leq F_1(y_1), \ldots, F_N(Y_N) \leq F(y_N)) \\
= P(U_1 \leq F_1(y_1), \ldots, U_N \leq F_N(y_N)) \\
= C(F_1(y_1), \ldots, F_N(y_N)).
\]

- The copula is the joint probability distribution \(C\) with univariate marginal distributions. (That is, it exists on the \(N\)-dimensional hypercube.)

- A copula \(C(u_1, \ldots, u_N)\) expresses the dependence among returns on the basis of quantiles. It gives an answer to the question what is the joint probability that the first return takes a value below the \(u_1\)-quantile, the second return takes a value below the \(u_2\)-quantile, etc. It is not necessary to specify the numerical values of these quantiles.

\(^1\)Sklar’s theorem (Sklar (1959)) expresses the importance of copulas in the modeling of multivariate random variables.
Copula Functions: Density

- Since the copula is just a probability distribution, it can be characterized by means of a CDF or PDF.

- Given a copula function $C$, the density is computed according to the formula

$$c(u_1, \ldots, u_n) = \frac{\partial^n C(u_1, \ldots, u_n)}{\partial u_1 \cdots \partial u_n}.$$ 

- That way, the density of the copula can be expressed via the density of the random variable:

$$c(F_{Y_1}(y_1), \ldots, F_{Y_n}(y_n)) = \frac{f_Y(y_1, \ldots, y_n)}{f_{Y_1}(y_1) \cdots f_{Y_n}(y_n)}.$$ 

- The numerator contains the density of the vector $Y$, while in the denominator is the density of $Y$ but under the assumption that the component returns are independent random variables.
Copula density of a bivariate normal distribution with mean

\[ \mu = (0, 0) \]

and covariance matrix

\[ \Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}. \]
Consider the last expression for the copula density (two slides ago).

If $Y$ has independent components, the density of the corresponding copula is a constant in the unit hypercube: $c_0(u_1, \ldots, u_n) = 1$ and the copula has the simple form

$$C_0(u_1, \ldots, u_n) = u_1 \cdots u_n.$$ 

This copula characterizes stochastic independence and is called *independence copula*. 
Consider again the copula density.

For each value of the vector of arguments \( y = (y_1, \ldots, y_n) \), the PDF expression provides information about the degree of dependence among the events that \( Y_i \) is simultaneously in a small neighborhood of \( y_i \) for \( i = 1, \ldots, n \).

That is, the copula density provides information about the \emph{local} structure of the dependence.

With respect to the PDF \( c_0 \) of the independence copula, the PDF of an arbitrary copula can have a value either above 1 or below 1.

From that value, we can infer about the degree of dependence of the corresponding \( n \) events, as follows on the next slide.
Suppose that for some value of $y$, the right-hand side of the expression for the copula PDF is close to zero. That is, the denominator is much larger than the numerator:

$$f_Y(y_1, \ldots, y_n) < f_{Y_1}(y_1) \cdots f_{Y_n}(y_n).$$

The joint probability of the events that $Y_i$ is in a small neighborhood of $y_i$ for $i = 1, \ldots, n$ is smaller than what it would have been if they were independent.

Therefore, there is a very small probability of the events occurring simultaneously.

Conversely, suppose that the copula density is larger than 1. This implies that the corresponding events are more likely to happen together than if they were independent.
As mentioned already, under the Gaussian copula, the probability of large negative returns occurring jointly together is severely underestimated, due to the asymptotic independence of events implied by the copula.

A better alternative is the Student’s $t$ copula. It models the probability of joint extreme events more accurately. Disadvantages of the classical $t$ copula:

- Symmetric tail dependence, resulting in the same probability for joint extreme positive and negative returns.
- Homogeneous tail dependence among assets.

It is possible to extend the $t$ copula in a way suggested by the subordinated representation of the $t$ distribution we discussed earlier. For maximal flexibility, the subordination in each component (or groups of components) of $Y$ is controlled by individual parameters, as we explain on the next slide.
The Subordinated Student’s \( t \) Copula

- In the multivariate case, the subordinated representation of the \( t \) distribution can be expressed in the following way:

\[
X = \begin{pmatrix}
\mu_1 + W_1 + \sqrt{W_1}Z_1 \\
\vdots \\
\mu_n + W_n + \sqrt{W_n}Z_n
\end{pmatrix},
\]

where \( Z = (Z_1, \ldots, Z_n)' \sim N_n(0, \Sigma) \) and \( W_i \sim \text{inv-gamma}(\frac{\nu_i}{2}, \frac{\nu_i}{2}) \), for \( i = 1, \ldots, n \).

- The \( t \)-copula based on this representation is known as “subordinated Student’s \( t \)-copula”. As can be seen from the expression above, that copula is based on the Gaussian copula but
  - The volatility and correlations are random.
  - The copula can be skewed in different directions.
Suppose the returns on $N$ risk factors are observed for $T$ periods. Denote the $T \times N$ data matrix by $R$.

**Copula Estimation**

- Calibrate a univariate $t$-distribution to each asset’s return (the columns of $R$). Denote the estimated degrees of freedom by $\hat{\nu}_1, \ldots, \hat{\nu}_N$.

- Calculate the marginal CDFs, using the degrees of freedom estimates:

$$
\begin{pmatrix}
F_1(R_{1,1}, \hat{\nu}_1) & \cdots & F_N(R_{1,N}, \hat{\nu}_N) \\
F_1(R_{2,1}, \hat{\nu}_1) & \cdots & F_N(R_{2,N}, \hat{\nu}_N) \\
\vdots & & \vdots \\
F_1(R_{T,1}, \hat{\nu}_1) & \cdots & F_N(R_{T,N}, \hat{\nu}_N)
\end{pmatrix}
$$
Scenario Generation from Student’s $t$ Copula

- The matrix of marginal CDF values above can be viewed as a matrix of copula observations. At time $t$, the observed copula value is given by

$$C_t = (F_1(R_{t,1}, \hat{\nu}_1) \cdots F_N(R_{t,N}, \hat{\nu}_N)).$$

- Assuming that $C_t$ has an $N$-dimensional $t$-distribution, fit that distribution and estimate its parameter $\nu^C$.

Notice that the DOF parameter $\nu^C$ is common for all assets. As discussed earlier, it is possible to calibrate a multivariate $t$-distribution, whose components’ subordination is controlled by individual parameters.
Scenario Generation from Student’s $t$ Copula

Scenario generation

- Simulate $M \ N$-dimensional vectors from the copula (that is, from the $N$-dimensional $t$-distribution with $\hat{\nu}^C$)

$$V = \begin{pmatrix}
V_{1,1} & \cdots & V_{1,N} \\
V_{2,1} & \cdots & V_{2,N} \\
\vdots & & \vdots \\
V_{M,1} & \cdots & V_{M,N}
\end{pmatrix}.$$  

Each row of the matrix $V$ shares the dependence introduced by the copula.

The columns represent $i.i.d$ realizations of univariate uniform $[0, 1]$ distribution (the copula’s margins).
Scenario Generation from Student’s $t$ Copula

- Map the copula scenarios to the sample space of the return random variables, through the inverse fitted marginal Student’s $t$ CDFs:

$$
\begin{pmatrix}
T_{1,1} & \cdots & T_{1,N} \\
T_{2,1} & \cdots & T_{2,N} \\
\vdots \\
T_{M,1} & \cdots & T_{M,N}
\end{pmatrix} = \begin{pmatrix}
F_{1}^{-1}(V_{1,1}, \hat{\nu}_1) & \cdots & F_{N}^{-1}(V_{1,N}, \hat{\nu}_N) \\
F_{1}^{-1}(V_{2,1}, \hat{\nu}_1) & \cdots & F_{N}^{-1}(V_{2,N}, \hat{\nu}_N) \\
\vdots \\
F_{1}^{-1}(V_{M,1}, \hat{\nu}_1) & \cdots & F_{N}^{-1}(V_{M,N}, \hat{\nu}_N)
\end{pmatrix}.
$$
Outline: Fat-Tailed Models’ Building Blocks

Univariate Fat-Tailed Models
- Student’s $t$ Distribution
- Stable Paretian Distributions
- Extreme Value Theory

Copula Functions
- Dependence and Correlation
- Modeling Multivariate Data
- Copula Functions
- Scenario Generation from $t$-copula

Estimation Methodologies
- Maximum Likelihood Estimation
- Hypothesis Tests and Confidence Intervals
- Fitting Stable Distributions
- Comparing Probability Distributions
To estimate a statistical model is to estimate its parameters from the sample data.

Suppose that a distribution with density $f$ is given and that $\theta$ is a parameter or a vector of parameters of that distribution.

Consider a sample of $T$ observations $X_t$, $t = 1, 2, \ldots, T$ extracted from a population with distribution $f$.

An estimator $\hat{\theta}$ of the parameter $\theta$ is a function $\hat{\theta} = g(X_1, \ldots, X_T)$ of the sample which produces numbers close to the parameter $\theta$ (in a sense that will become clear in a bit).

Any estimator is characterized by several important properties, two of which are fundamental:

- An estimator $\hat{\theta}$ is **unbiased** if the mean of the estimator equals the true parameter $\theta$: $E(\hat{\theta}) = \theta$ for any sample size.

- An estimator $\hat{\theta}$ is **consistent** if the estimator converges to the true parameter $\theta$, as the sample size expands to infinity: $\hat{\theta} \xrightarrow{p} \theta$. 
Maximum Likelihood Estimation

The method of maximum likelihood (ML) is a fundamental estimation method that could be applied in most practical applications and possesses several attractive theoretical properties.

**Example**

- Suppose you flip a coin 1,000 times and get 700 heads.
- Is the coin biased? Or is it fair and you have experienced a particularly unlikely stream of outcomes?
- It is reasonable to conclude that the coin is biased with a 70% probability of heads. That is, we rule out the occurrence of very unlikely events in practice.

The ML principle generalizes the above idea:

> The estimate of a distribution parameter should maximize the probability of occurrence of the observed sample, given the hypothesized distribution.
We now state the ML principle formally. For an independently and identically distributed sample of data $X = (X_1, \ldots, X_T)$ characterized by a probability density function $f$, the likelihood function is defined as

$$L(\theta | X_1, \ldots, X_T) = \prod_{t=1}^{T} f(X_t | \theta),$$

where $\theta$ is the distribution parameter.

The ML estimate of $\theta$ is the value that maximizes the likelihood:

$$\hat{\theta} = \text{argmax} (L(\theta)).$$

Since the log function is strictly monotone, it is usual to replace $L$ with the logarithm of $L$:

$$\hat{\theta} = \text{argmax} (\log L(\theta))$$

$$= \text{argmax} \sum_{t=1}^{T} \log f(X_t | \theta).$$
Any estimation process yields results that depend on the specific sample of data. As sample data are random variables, estimators (functions of the sample data) are also random variables.

The estimators are characterized by probability distributions, called *sampling distributions*. These are critical for testing and choosing hypotheses because, in general, we do not know if the model we are estimating is the correct model: Any model is only a scientific hypothesis.

The distribution of estimators clearly depends on the distribution of the sample data. In general, determining the sampling distribution is a difficult task. In simpler models, where it is “safe” to assume that variables have given distributions, for e.g., normal, it is possible to derive the sampling distributions.

We briefly consider an example on the next slide.
Sampling Distributions: Example

- Consider an \(i.i.d\). sample of \(T\) observations from \(N(\mu, \sigma^2)\) distribution. The sample mean and variance below are unbiased estimators of \(\mu\) and \(\sigma^2\), respectively,

\[
\bar{\mu} = \frac{1}{T} \sum_{t=1}^{T} X_t \quad \text{and} \quad \bar{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} (X_t - \bar{\mu})^2.
\]

- It is easy to show that the distribution of \(\bar{\mu}\) has a normal distribution, while \(\bar{\sigma}^2\) is distributed with the \(\chi^2\)-distribution with \(T\) degrees of freedom.
Hypothesis Tests

- A common statistical procedure is to run a hypothesis test to check whether a given hypothesis about a parameter can be considered true or can be rejected.

- A **test statistic** is a function of the data sample and is, therefore, a random variable.

- Suppose we want to test a hypothesis $H_0$. It is called the **null hypothesis** (or simply, the *null*). Consider a test statistic $k$ relative to $H_0$. For example, we want to test the null hypothesis that the mean is zero.

- Intuitively, we can reject the null that the mean is zero, if the estimation yields a value of the parameter sufficiently distant from zero. In other words, how can we decide if the distance is significant?

- Suppose we know the sampling distribution of the test statistic. Given that and an observed value of the statistic, the **p-value** is the probability in the tail beyond the observed statistic value.

- A small p-value is evidence against the null hypothesis. As a rule of thumb, a hypothesis test is considered significant (the null rejected) if the p-value is less than 0.05.
Confidence Intervals

- Each estimated parameter is associated with a confidence interval, defined as the interval within which the estimated parameter will be with a given probability.

- If we know the sampling distribution of the estimator, we can determine the confidence interval such that, if the model is specified correctly, the interval encloses the true parameter with a given probability.

- A confidence interval has the form

  $$\text{sample estimate} \pm \text{margin of error},$$

  where the margin of error depends on both the sampling distribution and the estimator variance.

- Confidence intervals are calculated in order to assess the precision with which a parameter is estimated. For instance, a 95% confidence interval $[2, 7]$ for a given parameter implies that the parameter estimation is more precise and (ultimately) useful than if the interval were $[-10, 10]$. 

Bootstrap Confidence Intervals

In more complicated models, it is possible that the sampling distribution of an estimator cannot be derived (and even that the estimator is not available as an analytical expression but can only be computed numerically).

How can then one calculate a confidence interval or test hypotheses about the parameter?

*Bootstrapping* is a simulation method that allows estimation of the sampling distribution of an estimator in most practical situations.¹

We describe the steps of building a confidence interval using two bootstrap methods: case resampling and parametric bootstrap.

---

¹See, for example, Ruppert (2004).
Bootstrap Confidence Intervals

Case Resampling

- This is a very general bootstrap method, involving resampling of the observed data sample.

- Denote the parameter of interest by $\theta$. Estimate the parameter and denote the estimate by $\hat{\theta}_1$.

- Resample the data with replacement. The new sample must be of equal size as the original one. Estimate the parameter using the new sample and denote the estimate by $\hat{\theta}_2$.

- Repeat the step above a sufficient number of times to obtain the so-called "bootstrap distribution" of $\hat{\theta}$. That distribution represents an approximation to $\hat{\theta}$'s sampling distribution.

- The bootstrap confidence interval is then derived as the interval enclosed between the appropriate percentiles of the bootstrap distribution. For e.g., a 95% bootstrap confidence interval is enclosed between the 2.5th and the 97.5th percentiles.
Parametric Bootstrap

- Suppose a parametric model has been fitted to the observed data sample and an estimate of $\theta$ calibrated.

- The parametric model is used to generate a new sample of the same size as the original data and a new parameter estimate obtained.

- The step above is repeated to obtain the bootstrap distribution of the estimator.

- The bootstrap confidence interval is derived as in the previous method.
Applying ML estimation to calibrating stable distributions requires a way to determine a numerical approximation of the stable density function. A fast and very efficient approach for stable estimation is based on the Fast Fourier Transform (FFT). It uses the stable characteristic function we used to define the stable distribution and inverts it numerically.¹

Let $f$ be a continuous PDF of a distribution with characteristic function $\varphi$. The two functions are related by the following equation:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t) dt.$$  

With this formula, we are able to (theoretically) calculate the value of $f$ at each given $x$. To perform the task, the FFT—an algorithm that can be used for simultaneous evaluation of the integral above for many different $x$-values—is employed.

¹See Rachev and Mitnik (2000) for details.
Calibrating a probability distribution to a set of observations must be followed by an analysis of how good the fit between the empirical and the theoretical distributions is.

A *probability metric* is a function that assigns distances to two given probability distributions. We will consider two distances, giving rise to two statistical tests:

- Kolmogorov-Smirnov (K-S) distance.
- Anderson-Darling (A-D) distance.
Consider two probability distributions $P$ and $Q$ and denote their CDFs by $F_P$ and $F_Q$, respectively.

The distance between $P$ and $Q$ can be assessed by calculating the highest distance between the values $F_P(x)$ and $F_Q(x)$ for varying $x$.

Mathematically, this means calculating the supremum distance between $F_P$ and $F_Q$,\(^1\)

$$d(P, Q) = \| F_P - F_Q \|_\infty = \sup_x |F_P(x) - F_Q(x)|.$$ 

\(^1\)In mathematics, the supremum is the least upper bound of a set.
K-S Goodness-of-Fit Test

- The empirical CDF $F_n$ of a sample of $n$ observations is given by

$$F_n(t) = \frac{1}{n} \#\{x_i | x_i \leq t\},$$

where $\#\{\cdot\}$ denotes the number of elements contained in $\{\cdot\}$.

- Null hypothesis: $F_n$ and $F$ are the same (i.e., the data are realizations from distribution $F$).

- Under the hypothesis that the sample was generated by a distribution with CDF $F$, the distribution of the K-S distance between $F$ and $F_n$ is tabulated. That is, depending on the values of $n$ and $d$, the p-value can be calculated and a decision for/against the null can be made.
Goodness of Fit: The Anderson-Darling Distance

- In financial applications, one may be interested in analyzing in detail the fit in the tails of the data.

- The K-S distance measures the uniform distance between the two CDFs, i.e., the maximum deviation, regardless of whether it occurs in the tail or the bulk of the data.

- An alternative distance that assigns a higher ”importance” on the tails is the A-D statistic, defined as

\[ AD = \sup_{x} \frac{|F_n(x) - F(x)|}{\sqrt{F(x)(1 - F(x))}}. \]

- The distance between the theoretical and empirical distributions in the numerator is rescaled: the value of the denominator is small for very large and very low \( x \)-values.

- Thus, a certain deviation between \( F_n \) and \( F \) is weighed more heavily if it occurs in the tails.
Testing for Normality

- The K-S test can be applied to answer a question such as: Are the data generated by a specific normal distribution, such as a standard normal? However, it cannot be used to answer: Are the data generated by any normal distribution?

- The Jargue-Bera test (J-B test) of normality addresses the latter question. The test statistic is given by

\[ JB = \frac{n}{6} \hat{s}^2 + \frac{n}{24} (\hat{k} - 3)^2, \]

where

\[ \hat{s} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{\hat{\sigma}} \right)^3 \]

and

\[ \hat{k} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{\hat{\sigma}} \right)^4 \]

are the sample skewness and sample kurtosis, respectively, and \( \hat{\sigma} \) is the sample standard deviation.

- The \( JB \) statistic has asymptotically the \( \chi^2 \) distribution with 2 degrees of freedom. For example, the 95%-quantile of the \( \chi^2(2) \) distribution is equal to 5.99.
References


