Statistical Methods for Finance

Tan Jun Yu

AY20/21 Semester 1

Compiled from lecture notes by Prof Xia Yingcun.

Contents

0 Introduction

Some stylized facts in financial data:

1. Extremely low signal-to-noise ratio: i.e. $\mu/\sigma \ll 1$ or $\mu \ll \sigma$
2. Price changes are less volatile in bull markets and more volatile in bear markets.
3. The autocorrelations of the changes/returns of an asset is very weak, thus prediction of the returns is not possible. However, autocorrelation of volatility is strong, and the volatility can be predicted (GARCH model).
4. Financial data usually have more extreme events/values than suspected (heavy tailed distribution).
5. Correlations between different assets at the extremes of the market are higher than other occasions (copula).
6. Past price changes are negatively correlated with future volatilities.
7. ...

0.1 Return

An asset is anything that is capable of being owned or controlled to produce value (e.g. a portfolio of stocks or an equipment). Usually, we use P_t (or p_t) to denote the price of an asset at time t.

Definition 0.1 (Profit)**.** The profit/revenue of an asset at time t is

$$v_t = P_t - P_{t-1}$$

□

We model the movement of the price P_t as a random walk by assuming that $v_t, t = 1, \dots$ are IID random variables with

$$\mathbf{E}(v_t) = \mu, \quad \mathrm{Cov}(v_1) = \sigma^2$$

Definition 0.2 (Random Walk)**.** Let P_0 be a point, and $P_t = P_0 + v_1 + \cdots + v_t$. Process $P_0, P_1, \dots$ is called a random walk and $v_1, v_2, \dots$ are its step sizes. We have

$$\mathbf{E}(P_t|P_0) = P_0 + t\mu, \quad \mathrm{Cov}(P_t|P_0) = \sigma^2 t$$

Parameter μ is called the drift and determines the general direction of the random walk. Parameter σ is the volatility and determines how much the random walk fluctuates about or depart from the conditional mean $P_0 + \mu t$. If $v_t, t = 1, 2, \dots$ are normally distributed, then the process is called a normal random walk. □

Definition 0.3 (Gross Return)**.** The gross return over k periods is

$$\begin{aligned} G_t(k) &= \frac{P_t}{P_{t-k}} \\ &= \frac{P_{t-k+1}}{P_{t-k}} \times \frac{P_{t-k+2}}{P_{t-k+1}} \times \cdots \times \frac{P_{t-1}}{P_{t-2}} \times \frac{P_t}{P_{t-1}} \\ &= G_{t-k+1}(1) \times G_{t-k+2}(1) \times \ldots \times G_t(1) \end{aligned}$$

Without specifying the number of period, gross return is a single period return, i.e. $G_t = G_t(1)$. □

Definition 0.4 (Net Return)**.** Assuming no dividend, the net return over k holding periods is

$$R_t(k) = \frac{P_t}{P_{t-k}} - 1 = \frac{P_t - P_{t-k}}{P_{t-k}}$$

Usually, $R_t(k) \geq 1$. □

Definition 0.5 (Log Return)**.** The log return over k periods is

$$r_t(k) = \log\left(\frac{P_t}{P_{t-k}}\right) = \log(1 + R_t(k)) = \log(P_t) - \log(P_{t-k})$$

□

Remark.

- When $|R_t(k)|$ is small, the difference between $r_t(k)$ and $R_t(k)$ is negligible, because $r_t(k) = \log(1 + R_t(k)) \approx R_t(k)$
- k-period log-return is simply the sum of the single-period log returns:

$$\begin{aligned} r_t(k) &= \log(P_t) - \log(P_{t-k}) \\ &= [\log(P_{t-k+1}) - \log(P_{t-k})] + [\log(P_{t-k+2}) - \log(P_{t-k+1})] + \cdots + [\log(P_t) - \log(P_{t-1})] \\ &= r_{t-k+1} + r_{t-k=2} + \cdots + r_t \end{aligned}$$

- Log-return is sometimes more convenient than net returns.

Definition 0.6 (Adjustment for Dividend)**.** If a dividend (or interest) D_t is paid prior to time t, so that the initial price of day t is $P_{t-1} - D_t$, then the gross return on day t is

$$1 + R_t = \frac{P_t}{P_{t-1} - D_t}$$

We need to make adjustment to the prices at time $t - 1$ and before, by a multiplier a

$$P'_{t-1} = P_{t-1} * a, \quad \ldots, \quad P'_{t-k} = P_{t-k} * a$$

such that the returns (for the past periods) remain unchanged, i.e.

$$\frac{P_t}{P_{t-1} - D_t} = \frac{P_t}{P'_{t-1}} = \frac{P_t}{P_{t-1} * a} \text{and} \frac{P_{t-k}}{P_{t-k-1}} = \frac{P'_{t-k}}{P'_{t-k-1}}$$

It is easy to see that

$$a = 1 - \frac{D_t}{P_{t-1}}$$

In the analysis of financial data over a long period of time, we should use adjusted prices. □

Definition 0.7 (Excess Return)**.** Excess return is the difference $r_t - r_t^*$ between the asset's log return r_i and the log return r_t^* on some reference asset (usually risk-free). □

Definition 0.8 (Returns of a Portfolio)**.** Suppose one has a portfolio consisting of p different assets. Let w_i be the weight, often expressed as a percentage, of the portfolio's value invested in asset i, i.e. $\sum_{i=1}^{p} w_i = 1$. Then, the value of the asset i is $w_i P_t$ when the total value of the portfolio is P_t.

Suppose R_{it} and r_{it} are the net return and log-return of asset i at time t, respectively. The value of the portfolio provided by the asset at time t is $w_i P_{t-1}(1 + R_{it})$, so the total value of the portfolio is

$$P_t = \sum_{i=1}^{p} w_i P_{t-1}(1 + R_{it}) = \left(1 + \sum_{i=1}^{p} w_i R_{it}\right) P_{t-1}$$

Therefore the overall net return R_t and the log return r_t of the portfolio are respectively

$$R_t = \frac{P_t}{P_{t-1}} - 1 = \sum_{i=1}^{p} w_i R_{it}$$

and

$$r_t = \log\left(1 + \sum_{i=1}^{p} w_i R_{it}\right) \approx \sum_{i=1}^{p} w_i R_{it} \approx \sum_{i=1}^{p} w_i r_{it}$$

□

Theorem 0.1 (Random Walk Model for Log Return)**.** *The random walk hypothesis states that the single-period log returns,* $r_t = \log(P_t) - \log(P_{t-1})$ *are independent. Thus,* $\log(P_t) - \log(P_0)$ *is a random walk if* r_i *are IID.*

Sometimes, we further assume

- $r_t \sim N(\mu, \sigma^2)$
- *If* $r_t, t = 1, 2, \ldots$ *are IID, then* $\log(P_t) - \log(P_{t-k}) \sim N(k\mu, k\sigma^2)$, *because* $\log(P_t) - \log(P_{t-k}) = r_t + r_{t-1} + \ldots + r_{t-k+1}$
- $\frac{P_t}{P_{t-k}}$ *will be lognormal.*

0.2 Geometric Random Walks

Recall that $\log\left(\frac{p_t}{p_{t-k}}\right) = r_t + \cdots + r_{t-k+1}$. Therefore

$$\frac{P_t}{P_{t-k}} = \exp(r_t + \cdots + r_{t-k+1})$$

So taking $k = t$, we have

$$P_t = P_0 \exp(r_t + r_{t-1} + \cdots + r_1)$$

We call such a process whose logarithm is a random walk a geometric random walk or an exponential random walk. If $r_1, r_2, \ldots$ are IID $N(\mu, \sigma^2)$, then P_t is lognormal for all t and the process is called a lognormal geometric random walk with parameters (μ, σ^2).

0.3 Interest Rate

- In the case of a "risk-free" asset (e.g. a Treasury bond), the rate of return is called interest rate.

- If the interest rate is a constant R and is compounded once per unit period (i.e. the interest is converted to the principal), then the value of the risk-free asset at time t is

$$P_t = P_0(1+R)^t$$

- Suppose for any small period of time, say $1/n$, the interest rate is r/n, after each time period the interest is converted to the principal. By doing this to time t, there are nt periods, and the asset's value is

$$P_t = P_0\left(1+\frac{r}{n}\right)^{nt}$$

- If the interest rate is continuously compounded at rate r, then by letting $n \to \infty$,

$$P_t = P_0\left(1+\frac{r}{n}\right)^{nd} \to P_0 e^{rt}, \quad \text{or,} \quad P_t = P_0 e^{rt},$$

  so $\log\left(\frac{P_t}{P_{t-1}}\right) = r$ and $\frac{\mathrm{d}P_t}{\mathrm{d}t} = rP_t$, or $\frac{\mathrm{d}P_t}{\mathrm{d}P_t} = r\,\mathrm{d}t$. This is an ordinary differential equation.

0.4 Asset Prices with Risk

A commonly used model for risky asset prices simply generalizes this ordinary differential equation to a stochastic differential equation of the form

$$\frac{\mathrm{d}P_t}{\mathrm{d}P_t} = \theta\,\mathrm{d}t + \underbrace{\sigma\,\mathrm{d}w_t}_{\text{risk}}$$

where $\{w_t, t \geq 0\}$ is Brownian motion, with

$$\mathrm{d}w_t \sim N(0, \mathrm{d}t)$$

The price process P_t is called geometric Brownian motion, with volatility σ and continuously compounded rate of return or instantaneous rate of return θ, and has the explicit representation

$$P_t = P_0 \exp\left\{\left(\theta - \frac{\sigma^2}{2}\right)t + \sigma w_t\right\}$$

1 Exploratory Data Analysis and Moments

1.1 Useful Distributions

- Normal distribution: $\mathrm{N}(\mu, \sigma^2)$

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- t-distribution: t_ν or $t(\nu)$

$$f(x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{v+1}{2}}$$

  where $\nu > 0$ is the degree of freedom.

- Log-normal distribution: $\log\mathrm{N}(\mu, \sigma)$

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{[\log(x)-\mu]^2}{2\sigma^2}\right\}, \quad x > 0$$

1.2 Moments

Definition 1.1 (Moment, Central Moment)**.** Let X be a random variable. The k-th moment of X is $\mathbf{E}(X^k)$, and the k-th central moment is defined as

$$\mu_1 = \mathbf{E}(X), \quad \mu_k = \mathbf{E}[(X-\mathbf{E}(X))^k], \quad k > 1$$

□

Remark.

- First moment is the mean, μ.
- Second central moment is the variance, σ^2.

$$\sigma^2 = \mu_2 = \mathrm{Cov}(X) = \mathbf{E}[(X-\mathbf{E}(X))^2]$$

  - σ or $\sigma(X)$ or σ_X is the standard deviation.
  - σ is usually related to the scale or dispersion of the distribution.
  - If X is the return, then $\sigma(X)$ measures the risk.

Example 1.1 (Useful Results of Moments)**.**

1. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then

$$\mathbf{E}(X) = \mu, \quad \mathrm{Cov}(X) = \sigma^2, \quad \mu_3 = 0, \quad \mu_4 = 3\sigma^4$$

2. If $X \sim t(\nu)$, then

$$\mathbf{E}(X) = 0, \quad \mathrm{Cov}(X) = \frac{\nu}{\nu-2} \text{ with } \nu > 2,$$
$$\mu_3 = 0,$$
$$\mathbf{E}(X^4) = \frac{3}{\nu-4}\frac{\nu^2}{\nu-2}, \quad \nu > 4$$

3. If $X \sim \text{logN}(\mu, \sigma^2)$, then

$$\mathbf{E}(X^s) = \exp\left(s\mu + \frac{1}{2}s^2\sigma^2\right)$$

□

Example 1.2. The price of an asset can be modelled by

$$P_t = P_0 \exp\left\{\left(\theta - \frac{\sigma^2}{2}\right)t + \sigma w_t\right\}$$

where $w_t \sim \text{N}(0, t)$. Then,

$$\frac{P_t}{P_0} \sim \text{logN}\left(\left(\theta - \frac{\sigma^2}{2}\right)t, \sigma^2 t\right)$$

By point 3 of Example 1.1, we have

$$\mathbf{E}(P_t) = P_0 \exp(t\theta)$$

□

Definition 1.2 (Sharpe Ratio)**.** For random variable X, the Sharpe ratio is defined as

$$SR = \frac{\mathbf{E}(X)}{\sigma(X)}$$

In finance, the Sharpe ratio of return R is

$$SR(R) = \frac{\mathbf{E}(R - r_f)}{\sigma(R)}$$

where r_f is the risk-free return. □

Remark.

- Coefficient of Variation, $CV = \frac{\sigma(X)}{\text{E}(X)}$, is a measure of relative variability. The lower the ratio, the better is the risk/return trade-off.
- The Sharpe Ratio is a measure that indicates the average return minus risk-free return divided by the standard deviation of return on an investment. It is a measure for calculating risk-adjusted return; the higher, the better.
- The Treynor Ratio, $\frac{\text{E}(R_p - r_f)}{\beta_p}$, where R_p is the portfolio return and β_p is the beta of the portfolio (rate of return due to overall market performance), is a performance metric for determining how much excess return was generated for each unit of risk taken on by a portfolio. This allows investors to adjust a portfolio's returns for systemic/market risk instead of total risk in Sharpe ratio.
- The Sharpe ratio along with Treynor ratio and Jensen's alpha, is often used to rank the performance of portfolios or investments.

Theorem 1.1 (Estimating Mean and Standard Deviation)**.** *With sample* $Y_1, \dots, Y_n$*,*

$$\hat{\mu} = \hat{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

If $Y_1, \dots, Y_n$ are IID with mean μ and standard deviation σ, then approximately

$$\begin{aligned}
\sqrt{n}(\hat{\mu} - \mu) &\sim \mathrm{N}(0, \sigma^2) \\
\sqrt{n}(\hat{\sigma}^2 - \sigma^2) &\sim \mathrm{N}(0, 2\sigma^4) \qquad (1) \\
\sqrt{n}(\hat{\sigma} - \sigma) &\sim \mathrm{N}\left(0, \frac{1}{2}\sigma^2\right) \qquad (2)
\end{aligned}$$

Proof of (1). Fact: Suppose $\hat{\theta} - \theta \sim \mathrm{N}(0, \frac{1}{2}\varsigma^2)$, then for any function $f(\theta)$ with derivative $|f'(\theta) < \infty|$,

$$f(\hat{\theta}) - f(\theta_0) \sim \mathrm{N}\left(0, \frac{1}{n}(f'(\theta_0))^2\varsigma^2\right)$$

(incomplete) ■

Proof of (2). Consider the Taylor expansion of $x^{\frac{1}{2}}$ at x_0:

$$x^{\frac{1}{2}} \approx x_0^{\frac{1}{2}} + \frac{1}{2x_0^{\frac{1}{2}}}(x - x_0)$$

Then

$$\begin{aligned}
\hat{\sigma} = (\hat{\sigma}^2)^{\frac{1}{2}} &\approx (\sigma^2)^{\frac{1}{2}} + \frac{1}{2(\sigma^2)^{\frac{1}{2}}}(\hat{\sigma}^2 - \sigma^2) \\
&= \sigma + \frac{1}{2\sigma}(\hat{\sigma}^2 - \sigma^2)
\end{aligned}$$

By (1), we have

$$\hat{\sigma} \sim \mathrm{N}\left(\sigma, \frac{1}{(2\sigma)^2}\left(\frac{1}{n}2\sigma^4\right)\right) = \mathrm{N}\left(\sigma, \frac{1}{2n}\sigma^2\right)$$

■

Remark. The 95% confidence interval for μ and σ are respectively

$$\hat{\mu} \pm 1.96\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \hat{\mu} \pm 1.96\frac{\hat{\sigma}}{\sqrt{n}}$$

$$\hat{\sigma} \pm 1.96\frac{\sigma}{\sqrt{2n}} \quad \text{or} \quad \hat{\sigma} \pm 1.96\frac{\hat{\sigma}}{\sqrt{2n}}$$

So if $0 \in \hat{\mu} \pm 1.96\frac{\hat{\sigma}}{\sqrt{2n}}$, we conclude that $H_0 : \mu = 0$ can be accepted with significance level $\alpha = 0.05$.

Theorem 1.2 (Estimating Sharpe Ratio)**.** *We can also estimate the Sharpe Ratio by*

$$\widehat{SR} = \frac{\hat{\mu}}{\hat{\sigma}}$$

Then,

$$\widehat{SR} - SR \sim \mathrm{N}\left(0, \frac{1}{n}\left(1 + \frac{1}{2}SR^2\right)\right) \qquad (3)$$

The 95% confidence interval for SR is

$$\widehat{SR} \pm 1.96\sqrt{\left(1 + \frac{1}{2}\widehat{SR}\right)/n}$$

Proof of (3). (incomplete) ■

1.3 Quantile, Value-at-Risk and Expected Shortfall

Definition 1.3 (Quantile)**.** Suppose $X \sim F(x)$, for any $0 < q < 1$, the q-th quantile of X is defined as

$$Q_q(X) = \max\{x : P(X < x) \leq q\}$$

If $F(x)$ is strictly increasing, then

$$P(X < Q_q(X)) = q, \quad \text{and} \quad Q_q(X) = F^{-1}(q)$$

□

Definition 1.4 (Value-at-Risk)**.** If X is the net or log return of an asset, then $-Q_q(X)$ is called the $100q\%$ value-at-risk (VaR).

$$\mathrm{VaR}_q(X) = -Q_q(X) = -\max\{v : F(v) \leq q\}$$

Note that q is usually very small (0.01, or 0.001, or 0.0001). □

Remark.

- Usually, there is real loss in extremely bad situations, so $\mathrm{VaR} > 0$.
- VaR is a measure of the risk of loss of an asset X: the minimum loss incurred in the $100q\%$-th worst cases (or minimum amount required reserves to cover losses that regularly occurs).
- $\mathrm{VaR}_q(X + c) = \mathrm{VaR}_q(X) - c$ for any constant c.
- If $X \leq Y$, then $\mathrm{VaR}_q(X) \geq \mathrm{VaR}_q(Y)$.
- $\mathrm{VaR}_q(\lambda X) = \lambda \mathrm{VaR}_q(X)$ for any constant $\lambda > 0$.

Definition 1.5 (Expected Shortfall)**.** The expected shortfall of X at level q is

$$\mathrm{ES}_q(X) = \frac{1}{q} \int_0^q \mathrm{VaR}_\alpha(X) \, \mathrm{d}\alpha$$

The expected shortfall at $100q\%$ level is the average loss of the asset in the worst $100q\%$ of the cases. □

Remark.

- $\mathrm{ES}_q(X + c) = \mathrm{ES}_q(X) - c$ for any constant c.
- If $X \leq Y$, then $\mathrm{ES}_q(X) \geq \mathrm{ES}_q(Y)$.
- $\mathrm{ES}_q(\lambda X) = \lambda \mathrm{ES}_q(X)$ for any constant $\lambda > 0$.

- ES is also called conditional value-at-risk (CVaR), average value-at-risk (AVaR), and expected tail loss (ETL).

$$\begin{aligned}
\mathrm{ES}_q(X) &= -\mathbf{E}(X|X < -\mathrm{VaR}_q(X)) \\
&= -\frac{1}{q}\int_{-\infty}^{-\mathrm{VaR}_q(X)} xf(x)\,\mathrm{d}x = -\frac{1}{q}\int_{-\infty}^{-\mathrm{VaR}_q} x\,\mathrm{d}F(x) \\
&= -\frac{1}{q}\int_0^q F^{-1}(\alpha)\,\mathrm{d}\alpha \qquad (\text{let } \alpha = F(x)) \\
&= \frac{1}{q}\int_0^q \mathrm{VaR}_\alpha(X)\,\mathrm{d}\alpha
\end{aligned}$$

where $f(x)$ is the density function of X.

1.4 Coherent Risk Measures

So far, we have 3 measures of risk: variance / standard deviation, VaR, and ES. Let ρ be a generic measure of risk that maps the riskiness of an asset to an amount of required reserves to cover losses that regularly occur. Let r, r_1, r_2 be random variables such as returns.

Definition 1.6 (Coherent Risk Measures)**.** A risk measure which satisfies the following four properties is coherent.

- Drift Invariance: $\rho(r + c) = \rho(r) - c$
- Homogeneity: $\rho(\lambda r) = \lambda\rho(r)$ for any $\lambda > 0$
- Monotonicity: for a pair of (r_1, r_2), if $r_1 \geq r_2$, then $\rho(r_1) \leq \rho(r_2)$
- Subadditivity: $\rho(r_1 + r_2) \leq \rho(r_1) + \rho(r_2)$, i.e. risks of combinations of two assets is less than the total risks of the two assets separately

□

Remark.

- VaR is not subadditive (see counterexample below).
- However if $r_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $r_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ are jointly normal, then

$$\mathrm{VaR}(r_1 + r_2) \leq \mathrm{VaR}(r_1) + \mathrm{VaR}_2 \quad \text{(subadditivity)} \tag{1}$$

Proof for (1). We have $r_1 + r_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_3^2)$ with $\sigma_3^2 = \mathrm{Cov}(r_1 + r_2) \leq (\sigma_1 + \sigma_2)^2$, i.e. $\sigma_3 \leq \sigma_1 + \sigma_2$. Let $r_0 \sim \mathrm{N}(0, 1)$. Thus,

$$\begin{aligned}
\mathrm{VaR}_\alpha(r_1) &= \sigma_1\mathrm{VaR}_\alpha(r_0)\mu_1 \\
\mathrm{VaR}_\alpha(r_2) &= \sigma_2\mathrm{VaR}_\alpha(r_0)\mu_2 \\
\mathrm{VaR}_\alpha(r_1 + r_2) &= \sigma_3\mathrm{VaR}_\alpha(r_0) - \mu_1 - \mu_2
\end{aligned}$$

and

$$\mathrm{VaR}_\alpha(r_1) + \mathrm{VaR}_\alpha(r_2) - \mathrm{VaR}_\alpha(r_1 + r_2) = (\sigma_1 + \sigma_2 - \sigma_3)\mathrm{VaR}(r_0) \geq 0$$

■

Remark. ES is coherent. We only ned to show the subadditivity. Consider two variables X and Y and n simultaneous realizations $\{(X_i, Y_i), i = 1, \ldots, n\}$.

$$\begin{aligned} \mathrm{ES}_\alpha(X+Y) &= -\frac{\sum_{i=1}^{n*\alpha}(X+Y)_{(i)}}{n*\alpha} \\ &\leq -\frac{\sum_{i=1}^{n*\alpha}(X_{(i)}+Y_{(i)})}{n*\alpha} \\ &= -\frac{\sum_{i=1}^{n*\alpha}X_{(i)}}{n*\alpha} - \frac{\sum_{i=1}^{n*\alpha}Y_{(i)}}{n*\alpha} \\ &= \mathrm{ES}_\alpha(X) + \mathrm{ES}_\alpha(Y) \end{aligned}$$

Example 1.3 (VaR is not subadditive).

x	0.887	-2.395	0.455	0.195	-1.843	0.896	0.998	1.926	0.127	1.213
y	0.245	0.535	-0.208	-0.534	0.789	-2.012	1.296	-0.457	1.122	-0.289
$x+y$	1.132	-1.861	0.247	-0.338	-1.054	-1.116	2.294	1.469	1.248	0.924

For VaR, $x_{(4)} = 0.1954, y_{(4)} = -0.2886$ but $(x+y)_{(4)} = -0.3382$, thus

$$x_{(4)} + y_{(4)} > (x+y)_{(4)}$$

or

$$\mathrm{VaR}_{0.4}(X) + \mathrm{VaR}_{0.4}(Y) < \mathrm{VaR}_{0.4}(Y+X)$$

However, for ES,

$$\mathrm{ES}_{0.4}(X) = 0.9790, \quad \mathrm{ES}_{0.4}(Y) = 0.8228, \quad \mathrm{ES}_{0.4}(X+Y) = 1.0921$$

thus

$$\mathrm{ES}_{0.4}(X) + \mathrm{ES}_{0.4}(Y) > \mathrm{ES}_{0.4}(X+Y)$$

□

1.5 Skewness & Kurtosis

Definition 1.7 (Skewness). The skewness coefficient X measures the degree of asymmetry, and is defined as

$$\mathrm{SK}(X) = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}}$$

□

Remark.

- $\mathrm{SK}(X) = 0$: symmetric (around any value) distributions e.g. normal distribution, t-distribution, binomial distribution with $p = 0.5$
- $\mathrm{SK}(X) > 0$: positive skewness (or right skewed) indicates a relatively long right tail compared to the left tail e.g. χ^2-distribution, F-distribution, and Beta(β_1, β_2) with $\beta_1 > \beta_2$, Binomial distribution with $p < 0.5$
- $\mathrm{SK}(X) < 0$: negative skewness (or left skewed) indicates the opposite, e.g. Beta(β_1, β_2) with $\beta_2 < \beta_1$, Binomial distribution with $p > 0.5$

- *Stylized fact*: In finance, it is believed that aggregate stock market returns display negative skewness, and firm-level stock returns display positive skewness.

Definition 1.8 (Kurtosis)**.** The kurtosis of random variable Y is defined as

$$\mathrm{Kur}(Y) = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\sigma_2^2}$$

The excess kurtosis is

$$\mathrm{Ex.Kur}(Y) = \mathrm{Kur}(Y) - 3$$

□

Remark.

- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. A distribution has heavy tail if kurtosis is greater than 3; normal distribution has kurtosis of 3.
- *Stylized fact*: In finance, the returns have heavy tails.

Example 1.4 (Kurtosis for commonly-used distributions)**.**
- $X \sim \mathrm{N}(\mu, \sigma^2)$: $\mathrm{Kur}(X) = 3$
- $X \sim t(v)$: $\mathrm{Kur}(X) = 3 + \frac{6}{v-4}$ for $v > 4$. Otherwise, the kurtosis does not exist.
- Lognormal with μ and σ: $\mathrm{Kur} = e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 3$

□

Theorem 1.3 (Estimation of Skewness and Kurtosis)**.** *Suppose $Y_1, \dots, Y_n$ are samples from a distribution. Let the sample mean and standard deviation be $\bar{Y}$ and s. Then the skewness and kurtosis are respectively*

$$\widehat{\mathrm{SK}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \bar{Y}}{s}\right)^3, \qquad \widehat{\mathrm{Kur}} = \frac{1}{n}\left(\frac{Y_i - \bar{Y}}{s}\right)^4$$

and

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^k$$

Theorem 1.4 (Test of Skewness)**.** *Given null hypothesis H_0 : X is symmetric, under H_0, we have*

$$\sqrt{n}\widehat{\mathrm{SK}} \sim \mathrm{N}(0, \sigma_{\mathrm{SK}}^2)$$

where

$$\sigma_{\mathrm{SK}}^2 \approx 9 + \frac{\hat{\mu}_6}{\hat{\mu}_2^3} - \frac{6\hat{\mu}_4}{\hat{\mu}_2^2}$$

If X has normal distribution, $\sigma_{\mathrm{SK}}^2 = 6$ With significance level $\alpha = 0.05$, the acceptance region for H_0 is $0 \pm 1.96\frac{\sigma_{\mathrm{SK}}}{\sqrt{n}}$. If X has normal distribution, then the acceptance region is $0 \pm 1.96\frac{\sqrt{6}}{\sqrt{n}}$.

Theorem 1.5 (Anscombe-Glynn Test for Kurtosis)**.** *Given null hypothesis H_0 : X has kurtosis 3, under H_0, we have*

$$\sqrt{n}(\widehat{\mathrm{Kur}} - 3) \sim \mathrm{N}(0, \sigma_{\mathrm{Kur}}^2)$$

where

$$\sigma^2_{\text{Kur}} \approx 24 + \frac{6\hat{\mu}_3^2}{\hat{\mu}_2^3} - \frac{8\hat{\mu}_3\hat{\mu}_5}{\hat{\mu}_2^4}$$

If X has normal distribution, $\sigma^2_{\text{Kur}} = 24$. With significance level $\alpha = 0.05$, the acceptance region for H_0 is $3 \pm 1.96\frac{\sigma_{\text{Kur}}}{\sqrt{n}}$. If X has normal distribution, then the acceptance region is $3 \pm 1.96\frac{\sqrt{24}}{\sqrt{n}}$.

Theorem 1.6 (Jarque-Bera Test of Normality)**.** *The Jarque-Bera test check whether data have the skewness and kurtosis matching a normal distribution.*

$$JB = \frac{n-k+1}{6}\left(\widehat{\text{SK}}^2 + \frac{1}{4}(\widehat{\text{Kur}} - 3)^2\right)$$

where n is the number of observations, and k is the number of regressors in a model if the data is the residuals, or $k = 0$ if it is the observed original data.

If the data comes from a normal distribution, then

$$JB \sim \chi^2(2)$$

1.6 Heavy Tailed Distributions

Definition 1.9 (Heavy-Tailed Distribution)**.** The distribution of a random variable X has a heavy right tail if

$$\lim_{x\to\infty} e^{\lambda x} P(X > x) = \infty, \quad \text{for all } \lambda > 0$$

or heavy left tail if

$$\lim_{x\to-\infty} e^{\lambda |x|} P(X < x) = \infty, \quad \text{for all } \lambda > 0$$

If

$$\lim_{x\to\infty} \frac{P(X > x)}{P(Y > x)} \to \infty \quad \text{or} \quad \lim_{x\to-\infty} \frac{P(X < x)}{P(Y < x)} \to \infty$$

we say X has heavier tail then Y. □

Definition 1.10 (Fat-Tailed Distribution)**.** A distribution is fat-tailed if

$$P(|X| > z) \propto z^{-\alpha} \quad \text{for some } \alpha > 0$$

□

Example 1.5.

- Normal distribution Z does not have heavy-tail

$$P(Z > z) \approx \frac{1}{\sqrt{2\pi}} \frac{e^{-z^2/2}}{z + 0.8e^{0.4z}}$$

- t-distributions obviously have heavy tails. When z is large,

$$P(Z > z) \approx \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\nu+1} z^{-\nu}$$

  which is polynomial tail.

- Lognormal has right heavy tail.

□

1.7 Maximum Likelihood Estimate of Distributions

Definition 1.11 (Likelihood)**.** Suppose $X \sim f(x|\theta)$, but with $\theta = (\theta_1, \ldots, \theta_p)$ unknown. For IID observations $x_1, x_2, \ldots, x_n$, the joint density is

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\theta) &= f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta) \\ &= \prod_{i=1}^{n} f(x_i|\theta) \end{aligned}$$

This function is the likelihood, denoted by $\mathcal{L}(\theta)$. The log-likelihood is

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

The maximum likelihood estimator (MLE) $\hat{\theta}$ satisfies

$$\frac{\partial \log \mathcal{L}(\hat{\theta})}{\partial \theta} = 0$$

□

Example 1.6. Suppose that $x_1, \ldots, x_n$ are IID $\mathrm{N}(\mu, \sigma^2)$ with (μ, σ^2) unknown. The log-likelihood for the unknown parameter (μ, σ^2) is

$$\log \mathcal{L}(\mu, \sigma^2) = -\frac{n}{2}\left[\log(\sigma^2) + \log(2\pi)\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Therefore,

$$\begin{aligned} \frac{\partial \log \mathcal{L}(\hat{\theta})}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \\ \frac{\partial \log \mathcal{L}(\hat{\theta})}{\partial \sigma^2} &= -\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

The MLE are respectively

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

□

Example 1.7. Suppose that $Y_1, \ldots, Y_n$ are IID, after standardizing, $(Y_i - m)/s \sim t(\nu)$. The pdf and log-likelihood for the unknown parameter (s, m, ν) are respectively

$$f(Y_i|m, s, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{((Y_i - m)/s)^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

and

$$\log \mathcal{L}(m, s, \nu) = n \log\left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\right) - \frac{\nu+1}{2}\sum_{i=1}^{n}\log\left(1 + \frac{((Y_i - m)/s)^2}{\nu}\right)$$

The MLE of (m, s, ν) are solution to

$$\frac{\partial \log \mathcal{L}(m, s, \nu)}{\partial m} = 0, \quad \frac{\partial \log \mathcal{L}(m, s, \nu)}{\partial s} = 0, \quad \frac{\partial \log \mathcal{L}(m, s, \nu)}{\partial \nu} = 0$$

There is no closed form for the estimator. □

1.8 AIC and BIC

Definition 1.12 (AIC, BIC)**.** Akaike's Information Criterion (AIC) is defined as

$$\text{AIC} = -2 \log \mathcal{L}(\hat{\theta}_{ML}) + 2p$$

Bayesian Information Criterion (BIC) is defined as

$$\text{BIC} = -2 \log \mathcal{L}(\hat{\theta}_{ML}) + \log(n)p$$

where $\log \mathcal{L}(\hat{\theta}_{ML})$ is the maximized value of the objective function, p is the number of parameters in the model/distribution, and n is the sample size. □

Remark.

- The terms $2p$ and $\log(n)p$ are called complexity penalties since they penalize larger distributions.
- AIC or BIC can be used for comparison of different distributions: for both criteria, a distribution with smaller AIC/BIC is preferred.

2 Multivariate Statistical Models and Portfolio Theory

Definition 2.1 (Multivariate Distribution)**.** Suppose $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_2)^\top$ are p-dimensional random variables. Its multivariate cumulative function $F(x_1, \ldots, x_p)$ is

$$F(x_1, \ldots, x_2) = P(\mathbf{x}_1 \leq x_1, \ldots, \mathbf{x}_p \leq x_p)$$

If there is a function $f(x_1, \ldots, x_p) \geq 0$ such that, for any $a_1, b_1, \ldots, a_p, b_p$,

$$P(a_1 < \mathbf{x}_1 \leq b_1, \ldots, a_p < \mathbf{x}_p \leq b_p) = \int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} f(x_1, \ldots, x_p)\, \mathrm{d}x_1 \cdots \mathrm{d}x_p$$

then

$$f(x_1, \ldots, x_p)$$

is the pdf. It follows that

$$f(x_1, \ldots, x_p) = \frac{\partial^p F(x_1, \ldots, x_p)}{\partial x_1 \cdots \partial x_p}$$

□

Definition 2.2 (Multivariate Normal: $\mathrm{N}(\mu, \Sigma)$)**.** The multivariate normal distribution has pdf

$$f_X(x_1, \ldots, x_p) = \frac{1}{(2\pi)^{k/2} |\mathbf{\Sigma}|^{1/2}} \exp\biggl(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top\biggr) \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

where $|\mathbf{\Sigma}|$ is the determinant of

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix},$$

$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top, \mathbf{x} = (x_1, \ldots, x_p)^\top$. We also write $\sigma_{kk} = \sigma_k^2$ and

$$\sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$$

where ρ_{ij} is the correlation coefficient between $\mathbf{x}_i$ and $\mathbf{x}_j$.

The marginal distributions are

$$\mathbf{x}_k \sim \mathrm{N}(\mu_k, \sigma_{kk})$$

The bivariate (X, Y) case has pdf

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\biggl(-\frac{1}{2(1-\rho^2)}\biggl[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\biggr]\biggr)$$

where ρ is the correlation between X and Y, $\sigma_x > 0, \sigma_y > 0$. In this case,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

□

Definition 2.3 (Multivariate t: $t(\nu, \mu, \Sigma)$)**.** The multivariate t distribution has pdf

$$f_{\mathbf{x}}(x_1, \dots, x_p) = \frac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)(\nu\pi)^{p/2}} |\mathbf{\Sigma}|^{-1/2} \left[1 + \frac{1}{\nu}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-(\nu+p)/2}$$

The bivariate case (X, Y) has pdf

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \left\{1 + \frac{1}{\nu(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\rho_y)^2}{\sigma_x\sigma_y}\right]\right\}^{-(\upsilon+2)/2}$$

where μ_x, μ_y are the expectations of X and Y, but σ_x and σ_y are not the standardized deviations, and ρ is not the correlation coefficient. Note that $\frac{\nu}{\nu-2}\sigma_x^2 = \mathrm{Cov}(x)$. □

Definition 2.4 (Expectation, Covariance and Correlation Coefficient Matrix)**.** Suppose $Y = (\mathbf{y}_1, \dots, \mathbf{y}_p)^\top$ are p-dimensional random variables. The population expectation is

$$\mathbf{E}(Y) = \begin{pmatrix} \mathbf{E}(\mathbf{y}_1) \\ \mathbf{E}(\mathbf{y}_2) \\ \vdots \\ \mathbf{E}(\mathbf{y}_p) \end{pmatrix} = \begin{pmatrix} \mu_{\mathbf{y}_1} \\ \mu_{\mathbf{y}_2} \\ \vdots \\ \mu_{\mathbf{y}_p} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

The population variance between $\mathbf{y}_i$ and $\mathbf{y}_j$ is

$$\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{E}[(\mathbf{y}_i - \mu_{\mathbf{y}_i})(\mathbf{y}_j - \mu_{\mathbf{y}_j})] = \sigma_{ij}$$

The population correlation coefficient between $\mathbf{y}_i$ and $\mathbf{y}_j$ is

$$\mathrm{Cor}(\mathbf{y}_i, \mathbf{y}_j) = \frac{\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j)}{\sigma_{\mathbf{y}_i}\sigma_{\mathbf{y}_j}} = \frac{\mathbf{E}[(\mathbf{y}_i - \mu_{\mathbf{y}_i})(\mathbf{y}_j - \mu_{\mathbf{y}_j})]}{\sigma_{\mathbf{y}_i}\sigma_{\mathbf{y}_j}} = \rho_{ij} = \rho_{\mathbf{y}_i, \mathbf{y}_j}$$

For $Y = (\mathbf{y}_1, \cdots, \mathbf{y}_p)^\top$, the covariance matrix is

$$\begin{aligned} \Sigma &= \mathrm{Cov}(Y) \\ &= \begin{pmatrix} \mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_1) & \mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_2) & \cdots & \mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_p) \\ \mathrm{Cov}(\mathbf{y}_2, \mathbf{y}_1) & \mathrm{Cov}(\mathbf{y}_2, \mathbf{y}_2) & \cdots & \mathrm{Cov}(\mathbf{y}_2, \mathbf{y}_p) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(\mathbf{y}_p, \mathbf{y}_1) & \mathrm{Cov}(\mathbf{y}_p, \mathbf{y}_2) & \cdots & \mathrm{Cov}(\mathbf{y}_p, \mathbf{y}_p) \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \\ &= \mathbf{E}\{[Y - \mathbf{E}(Y)][Y - \mathbf{E}(Y)]^\top\} \end{aligned}$$

and the correlation matrix is defined as

$$\mathrm{Cor}(Y) = \begin{pmatrix} 1 & \mathrm{Cor}(\mathbf{y}_1, \mathbf{y}_2) & \cdots & \mathrm{Cor}(\mathbf{y}_1, \mathbf{y}_p) \\ \mathrm{Cor}(\mathbf{y}_2, \mathbf{y}_1) & 1 & \cdots & \mathrm{Cor}(\mathbf{y}_2, \mathbf{y}_p) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cor}(\mathbf{y}_p, \mathbf{y}_2) & \mathrm{Cor}(\mathbf{y}_p, \mathbf{y}_2) & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

Both $\mathrm{Cov}(Y)$ and $\mathrm{Cor}(Y)$ are positive semi-definite. □

Remark.

- If $X \sim \mathrm{N}(\mu, \Sigma)$, then $\mathbf{E}(X) = \mu, \mathrm{Cov}(X) = \Sigma$.

- If $X \sim t(\nu, \mu, \Sigma)$, then $\mathbf{E}(X) = \mu, \mathrm{Cov}(X) = c\Sigma$.
- For any $p \times q$ matrix A, $Z = AY$ is a $q \times 1$ random vector.
  - $\mathbf{E}(AY) = A\mathbf{E}(Y)$
  - $\mathrm{Cov}(AY) = A\mathrm{Cov}(Y)A^\top$
  - For details of special cases, refer to notes.

Definition 2.5 (Sample Mean, Covariance, and Correlation Coefficient)**.** For sample $\{Y_\ell = (\mathbf{y}_{1\ell}, \mathbf{y}_{2\ell}, \ldots, \mathbf{y}_{p\ell})\}$, the sample mean is

$$\bar{Y} = \frac{1}{n}\sum_{\ell=1}^{n} Y_\ell$$

The sample variance and covariance are

$$\begin{aligned} s_{ij} &= \frac{1}{n}\sum_{\ell=1}^{n}(\mathbf{y}_{i\ell} - \bar{\mathbf{y}}_i)(\mathbf{y}_{j\ell} - \bar{\mathbf{y}}_j), && \text{(biased, from MLE), or} \\ s_{ij} &= \frac{1}{n-1}\sum_{\ell=1}^{n}(\mathbf{y}_{i\ell} - \bar{\mathbf{y}}_i)(\mathbf{y}_{j\ell} - \bar{\mathbf{y}}_j), && \text{(unbiased)} \end{aligned}$$

where $\bar{\mathbf{y}}_i$ is the sample mean

$$\bar{\mathbf{y}}_i = \frac{1}{n}\sum_{\ell=1}^{n} \mathbf{y}_{i\ell}$$

The correlation is calculated by

$$r_{\mathbf{y}_i,\mathbf{y}_j} = \frac{\sum_{\ell=1}^{n}(\mathbf{y}_{i\ell} - \bar{\mathbf{y}}_i)(\mathbf{y}_{j\ell} - \bar{\mathbf{y}}_j)}{\sqrt{\sum_{\ell=1}^{n}(\mathbf{y}_{i\ell} - \bar{\mathbf{y}}_i)^2}\sqrt{\sum_{\ell=1}^{n}(\mathbf{y}_{j\ell} - \bar{\mathbf{y}}_j)^2}}$$

The sample covariance matrix is

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

and the sample correlation matrix is

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

□

Theorem 2.1 (Hypothesis on Correlation)**.** *Given null hypothesis $H_0 : \rho_{X,Y} = 0$. Under H_0,*

$$\sqrt{n-3} \times r_{X,Y} \sim \mathrm{N}(0, 1)$$

In other words,

- *if* $|r_{X,Y}| > z_{1-\alpha/2}\frac{1}{\sqrt{n-3}}$, *then reject* H_0 *with significance level* α.
- *if* $|r_{X,Y}| \leq z_{1-\alpha/2}\frac{1}{\sqrt{n-3}}$, *do not reject* H_0 *with significance level* α.

Remark (Estimation of Covariance Matrix)**.** The sample covariance matrix

$$S = \frac{1}{n}\sum_{\ell=1}^{n}(Y_\ell - \bar{y})(Y_\ell - \bar{Y})^\top$$

can be shrunk towards a biased estimator with lower variance,

$$S^* = (1-\lambda)S + \lambda I$$

where I is the identity matrix. This improves performance in high-dimensional setting. The choice of λ can be selected either theoretically or cross-validation.

Definition 2.6 (Portfolio)**.** Suppose there are p assets with returns R_i where $i = 1, \ldots, p$ respectively. $R = (R_1, \ldots, R_p)^\top$. Their expected returns and risks are respectively

$$r_i = \mathbf{E}(R_i), \quad \sigma_i = \sigma_{R_i}, \quad \operatorname{Cov}(R) = \Sigma = (\sigma_{ij})_{1\leq i,j\leq p} \quad (\sigma_i = \sqrt{\sigma_{ii}})$$

A portfolio is a new asset consisting of the existing assets

$$R_N = \sum_{i=1}^{p} R_p = w^\top R$$

where w_i is the weight of the i-th asset and satisfies $\sum_{i=1}^{p} w_i = 1$. The expected return is

$$\mathbf{E}(R_N) = \sum_{i=1} w_i \mathbf{E}(R_i) = w^\top \mathbf{E}(R)$$

The variance or risk of the portfolio return is

$$\sigma_N^2 = \sum_{i=1} w_i^2 \sigma_i^2 + \sum_{i=1}\sum_{j\neq i} w_i w_j \sigma_{ij}$$

where σ_{ij} is the covariance between the returns on assets i and j. Alternatively,

$$\sigma_N^2 = \sum_{i=1}\sum_{j=1} w_i w_j \sigma_{ij} = w^\top \Sigma w$$

The Sharpe ratio to be maximized is

$$\frac{\mathbf{E}(R_N - r_f)}{\sigma_N} = \frac{\sum_{i=1} w_i \mathbf{E}(R_i - r_f)}{\left(\sum_{i=1}\sum_{j=1} w_i w_j \sigma_{ij}\right)^{\frac{1}{2}}}$$

with respect to $w_i, \ldots, w_p$. □

Example 2.1 (Investment in One Risky Asset and One Risk-Free Asset)**.**

$$R_N = wR_1 + (1-w)r_f = w(R_1 - rf) + rf$$

where R_1 has expected return r_1 and risk σ_1, and r_f is the return of the risk-free asset.

- Expected return: $r_N = \mathbf{E}(R_N) = wr_1 + (1-w)r_f$
- Risk: $\sigma_N = w\sigma_1$, because $\text{Cov}(R_N) = w^2\text{Cov}(R_1) = w^2\sigma_1^2$
- Sharpe ratio: $\frac{\text{E}(R_N)-r_f}{\sigma_N} = \frac{r_1-r_f}{\sigma_1}$
- Excess return: $\mathbf{E}(R_N) - r_f = \frac{r_1-r_f}{\sigma_1}\sigma_N$ (proportional to risk with a fixed coefficient)
- The plot of (σ_N, r_N) is a straight line, called the capital market line.

□

Example 2.2 (Investment in Two Risky Assets)**.**

$$R_N = wR_1 + (1-w)R_2$$

with $\mathbf{E}(R_i) = r_i, \quad \text{Cov}(R_i) = \sigma_i^2, \quad \text{Cov}(R_1, R_2) = \sigma_{12}$.

- Expected return: $r_N = \mathbf{E}(R_N) = wr_1 + (1-w)r_2$
- Risk: $\sigma_N = (w^2\sigma_1^2 + 2w(1-w)\sigma_{12} + (1-w)^2\sigma_2^2)^{\frac{1}{2}}$
- The plot of $(\sigma_N(w), r_N(w))$ for all $0 \le w \le 1$ is a curve called the efficient frontier.

□

Example 2.3 (Investment in Three or more Risky Assets)**.**

$$R_N = w_1R_1 + w_2R_2 + \cdots + w_pR_p$$
$$\text{subject to } w_i \ge 0, \quad \sum_{i=1}^{p} w_i = 1$$

with $\mathbf{E}(R_i) = r_i$, $\text{Cov}(R_i) = \sigma_i^2$, $\text{Cov}(R_i, R_j) = \sigma_{ij}$.

- Expected return: $r_N = \sum_{i=1}^n w_i r_i$
- Risk: $\sigma_N = \{\sum_{i=1}^p w_i^2\sigma_i^2 + \sum_{j=1}^p \sum_{i\neq j}^p w_i w_j \sigma_{ij}\}^{\frac{1}{2}}$
- The plot of $(\sigma_N(w_1,\ldots,w_p), r_N(w_1,\ldots,w_p))$ form a region.
- For each r_N there is (usually) only one set of $w_1,\ldots,w_p$ that has the smallest σ_N. For each σ_N there is only one set $w_1,\ldots,w_p$ that has the highest r_N.
- The upper outer surface of the region is the efficient frontier.

□

Theorem 2.2 (Global Minimum Variance Portfolio)**.** *The global minimum variance portfolio is attained at*

$$w = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}}$$

where Σ is the covariance matrix.

Proof.

$$\min_w \sigma_N^2 = w^\top\Sigma w$$
$$\text{s.t. } w^\top\mathbf{1} = \mathbf{1}$$

The Lagrangian is

$$L(w,\lambda) = w^\top \Sigma w + \lambda(w^\top \mathbf{1} - 1)$$

The first order conditions are

$$\begin{aligned}
\frac{\partial L(w,\lambda)}{\partial w} &= \frac{\partial w^\top \Sigma w}{\partial w} + \frac{\partial}{\partial w}\lambda(w^\top \mathbf{1} - 1) = 2\Sigma w + \lambda \mathbf{1} = \mathbf{0} \\
\frac{\partial L(w,\lambda)}{\partial \lambda} &= \frac{\partial w^\top \Sigma w}{\partial \lambda} + \frac{\partial}{\partial \lambda}\lambda(w^\top \mathbf{1} - 1) = w^\top \mathbf{1} - 1 = 0
\end{aligned}$$

From the first equation, we have that

$$w = -\frac{1}{2}\lambda \Sigma^{-1}\mathbf{1}$$

Multiplying both sides by $\mathbf{1}^\top$ and using the second equation,

$$1 = \mathbf{1}^\top w = -\frac{1}{2}\lambda \mathbf{1}^\top \Sigma^{-1}\mathbf{1} \implies \lambda = -2\frac{1}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}}$$

Substituting the value of λ in the equation for w,

$$w = -\frac{1}{2}(-2)\frac{1}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}}\Sigma^{-1}\mathbf{1} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}}$$

■

Definition 2.7 (Tangency Portfolio)**.** The portfolio with the highest Sharpe ratio

$$\frac{r_N - r_f}{\sigma_N}$$

is the tangency portfolio. The risk-free capital market line crosses the efficient frontier at this point. □

Theorem 2.3 (Tangency Portfolio)**.** *The tangency portfolio is attained at*

$$w = \frac{\Sigma^{-1}(r - r_f \cdot \mathbf{1})}{\mathbf{1}^\top \Sigma^{-1}(r - r_f \cdot \mathbf{1})}$$

3 Copula

3.1 Definition of Copula and its Properties

Definition 3.1 (Copula)**.** A copula is a special multivariate cumulative distribution function.

$$C(u_1, u_2, \ldots, u_p)$$

whose univariate marginal distributions are all Uniform(0,1). □

Remark (Properties of Copula)**.**

- $(u_1, \ldots, u_p) \in [0, 1]^p$
- $C(u_1, \ldots, u_p)$ is increasing on $[0, 1]^p$
- $C(u_1, \ldots, u_p)$ has marginal CDF as

  $$C_k(u) = C(1, \ldots, 1, u, 1, \ldots, 1) = u$$

  for all $u \in [0, 1]$.
- The density of a copula is

  $$c(u_1, \ldots, u_p) = \frac{\partial^p C(u_1, \ldots, u_p)}{\partial u_1 \cdots \partial u_p}$$

Theorem 3.1 (Sklar's Theorem)**.** *Any continuous random vector $X = (\mathbf{x}_1, \ldots, \mathbf{x}_2)^\top$ has a copula.*

- *If $\mathbf{x}$ is continuous univariate with cumulative distribution $F_{\mathbf{x}}(x)$, then*

  $$F_{\mathbf{x}}(\mathbf{x}) \sim \text{Uniform}[0, 1]$$

- *For any random vector $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^\top$ with joint CDF $F(x_1, x_2, \ldots, x_p)$ and marginal CDF $F_1(x_1), F_2(x_2), \ldots, F_p(x_p)$, let $\mathbf{u}_i = F_i(\mathbf{x}_i)$, and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)^\top$. Then*

  $$\begin{aligned} C^*(u_1, \ldots, u_p) &:= P(\mathbf{u}_1 \leq u_1, \mathbf{u}_2 \leq u_2, \ldots, \mathbf{u}_p \leq u_p) \\ &= P(\mathbf{x}_1 \leq F_1^{-1}(u_1), \mathbf{x}_2 \leq F_2^{-1}(u_2), \ldots, \mathbf{x}_p \leq F_p^{-1}(u_p)) \\ &= F(F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots, F_p^{-1}(u)) \end{aligned}$$

  is a copula. This copula is inherently embedded in the joint distribution. If we let $u_k = F_k(x_k)$, then

  $$F(x_1, \ldots, x_p) = C^*(F_1(x_1), F_2(x_2), \ldots, F_p(x_p))$$

- *For the copula inherently embedded, its density is*

  $$c^*(u_1, u_2, \ldots, u_p) = \frac{f(F_1^{-1}(u_1), \ldots, F_p^{-1}(u_1))}{f_1(F_1^{-1}(u_1)) \times \cdots \times f_p(F_p^{-1}(u_p))}$$

- *For any marginal distribution $F_1(u_1), \ldots, F_p(u_p)$, we can introduce a copula $C_{new}(u_1, \ldots, u_p)$ to combine them and generate a joint distribution*

  $$F_{new}(y_1, \ldots, y_p) = C_{new}\{F_{\mathbf{y}_1}(y_1), \ldots, F_{\mathbf{y}_p}(y_p)\}$$

and density

$$f_{new}(y_1, \dots, y_p) = c_{new}\{F_{\mathbf{y}_1}(y_1), \dots, F_{\mathbf{y}_p}(y_p)\} f_{\mathbf{y}_1}(y_1) \times \cdots \times f_{\mathbf{y}_p}(y_p)$$

- *X and Y are independent if and only if their copula is*

$$C(u, v) = uv, \quad 0 \le u, v \le 1$$

or

$$c(u, v) = 1, \quad 0 \le u, v \le 1$$

3.2 Commonly Used Copulas

- Independence copula: $C(u_1, \dots, u_p) = u_1 \cdots u_p$
- Co-monotonocity copula: $M(u_1, \dots, u_p) = \min(u_1, \dots, u_p)$
- Counter-monotonicity copula: $W(u_1, \dots, u_p) = \max\{1 - p + \sum_{i=1}^{p} u_i, 0\}$

Remark. For any copula $C(u_1, \dots, u_p)$,

$$W(u_1, \dots, u_p) \le C(u_1, \dots, u_p) \le M(u_1, \dots, u_p)$$

Definition 3.2 (Gaussian Copula)**.** For a given covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$, the Gaussian copula with parameter matrix Σ and expectation μ, can be written as

$$C_{\Sigma}^{\text{Gauss}}(u) = \Phi_{\Sigma}(\Phi_1^{-1}(u_1), \dots, \Phi_p^{-1}(u_p))$$

where Φ_k^{-1} is the inverse CDF of $\text{N}(0, \sigma_{kk})$, and Φ_{Σ} is the joint CDF of a multivariate normal distribution with mean vector 0 and covariance matrix equal to the correlation matrix, $\Sigma = R$. For the bivariate case, we have

$$C_{\rho}^{\text{Gauss}}(u_1, u_2) = \Phi_{\rho}(\Phi_1^{-1}(u_1), \Phi_2^{-1}(u_2))$$

or equivalently

$$C(u_1, u_p; \rho) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left\{-\frac{s_1^2 - 2\rho s_1 s_2 + s_2^2}{2(1-\rho^2)}\right\} \mathrm{d}s_1 \, \mathrm{d}s_2$$

where $\Phi^{-1}(\cdot)$ is the quantile or inverse function of the standard normal. Note that this copula is jointly normal, even if the distribution is not normal. □

Definition 3.3 (*t* Copula)**.**

$$C_{\nu,\Sigma}(u_1, \dots, u_p) = \int_{-\infty}^{t_\nu^{-1}(u_1)} \cdots \int_{-\infty}^{t_\nu^{-1}(u_p)} \frac{\Gamma\frac{\nu+p}{2}}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{p}{2}}} |\Sigma|^{\frac{1}{2}} \left[1 + \frac{1}{\nu} X^\top \Sigma^{-1} X\right]^{-\frac{\nu+p}{2}} \mathrm{d}x_1 \cdots \mathrm{d}x_p$$

where t_ν^{-1} denotes the quantile function of a standard univariate t_ν distribution, and $X = (x_1, \dots, x_p)^\top$. For the bivariate case, we have

$$C(u_1, u_2; \nu, \rho) = \int_{-\infty}^{t_\nu^{-1}(u_1)} \cdots \int_{-\infty}^{t_\nu^{-1}(u_p)} \frac{1}{2\pi\sqrt{(1-\rho^2)}} \left[1 + \frac{s_1^2 - 2\rho s_1 s_2 + s_2^2}{\nu(1-\rho^2)}\right]^{-\frac{\nu+2}{2}} \mathrm{d}x_1 \, \mathrm{d}x_2$$

□

Theorem 3.2. *The copula generated by* $X = (\mathbf{x}_1, \dots, \mathbf{x}_p)^\top$ *is the same as that generated by* $\tilde{X} = (\frac{\mathbf{x}_1 - \mu_1}{\sigma_1}, \dots, \frac{\mathbf{x}_p - \mu_p}{\sigma_p})^\top$ *if* $\sigma_1 > 0, \dots, \sigma_p > 0$.

Theorem 3.3 (Archimedean Copula)**.** *The Archimedean copula is*

$$C(u_1, \dots, u_p) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_p))$$

where ψ *is the generator. The following table are the most popular copulas generated by the Archimedian copula.*

Name	$\psi(t)$	*Parameter*
Ali-Mikhail-Haq	$\frac{1-\theta}{\exp(t)-\theta}$	$\theta \in [0, 1)$
Clayton	$(1 + \theta t)^{-1/\theta}$	$\theta \in [0, \infty)$
Frank	$-\frac{1}{\theta}\log(1 - (1 - \exp(-\theta))\exp(-t))$	$\theta \in (0, \infty)$
Gumbel	$\exp\left(-t^{1/\theta}\right)$	$\theta \in [1, \infty)$
Independence	$\exp(-t)$	
Joe	$1 - (1 - \exp(-t))^{1/\theta}$	$\theta \in [1, \infty)$

Usually, the bigger θ *is, the stronger is the dependence between the variables.*

Theorem 3.4 (Mixture of Copulas)**.** *Suppose* $C(u_1, \dots, u_p)$ *and* $D(u_1, \dots, u_p)$ *are two copulas, then for any* $w : 0 \le w \le 1$,

$$wC(u_1, \dots, u_p) + (1 - w)D(u_1, \dots, u_p)$$

is also a copula.

3.3 Tail Dependence of Bivariate Copula

Let $I_i = 1(\mathbf{y}_i \le F_i^{-1}(u)), i = 1, 2$. We can consider I_i to be the indicator of default for the i-th bond/asset. Then

$$\begin{aligned} E(I_1 I_2) &= P(I_i = 1 \cap I_2 = 1) \\ &= P(\mathbf{y}_1 \le F_1^{-1}(u) \text{ and } \mathbf{y}_2 \le F_2^{-1}(u)) \\ &= F(F_1^{-1}(u), F_2^{-1}(u_2)) \\ &= C(u, u) \end{aligned}$$

and the correlation coefficient is

$$\mathrm{Cor}(I_1 I_2) = \frac{\mathbf{E}(I_1 I_2) - \mathbf{E}(I_1)E(I_2)}{\sqrt{[\mathbf{E}(I_1^2) - \mathbf{E}(I_1)^2][\mathbf{E}(I_2^2) - \mathbf{E}(I_2)^2]}} = \frac{C(u, u) - u^2}{u - u^2}$$

So the lower tail correlation of $\mathbf{y}_1$ and $\mathbf{y}_2$ is

$$\lim_{u \to 0} \mathrm{Cor}(I_1, I_2) = \lim_{u \to 0} \frac{C(u, u)}{u}$$

Definition 3.4 (Lower and Upper Tail Dependence)**.** Lower tail dependence λ_L is defined as

$$\lambda_L := \lim_{q \to 0^+} P(\mathbf{y}_2 \le F_{\mathbf{y}_2}^{-1}(q) | \mathbf{y}_1 \le F_{\mathbf{y}_1}^{-1}(q)) = \lim_{q \to 0^+} \frac{C_Y(q, q)}{q}$$

Upper tail dependence λ_U is defined as

$$\lambda_L := \lim_{q \to 1^-} P(\mathbf{y}_2 \geq F_{\mathbf{y}_2}^{-1}(q) | \mathbf{y}_1 \geq F_{\mathbf{y}_1}^{-1}(q)) = \lim_{q \to 1^-} \frac{1 - 2q + C_Y(q,q)}{1-q}$$

□

Check lecture notes for tail dependence of commonly-used copulas.

3.4 Empirical Copula

Consider only bivariate case $Z = (\mathbf{x}, \mathbf{y})$, their distribution (either marginal or joint) is unknown and needs to be estimated. For observations $(x_i, y_i), i = 1, \dots, n$, the empirical CDF of Z is

$$\hat{F}(x,y) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x, y_i \leq y) \quad \text{or} \quad \frac{1}{n+1} \sum_{i=1}^{n} I(x_i \leq x, y_i \leq y)$$

and the marginal CDF is

$$\hat{F}_{\mathbf{x}}(x) = \frac{1}{n+1} \sum_{i=1}^{n} I(x_i \leq x) = \frac{\#\{x_i : x_i \leq x\}}{n+1}$$
$$\hat{F}_{\mathbf{y}}(y) = \frac{1}{n+1} \sum_{i=1}^{n} I(y_i \leq y) = \frac{\#\{y_i : y_i \leq y\}}{n+1}$$

where # is the number of elements in a set. We define

$$u_j = \hat{F}_{\mathbf{x}}(x_j) = \frac{\#\{x_i : x_i \leq x\}}{n+1} = \frac{\text{rank of } x_j \text{ in } x_1, \dots, x_n}{n+1}$$

and

$$v_j = \hat{F}_{\mathbf{y}}(y_j) = \frac{\#\{y_i : y_i \leq y\}}{n+1} = \frac{\text{rank of } y_j \text{ in } y_1, \dots, y_n}{n+1}$$

and they are observations from $\mathbf{u}_{\mathbf{x}} = F_{\mathbf{x}}(\mathbf{x})$ and $\mathbf{u}_{\mathbf{y}} = F_{\mathbf{y}}(\mathbf{y})$. $\{u_i, i = 1, \dots, n\}$ and $\{v_i, i = 1, \dots, n\}$ are uniformly distributed on $[0,1]$. Recall that the copula of $(\mathbf{x}, \mathbf{y})$ is the CDF of $\mathbf{u}_{\mathbf{x}} = F_{\mathbf{x}}(\mathbf{x})$ and $\mathbf{u}_{\mathbf{y}} = F_{\mathbf{y}}(\mathbf{y})$, therefore the empirical copula is

$$\hat{C}(u_1, u_2) = \frac{1}{n} \sum_{i=1}^{n} I(u_i \leq u_1, v_i \leq u_2)$$

3.5 Fitting Copula

To model a joint distribution of $(\mathbf{y}_1, \dots, \mathbf{y}_p)$, we only need to model the marginal distribution of each variable $F_k(y_k|\theta_k), k = 1, \dots, p$, as well as the copula density $c(u_1, \dots, u_p|\theta_C)$. For example, we can have $F(x,y) = C_t\Big(t_{v_1}(\frac{x-\mu_1}{s_1}), t_{v_2}(\frac{x-\mu_2}{s_2})|\rho, \nu\Big)$. More generally, the joint density is

$$c(F_1(y_1|\theta_1), \dots, F_p(y_p|\theta_p)|\theta_C) \times f_1(y_1|\theta_1) \cdots f_p(y_p|\theta_p)$$

For samples $Y_1 = (y_{11}, \dots, y_{1p})^\top, \dots, Y_n = (y_{n1}, \dots, y_{np})^\top$, the log-likelihood function is

$$\log\{L\} = \sum_{i=1}^{n} \log[c(F_1(y_{i1}|\theta_1), \dots, F_p(y_{ip}|\theta_p)|\theta_C) f_1(y_{i1}|\theta_1) \cdots f_p(y_{ip}|\theta_p)]$$

The maximum values for the parameters are the MLE of the parameters $(\theta_1, \ldots, \theta_p; \theta_C)$. Note that $F_k(y_{ik}|\theta_k) \approx u_{ik}$, so we can estimate θ_C directly by maximizing

$$\log\{L\} = \sum_{i=1}^{n} [c(u_{ik}, \ldots, u_{ik}|\theta_C)]$$

If we need to estimate the distribution function, we then only need to estimate each marginal density separately, by maximizing

$$\sum_{i=1}^{n} \log[f_k(y_{ik}|\theta_k)]$$

We can also use AIC or BIC to select among different copulas, where

$$\text{AIC} = -2\log(L) + 2n_p, \qquad \text{BIC} = -2\log(L) + \log(n) \times n_p$$

We prefer the copula with smallest AIC or BIC.

In summary, the steps to fit the joint distribution of two random variables are:

1. Fit a distribution to each variable separately, denote them $\hat{F}_1(x_1)$ and $\hat{F}_2(x_2)$ respectively.
2. Fit a copula to the ranks of the joint data, denoted by $\hat{C}(u_1, u_2)$.
3. The fitted joint distribution is given by

$$\hat{C}(\hat{F}_1(x_1), \hat{F}_2(x_2))$$

The t-copula is commonly used in financial data.

3.6 Monte Carlo Simulation

Suppose $G(u), F(x_1, \ldots, x_p)$ are strictly continuous increasing CDFs, and $F_i(x_i), i = 1, \ldots, p$ are the marginal CDF of $F(x_1, \ldots, x_p)$. Suppose $\mathbf{z} \sim G(x)$, and let $\mathbf{u} = G(\mathbf{z})$, then $\mathbf{u} \sim$ Uniform$(0, 1)$. If $\mathbf{u} \sim$ Uniform$(0, 1)$, let $\mathbf{z} = G^{-1}(\mathbf{u})$, then $\mathbf{z} \sim G(z)$.

Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \sim F(x_1, \ldots, x_p)$, then

- Marginal $\mathbf{x}_i \sim F_i(x_i) = F(\infty, \ldots, x_i, \ldots, \infty)$
- Let $U \sim \big(F_1^{-1}(\mathbf{x}_1), \ldots, F_p^{-1}(\mathbf{x}_p)\big)$, then $U \sim F\big(F_1^{-1}(u_1), \ldots, F_p^{-1}(u_p)\big) = C(u_1, \ldots, u_p)$

Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_p) \sim C(u_1, \ldots, u_p)$, and let $X = \big(F_1^{-1}(\mathbf{u}_1), \ldots, F_1^{-1}(\mathbf{u}_p)\big)$, then

$$X \sim C(F_1(x_1), \ldots, F_p(x_p))$$

For a general distribution $\mathbf{x} \sim F(x)$, we can generate random numbers from it:

$$\mathbf{x}_i = F^{-1}(\mathbf{u}_i), \qquad \mathbf{u}_i \sim \text{Uniform}(0, 1), \quad i = 1, \ldots, N$$

As a result,

$$\mathbf{E}(\mathbf{x}^k) = \int x^k \, \mathrm{d}F(x) = \lim_{N \to \infty} \frac{\sum_{i=1}^{N} \mathbf{x}_i^k}{N}$$

and

$$\text{VaR}_q(X) = -Q_q(\mathbf{x}) = -\lim_{N \to \infty} \text{quantile}(\{x_i : i = 1, \ldots, N\}, \text{prob} = q)$$

$$\text{ES}_q(\mathbf{x}) = \frac{1}{q} \int_{-\infty}^{Q_q(x)} x \, \mathrm{d}F(x) = -\lim_{N \to \infty} \frac{\sum_{\mathbf{x}_i : \mathbf{x}_i < Q_q(\mathbf{x})} \mathbf{x}_i}{\#\{\mathbf{x}_i : \mathbf{x}_i < Q_q(\mathbf{x})\}}$$

4 Capital Asset Pricing Model and Factor Models

4.1 Sharpe's Single-Index Model

Suppose there are N assets, denoted by $i = 1, \dots, N$, with returns in period t, $R_{it} = 1, \dots, T$,

$$R_{it} = \alpha_i + \beta_i R_{mt} + \epsilon_{it}, \qquad i = 1, \dots, N, \quad t = 1, \dots, T$$

where α_i, β_i are constant over time, R_{mt} is return on diversified market index, and ϵ_{it} are random error terms (idiosyncratic error) assumed uncorrelated with R_{mt}. We also assume

- $\mathrm{Cov}(R_{mt}, \epsilon_{is}) = 0, \quad \forall i, t, s$ (market returns independent from noise)
- $\mathrm{Cov}(\epsilon_{is}, \epsilon_{jt}) = 0, \quad \forall i \neq j, t, s$ (noise of each stock independent from each other)
- $\epsilon_{it} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2_{\epsilon,i})$ (residuals have mean 0)
- $R_{m,t} \overset{\text{iid}}{\sim} \mathrm{N}(\mu_m, \sigma^2_m)$ (market is random)

By manipulating the numerator,

$$\beta_i = \frac{\mathrm{Cov}(R_{it}, R_{mt})}{\mathrm{Cov}(R_{mt})} = \frac{\sigma_{iM}}{\sigma^2_m}$$

β_i captures the contribution of asset i to the volatility of the market index. Also,

$$\epsilon_{it} = R_{it} - \alpha_i - \beta_i R_{mt}$$

The return of the market, R_{mt} captures common "market-wide" news. β_i measures sensitivity to "market-wide" news. Random error ϵ_{it} captures "firm-specific" news unrelated to market-wide news. Returns are correlated only through their exposure to common market-wide news captured by β_i.

Theorem 4.1 (Statistical Properties of Single-Index Model).

$$R_{it} = \alpha_i + \beta_i R_{mt} + \epsilon_{it}$$

- $\mu_i = \mathbf{E}(R_{it}) = \alpha_i + \beta_i \mu_m$
- $\sigma^2_i = \mathrm{Cov}(R_{it}) = \beta^2_i \sigma^2_m + \sigma^2_{\epsilon,i}$
- $\sigma_{ij} = \mathrm{Cov}(R_{it}, R_{jt}) = \sigma^2_m \beta_i \beta_j$
- $R_{it} \sim \mathrm{N}(\mu_i, \sigma^2_i) = \mathrm{N}(\alpha_i + \beta_i \mu_m, \beta^2_i \sigma^2_m + \sigma^2_{\epsilon,i})$

Remark (Implications).

- $\sigma_{ij} = 0$ if $\beta_i = 0$ or $\beta_j = 0$ (asset i or asset j do not respond to market news)
- $\sigma_{ij} > 0$ if $\beta_i, \beta_j > 0$ or $\beta_i, \beta_j < 0$ (asset i and asset j respond to market news in the same direction)
- $\sigma_{ij} < 0$ if $\beta_i > 0$ and $\beta_j < 0$ or if $\beta_i < 0$ and $\beta_j > 0$ (asset i and asset j respond to market news in the opposite direction)

Theorem 4.2 (Decomposition of Total Variance).

$$R_{it} = \alpha_i + \beta_i R_{mt} + \epsilon_{it}$$

$$\underbrace{\sigma_i^2}_{\textit{total variance}} = \text{Cov}(R_{it}) = \underbrace{\beta_i^2 \sigma_m^2}_{\textit{market variance}} + \underbrace{\sigma_{\epsilon,i}^2}_{\textit{non-market variance}}$$

Dividing both sides by σ_i^2,

$$\begin{aligned} 1 &= \frac{\beta_i^2 \sigma_m^2}{\sigma_i^2} + \frac{\sigma_{\epsilon,i}^2}{\sigma_i^2} \\ &= R_i^2 + 1 - R_i^2 \end{aligned}$$

where $R_i^2 = \frac{\beta_i^2 \sigma_m^2}{\sigma_i^2}$ is the proportion of market variance, and $1 - R_i^2$ is the proportion of non-market variance. Similarly, the covariance matrix can be decomposed as

$$\Sigma = \sigma_m^2 \beta^\top \beta + \text{diag}(\sigma_{\epsilon,i}^2, \dots, \sigma_{\epsilon,n}^2)$$

Remark (Sharpe's Rule of Thumb)**.** A typical stock has $R_i^2 = 30\%$, i.e. proportion of market variance in a typical stock is 30% of total variance.

4.2 Capital Asset Pricing Model

The Capital Asset Pricing Model is

$$\mathbf{E}(R_i) = r_f + \beta_i(\mathbf{E}(R_m) - r_f)$$

or

$$R_i = r_f + \beta_i(R_m - r_f) + \epsilon_i$$

where

- R_m is the return of the market
- $\mathbf{E}(R_m)$ is known as the market premium
- $\mathbf{E}(R_i) - r_f$ is known as the risk premium. The model can be written as

$$\mathbf{E}(R_i) - r_f = \beta_i(\mathbf{E}(R_m) - r_f)$$

- β_i is the sensitivity of expected excess asset returns to the expected excess market returns, or

$$\beta_i = \frac{\text{Cov}(R_i, R_m)}{\text{Cov}(R_m)}$$

- We can also write β_i as

$$\beta_i = \rho_{R_i,R_m} \frac{\sigma_{R_i}}{\sigma_{R_m}}$$

  where ρ_{R_i,R_m} is the correlation coefficient.
- The expected market return is usually estimated by measuring the log-return of the historical returns on a market portfolio (e.g. S&P500).

4.3 Asset Pricing

Assuming that the CAPM is correct, an asset is correctly priced when its estimated price is the same as the present value of future price of the asset, discounted at the rate suggested by CAPM. If the estimated price is higher than the CAPM valuation, then the asset is undervalued, and vice versa.

For any time in the future T, we have return (of T period)

$$R_T = \frac{P_T - P_0}{P_0}, \quad \text{and} \quad \mathbf{E}(R_T) = \frac{\mathbf{E}(P_T) - P_0}{P_0}$$

Here, P_0 is fixed but P_T is random. By CAPM,

$$\mathbf{E}(R_T) = r_f + \frac{\text{Cov}(R_T, R_m)}{\text{Cov}(R_m)}(\mathbf{E}(R_m) - r_f)$$

Suppose we can predict the expected price P_T and R_m at time T. Then we can determine the price P_0.

$$P_0 = \frac{1}{1 + r_f}\left[\mathbf{E}(P_T) - \frac{\text{Cov}(P_T, R_m)}{\text{Cov}(R_m)}(\mathbf{E}(R_m) - r_f)\right]$$

where P_T is the expected price of the asset or portfolio.

4.4 Derivation of CAPM

Suppose there are p risky assets R_i with expected return r_i, risk $\sigma_i, i = 1, 2, \ldots, p$. Let r_f be the risk-free return. The optimal portfolio

$$R_m = \tilde{w}_1 R_1 + \cdots + \tilde{w}_p R_p$$

is the tangency portolio, or at the capital market line, with expected return, risk and covariance as

$$r_m = \mathbf{E}(R_m), \quad \sigma_m^2 = \text{Cov}(R_m), \quad \sigma_{i,m} = \text{Cov}(R_i, R_m)$$

Consider a portfolio

$$R_N = \sum_{i=1}^{p} w_i R_i, \quad \sum_{i=1}^{p} w_i = 1, \quad w_i \geq 0$$

with expected return

$$r_N = \mathbf{E}(R_N) = \sum_{i=1}^{p} w_i \mathbf{E}(R_i) = \sum_{i=1}^{p} w_i \mathbf{E}(R_i) + \underbrace{\left(1 - \sum_{i=1}^{p} w_i\right) r_f}_{=0} \tag{1}$$

We want to minimize the risk

$$\sigma_N^2 = \text{Cov}(R_N) = \sum_{i=1}^{p} w_i^2 \sigma_i^2 + \sum_{i=1}^{p} \sum_{j=1, j\neq i}^{p} w_i w_j \text{Cov}(R_i, R_j)$$

Define function C:

$$C = \sigma_N^1 + \lambda\left[r_N - \sum_{i=1}^{p} w_i \mathbf{E}(R_i) - \left(1 - \sum_{i=1}^{p} w_i\right) r_f\right]$$

where λ is a Lagrangian multiplier due to constraint (1). To minimize the risk of the portfolio,

$$\begin{aligned} \frac{\partial C}{\partial w_i} &= \frac{1}{2}\sigma_N^{-1}\left[2w_i\sigma_i^2 + 2\sum_{j=1,j\neq i}^{p} w_j \mathrm{Cov}(R_1, R_j)\right] - \lambda(\mathbf{E}(R_i) - r_f) = 0 \\ \frac{\partial C}{\partial \lambda} &= r_N - \sum_{i=1}^{p} w_i \mathbf{E}(R_i) - \left(1 - \sum_{i=1}^{p} w_i\right) r_f = 0 \end{aligned} \tag{2}$$

Taking the first p equations and multiplying them by $w_1, w_2, \ldots, w_p$, we obtain

$$\sigma_N^{-1}\left[w_i^2\sigma_i^2 + \sum_{j=1,j\neq i}^{p} w_i w_j \mathrm{Cov}(R_i, R_j)\right] - \lambda w_i(\mathbf{E}(R_i) - r_f) = 0, \qquad i = 1, \ldots, p$$

Summing them up over $i = 1, \ldots, p$, we have

$$\sigma_N^{-1}\left[\sum_{i=1}^{p} w_i^2\sigma_i^2 + \sum_{i=1}^{p}\sum_{j=1,j\neq i}^{p} w_i w_j \mathrm{Cov}(R_i, R_j)\right] - \lambda \sum_{i=1}^{p} w_i(\mathbf{E}(R_i) - r_f) = 0$$

which is equivalent to

$$\sigma_N^{-1}\sigma_N^2 = \lambda \sum_{i=1}^{p} w_i(\mathbf{E}(R_i) - r_f)$$

or

$$\sigma_N = \lambda(r_N - r_f)$$

Note that this is true for all efficient portfolios. If the market is efficient, the market portfolio satisfies the above equations and is a tangency portfolio, i.e.

$$w_i = \tilde{w}_i, i = 1, \ldots, p, \quad \sum_{i=1}^{p} \tilde{w}_i = 1, \quad r_N = r_m$$

Then

$$\sigma_m = \lambda(r_m - r_f)$$

and thus

$$\frac{1}{\lambda} = \frac{r_m - r_f}{\sigma_m} \tag{3}$$

For R_m, $\sigma_N = \sigma_m$. Rewriting the equation (2),

$$\sigma_m^{-1}\left[\tilde{w}_i\sigma_i^2 + \sum_{j=1,j\neq i}^{p} \tilde{w}_j \mathrm{Cov}(R_i, R_j)\right] - \lambda(\mathbf{E}(R_i) - r_f) = 0$$

Solving, we have

$$\begin{aligned} \mathbf{E}(R_i) &= r_f + \frac{1}{\lambda\sigma_m}\left[\tilde{w}_i\sigma_i^2 + \sum_{j=1,j\neq i}^{p} \tilde{w}_j \mathrm{Cov}(R_i, R_j)\right] \\ &= r_f + \frac{1}{\lambda\sigma_m}\mathrm{Cov}\left(R_i, \sum_{j=1}^{p} \tilde{w}_j R_j\right) \\ &= r_f + \frac{1}{\lambda\sigma_m}\mathrm{Cov}(R_i, R_m) \end{aligned}$$

Substituting λ with (3), we obtain

$$\mathbf{E}(R_i) = r_f + \frac{r_m - r_f}{\sigma_m^2}\text{Cov}(R_i, R_m)$$

Thus

$$\mathbf{E}(R_i) = r_f + \underbrace{\frac{\text{Cov}(R_i, R_m)}{\sigma_m^2}}_{\beta_i}(r_m - r_f)$$

4.5 Multi-Factor Models

Suppose that asset return R_i is driven by a K common factors $f_1, \dots, f_K$ and idiosyncratic noise u_i. A multi-factor regression model is

$$R_i = \alpha_i + \sum_{j=1}^{K} \beta_{ij} f_j + u_i, \qquad i = 1, 2, \dots, n$$

for n assets, where

- α_i is the regression intercept for return of asset i.
- $f_1, \dots, f_K$ are common factors driving all asset returns, with $\text{Cov}(f_i) = \lambda_i$ and $\text{Cov}(f_i, f_j) = 0$ for $i \neq j$.
- β_{ij} gives how sensitive the return of asset i with respect to the j-th factor, which is called the factor loading of asset i on factor f_j.
- u_i is the idiosyncratic component in asset i's return that is unrelated to other asset returns

$$\mathbf{E}(u_i) = 0, \quad \text{Cov}(u_i) = \sigma_{u_i}^2, \quad \text{Cov}(u_i, u_j) = 0 \text{ for } i \neq j$$

- u_i is independent of $f_1, \dots, f_K$.

Theorem 4.3. *If the market is efficient, and the factors are indeed complete, then the model can be further written as*

$$R_i - r_f = \sum_{j=1}^{K} \beta_{ij} f'_j + u_i$$

where $f'_j = f_j - r_f$ or $f'_j = f_j$ depending on whether f_j is an asset or difference of two assets.

Theorem 4.4 (Expectation, Variance and Covariance of R_i)**.**

$$\mathbf{E}(R_i) - r_f = \sum_{j=1}^{K} \beta_{ij}\mathbf{E}(f'_j)$$

$$\text{Cov}(R_i) = \sum_{j=1}^{K} \beta_{ij}^2 \text{Cov}(f_j) + \sigma_{u_i}^2$$

$$\text{Cov}(R_i, R_j) = \sum_{k=1} \beta_{ik}\beta_{jk}\text{Cov}(f_k)$$

Definition 4.1 (Fama-French Three Factor Model)**.**

$$R_{i,t} = r_{f,t} + \beta_{i,m}(R_{t,m} - r_{f,t}) + \beta_{i,s}\mathrm{SMB}_t + \beta_{i,v}\mathrm{HML}_t + \alpha_i + \epsilon_{i,t}$$

where

- $R_{i,t}$ is the stock return during period t.
- $r_{f,t}$ is the return on the risk-free asset.
- α_i is zero under some assumptions.
- β's are the betas for each factor.
- $\epsilon_{i,t}$ is regression error.
- SMB is the return of small capital firms minus that of big capital firms.
- HML is the return of high (book value)/(market value) ratio minus that of low.

□

Remark (The Fourth Factor)**.** The fourth factor (proposed by Carhart, 1997) is momentum: the difference between returns on diversified portfolios of stocks that perform well and poorly in the short term. Momentum is short-lived and therefore not useful to estimate the cost of capital although it does explain stock returns. However, momentum is useful in trading (momentum investing), where large increase in the price of a security will be followed by additional gains and vice versa for declining values (in a short term). Statistical time series models are useful for momentum.

Remark (Selecting Factors)**.**

- Macroeconomic factors: treat factors as observable and specify the f_j directly. Commonly used macroeconomic factors are market portfolio, inflation, interest rates, credit spreads, and business cycle variables. We can estimate the loadings via regression if these variables will be sufficient to capture the systemic risk in the economy.
- Factors based on characteristics: construct indices of some firm characteristics (e.g. B/M), and treat these risk indices as sensitivities to the factors associated with those characteristics (such as a B/M factor). An example is the Fama-French Three Factor Model.
- Statistical factors: principal component analysis. We treat both the factors and the loadings as unobservable.

Definition 4.2 (Arbitrage Pricing Model)**.** The Arbitrage Pricing Theory states that if asset returns follow a factor structure, then the following relation exists between expected returns and the factor sensitivities.

$$\mathbf{E}(r_j) = r_f + \beta_{j1}\mathrm{RP}_1 + \beta_{j2}\mathrm{RP}_2 + \cdots + \beta_{jn}\mathrm{RP}_K$$

where RP_k is the risk premium of the factor, usually $\mathrm{RP}_k = \mathbf{E}(f_k - r_f)$ or $\mathrm{RP}_k = \mathbf{E}(f_k)$. □

4.6 Security Market Line

The CAPM can be rearranged and expressed in terms of the the security market line (SML).

$$\mathbf{E}(R_i) = r_f + (\mathbf{E}(R_m) - r_f)\beta_i, \quad i = 1, \ldots, n$$

Note that for different asset i, $\mathbf{E}(R_m) - r$ is fixed. Here, $\mathbf{E}(R_i)$ is the expected return, β_i is the sensitivity (smaller sensitivity, smaller risk; bigger sensitivity, bigger return), and $r_f + (\mathbf{E}(R_m) - r_f)\beta_i$ is the required rate of return (minimum return for such sensitivity).

Since r_f and $\mathbf{E}(R_m) - r$ are fixed for all i assets, we consider regression of

$$\mathbf{E}(R_i) = c_0 + c_1\beta_i$$

or

$$R_i = c_0 + c_1\beta_i + \epsilon_i$$

and estimate c_0 and c_1. c_1 is re-estimator of the market premium.

If CAPM is correct, then all securities should lie on the SML. The SML can be used to pick underpriced or overpriced assets.

4.7 Principal Component Analysis

Given p-dimension random variable $X = (\mathbf{x}_1, \dots, \mathbf{x}_p)^\top$ with $\mathbf{E}(X) = \mu$, $\Sigma = \mathrm{Cov}(X)$, consider the eigendecomposition of Σ:

$$\Sigma = \begin{pmatrix} w_{11} & w_{21} & \cdots & w_{p1} \\ w_{12} & w_{22} & \cdots & w_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1p} & w_{2p} & \cdots & w_{pp} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \begin{pmatrix} w_{11} & w_{21} & \cdots & w_{p1} \\ w_{12} & w_{22} & \cdots & w_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1p} & w_{2p} & \cdots & w_{pp} \end{pmatrix}^\top$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

$$W = \begin{pmatrix} w_{11} & w_{21} & \cdots & w_{p1} \\ w_{12} & w_{22} & \cdots & w_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1p} & w_{2p} & \cdots & w_{pp} \end{pmatrix} = (\mathbf{w}_1, \dots, \mathbf{w}_p)$$

is an orthogonal matrix, i.e. $WW^\top = W^\top W = I_p$, and $\mathbf{w}_i^\top \mathbf{w}_j = 1$ if $i = j$, 0 otherwise. $\lambda_1, \dots, \lambda_p$ are the p eigenvalues and $\mathbf{w}_1, \dots, \mathbf{w}_p$ are the p eigenvectors, so we have

$$\Sigma \mathbf{w}_i = \lambda_i \mathbf{w}_i, \quad i = 1, \dots, p$$

We can write

$$\Sigma = \lambda_1 \mathbf{w}_1 \mathbf{w}_1^\top + \cdots + \lambda_p \mathbf{w}_p \mathbf{w}_p^\top \quad \text{and} \quad \mathbf{w}_i^\top \Sigma \mathbf{w}_i = \lambda_i, \quad i = 1, \dots, p$$

We call $\mathbf{z}_i = \mathbf{w}_i^\top X$ the i-th principal component (PC), where

$$\mathrm{Cov}(\mathbf{z}_i) = \lambda_i \quad \text{and} \quad \mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j) = 0 \text{ if } i \neq j$$

The first PC of X is the linear combination $\mathbf{z}_1 = \mathbf{w}_1^\top X$ that maximizes $\mathrm{Cov}(\mathbf{z}_1)$, subject to the constraint $\mathbf{w}_1^\top \mathbf{w}_1 = \sum_{j=1}^p w_{1j}^2 = 1$, such that $\mathbf{z}_1$ contains the most information (or to make $\mathbf{z}_1$ explain the largest variation in X).

The second PC $\mathbf{z}_2 = \mathbf{w}_2^\top X$ maximizes $\mathrm{Cov}(\mathbf{z}_2)$ subject to the constraints $\mathbf{w}_2^\top \mathbf{w}_2 = 1$ and $\mathrm{Cov}(\mathbf{z}_2, \mathbf{z}_1) = 0$. In general, the i-th PC $\mathbf{z}_i = \mathbf{w}_i^\top X$ maximizes $\mathrm{Cov}(\mathbf{z}_i)$ subject to the constraints $\mathbf{w}_i^\top \mathbf{w}_i = 1$ and $\mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j) = 0$ for $j = 1, \dots, i-1$.

We can write $(\mathbf{z}_1, \dots, \mathbf{z}_p)^\top = (\mathbf{w}_1, \dots, \mathbf{w}_p)^\top X$, and thus

$$(\mathbf{x}_1, \dots, \mathbf{x}_p)^\top = (\mathbf{w}_1, \dots, \mathbf{w}_p)(\mathbf{z}_1, \dots, \mathbf{z}_p)^\top$$

and w_{ji} are loadings.

For the total variance $\sum_{i=1}^p \mathrm{Cov}(\mathbf{x}_i)$, we have

$$\begin{aligned}\sum_{i=1}^p \mathrm{Cov}(\mathbf{x}_i) = \mathrm{tr}(\Sigma) &= \mathrm{tr}\left(W \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} W^\top \right) \\ &= \mathrm{tr}\left(\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} W^\top W \right) = \mathrm{tr}\left(\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \right) \\ &= \sum_{i=1}^p \lambda_i\end{aligned}$$

$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}$ is the proportion of total variance in X explained by the i-th PC. The cumulative proportion of total variance explain by the first i PCs is

$$\frac{\sum_{j=1}^i \lambda_j}{\sum_{j=1}^p \lambda_j}$$

Suppose we define a common factor $\mathbf{z} = w^\top X$ and use it to predict/explain all variables $\mathbf{x}_1, \dots, \mathbf{x}_p$ by model

$$\mathbf{x}_i = \alpha_i + \beta_i \mathbf{z} + \epsilon_i$$

where $\mathbf{E}(\epsilon_i) = 0$ and $\mathrm{Cov}(\mathbf{z}, \epsilon_i) = 0$. The prediction error for this stock i is defined as

$$\epsilon_i = \mathbf{x}_i - \alpha_i - \beta_i \mathbf{z}$$

The first PC has the highest prediction ability. In other words, if we minimize

$$\min_w \mathbf{E}\left(\sum_{i=1}^p \epsilon_i^2 \right)$$

then the solution is $w = \mathbf{w}_1$ and the common factor is $\mathbf{z} = \mathbf{w}_1^\top X$, where $\mathbf{w}_1$ is the first PC.

We interpret the factors $\mathbf{z}_1, \mathbf{z}_2, \dots$ by two methods:

- Check the coefficients (loadings) w_{ij}'s in $\mathbf{z}_1, \mathbf{z}_2, \dots$
- Calculate the correlation coefficient between $\mathbf{z}_1, \mathbf{z}_2, \dots$ with a known variable. If the correlation coefficient is very big, we can conclude that the variable is a factor.

For empirical PCA, we simply replace Σ by

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top, \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

4.8 Statistical Factor Models

For the first principal components,

$$\begin{pmatrix}\mathbf{z}_1\\ \mathbf{z}_2\\ \vdots\\ \mathbf{z}_p\end{pmatrix} = \begin{pmatrix}\mathbf{w}_1^\top X\\ \mathbf{w}_2^\top X\\ \vdots\\ \mathbf{w}_p^\top X\end{pmatrix} = (\mathbf{w}_1,\dots,\mathbf{w}_p)^\top X = (\mathbf{w}_1,\dots,\mathbf{w}_p)^\top \begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_p\end{pmatrix}$$

Therefore

$$\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_p\end{pmatrix} = (\mathbf{w}_1,\dots,\mathbf{w}_p) \begin{pmatrix}\mathbf{z}_1\\ \mathbf{z}_2\\ \vdots\\ \mathbf{z}_p\end{pmatrix}$$

or

$$\begin{aligned}
\mathbf{x}_1 &= w_{11}\mathbf{z}_1 + w_{21}\mathbf{z}_2 + \cdots + w_{K1}\mathbf{z}_K + w_{K+1,1}\mathbf{z}_{K+1} + \cdots + w_{p1}\mathbf{z}_p\\
\mathbf{x}_2 &= w_{12}\mathbf{z}_1 + w_{22}\mathbf{z}_2 + \cdots + w_{K2}\mathbf{z}_K + w_{K+1,2}\mathbf{z}_{K+1} + \cdots + w_{p2}\mathbf{z}_p\\
&\vdots\\
\mathbf{x}_p &= w_{1p}\mathbf{z}_1 + w_{2p}\mathbf{z}_2 + \cdots + w_{Kp}\mathbf{z}_K + w_{K+1,p}\mathbf{z}_{K+1} + \cdots + w_{pp}\mathbf{z}_p
\end{aligned}$$

If $\lambda_k = 0$ for all $k \geq K+1$, then $\mathbf{z}_k = 0$ or μ_k (constant). The above model becomes

$$\mathbf{x}_i = w_{1i}\mathbf{z}_1 + w_{2i}\mathbf{z}_2 + \cdots + w_{Ki}\mathbf{z}_K, \quad i = 1,\dots,p$$

If not, we have

$$\mathrm{Cov}(\mathbf{x}_i) = w_{1i}^2\lambda_1 + w_{2i}^2\lambda_2 + \cdots + w_{Ki}^2\lambda_K + \underbrace{w_{K+1,i}^2\lambda_{K+1} + \cdots + w_{pi}^2\lambda_p}_{\text{small when } K \text{ is large}}$$

By summing over $i = 1,\dots,p$, we have

$$\begin{aligned}
\mathrm{Cov}(\mathbf{x}_1) + \cdots + \mathrm{Cov}(\mathbf{x}_p) &= \sum_{i=1}^p w_{1i}^2\lambda_i + \sum_{i=1}^p w_{2i}^2\lambda_2 + \cdots + \sum_{i=1}^p w_{Ki}^2\lambda_K\\
&\quad + \underbrace{\sum_{i=1}^p w_{K+1,i}^2\lambda_{K+1} + \cdots + \sum_{i=1}^p w_{pi}^2\lambda_p}_{\text{small}}\\
&= \lambda_1 + \cdots + \lambda_K + \underbrace{\lambda_{K+1} + \cdots + \lambda_p}_{\text{small}}
\end{aligned}$$

The last few terms are very small if $\frac{\lambda_1+\cdots+\lambda_K}{\lambda_1+\cdots+\lambda_p} \approx 1$. Thus we can write

$$\mathbf{x}_i = w_{1i}\mathbf{z}_1 + w_{2i}\mathbf{z}_2 + \cdots + w_{Ki}\mathbf{z}_K + u_i, \quad i = 1,\dots,p$$

where $u_i = w_{K+1,i}\mathbf{z}_{K+1} + \cdots + w_{pi}\mathbf{z}_p$ with $\mathrm{Cov}(u_i)$ being small. This is the factor model, and $\mathbf{z}_1,\dots,\mathbf{z}_K$ are factors.

5 Financial Time Series

5.1 Theoretical Definitions

Definition 5.1 (Time Series)**.** A time series $\{Z_t\}$ is a sequence of random variables indexed by time t.

$$\{\ldots, Z_1, Z_2, \ldots, Z_{t-1}, Z_t, Z_{t+1}, \ldots\}$$

A realization of a stochastic process (or time series) is the sequence of observed data

$$\{\ldots, Z_1 = z_1, Z_2 = z_2, \ldots, Z_{t-1} = z_{t-1}, Z_t = z_t, Z_{t+1} = z_{t+1}, \ldots\}$$

Hereafter, we use $\{z_t\}$ for both theoretical time series and its observations. □

Definition 5.2 (Mean, Variance, Covariance Functions)**.** Let $\{z_t\}$ be a time series (a sequence of random variables) with $\mathbf{E}(z_t^2) < \infty$. The mean function of $\{z_t\}$ is

$$\mu_t = \mathbf{E}(z_t) \qquad \text{or denoted } \mu_z(t)$$

The variance function of $\{z_t\}$ is

$$\sigma_t^2 = \mathrm{Cov}(z_t) = \mathbf{E}(z_t - \mu_t)^2$$

The covariance function of $\{z_t\}$ is

$$\gamma(r, s) = \mathrm{Cov}(z_r, z_s) = \mathbf{E}[(z_r - \mu_r)(z_s - \mu_s)] \qquad \text{or denoted } \gamma_z(r, s)$$

□

Remark.

One important assumption behind time series analysis is that "history repeats itself". Patterns must repeatedly appear so that

- We can observe the pattern
- We can make inferences about the pattern
- We can make predictions of the time series based on the pattern

Stationarity is defined and assumed in most time series analyses.

Definition 5.3 (Strictly Stationary)**.** $\{z_t\}$ is strictly stationary if for any given finite integer k, and for any subset of subscripts $t_1, t_2, \ldots, t_k$, the joint distribution of $z_{t_1}, z_{t_2}, \ldots, z_{t_k}$ depends only on $t_1 - t_2, t_2 - t_3, \ldots, t_{k-1} - t_k$ but not directly on $t_1, \ldots, t_k$. □

Remark.

- If $\{Z_t\}$ is stationary, then (Z_6, Z_9, Z_20) and (Z_{21}, Z_{24}, Z_{25}) have the same distribution.
- If $\{Z_t\}$ is an IID sequence, then it is strictly stationary.
- Let $\{Z_t\}$ be an IID sequence and X independent of $\{Z_t\}$. Let $Y_t = Z_t + X$. Then the sequence $\{Y_i\}$ is also stationary.
- For any function g, $\{g(z_t)\}$ is also strictly stationary.

Definition 5.4 (Weakly Stationary)**.** $\{z_t\}$ is weakly stationary (or covariance stationary) if

- μ_t (denoted by μ) does not depend on t (the mean of random variables at different time points do not depend on t).
- $\gamma(r, s) = \gamma(|r - s|)$ (the covariance of any two random variables depends only on the time difference of their observations).

□

Remark. If $\{Z_t\}$ is strictly stationary and $\text{Cov}(Z_t) < \infty$, then $\{Z_t\}$ is also weakly stationary. However, a weakly stationary time series is usually not strictly stationary.

Definition 5.5 (Autocovariance and Autocorrelation)**.** If $\{z_t\}$ is stationary, then

- $\gamma(h) = \text{Cov}(z_{t+h}, z_t)$ is called the autocovariance function (ACVF) of $\{z_t\}$.
- $$\rho_z(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\text{Cov}(z_{t+h})}{\text{Cov}(z - t)} = \text{Cor}(z_{t+h}, z_t)$$

  is called the autocorrelation function (ACF) at lag h of $\{z_t\}$.

□

Remark (Basic Properties of $\gamma(\cdot)$)**.**

- $\gamma(0) \geq 0$
- $|\gamma(h)| \leq \gamma(0)$ for all h
- $\gamma(h) = \gamma(-h)$

Definition 5.6 (White Noise)**.** If $\{z_t\}$ is stationary with $\mathbf{E}(z_t) = 0$, $\gamma(0) = \sigma^2$, and

$$\gamma(h) = 0, \quad \forall h \neq 0$$

we call $\{z_t\}$ a white noise sequence, denoted by $z_t \sim \text{WN}(0, \sigma_z^2)$. □

Example 5.1. Suppose $\{y_t\}$ is a stationary time series, and $\{\epsilon_t\} \sim \text{WN}(0, \sigma^2)$. If they are independent (which means $\mathbf{E}(y_t \epsilon_{t+k}) = \mathbf{E}(y_t)\mathbf{E}(\epsilon_{t+k})$ for all t and k), then

$$z_t = \{a + by_t^2\}^{\frac{1}{2}} \epsilon_t, \quad a \geq 0, b \geq 0$$

is also a white noise. □

Example 5.2. Suppose $\{\epsilon_t\}$ is a sequence of IID $N(0, \sigma^2)$ (thus it is a white noise). Let

$$z_t = \{a + bz_{t-1}^2\}^{\frac{1}{2}} \epsilon_t, \quad a \geq 0, b \geq 0$$

Then $\{z_t\}$ is also a white noise. This is an ARCH model. □

Remark. We can build stationary time series from white noise sequences.

Theorem 5.1 (Wold's Decomposition Theorem)**.** *Any weakly stationary time series $\{Y_t\}$ can be represented in the form*

$$Y_t = \mu + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots = \mu + \sum_{k=0}^{\infty} \phi_k \epsilon_{t-k}$$

where $\phi_0 = 1$, $\sum_{k=1}^{\infty} \phi_k^2 < \infty$ and $\{\epsilon_t\} \sim \text{WN}(0, \sigma^2)$.

Remark (Properties)**.**

- $\mathbf{E}(Y_t) = \mu$
- $\gamma(0) = \text{Cov}(Y_t) = \sigma^2 \sum_{k=0}^{\infty} \phi_k^2$
- When $h > 1$, $\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \phi_j \phi_{h+j}$

Remark. In finance data, stationary time series can usually be obtained from non-stationary time series by considering the first-difference

$$z_t = x_t - x_{t-1}$$

If z_t is stationary, then x_t is integrated of order 1. We can also consider any k-th difference.

5.2 Sample ACVF and ACF

Definition 5.7 (Sample Mean, Sample Variance, SACVF, SACF)**.** Suppose we have observations $z_1, z_2, \ldots, z_n$. Suppose that $\{z_t\}$ is stationary, then we have to estimate μ_z, $\gamma(h)$ and $\rho_z(h)$ with $h = 0, 1, 2, \ldots$.

- Sample mean: $\bar{z} = \frac{1}{n} \sum_{t=1}^{n} z_t$
- Sample variance: $\hat{\sigma}_z^2 = \frac{1}{n} \sum_{t=1}^{n} (z_t - \bar{z})^2 = \hat{\gamma}(0)$
- Sample autocovariance function at lag h (SACVF):

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (z_t - \bar{z})(z_{t+h} - \bar{z})$$

- Sample autocorrelation coefficient function at lag h (SACF):

$$\hat{\rho}(h)(\text{or } r_h) = \frac{\sum_{t=1}^{n-h} (z_t - \bar{z})(z_{t+h} - \bar{z})}{\sum_{t=1}^{n} (z_t - \bar{z})^2} = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

□

Theorem 5.2 (Testing for $\rho(h) = 0, h \geq 1$)**.** *Suppose that $z_1, \ldots, z_n$ are realizations from a stationary time series. The SACF at lag h is*

$$r_h = \frac{\sum_{t=1}^{n-h} (z_t - \bar{z})(z_{t+h} - \bar{z})}{\sum_{t=1}^{n} (z_t - \bar{z})^2} \qquad \textit{where } \bar{z} = \frac{1}{n} \sum_{t=1}^{n} z_t$$

Under $H_0' : \rho(h) = 0$ for all $h > 0$,

$$s_{r_h} = \left(\frac{1}{n}\right)^{\frac{1}{2}}$$

If $|r_h| < 2s_{r_h}$, do not reject H_0, otherwise reject H_0.

Theorem 5.3 (Test for White Noise)**.** *The Ljung-Box test is a statistical test for whether a number of autocorrelations of a time series are different from 0.*

$$H_0 : \rho(k) = 0, k = 1, \ldots, h$$

Suppose the time series is $\{y_1, y_2, \ldots, y_n\}$ with sample autocorrelation coefficients $r(k)$. The test statistic is

$$Q(h) = n(n+2)\sum_{k=1}^{h} \frac{\{r(k)\}^2}{n-k}$$

where n is the sample size, $r(k)$ is the sample autocorrelation at lag k, and h is the number of lags being tested. Under H_0,

$$Q(h) \sim \chi^2(h)$$

For significance level α, the critical region for rejection of the hypothesis of randomness is rejected if the calculated value (Ljung-Box statistic)

$$Q^*(h) > \chi^2_{1-\alpha}(h)$$

where $\chi^2_{1-\alpha}(h)$ is the $(1-\alpha)$-quantile of the χ^2 distribution with h degrees of freedom. The bigger the probability $\chi^2(h) > Q^*(h)$ is, the smaller $Q^*(h)$ is, the more likely H_0 is correct.

$$p\text{-}value = P(\chi^2(h) > Q^*(h))$$

We reject H_0 is p-value $< \alpha$.

5.3 Moving Average Model

For any integer $q \geq 1$, the moving average model of order q (MA(q)) is

$$x_t = \mu + a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}$$

More generally, let

$$x_t = \mu + \sum_{j=0}^{\infty} \psi_j a_{t-j}$$

where $\psi_0 = 1$ and $\{a_t\} \sim \mathrm{WN}(0, \sigma^2)$ and

$$\sum_{j=0}^{\infty} |\psi_j| < \infty \quad (\text{or } \sum_{j=0}^{\infty} \psi_j^2 < \infty)$$

$\{x_t\}$ is called a stationary general linear process or MA(∞) process. For MA(q), we have

- $\mathbf{E}(x_t) = \mu$
- $\mathrm{Cov}(x_t) = (1 + \theta_1^2 + \cdots + \theta_q^2)\sigma^2$
- If $|h| \leq q$,

$$\mathrm{Cov}(x_t, x_{t-h}) = \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|}$$

- $\mathrm{Cov}(x_t, x_{t-h}) = 0$ if $|h| > q$

5.4 Autoregressive Model

The autoregressive model of order p (or AR(p)) is

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t$$

where $a_t \sim \mathrm{WN}(0, \sigma^2)$. It follows that

$$\mathbf{E}(x_t) = \phi_0 + \phi_1 \mathbf{E}(x_{t-1}) + \cdots + \phi_p \mathbf{E}(x_{t-p})$$

If it is stationary, let $\mu = \mathbf{E}(x_t)$. We have

$$\begin{aligned}(1 - \phi_1 - \cdots - \phi_p)\mu &= \phi_0 \\ \mu &= \frac{\phi_0}{1 - \phi_1 - \ldots - \phi_p}\end{aligned}$$

So AR(p) can be written as

$$x_t = (1 - \phi_1 - \cdots - \phi_p)\mu + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t$$

which is equivalent to

$$(x_t - \mu) = \phi_1(x_{t-1} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + a_t$$

Let $y_t = x_t - \mu$, we have

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + a_t$$

Therefore, in theoretical analyses, we only need to consider the case $\mu = 0$.

Consider AR(1), where $x_t = \phi_1 x_{t-1} + a_t$. If $|\phi_1| < 1$, then

$$x_t = a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + \cdots = \sum_{k=0}^{\infty} \phi_1^k a_{t-k}$$

which is a general linear process with

$$\sum_{k=0}^{\infty} |\phi_1^m| < \infty$$

Therefore,

- If $|\phi_1| < 1$, AR(1) is stationary.
- If $|\phi_1| > 1$, the time series will diverge to ∞.
- If $|\phi_1| = 1$, the time series will stay in a reasonable region, and $x_t = x_{t-1} + a_t$, which is also not stationary, but is a random walk.

Definition 5.8 (Backshift Operator)**.** For any time series y_t, define

$$Ly_t = y_{t-1}, \quad L^k y_t = y_{t-k}, \quad k \geq 1$$

For example, $Ly_{20} = y_{19}$, and $L^{12} y_{20} = y_8$. □

Remark.

- A general AR(p) model
$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + a_t$$
can be written as
$$\phi_p(L) y_t = a_t$$
where $\phi_p(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$.
- A general MA(q) model
$$y_j = a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}$$
can be written as
$$y_t = \theta_q(L) a_t$$
where $\theta_q(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$.
- Another case is $(1-L)X_t = X_t - X_{t-1}$. This is the difference operator. We can also consider higher order difference $(1-L)^d X_t$.

The AR(1) model is written as $(1 - \phi_1 L)x_t = a_t$. We have proved that, if $|\phi_1| < 1$, then x_t is stationary. Equivalently, if the root of

$$1 - \phi_1 L = 0$$

is outside of the unit circle $|z|^2 = 1$ (where z is a complex number), i.e. $|L| = |1/\phi_1| > 1$, then the AR(1) model is stationary.

For the AR(p) model, if all the roots of $\phi_p(L) = 0$ are outside the unit circle, then the time series is stationary.

Generally, if AR(p)
$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + a_t$$
is stationary, then it can be written as

$$y_t = \tilde{\mu} + a_t + \varphi_1 a_{t-1} + \varphi_2 a_{t-2} + \cdots \quad \text{where } \sum_{k=1}^{\infty} |\varphi_k| < \infty$$

Thus,

- $\text{Cov}(y_t, a_t) = \sigma^2$ and $\text{Cov}(y_t, a_{t+k}) = 0$ for any $h > 0$
- $\text{Cov}(y_t, y_{t-h}) = \phi_1 \text{Cov}(y_{t-1}, y_{t-h}) + \cdots + \phi_p \text{Cov}(y_{t-p}, y_{t-h})$ when $h > 0$
- $\text{Cov}(y_t, y_t) = \phi_1 \text{Cov}(y_{t-1}, y_t) + \cdots + \phi_p \text{Cov}(y_{t-1}, y_t) + \sigma^2$

Definition 5.9 (ARMA, ARIMA)**.** The ARMA(p, q) model is

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t + \psi_1 a_{t-1} + \cdots + \psi_q a_{t-q}$$

Again, by letting $\mu = \mathbf{E}(X_t)$, the model can be written as

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + a_t + \psi_1 a_{t-1} + \cdots + \psi_q a_{t-q}$$

Sometimes we consider $Z_t = (1-L)^d X_t$. If Z_t follows ARMA(p, q), then we say X_t follows ARIMA(p,d,q). □

5.5 Parameter Estimation and Order Determination

The ARMA(p, q) model has parameter

$$\theta = (\mu, \phi_1, \dots, \phi_p, \psi_1, \dots, \psi_q, \sigma^2)$$

If p, q are known, we can use maximum log-likelihood. If p, q are unknown, we choose p, q with lowest AIC or BIC:

$$\begin{aligned} \text{AIC}(D) &= -2\ell(\hat{\theta}) + 2D \\ \text{BIC}(D) &= -2\ell(\hat{\theta}) + D \log n \end{aligned}$$

where n is the sample size, $D = p + q + 1$, and p, q and 1 are the numbers of parameters in AR, MA and constant μ respectively.

Usually, we check the residuals (innovation). If the model is appropriate, the innovation should be white noise, which means there is no information left in the innovations (for linear prediction). The residuals are

$$\hat{a}_t = x_t - \{\hat{\mu} + \hat{\phi}_1(x_{t-1} - \hat{\mu}) + \cdots + \hat{\phi}_p(x_{t-p} - \hat{\mu}) + \hat{\psi}_1\hat{a}_{t-1} + \cdots + \hat{\psi}_q\hat{a}_{t-q}\}$$

And we can use Ljung-Box test.

5.6 Prediction of Stationary Process

Let $\{y_t\}$ be a stationary process, e.g. an ARMA(p, q) with Wold representation

$$y_t = \mu + a_t + \phi_1 a_{t-1} + \phi_2 a_{t-2} + \cdots$$

For prediction without any information, the prediction of y_{t+1} is

$$\mathbf{E}(y_{t+1}) = \mu + \mathbf{E}(a_{t+1}) + \phi_1\mathbf{E}(a_{t+1-1}) + \phi_2\mathbf{E}(a_{t+1-2}) + \cdots = \mu$$

For AR(p) model,

$$\begin{aligned} \mathbf{E}(y_{t+1}) &= \phi_0 + \phi_1\mathbf{E}(y_{t+1-1}) + \cdots + \phi_2\mathbf{E}(y_{t+1-2}) + \cdots + \phi_p\mathbf{E}(y_{t+1-p}) + \mathbf{E}(a_{t+1}) \\ &= \phi_0 + \phi_1\mu + \cdots + \phi_p\mu = \mu \end{aligned}$$

For prediction with information up to time t, let $I_t = \{y_t, y_{t-1}, \dots\}$ denote the information available up to time t. Our purpose is to predict $y_{t+h}, h \geq 1$ based on I_t. Suppose we have observations $y_1, \dots, y_T$ and thus $a_T, a_{T-1}, a_{T-2}, \dots$. A linear predictor of $y_{T+h|T}$ is a linear function of the variables in $I_T = \{a_T, a_{T-1}, a_{T-2}, \dots\}$.

$$\begin{aligned} &\mathbf{E}(y_{T+h|t} | a_T, a_{T-1}, \dots) \\ &= \mu + \mathbf{E}(a_{T+h}) + \phi_1\mathbf{E}(a_{T+h-1}) + \cdots + \phi_{h-1}\mathbf{E}(a_{T+h-(h-1)}) + \phi_h\mathbf{E}(a_T) + \phi_{h+1}\mathbf{E}(a_{T-1}) + \cdots \\ &= \mu + \phi_k a_T + \phi_{h+1}a_{T-1} + \cdots \end{aligned}$$

Definition 5.10 (Forecast Error). Define $y_{t+h|t}$ as the forecast of y_{t+h} based on I_t. The forecast error is

$$e_{t+h|t} = y_{t+h} - y_{t+h|t}$$

The mean squared error (MSE) of the forecast is

$$\mathrm{MSE}(e_{t+h|t}) = \mathbf{E}[(y_{t+h} - y_{t+h|t})^2]$$

The forecast error of the linear predictor is

$$\begin{aligned} e_{t+h|t} &= y_{t+h} - y_{t+h|t} \\ &= a_{t+h} + \phi_1 a_{t+h-1} + \cdots + \phi_{h-1} a_{t+1} \end{aligned}$$

The MSE of the forecast error is

$$\mathrm{MSE}(e_{t+h|t}) = \sigma^2(1 + \phi_1^2 + \cdots + \phi_{h-1}^2)$$

If $\{a_t\}$ is Gaussian, then

$$y_{t+h|t} \sim \mathrm{N}\big(y_{t+h}, \sigma^2(1 + \phi_1^2 + \cdots + \phi_{h-1}^2)\big)$$

A 95% confidence interval for the h-step ahead prediction has the form

$$y_{t+h|t} \pm 1.96\sqrt{\sigma^2(1 + \phi_1^2 + \cdots + \phi_{h-1}^2)}$$

The standard error of prediction is

$$\sqrt{\sigma^2(1 + \phi_1^2 + \cdots + \phi_{h-1}^2)}$$

□

6 Conditional Heteroscedastic Models

For time series $\{y_t\}$, denote information up to time t by $I_t = \sigma\{y_t, y_{t-1}, y_{t-2}, \dots\}$. The conditional mean $\mu_t = \mathbf{E}(y_t|I_{t-1})$ can be used to predict y_t based on past information I_{t-1}. Note that the autocorrelation coefficients in y_t are almost zero, so prediction is almost impossible. However, we can still calculate conditional variance (volatility),

$$\sigma_t^2 = \mathrm{Cov}(y_t|I_{t-1}) = \mathbf{E}[(y_t - \mu_t)^2|I_{t-1}]$$

from high-frequency data, implied volatility, or from weighted-average.

Remark (Importance of σ_t)**.**

- Options (derivatives) pricing e.g. Black-Scholes formula
- Risk management (e.g. VaR)
- Asset allocation (e.g. minimum-variance portfolio)
- Interval forecasts
- Explanation of special patterns in financial data (e.g. heavy tails, volatility clustering, co-movement)

Theorem 6.1 (Weighted Average Approach)**.** *Suppose the returns u_i have mean 0, otherwise consider $u_i - \mathbf{E}(u_i)$. We estimate the variance by*

$$\sigma_t^2 = \sum_{i=1}^{m} w_i u_t^2 \text{ (for prediction)}, \qquad \text{or} \qquad \sigma_t^2 = \sum_{i=1}^{m-1} w_i u_{t-1}^2 \text{ (for fitting)}$$

where $w_1 + w_2 + \cdots + w_m = 1$. Usually, w_i decreases with i. m can be infinity providing $w_1 + w_2 + \ldots = 1$, a special case is

$$w_k = (1-\lambda)\lambda^{k-1}, \quad k = 1, \dots$$
$$\sigma_t^2 = (1-\lambda)\sum_{i=1}^{\infty} \lambda^{i-1} u_{t-1}^2, \quad m = \infty$$

This leads to the recurrence formula

$$\sigma_t^2 = (1-\lambda)u_{t-1}^2 + \lambda\sigma_{t-1}^2$$

where popular choices of λ are close to 1.

Remark. In the above model, the sum of coefficients σ_{t-1}^2 and u_{t-1}^2 are respectively λ and $1-\lambda$. We can make it more general, which leads to the GARCH model

$$\sigma_t = \omega + \alpha u_{t-1}^2 + \beta\sigma_{t-1}^2$$

with $\omega \geq 0, \alpha \geq 0, \beta \geq 0$ but $\alpha + \beta \leq 1$.

Theorem 6.2 (Econometric Modeling of Conditional Mean and Conditional Variance)**.** *For return r_t, denote the conditional mean by u_t:*

$$r_t = \mu_t + u_t$$

where μ_t is usually modeled by ARMA(p, q),

$$\mu_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} + \sum_{i=1}^{p} \psi_i u_{t-i}$$

Volatility models are concerned with the modeling of

$$\begin{aligned} \sigma_t^2 &= \mathrm{Cov}(r_t | r_{t-1}, r_{t-2}, \ldots) \\ &= \mathrm{Cov}(u_t | r_{t-1}, r_{t-2}, \ldots) \end{aligned}$$

6.1 ARCH and GARCH

Definition 6.1 (ARCH(m) Model)**.**

$$\begin{aligned} r_t &= \mu_t + u_t \\ u_t &= \sigma_t \epsilon_t \\ \sigma_t^2 &= \omega + \alpha_1 u_{t-1}^2 + \cdots + \alpha_m u_{t-m}^2 \end{aligned}$$

where $\{\epsilon_t\}$ is a sequence of IID random variables with

$$\mathbf{E}(\epsilon_t) = 0, \qquad \mathrm{Cov}(\epsilon_t) = 1 \text{ (if it exists)}$$

and $\omega > 0, \alpha_i > 0$ for $i > 0$.

$$\epsilon_t \perp\!\!\!\perp \{\sigma_s, s \leq t\}$$

and we call $u_t = r_t - \mu_t$ residuals and ϵ_t standardized residuals (or innovation). □

Remark. Commonly used distributions of ϵ_t are the standard normal and standardized Student's t distribution.

Example 6.1 (ARCH(1) Model)**.** The ARCH(1) model is

$$r_t = \mu_t + u_t, \qquad u_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \omega + u_{t-1}^2$$

where $\omega > 0, \alpha_1 > 0$.

- $\mathbf{E}(u_t) = 0$
- u_t is called residuals, ϵ_t is called standardized residuals.
- The model can be written as
  $$u_t^2 = \omega + \alpha_1 u_{t-1}^2 + \eta_t$$
  where $\eta_t = \sigma_t^2(\epsilon_t^2 - 1)$, which is a white noise.
- If $\alpha_1 < 1$, $\mathbf{E}(u_t^2)$ is constant, and u_t^2 is also stationary, but u_t^2 (or $u_t^2 - \mathbf{E}(u_t^2)$) is not a white noise.
- Assuming $\mathrm{Cov}(u_t)$ is constant for all t,
  $$\mathrm{Cov}(u_t) = \frac{\omega}{1 - \alpha_1}, \qquad \text{if } 0 < \alpha_1 < 1$$

- Under normality, and assuming the moments are all constants for t, we have

  $$\mu_4 = \frac{3\omega^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}$$

  provided $0 < \alpha_1^2 < \frac{1}{3}$. However, the fourth moment does not exist if $\alpha_1^2 > \frac{1}{3}$.
- The kurtosis is

  $$\frac{\mu_4}{\mu_2^2} = 3\frac{1-\alpha_1^2}{1-3\alpha_1^2} > 3$$

  which explains why the data has heavy tails if it indeed follows an ARCH model, even if the innovation is indeed normally distributed.

□

We can use MLE to estimate ω and $\alpha_i, \imath = 1, \ldots, m$ and use AIC or BIC to choose a suitable m.

Prediction of σ_{t+h}^2 is based on $u_1, \ldots, u_t$:

$$\begin{aligned}
\hat{\sigma}_{t+1}^2 &= \omega + \alpha_1 u_{t+1-1}^2 + \cdots + \alpha_m u_{t+1-m}^2 \\
\hat{\sigma}_{t+2}^2 &= \omega + \alpha_1 \hat{\sigma}_{t+2-1}^2 + \alpha_2 u_{t+2-2}^2 + \cdots + \alpha_m u_{t+2-m}^2 \\
\hat{\sigma}_{t+3}^2 &= \omega + \alpha_1 \hat{\sigma}_{t+3-1}^2 + \alpha_2 \hat{\sigma}_{t+3-2}^2 + \cdots + \alpha_m u_{t+3-m}^2 \\
&\vdots
\end{aligned}$$

A general ARMA(p,q) + GARCH(m,s) model is

$$r_t = \mu_t + u_t, \quad u_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \omega + \sum_{i=1}^{m} \alpha_i u_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2$$

where μ_t is the ARMA(p,q) model

$$\mu_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} + \sum_{i=1}^{q} \psi_i u_{t-i}$$

ϵ_t are IID with $\mathbf{E}(\epsilon_t) = 0, \mathbf{E}(\epsilon_t^2) = 1, \omega > 0, \alpha_i \geq 0, \beta_j \geq 0$, and

$$\sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_i) < 1$$

Let $\eta_t = u_t^2 - \sigma_t^2$, then $\{\eta_t\}$ is an uncorrelated series (white noise). The GARCH model becomes

$$u_t^2 = \omega + \sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_i) u_{t-i}^2 + \eta_t - \sum_{j=1}^{s} \beta_j \eta_{t-j}$$

This is an ARMA form for the squared series u_t^2. It can be used to understand properties of GARCH models (e.g. moment equations, forecasting, etc).

Remark (Probability Properties of GARCH(1,1))**.**

$$\sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

- Weakly stationary: $0 \leq \alpha_1, \beta \leq 1, (\alpha_1 + \beta_1 < 1)$
- Volatility clustering
- Unconditional variance

$$\mathbf{E}(u_t^2) = \frac{\omega}{1 - \alpha_1 - \beta_1}$$

- Heavy tails: if $\epsilon_t \sim N(0,1)$ and $1 - 2\alpha_1^2 - (\alpha_1 - \beta_1)^2 > 0$, then

$$\text{Kur} = \frac{\mathbf{E}(u_t^4)}{[\mathbf{E}(u_t^2)]^2} = \frac{3[1 - (\alpha_1 + \beta_1)^2]}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha_1^2} > 3$$

The prediction of volatility by using the ARMA(p,q) + GARCH(m,s) is given by:

$$\begin{aligned}
\hat{\sigma}_{t+1}^2 &= \omega + \alpha_1 u_t^2 + \cdots + \alpha_m u_{t+1-m}^2 + \beta_1 \sigma_t^2 + \cdots + \beta_s \sigma_{t+1-s}^2 \\
\hat{\sigma}_{t+2}^2 &= \omega + \alpha_1 \hat{\sigma}_{t+1}^2 + \alpha_2 u_t^2 + \cdots + \alpha_m u_{t+2-m}^2 + \beta_1 \hat{\sigma}_{t+1}^2 + \beta_2 \sigma_t^2 + \cdots + \beta_s \sigma_{t+2-s}^2 \\
\hat{\sigma}_{t+3}^2 &= \omega + \alpha_1 \hat{\sigma}_{t+2}^2 + \alpha_2 \hat{\sigma}_{t+1}^2 + \cdots + \alpha_m u_{t+3-m}^2 + \beta_1 \hat{\sigma}_{t+2}^2 + \beta_2 \hat{\sigma}_{t+1}^2 \cdots + \beta_s \sigma_{t+3-s}^2 \\
&\vdots
\end{aligned}$$

Definition 6.2 (GARCH-M Model)**.** The GARCH-M (or GARCH-in-Mean) model is given by:

$$r_t = \mu_t + c\sigma_t + u_t, \quad u_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

where c is referred to as risk premium, which is expected to be positive. □

Definition 6.3 (APARCH Model)**.** In some financial time series, large negative returns appear to increase volatility more than positive returns of the same magnitude. The APARCH(p,q) model for the conditional standard deviation is

$$\sigma_t^\delta = \omega + \sum_{i=1}^{p} \alpha_i (|u_{t-1}| - \gamma_i u_{t-1})^\delta + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^\delta$$

where $\delta > 0$ and $-1 < \gamma_j < 1, j = 1, \ldots, p$. Note that $\delta = 2$ and $\gamma_1 = \cdots = \gamma_p = 0$ gives a standard GARCH model. □

6.2 Value-at-Risk Conditional on the Past

Given past information $r_1, \ldots, r_{t-1}$, for a small $\alpha > 0$, we define the value-at-risk as

$$\text{VaR}_t(\alpha) = -\max\{v : P(r_t \leq v | r_{t-1}, r_{t-2}, \ldots) \leq \alpha\}$$

In other words, Var_t with level α means

$$P(u_t > -\text{VaR}_t(\alpha)) = 100(1 - \alpha)\%$$

If we assume $r_t = \mu_t + \sigma_t \epsilon_t$, and $\epsilon_t \perp\!\!\!\perp \sigma_t$, then

$$\text{VaR}_t(\alpha) = -\mu_t - \sigma_t z_\alpha = -\mu_t + \sigma_t \text{VaR}_\alpha(\epsilon)$$

where z_α is the α-th percentile of ϵ_t. If $\epsilon_t \sim N(0,1)$, then $z_\alpha = \Phi^{-1}(\alpha)$, the α quantile of the standard normal.

If t-distribution is considered for the residuals with degree of freedom ν, because ϵ_t is standardized (i.e. $\epsilon_t = \epsilon_t'/\sigma$ where ϵ_t' follows $t(\nu)$), and $\sigma = \sqrt{\frac{\nu}{\nu-2}}$, thus

$$z_\alpha = \frac{t_\nu^{-1}(\alpha)}{\sqrt{\frac{\nu}{\nu-2}}}$$

where $t_v^{-1}(\alpha)$ is the α quantile of $t(\nu)$.

To validate the calculation of $\text{VaR}_t(\alpha)$, the estimated probability

$$\hat{p} = \frac{\#\{u_t > -\text{VaR}_t(\alpha), t = 1, \ldots, T\}}{T}$$

should be approximately $100(1-\alpha)\%$. If $\hat{p} > 100(1-\alpha)\%$, then the model is too conservative, otherwise it is too aggressive. It may be preferable for models to be a bit conservative rather than too aggresive.

For prediction, we first predict σ_{T+h}, then together with the distribution, calculate the VaR at day $T+h$ at level α:

$$-\mu_{T+h|T} - \sigma_{T+h|T} Q(\alpha)$$

where $Q(\alpha)$ is the quantile for the standardized residuals, $\mu_{T+h|T}$ is the prediction of the conditional mean, and $\sigma_{T+h|T}$ is the prediction of the conditional standard deviation of the time series.