ORIE 5640

Statistics and Data Analysis for Financial Engineering

Course Summary with Example Solutions

Sam Rittenhouse

Contents

1 Returns

The **Net Return** over the holding period from time $t-1$ to t is

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

A simple **Gross Return** is

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

A k-Period **Gross Return** is

$$\begin{aligned} 1 + R_t(k) = \frac{P_t}{P_{t-k}} &= \left(\frac{P_t}{P_{t-1}}\right)\left(\frac{P_{t-1}}{P_{t-2}}\right)\cdots\left(\frac{P_{t-k+1}}{P_{t-k}}\right) && (1)\\ &= (1+R_t)\cdots(1+R_{t-k+1}) && (2) \end{aligned}$$

A k-Period **log return** is defined as

$$\begin{aligned} r_t(k) &= \log\{1 + R_t(k)\} && (3)\\ &= \log\{(1+R_t)\cdots(1+R_{t-k+1})\} && (4)\\ &= \log(1+R_t) + \cdots + \log(1+R_{t-k+1}) && (5)\\ &= r_t + r_{t-1} + \cdots + r_{t-k+1} && (6) \end{aligned}$$

1.1 Adjusting for Dividends

If a dividend (or interest) D_t is paid prior to time t, then the gross return at time t is defined as

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

Consequentially, a k-Period **gross return** and a k-Period **log return** are defined as follows:

$$\begin{aligned} 1 + R_t(k) &= \left(\frac{P_t + D_t}{P_{t-1}}\right)\left(\frac{P_{t-1} + D_{t-1}}{P_{t-2}}\right)\cdots\left(\frac{P_{t-k+1} + D_{t-k+1}}{P_{t-k}}\right) && (7)\\ &= (1+R_t)(1+R_{t-1})\cdots(1+R_{t-k+1}) && (8)\\ r_t(k) &= \log\{1+R_t(k)\} = \log(1+R_t) + \cdots + \log(1+R_{t-k+1}) && (9)\\ &= \log\left(\frac{P_t + D_t}{P_{t-1}}\right) + \cdots + \log\left(\frac{P_{t-k+1} + D_{t-k+1}}{P_{t-k}}\right) && (10) \end{aligned}$$

1.2 The Random Walk Model

The random walk hypothesis states that the single-period log returns are independent (stems from additive nature). Let $Z_1, Z_2, ...$ be i.i.d. with mean μ and variance σ^2. Let S_0 be an arbitrary starting point and

$$S_t = S_0 + Z_1 + \cdots + Z_t, t \geq 1$$

The process $\{S_t | t \in \mathbb{Z}^+\}$ is known as a **random walk** where each **step** $Z_i \in \{Z_i | i \in \mathbb{Z}^+\}$. If $Z \sim N(\mu, \sigma^2)$, then the process is a **normal random walk** where μ is called the **drift** parameter and σ is the **volatility** parameter.

1.2.1 The Geometric Random Walk

Using result (6),

$$\log\{1+R_t(k)\} = r_t + \cdots + r_{t-k+1} \leftrightarrow \frac{P_t}{P_{t-k}} = 1 + R_t(k) = \exp(r_t + \cdots + r_{t-k+1}) \tag{11}$$

$$\leftrightarrow P_t = P_0 \exp(r_t + \cdots + r_{t-k+1}) \tag{12}$$

We call such a process whose logarithm is a random walk a **geometric random walk**.

1.3 Exercises with Solutions

Exercise 2.8

Suppose that $X_1, X_2, ...$ is a lognormal geometric random walk with parameters (μ, σ^2) such that $X_k = X_0 \exp(r_1 + \cdots + r_k)$.

a.) Find $P(X_2 > 1.3X_0)$

Here, $r = r_1 + r_2 \sim N(2\mu, 2\sigma^2)$ is utilized.

$$\begin{aligned} P(X_2 > 1.3X_0) &= P(X_0 e^{r_1+r_2} > 1.3X_0) && (13)\\ &= P(e^{r_1+r_2} > 1.3) && (14)\\ &= P(r_1 + r_2 > \log(1.3)) && (15)\\ &= P(r > \log(1.3)) && (16)\\ &= 1 - P(r \leq \log(1.3)) && (17)\\ &= 1 - \Phi\left(\frac{\log(1.3) - 2\mu}{\sqrt{2}\sigma}\right) && (18)\\ & && (19) \end{aligned}$$

b.) Find the density of X_1

Letting $Y = X_1$ and $X = R$ we have

$$f_Y(y) = f_R(h(y))\big|h'(y)\big| \tag{20}$$

where $f_R(r) = \frac{1}{\sigma}\phi\big(\frac{r-\mu}{\sigma}\big)$ and $Y = X_0 e^R \rightarrow R = h(y) = \log(y/X_0)$ and hence $h'(y) = \frac{1}{y}$. Thus plugging in for each equation we have the density of $Y = X_1$ as

$$f_Y(y) = \frac{1}{y\sigma}\phi\Big(\frac{\log(y/X_0) - \mu}{\sigma}\Big) \tag{21}$$

c.) Find the formula for the 0.9 quantile of X_k

The 0.9 quantile, which we will call q, satisfies

$$0.9 = P(\log(X_k) \leq \log(q)) = \Phi\left(\frac{\log(q) - k\mu - log(X_0)}{\sqrt{k}\sigma}\right) \tag{22}$$

Therefore,

$$\log(q) = k\mu + \log(X_0) + \sqrt{k}\sigma\Phi^{-1}(0.9) \tag{23}$$

$$\leftrightarrow q = \exp\{k\mu + \log(X_0) + \sqrt{k}\sigma\Phi^{-1}(0.9)\} \tag{24}$$

d.) Find $E[X_k^2]$

$$E[X_k^2] = X_0^2 E[\exp(2(r_1 + ... + r_k)] \tag{25}$$

and $2(r_1 + ... + r_k)$ is normally distributed with mean $2r\mu$ and variance $4r\sigma^2$, so

$$E[X_k^2] = X_0^2 \exp(2r\mu + 2r\sigma^2) \tag{26}$$

since $E[Y] = \exp(\mu + \sigma^2/2)$ when $\log(Y) \sim N(\mu, \sigma^2)$ (Pg. 677).

e.) Find the variance of X_k

Using the results and logic from part d, we see that the variance is

$$\text{Var}(X_k) = X_0 \exp\{r(\mu + \sigma^2)\}(\exp\{r(\mu + \sigma^2)\} + 1)$$

2 Fixed Income Securities

A **Zero-Coupon Bond** is a fixed income security that pays nothing until maturity. The price of a zero is given by

$$PRICE = PAR(1 + r/k)^{-kT}$$

where T is the time to maturity in years, r is the interest rate, and k is the number of compounds per year. Note that if the interest rate increases after you buy a zero, the value of your decreases and vice versa.

Coupon Bonds make regular interest payments. At maturity, one receives a principal payment equal to the par value of the bond and the final interest payment. If a bond with a par value of PAR matures in T years and makes semiannual coupon payments of C and the yield (rate of interest) is r per half-year, then the value of the bond when it is issued is

$$\sum_{t=1}^{2T} \frac{C}{(1+r)^T} + \frac{PAR}{(1+r)^{2T}} = \frac{C}{r} + \left\{PAR + \frac{C}{r}\right\}(1+r)^{-2T}$$

Let us suppose that a coupon bond pays semiannual coupon payments of C, has a par value of PAR, and has T years until maturity. Let $y_1, y_2, ..., y_{2T}$ by the half-year spot rates for the zero's of maturities $1/2, 1, 3/2, ..., T$ years. Then the **yield to maturity** (on a half-year basis) is the value of y which solves

$$\sum_{i=1}^{2T-1} \frac{C}{(1+y_i)^i} + \frac{PAR + C}{(1+y_n)^{2T}} = \sum_{i=1}^{2T-1} \frac{C}{(1+y)^i} + \frac{PAR + C}{(1+y)^{2T}}$$

The **Term Structure** of a bond can be described by the prices of the zero coupon bonds, the spot rates, or the forward rates. You can compute any of the previous three if you have the other 2 (described below):

$$P(n) = \frac{PAR}{(1+r_1)\cdots(1+r_n)} = \frac{PAR}{(1+y_n)^n}$$

$$y_n = \left\{\frac{PAR}{P(n)}\right\}^{1/n} - 1 = \left\{(1+r_1)\cdots(1+r_n)\right\}^{1/n} - 1$$

$$r_1 = y_1 \quad r_n = \frac{(1+y_n)^n}{(1+y_{n-1})^{n-1}} \quad n = 2, 3, \cdots$$

Coupon Bond Relationships

- price > par → coupon rate > current yield > yield to maturity
- price < par → coupon rate < current yield < yield to maturity

2.1 Continuous Compounding and Forward Rates

Term structure relationships for zero's with forward rates $r_1, ..., r_n$:

$$P(n) = \frac{PAR}{\exp(r_1 + \cdots + r_n)} \leftrightarrow \frac{P(n-1)}{P(n)} = \exp(r_n) \leftrightarrow \log\left\{\frac{P(n-1)}{P(n)}\right\} = r_n$$

$$P(n) = \frac{PAR}{\exp(ny_n)} \text{ where } y_n = n^{-1}\sum r_i$$

$$r_1 = y_n \text{ and } r_n = ny_n - (n-1)y_{n-1}, n > 1$$

To specify term structure in a realistic way, we assume that there is a function $r(t)$ called the **forward-rate function** such that the current price of a zero of maturity T with par value 1 is given by

$$D(T) = \exp\left\{-\int_0^T r(t)dt\right\}$$

$D(T)$ is known as the **discount function**. Other important relationships are listed below.

$$P(T) = PAR \times D(T) \qquad y_T = \frac{1}{T}\int_0^T r(t)dt$$

$$-\frac{d}{dT}\log P(T) = r(T)\ \forall T \qquad D(T) = \exp\{-Ty_T\}$$

2.2 Sensitivity and Duration

For a zero of maturity T,

$$\frac{\text{change in bond price}}{\text{bond price}} \approx -T \times \text{change in yield}$$

In the above equation, the minus sign shows us what we already knew, that bond prices move in the opposite direction to interest rates. Additionally, the relative change in bond price is proportional to T, which quantifies the principle that longer-term bonds have higher interest-rate risks than short-term bonds.

Assume that all yields of a coupon bond change by a constant amount δ, that is, y_T changes to $y_T + \delta$ for all T. Then we can write

$$\frac{\text{change in bond price}}{\text{bond price}} \approx -DUR \times \delta \leftrightarrow DUR \approx \frac{-1}{\text{price}} \times \frac{\text{change in price}}{\text{change in yield}}$$

where the right hand side is used as the definition of **duration**.

2.3 Exercises with Solutions

2.3.1 Exercise 1 (Pg. 40)

Suppose that the forward rate is $r(t) = 0.028 + 0.00042t$.

a.) What is the yield to maturity of a 20 year bond?

$$y_{20} = \frac{1}{20}\int_0^{20}(0.028 + 0.00042t)dt = \frac{1}{20}\left[0.028t + \frac{0.00042}{2}t^2\right]_0^{20} = 0.0322$$

b.) What is the price of a par $1,000 zero maturing in 15 years?

$$P(T) = \text{PAR} \times D(T) = (1000)\exp\{-15y_{15}\} = 1000\exp\{-15(0.03115)\} = \$626.72$$

2.3.2 Exercise 3

A coupon bond has a coupon rate of 3% and a current yield of 2.8%.

a.) Is the bond selling above or below par?

The bond is selling above par value, for the price of the bond is proportional to exp(−yield); thus when the yield decreases below the coupon rate, the price of the bond increases.

b.) Is the yield to maturity above or below 2.8%?

The yield to maturity is below 2.8% as the yield to maturity takes into account the loss of capital when buying the bond above par price and only maturing the par value.

2.3.3 Exercise 14

2.3.4 Exercise 15

3 Data Analysis

When estimating a density with a **kernel density estimator** (and bandwidth b),

$$\widehat{f}(y) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{b}\right)$$

one uses the **adjust** parameter in R to fine-tune the bandwidth after automatic selection that may have over- or under-fit the data. To **overfit** the data means that the density estimate adheres too closely and so is unduly influenced by random variation. To **underfit** the data means that the density estimate does not adhere closely enough and misses features in the true density.

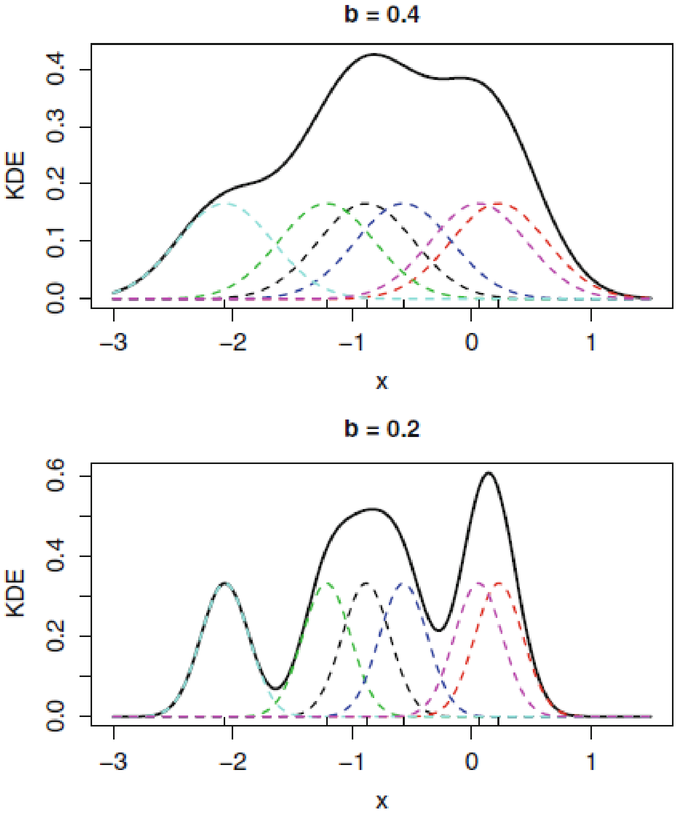


Fig. 4.5. *Illustration of kernel density estimates using a sample of size* 6 *and two bandwidths. The six dashed curves are the kernels centered at the data points, which are indicated by vertical lines at the bottom. The solid curve is the kernel density estimate created by adding together the six kernels. Although the same data are used in the top and bottom panels, the density estimates are different because of the different bandwidths.*

3.1 Order Statistics, the Sample CDF, and Sample Quantiles

The empirical CDF $F_n(y)$ is defined as

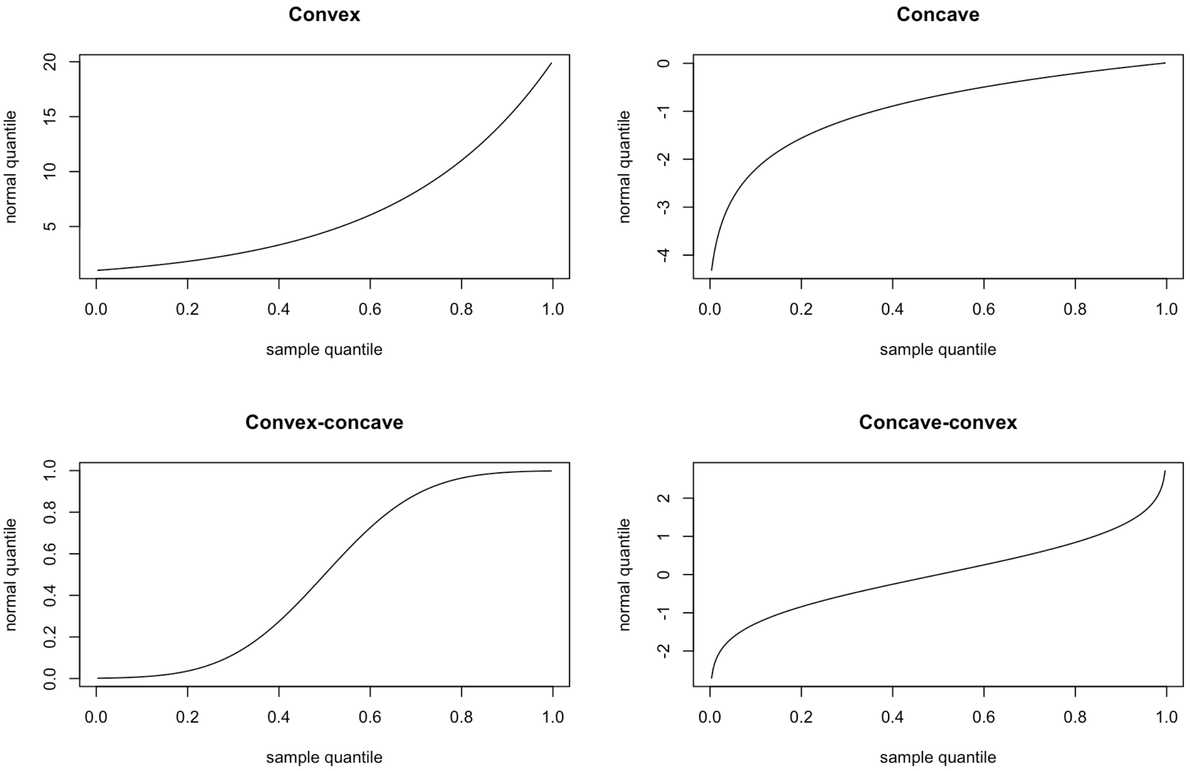
$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I\{Y_i \leq y\}$$

The only reason the EDF differs from the CDF is due to random variation.

Let $Y_1, ..., Y_n$ be an i.i.d. sample with a CDF F. Suppose that F has a density f that is continuous and positive at $F^{-1}(q), 0 < q < 1$. Then for large n, the q^{th} sample quantile is approximately normally distributed with mean equal to the population quantile $F^{-1}(q)$ and variance equal to

$$\frac{q(1-q)}{n[f\{F^{-1}(q)\}]^2}$$

If a normality assumption holds true, than the q^{th} sample quantile will be $\approx \mu + \sigma\Phi^{-1}(q)$, which is the population quantile. Therefore, a plot of sample quantiles versus Φ^{-1} will be linear.



Above, The four QQ plots illustrate, respectively, the four different cases our data can resemble: left skew, right skew, heavy tails, and light tails.

3.2 Tests for Normality

The **Shapiro-Wilk Test** tests the association between sample order statistics $Y_{(i)}$ and the expected normal order statistics, which, for large samples, are close to $\Phi^{-1}\{i/(n+1)\}$, the quantiles of the standard normal distribution. The test is implemented in R using the shapiro.test() function.

Define the **covariance** between two random variables X and Y as

$$Cov(X,Y) = \sigma_{XY} = E\big[\{X - E(X)\}\{Y - E(Y)\}\big]$$

The **Pearson correlation coefficient** between X and Y is

$$Corr(X,Y) = \rho_{XY} = \sigma_{XY}/\sigma_X\sigma_Y$$

3.3 Box Plots and Data Transformation

The logarithm transformation is the most widely used transformation as it stabilizes the variance of a variable whose conditional standard deviation is proportional to its conditional mean. The log transformation is sometimes embedded into the power transformation family by using the so-called **Box-Cox power transformation**

$$y^{(\alpha)} = \begin{cases} \frac{y^\alpha - 1}{\alpha}, & \alpha \neq 0 \\ \log(y), & \alpha = 0 \end{cases}$$

3.3.1 Transformation Kernel Density Estimation

Computation procedure:

1. start with data $Y_1, ..., Y_n$;
2. transform the data to $X_1 = g(Y_1), ..., X_n = g(Y_n)$;
3. Let $\widehat{f_X}$ be the usual KDE calculated on a grid $x_1, ..., x_m$ using $X_1, ..., X_n$;
4. plot the pairs $\big(g^{-1}(x_j)\text{, } \widehat{f_X}(x_j)|g'\{g^{-1}(x_j)\}|\big)$, $j = 1, ..., m$

If $X = g(Y)$, where g is monotonic and f_X and f_Y are the densities of X and Y, respectively, then

$$f_Y(y) = f_X\{g(y)\}|g'(y)|$$

3.4 Exercises with Solutions

3.4.1 Exercise 1

3.4.2 Exercise 7

Suppose that $Y_1, ..., Y_n$ are *i.i.d.* uniform on the interval $(0, 1)$, with density f and distribution F defined as

$$f(x) = \begin{cases} 1 & \text{if } x \in (0,1) \\ 0 & \text{otherwise} \end{cases} \text{ and } F(x) = \begin{cases} 0 & \text{if } x \le 0 \\ x & \text{if } x \in (0,1) \\ 1 & \text{if } x \ge 1 \end{cases}$$

Which sample quantile q will have the smallest variance?

Using the formula for approximate variance,

$$\frac{q(1-q)}{n[f\{F^{-1}(q)\}]^2}$$

we note that the argument in brackets in the denominator evaluates to 1, so we can isolate our attention to minimizing the function $q(1 - q)/n$

$$\frac{d}{dq}\frac{q}{n}(1-q) = \frac{1}{n}(1-2q) = 0 \to q = \frac{1}{2}$$

Thus we see $q = 1/2$ minimizes and therefore has the smallest variance.

4 Modeling Univariate Distributions

4.1 Parsimony

A model should only have as many parameters as needed to capture the important features of the data. Each unknown parameter is another quantity to estimate and another source of estimation error. Estimation error, among other things, increases the uncertainty when one forecasts future observations. On the other hand, a statistical model must have enough parameters to adequately describe the behavior of the data. A model with too few parameters can create biases because the model does not fit he data well. A statistical model with little bias, but without excess parameters, is called **parsimonious** and achieves a good tradeoff between bias and variance. Finding one or a few parsimonious models is an important part of data analysis.

4.2 Location, Scale, and Shape

A **location parameter** shifts a distribution to the right or left without changing the distribution's shape or variability. Scale parameters quantify dispersion. A parameter is a **scale parameter** for a univariate sample if the parameter is increased by the amount $|a|$ when the data are multiplied by a. Thus, if $\sigma(X)$ is a scale parameter for a random variable X, then $\sigma(aX) = |a|\sigma(X)$.

If $f(y)$ is any fixed density, then $f(y - \mu)$ is a family of distributions with location parameter μ; $\theta^{-1}f(y/\theta), \theta > 0$, is a family of distributions with a scale parameter θ; and $\theta^{-1}f\{\theta^{-1}(y-\mu)\}$ is a family of distributions with location parameter μ and scale parameter θ. In other words, $Y \sim f(y - \mu)$, $\theta Y \sim \theta^{-1}f(y/\theta)$, and $\theta Y + \mu \sim \theta^{-1}f\{\theta^{-1}(y - \mu)\}$.

A **shape parameter** is defined as any parameter that is not changed by location and scale changes. More precisely, for any $f(y)$, μ, and $\theta > 0$, the value of a shape parameter for the density $f(y)$ will equal the value of that shape parameter for $\theta^{-1}f\{\theta^{1}(y - \mu)\}$.

4.3 Skewness and Kurtosis

The **skewness** of a random variable Y is

$$\text{Sk} = E\left\{\frac{Y - E(Y)}{\sigma}\right\}^3$$

Positive skewness is also called right skewness and negative skewness is called left skewness. A distribution is **symmetric** about a point θ if $P(Y > \theta + y) = P(Y < \theta - y)$ for all $y > 0$. In this case, θ is a location parameter and equals $E(Y)$, provided that $E(Y)$ exists. The **kurtosis** of a random variable Y is

$$\text{Kur} = E\left\{\frac{Y - E(Y)}{\sigma}\right\}^4$$

It is difficult to interpret the kurtosis of an asymmetric distribution because it may measure both asymmetry and tail weight. Every normal distribution has a skewness coefficient of 0 and a kurtosis of 3. The skewness and kurtosis must be the same for all normal distributions, because the normal distribution has only location and scale parameters, no shape parameters.

Let the sample mean and standard deviation of a sample $Y_1, ..., Y_n$ be $\bar{Y}$ and s. Then the sample skewness denoted by $\widehat{Sk}$ is

$$\widehat{\text{Sk}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \bar{Y}}{s}\right)^3$$

and the sample kurtosis, denoted $\widehat{Kur}$ is

$$\widehat{\text{Kur}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \bar{Y}}{s}\right)^4$$

Skewness and kurtosis are highly sensitive to outliers. Sometimes outliers are due to **contaminants**, that is, bad data not from the population being sampled. An example would be a data recording error. A sample from a normal distribution with even a single contaminant that is sufficiently outlying will appear highly nonnormal according to the sample skewness and kurtosis. In such a case, a normal plot **will** look linear, except that the single contaminant will stick out.

4.3.1 Jarque-Bera Test

The Jarque-Bera test of normality compares the sample skewness and kurtosis to 0 and 3, their values under normality. The test statistic is

$$JB = n\big[\widehat{Sk}^2/6 + (\widehat{Kur} - 3)^2/24\big]$$

A large sample is used to compute a p-value. Under the null hypothesis, JB converges to the chi-square distribution with 2 degrees of freedom (χ_2^2) as the sample size becomes infinite, so the approximate p-value is $1 - F_{\chi_2^2}(JB)$, where $F_{\chi_2^2}$ is the CDF of the χ_2^2-distribution.

4.3.2 Moments

Let X be a random variable. The k^{th} moment of X is $E(X^k)$, and the k^{th} absolute moment is $E|X|^k$. The k^{th} central moment is

$$\mu_k = E[\{X - E(X)\}^k]$$

4.4 Heavy-Tailed Distributions

Because kurtosis is particularly sensitive to tail weight, high kurtosis is nearly synonymous with having a heavy tailed distribution. Heavy-tailed distributions are important models in finance, because equity returns and other changes in market prices usually have heavy tails.

To achieve a very heavy right tail, the density must be such that

$$f(y) \sim Ay^{-(a+1)} \text{ as } y \to \infty$$

for some $A > 0$ and $a > 0$, which will be called a **right polynomial tail**, in contrast to

$$f(y) \sim A\exp(-y/\theta) \text{ as } y \to \infty$$

for some $A > 0$ and $a > 0$, which will be called a **right exponential tail**.

4.5 t-Distributions

If Z is $N(0,1)$, W is chiq-squared with ν degrees of freedom, and Z and W are independent, then the distribution of $Z/\sqrt{W/\nu}$ is called the ***t*-distribution** with ν degrees of freedom and denoted t_ν. The α-upper quantile of the t_ν-distribution is denoted $t_{\alpha,\nu}$. The density is

$$f_{t,\nu}(y) = \left[\frac{\Gamma\{(\nu+1)/2\}}{(\pi\nu)^{1/2}\Gamma(\nu/2)}\right]\frac{1}{\{1+(y^2/\nu)\}^{(\nu+1)/2}}$$

Here, Γ is the **gamma function** defined by

$$\Gamma(t) = \int_0^\infty x^{t-1}\exp(-x)dx, \quad t > 0$$

The variance of a t_ν is finite and equal to $\nu/(\nu-2)$ if $\nu > 2$. If $0 < \nu \leq 1$, then the expected valued does not exist and the variance is not defined. If $1 < \nu \leq 2$, then the expected value is 0 and the variance is infinite. If Y has a t_ν-distribution, then $\mu + \lambda Y$ is said to have a $t_\nu(\mu, \lambda^2)$ distribution, and λ will be called the **scale parameter**. With this notation, the t_ν and $t_\nu(0,1)$ distributions are the same. If $\nu > 1$, then the $t_\nu(\mu, \lambda^2)$ distribution has a mean of μ and if $\nu > 2$, then the variance is $\lambda^2\nu/(\nu-2)$.

4.5.1 Standardized t-Distributions

For $\nu > 2$, define the $t_\nu^{std}(\mu, \sigma^2)$ distribution to be equal to the $t_\nu[\mu, \{(\nu-2)\nu\}\sigma^2]$ distribution, so that μ and σ^2 are the mean and variance of the $t_\nu^{std}(\mu, \sigma^2)$ distribution. The advantage in using the $t_\nu^{std}(\mu, \sigma^2)$ distributions is that σ^2 is the variance, whereas for the $t_\nu(\mu, \lambda^2)$ distribution, λ^2 is not the variance but instead λ^2 is the variance times $(\nu-2)/\nu$.

4.5.2 Mixtures

A **normal scale mixture** is the distribution of the random variable $\mu + \sqrt{U}Z$ where μ is a a constant mean, Z is $N(0,1)$, U is a positive random variable giving the variable of each component, and Z and U are independent. A t-distribution is a continuous normal scale mixture with $\mu = 0$ and $U = \nu/W$, where ν and W are defined above.

4.6 Generalized Error Distributions

The **standardized generalized error distribution**, or **GED**, has density

$$f_{ged}^{std}(y|\nu) = \kappa(\nu)\exp\left\{-\frac{1}{2}\left|\frac{y}{\lambda_\nu}\right|^\nu\right\}, \quad -\infty < y < \infty$$

where $\kappa(\nu)$ and λ_ν are constants given by

$$\lambda_\nu = \sqrt{\frac{2^{-\frac{2}{\nu}}\Gamma(1/\nu)}{\Gamma(3/\nu)}} \text{ and } \kappa(\nu) = \frac{\nu}{\lambda_\nu 2^{(1+1/\nu)}\Gamma(1/\nu)}$$

and chosen so that the function integrates to 1. The shape parameter $\nu > 0$ determines the tail weight, with smaller values of ν giving greater weight.

The $f_{ged}^{std}(y|\nu)$ density is symmetric about 0, which is its mean, median, and mode, and has a variance equal to 1. However, it can be shifted and rescaled to create a location-scale family. The GED distribution with mean μ, variance σ^2 , and shape parameter has density

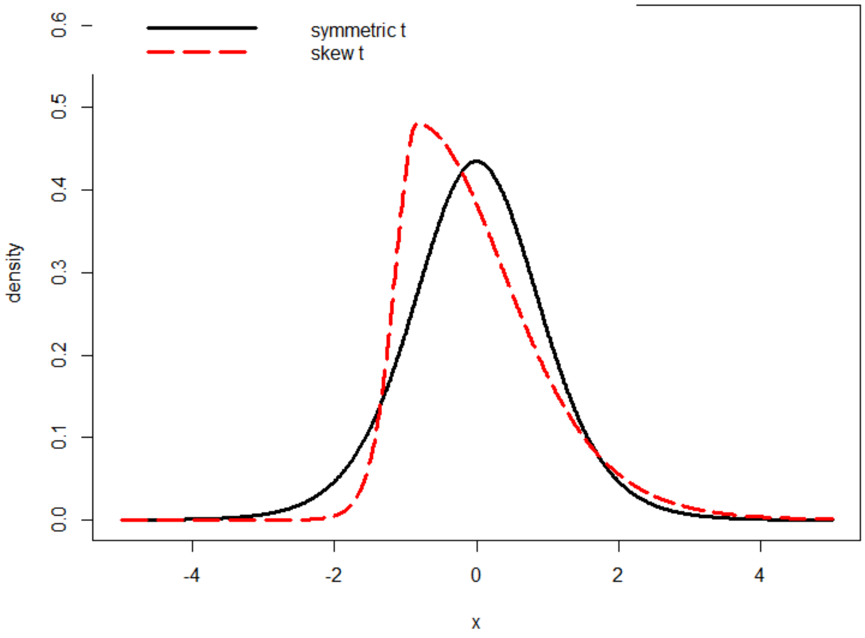
$$f_{ged}^{std}(y|\mu, \sigma^2, \nu) := f_{ged}^{std}\{(y-\mu)/\sigma|\nu\}/\sigma$$

4.7 Creating Skewed from Symmetric Distributions

Let ξ by a positive constant and f a density that is symmetric about 0. Define $f^*(y|\xi)$ to be

$$f^*(y|\xi) = \begin{cases} f(y\xi) & \text{if } y < 0, \\ f(y/\xi) & \text{if } y \geq 0 \end{cases}$$

Since $f^*(y|\xi)$ integrates to $(\xi^{-1} + \xi)/2$, $f^*(y|\xi)$ is divided by this constant to create a density.
If $\xi > 1$, then the right half of $f(y|\xi)$ is elongated relative to the left half, which induces right skewness.
If $\xi < 1$, then similarly left skewness is induced.
If f has a t-distribution, then $f(y|\xi)$ is called a skewed t-distribution (F-S skewed distribution as well). Below is an example of two distributions with $\xi = 2$. Note that the mode is to the left of the mean, typical of right-skewed densities.



4.8 Quantile-Based Location, Scale, and Shape Parameters

An alternative to the actual moments of a distribution is to use parameters based on quantiles. Any quantile $F^{-1}(p), 0 < p < 1$, is a location parameter. A positive weighted average of quantiles is also a location parameter. A scale parameter can be obtained from the difference between two quantiles:

$$s(p_1, p_2) = \frac{F^{-1}(p_2) - F^{-1}(p_1)}{a}$$

where $0 < p_1 < p_2 < 1$ and a is a positive constant.

A quantile-based shape parameter that quantifies skewness is a ratio with the numerator the difference between two scale parameters and the denominator a scale parameter:

$$\frac{s(1/2, p_2) - s(1/2, p_1)}{s(p_3, p_4)}$$

where $p_1 < 1/2$ and $p_2 = 1 - p_1$, and $0 < p_3 < p_4 < 1$.

A quantile-based shape parameter that quantifies tail weight is the ratio of the two scale parameters:

$$\frac{s(p_1, 1 - p_1)}{s(p_2, 1 - p_2)}$$

where $0 < p_1 < p_2 < 1/2$.

4.9 Maximum Likelihood Estimation (MLE)

Maximum likelihood is the most important and widespread method of estimation. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^\top$ be a vector of data and let $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^\top$ be a vector of parameters. Let $f(\boldsymbol{Y}, \boldsymbol{\theta})$ be the density of $\boldsymbol{Y}$, which depends on the parameters. Define the **likelihood function** to be

$$L(\boldsymbol{\theta}) = f(\boldsymbol{Y}|\boldsymbol{\theta})$$

This tells us the likelihood of the sample that was actually observed. The **maximum likelihood estimator** is the value of $\boldsymbol{\theta}$ that maximizes the likelihood function. The MLE is denoted by $\widehat{\boldsymbol{\theta}}_{ML}$. Often in practice it is easier and more useful to maximize the **log-likelihood function**, $l(\boldsymbol{\theta})$.

4.10 Fisher Information and CLT for MLE

Standard errors are essential for measuring the accuracy of estimators, and fortunately calculating them is fairly simple. We assume for now that θ is one-dimensional. The **Fisher information** is defined to be minus the expected second derivative of the log-likelihood, so if $\mathcal{I}(\theta)$ denote the Fisher information, then

$$\mathcal{I}(\theta) = -E\left[\frac{d^2}{d\theta^2}\log\{L(\theta)\}\right]$$

Result from CLT: Under suitable assumptions, for large enough sample sizes, the maximum likelihood estimator is approximately normally distributed with mean equal to the true parameter and with variance equal to the inverse of the Fisher information. Thus

$$s_{\widehat{\theta}} = \frac{1}{\sqrt{\mathcal{I}(\widehat{\theta})}} \text{ and } \widehat{\theta} \pm s_{\widehat{\theta}} z_{\alpha/2}$$

hold as respective formulas for the standard error and the confidence interval. The **observed Fisher information** is

$$\mathcal{I}^{obs}(\theta) = -\frac{d^2}{d\theta^2}\log\{L(\theta)\}$$

By the law of large numbers, this will be close to the expectation and may be useful if taking the expectation proves cumbersome. There is theory suggesting that using the observed Fisher information will result in a more accurate confidence interval, that is, an interval with the true coverage probability closer to the nominal value of $1-\alpha$, so observed Fisher information can be justified by more than mere convenience.

In the multivariate case, the second derivative is replaced by the Hessian matrix of second derivatives, and the result is known as the **Fisher information matrix**. The **Hessian matrix** of a function $f(x_1, ..., x_m)$ of m variables is the $m \times m$ matrix whose i, j^{th} entry is the second partial derivative of f with respect to x_i and x_j.The covariance matrix of the MLE can be estimated by the inverse of the observed Fisher information matrix.

4.10.1 Bias and Standard Deviation of MLE

In many examples, the MLE has a small bias that decreases to 0 at the rate n^{-1} as $n \to \infty$. more precisely,

$$\text{BIAS}(\widehat{\theta}_{ML}) = E(\widehat{\theta}_{ML}) - \theta \sim \frac{A}{n}, \text{ as } n \to \infty$$

for some constant A.

Although this bias can be corrected in some special problems, such as estimation of a normal variance, usually the bias is ignored since l usually is the sum of n terms and thus grows at rate n as well as the Fisher information. Therefore, the variance of the MLE decreases at rate n^{-1}, that is,

$$Var(\widehat{\theta}_{ML}) \sim \frac{B}{n}, \text{ as } n \to \infty$$

for some $B > 0$. Variability, obviously, however, should be measured by the standard deviation, and not the variance, meaning $SD(\widehat{\theta}_{ML}) \sim \sqrt{B}/\sqrt{n}$ as $n \to \infty$.

At this point, we note that the convergence of the standard deviation estimation implies that the bias becomes negligible as n gets large. This is especially important with financial markets data, where sample sizes tend to be big.

Additionally, the bias is ignored because if the MLE of a parameter θ is unbiased, the same is not true for a nonlinear function of θ. The reason for this is that for a nonlinear function g, in general,

$$E\{g(\widehat{\theta}) \neq g\{E(\widehat{\theta})\}$$

Thus, it is impossible to correct for all biases.

4.10.2 Likelihood Ratio Tests

Likelihood ratio tests, like MLE, are based upon the likelihood. Suppose that $\boldsymbol{\theta}$ is a parameter vector and that the null hypothesis puts m equality constraints on $\boldsymbol{\theta}$. More precisely, there are m functions $g_1, ..., g_m$ and the null hypothesis is that $g_i(\boldsymbol{\theta}) = 0$ for $i = 1, ..., m$. It is also assumed that none of the constraints are redundant; however, redundancies aren't necessarily easy to detect. One way to check is that the $m \times dim(\boldsymbol{\theta})$ matrix

$$\begin{bmatrix} \nabla g_1(\boldsymbol{\theta}) \\ \vdots \\ \nabla g_m(\boldsymbol{\theta}) \end{bmatrix}$$

must have rank m.

Let $\widehat{\boldsymbol{\theta}}_{ML}$ be the MLE without restrictions and let $\widehat{\boldsymbol{\theta}}_{0,ML}$ be the value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$ subject to the restrictions of the null hypothesis. The test rejects H_0 if

$$2\big[\log\{L(\widehat{\boldsymbol{\theta}}_{ML}) - \log\{L(\widehat{\boldsymbol{\theta}}_{0,ML})\}\big] \geq c$$

where c is the critical value. The left-hand side of the relation is twice the log of the likelihood ratio $L(\widehat{\boldsymbol{\theta}}_{ML})/L(\widehat{\boldsymbol{\theta}}_{0,ML})$, hence the name. A **critical value** can be found that gives a level which is exactly equal to α. The usual choice when the value is unknown is

$$c = \chi^2_{\alpha,m}$$

which is the α-upper quantile of the chi-square-distribution with m degrees of freedom.

4.11 AIC and BIC

To find a parsimonious model one needs a good tradeoff between maximizing fit and minimizing model complexity. AIC and BIC both achieve a good tradeoff and are defined

$$AIC = -2\log\{L(\widehat{\boldsymbol{\theta}}_{ML})\} + 2p \tag{27}$$

$$BIC = -2\log\{L(\widehat{\boldsymbol{\theta}}_{ML})\} + \log(n)p \tag{28}$$

where p is the number of parameters and n is the sample size. The smaller the better for these values.

The term **deviance** is often used to represent minus twice the log-likelihood, so AIC and BIC can be seen as the deviance plus a complexity penalty.

4.12 Profile Likelihood

Profile likelihood used likelihood ratio tests to create confidence intervals which makes finding the MLE convenient. Suppose the parameter vector is $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2)$. where θ_1 is a scalar parameter and the vector $\boldsymbol{\theta}_2$ contains the other parameters in the model. The profile likelihood for θ_1 is

$$L_{max}(\theta_1) = \max_{\boldsymbol{\theta}_2} L(\theta_1, \boldsymbol{\theta}_2)$$

Define $\widehat{\boldsymbol{\theta}_2}(\theta_1)$ as the value of $\boldsymbol{\theta}_2$ that maximizes the right hand side.

Let $\theta_{0,1}$ be a hypothesized value of θ_1. by the theory of likelihood ratio tests, one accepts $H_0 : \theta_1 = \theta_{0,1}$ if

$$L_{max}(\theta_{0,1}) > L_{max}(\widehat{\theta}_1) - \frac{1}{2}\chi^2_{\alpha,1}$$

The profile likelihood confidence region for θ_1 is the set of all null values that would be accepted, that is,

$$\left\{\theta_1 : L_{max}(\theta_{0,1}) > L_{max}(\widehat{\theta}_1) - \frac{1}{2}\chi^2_{\alpha,1}\right\}$$

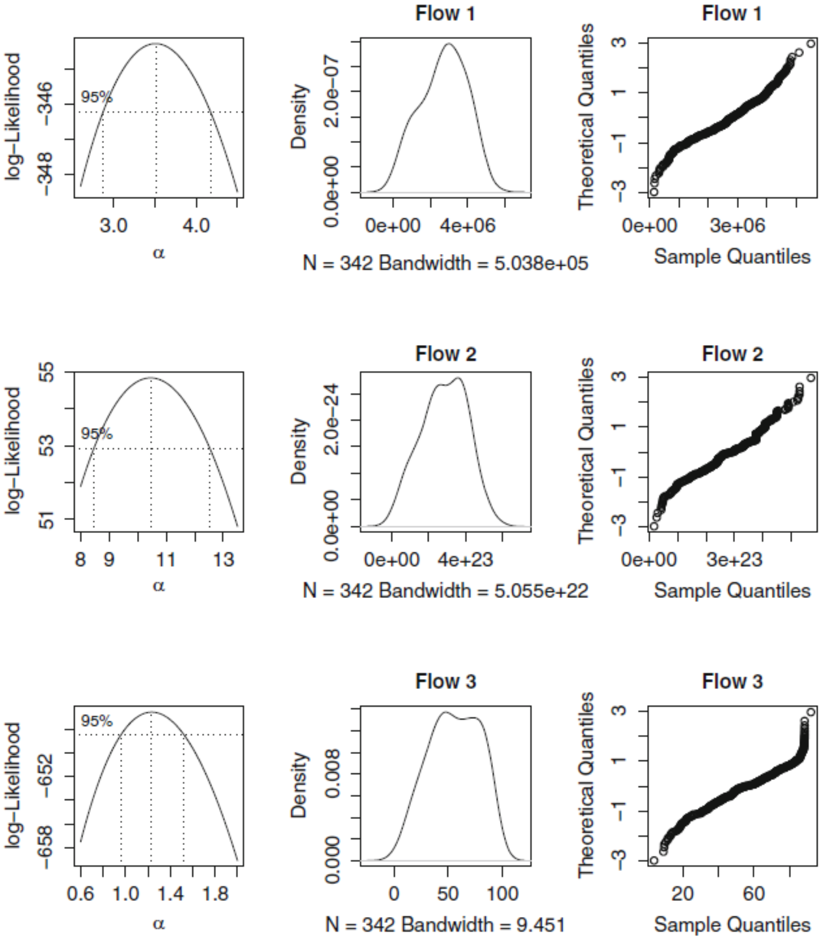


Fig. 5.12. *Profile log-likelihoods and 95 % confidence intervals for the parameter α of the Box–Cox transformation (left), KDEs of the transformed data (middle column), and normal plots of the transformed data (right).*

4.13 Exercises with Solutions

4.13.1 Exercise 2

Suppose that $Y_1, ..., Y_n$ are i.i.d. $N(\mu, \sigma^2)$, where μ is known. Find the MLE of σ^2.

$$\begin{aligned} l(\sigma^2) = \log\{L(\sigma^2)\} &= \log\left[\prod_{i=1}^{n} f_Y(Y_i|\sigma^2)\right] \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{Y_i-\mu}{\sigma}\right)^2\right\}\right) \\ \frac{d}{d\sigma^2} l(\sigma^2) &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(Y_i-\mu)^2 = 0 \\ \widehat{\sigma}^2_{ML} &= \frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)^2 \end{aligned}$$

4.13.2 Exercise 4

Let X be a random variable with mean μ and standard deviation σ.

a.) Show that the kurtosis of X is equal to 1 plus the variance of $\{X - \mu)/\sigma\}^2$.

Let $Z = \frac{X-\mu}{\sigma}$. It follows that

$$\text{Kurt}(X) = \text{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \text{E}[Z^4] = \text{Var}(Z^2) + \left[E[Z^2]\right]^2 = \text{Var}(Z^2) + 1 = \text{Var}\left[\left(\frac{X-\mu}{\sigma}\right)^2\right] + 1$$

b.) Show that the kurtosis of any random variable is at least 1.

Referencing Moors' interpretation of kurtosis above, we note that X was an arbitrary random variable with mean μ and variance σ^2, and since $Z^2 \geq 0 \ \forall \ z \in Z$, taking $\text{Var}[Z^2]$ will only return values ≥ 0. Thus, due to the non-negativity of variance, $\text{Kurt}[X] \geq 1$ for any random variable X with mean μ and variance σ^2.

c.) Show that a random variable X has a kurtosis equal to 1 if and only if $P(X = a) = P(X = b) = 1/2$ for some $a \neq b$.

$(\rightarrow)$:

$$\text{Kurt}[X] = 1 \leftrightarrow \text{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = 1 \tag{29}$$

$$\leftrightarrow \left(\frac{x-\mu}{\sigma}\right)^4 \int_{-\infty}^{\infty} f_X(x)dx = 1 \text{ if } X \text{ continuous,} \tag{30}$$

$$\text{or } \left(\frac{x-\mu}{\sigma}\right)^4 \sum_{x\in X} P(X = x) = 1 \text{ if } X \text{ discrete} \tag{31}$$

Regardless of whether X has a discrete or a continuous distribution,
$\int_{-\infty}^{\infty} f_X(x)dx = 1$ and $\sum_{x\in S(X)} P(X = x) = 1$ due to property of density functions integrating/evaluating to 1 over their respectively defined support sets. Thus, continuing the proof, we see that

the equation simplifies:

$$\begin{aligned}
&\leftrightarrow (x-\mu)^4 = \sigma^4 \tag{32}\\
&\leftrightarrow (x-\mu=\sigma) \vee (x-\mu=-\sigma) \tag{33}\\
&\leftrightarrow x \in \{\mu-\sigma, \mu+\sigma\} = X \tag{34}
\end{aligned}$$

Now we know that even though we didn't specify whether X was continuous or discrete earlier, $\text{Kurt}[X] = 1$ implies that $X \sim \text{Bernoulli}(p)$ where $p = P(X = \mu + \sigma)$ and $1 - p = P(X = \mu - \sigma)$. From here, we utilize the fact that the kurtosis of a Bernoulli random variable is $\frac{1}{1-p} + \frac{1}{p} - 3$ and set it equal to 1 and solve for p.

$$\begin{aligned}
\frac{1}{1-p} + \frac{1}{p} - 3 = 1 &\leftrightarrow \frac{1 + (1-p) - 3p(1-p)}{p(1-p)} = 1 \tag{35}\\
&\leftrightarrow 1 - 3p(1-p) = 1 \tag{36}\\
&\leftrightarrow -p^2 + p - \frac{1}{4} = 0 \tag{37}\\
&\leftrightarrow p = \frac{-(1) \pm \sqrt{(1)^2 - 4(-1)(-1/4)}}{2(-1)} = \frac{1}{2} \tag{38}\\
&\leftrightarrow P(X = \mu + \sigma) = \frac{1}{2} \tag{39}\\
&\leftrightarrow 1 - p = P(X = \mu - \sigma) = \frac{1}{2} \tag{40}
\end{aligned}$$

Setting $a = \mu+\sigma$ and $b = \mu-\sigma$, we see that $P(X = a) = P(X = b) = \frac{1}{2}$ for some $a \neq b$. Therefore, we have shown that $Kurt[X] = 1 \rightarrow P(X = a) = P(X = b) = \frac{1}{2}$ for some $a \neq b$.We now prove the other direction of this relation below.

($\leftarrow$):
Let X be a random variable with mean μ and variance σ^2, and let $a = \mu + \sigma$ and $b = \mu - \sigma$ such that $a \neq b$ (meaning $\sigma \neq 0$). Let us further suppose that $P(X = a) = P(X = b) = \frac{1}{2}$. Since $\sum_{x\in\{a,b\}} P(X = x) = 1$, we know that X can only take on a value of either a or b, which tells us that $X \sim \text{Bernoulli}(p)$ where $p = P(X = a)$ and $1 - p = P(X = b)$. Using the formula for the kurtosis of a Bernoulli random variable, computing $\text{Kurt}[X]$ gives:

$$\begin{aligned}
\frac{1}{1-p} + \frac{1}{p} - 3 &= \frac{1}{1-\frac{1}{2}} + \frac{1}{\frac{1}{2}} - 3 \tag{41}\\
&= 2 + 2 - 3 \tag{42}\\
&= 1 \tag{43}
\end{aligned}$$

Therefore, we find that

$$P(X = a) = P(X = b) = \frac{1}{2} \text{ for some } a \neq b \rightarrow \text{Kurt}[X] = 1 \tag{44}$$

Thus, we have shown that both directions of the relation hold and consequentially, $\text{Kurt}[X] = 1$ if and only if X represents a two-point, symmetric distribution; stated logically:

$$\text{Kurt}[X] = 1 \leftrightarrow P(X = a) = P(X = b) = \frac{1}{2} \text{ for some } a \neq b \tag{45}$$

4.13.3 Exercise 8

Show that $MSE(\widehat{\theta}) = \{\text{Bias}(\widehat{\theta})\}^2 + \text{Var}(\widehat{\theta})$.

$$
\begin{aligned}
\text{MSE}(\hat{\theta}) &= \text{E}((\hat{\theta} - \theta)^2) \qquad (46)\\
&= \text{E}((\hat{\theta} - \text{E}(\hat{\theta} + \text{E}(\hat{\theta}) - \theta)^2) \qquad (47)\\
&= \text{E}((\hat{\theta} - \text{E}(\hat{\theta}))^2 + 2((\hat{\theta} - \text{E}(\hat{\theta}))(\text{E}(\hat{\theta}) - \theta)) + (\text{E}(\hat{\theta} - \theta)^2)) \qquad (48)\\
&= \text{E}((\hat{\theta} - \text{E}(\hat{\theta}))^2 + 2(\text{E}(\hat{\theta}) - \theta)\text{E}(\hat{\theta} - \text{E}(\hat{\theta})) + (\text{E}(\hat{\theta}) - \theta)^2 \qquad (49)\\
&= \text{E}((\hat{\theta} - \text{E}(\hat{\theta})^2) + (\text{E}(\hat{\theta}) - \theta)^2 \qquad (50)\\
&= \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2 \qquad (51)
\end{aligned}
$$

4.13.4 Exercise 11

4.13.5 Exercise 12

4.13.6 Exercise 13

5 Resampling

Uncertainty in estimates of parameters in a given statistical model need also be considered. In order to assess the accuracy of our estimates, we generally need to employ a technique known as **resampling**. Resampling is the process by which one simulates a sample from the population by sampling from the sample. Each resample has the same sample size n as the original sample in order to try to represent the original sample as best as possible. There are two basic resampling methods, model-free and model-based.

In **model-free resampling**, the resamples are drawn with replacement from the original sample since we need independent observations.

In **model-based resampling**, one assumes that the original sample was drawn i.i.d. from a density in the parametric family, $\{f(\boldsymbol{y}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, so, for an unknown value of $\boldsymbol{\theta}$, $f(\boldsymbol{y}|\boldsymbol{\theta})$ is the population density.

5.1 Bootstrap Estimates of Bias, SD, and MSE

Let θ be a 1-D parameter, let $\widehat{\theta}$ be its estimate from the sample, and let $\widehat{\theta}_1^*, ..., \widehat{\theta}_B^*$ be estimates from B resamples. Additionally, define $\overline{\widehat{\theta}^*}$ to be the mean of $\widehat{\theta}_1^*, ..., \widehat{\theta}_B^*$. The asterisk indicates a statistics is calculated from a resample. The bootstrap estimate of bias is

$$\text{BIAS}_{boot}(\widehat{\theta}) = \overline{\widehat{\theta}^*} - \widehat{\theta}$$

The bootstrap standard error for $\widehat{\theta}$ is the sample standard deviation of $\widehat{\theta}_1^*, ..., \widehat{\theta}_B^*$, that is,

$$s_{boot}(\widehat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\theta}_b^* - \overline{\widehat{\theta}^*})^2}$$

where B is the number of resamples. The MSE is estimated by

$$\text{MSE}_{boot}(\widehat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\theta}_b^* - \widehat{\theta})^2$$

and, in fact,

$$\text{MSE}_{boot}(\widehat{\theta}) \approx \text{BIAS}^2_{boot}(\widehat{\theta}) + s^2_{boot}(\widehat{\theta})$$

The reason we do not have equality in the above relation is that we used B and not $B-1$, making our estimates slightly biased but usually only results in small approximation error.

5.2 Bootstrap Confidence Intervals

When a confidence interval uses an approximation, there are two coverage probabilities, the nominal one stated and the actual one that is unknown. Only for exact confidence intervals will these two probabilities be equal. The normal theory confidence interval for an estimate of θ is

$$\widehat{\theta} \pm s_{boot}(\widehat{\theta})z_{\alpha/2}$$

To avoid confusion, it should be emphasized that the normal approximation does not assume that the population is normally distributed by only that $\widehat{\theta}$ is normally distributed by CLT.

If we are not sampling from a normal distribution, then the distribution is presumed unknown, which leads us to two problems. First, we do not know the distribution of the population. Second, even if we did know, it is a difficult, usually intractable, probability calculation to get the distribution of the t-statistics from the distribution of the population. Luckily, we can remedy this problem with resampling.

The method of constructing a t-confidence interval for μ can be generalized to other parameters. Let $\widehat{\theta}$ and $s(\widehat{\theta})$ be the estimate of θ and its standard error calculated from the sample. Let $\widehat{\theta}_b^*$ and $s_b(\widehat{\theta})$ be the same quantities from the b^{th} bootstrapped sample. Then the b^{th} bootstrap t-statistic is

$$t_{boot,b} = \frac{\widehat{\theta} - \widehat{\theta}_b^*}{s_b(\widehat{\theta})}$$

Let t_L and t_U be the $\alpha/2$-lower and $\alpha/2$-upper sample quantiles of these t-statistics. Then the confidence interval is

$$\big(\widehat{\theta} + t_L s(\widehat{\theta}), \widehat{\theta} + t_U s(\widehat{\theta})\big)$$

5.2.1 Basic Bootstrap Interval

Let q_L and q_U be the $\alpha/2$-lower and -upper sample quantiles of $\widehat{\theta}_1^*, ..., \widehat{\theta}_B^*$. The fraction of estimates which satisfy

$$q_L \le \widehat{\theta}_b^* \le q_U$$

is $1-\alpha$. Algebraically, we can rewrite as

$$\widehat{\theta} - q_L \le \widehat{\theta} - \widehat{\theta}_b^* \le \widehat{\theta} - q_U$$

so that $\widehat{\theta} - q_U$ and $\widehat{\theta} - q_L$ are lower and upper quantiles for the distribution of $\widehat{\theta} - \widehat{\theta}_b^*$. Using the bootstrap approximation, these boundaries can act as lower and upper quantiles for the distribution of $\theta - \widehat{\theta}$. Thus adding $\widehat{\theta}$ back to the inequality, we see that the resulting bootstrap interval for θ which contains $1-\alpha$ of the samples is

$$\big(2\widehat{\theta} - q_U, 2\widehat{\theta} - q_L\big)$$

5.2.2 Percentile Interval

The basic percentile interval is simply (q_L, q_U) where each is the $\alpha/2$ quantile boundary for $\widehat{\theta}_1^*, ..., \widehat{\theta}_B^*$. Assume that for some monotonically increasing function g, $g(\widehat{\theta}^*)$ is symmetrically distributed about $g(\widehat{\theta})$. Because g is monotonically strictly increasing and the quantiles are transformation-respecting, $g(q_L)$ and $g(q_U)$ are lower and upper-$\alpha/2$ quantiles of $g(\widehat{\theta}_1^*), ..., g(\widehat{\theta}_B^*)$, and the basic percentile confidence interval for $g(\theta)$ is

$$\{g(q_L), g(q_U)\}$$

Now, if the above interval has coverage probability $1 - \alpha$ for $g(\theta)$, then, since g is strictly monotonically increasing, (q_L, q_U) has coverage probability $1 - \alpha$ for θ.

5.3 Exercises with Solutions

5.3.1 Exercise 1

To estimate the risk of a stock, a sample of 50 log returns was taken and s was 0.31. To get a confidence interval for σ, 10,000 resamples were take. Let $s_{b,boot}$ be the sample standard deviation of the bth resample. The 10,000 value of $s_{b,boot}/s$ were tanked from smallest to largest.

Rank	Value of $s_{b,boot}/s$
250	0.71
500	0.71
1,000	0.85
9,000	1.34
9,500	1.67
9,750	2.19

Find a 90% confidence interval for σ.

6 Multivariate Statistics

6.1 Covariance and Correlation Matrices

Let $\boldsymbol{Y} = (Y_1, ..., Y_d)^\top$ be a random vector. We define the expectation vector of $\boldsymbol{Y}$ to be

$$E(\boldsymbol{Y}) = \begin{bmatrix} E(Y_1) \\ \vdots \\ E(Y_d) \end{bmatrix}$$

The **covariance matrix** of $\boldsymbol{Y}$ is the matrix whose (i, j)-th entry is $\text{Cov}(Y_i, Y_j)$ for $i, j = 1, ..., N$. That is,

$$COV(\boldsymbol{Y}) = \begin{bmatrix} Var(Y_1) & Cov(Y_1, Y_2) & \cdots & Cov(Y_1, Y_d) \\ Cov(Y_2, Y_1) & Var(Y_2) & \cdots & Cov(Y_2, Y_d) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_d, Y_1) & Cov(Y_d, Y_2) & \cdots & Var(Y_d) \end{bmatrix}$$

The **correlation matrix** is denoted $\text{CORR}(\boldsymbol{Y})$ and is defined analogously. Note that the diagonal in the correlation matrix must be equal to 1 in each entry. The covariance matrix can be written as

$$\text{COV}(\boldsymbol{Y}) = E\big[\{\boldsymbol{Y} - E(\boldsymbol{Y})\}\{\boldsymbol{Y} - E(\boldsymbol{Y})\}^\top\big]$$

6.2 Linear Functions of Random Variables

If X and Y are random variables and w_1 and w_2 are constants, then

$$E(w_1X + w_2Y) = w_1E(X) + w_2E(Y)$$

and

$$Var(w_1X + w_2Y) = w_1^2Var(X) + 2w_1w_2Cov(X, Y) + w_2^2Var(Y)$$

Let $\boldsymbol{w} = (w_1, ..., w_d)^\top$ be a vector of weights such that

$$\boldsymbol{w}^\top \boldsymbol{Y} = \sum_{i=1}^{N} w_iY_i$$

is a weighted average of the components of $\boldsymbol{Y}$. Then,

$$E(\boldsymbol{w}^\top \boldsymbol{Y}) = \boldsymbol{w}^\top\{E(\boldsymbol{Y})\}$$

$$Var(\boldsymbol{w}^\top \boldsymbol{Y}) = \sum_{i=1}^{N}\sum_{j=1}^{N} w_iw_j\text{Cov}(Y_i, Y_j) = \boldsymbol{w}^\top\text{COV}(\boldsymbol{Y})\boldsymbol{w}$$

An important property of covariance and correlation matrices is that they are symmetric and positive semidefinite. A matrix $\boldsymbol{A}$ is said to be positive semidefinite (definite) if $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0 \ \ (> 0)$ for all vectors $\boldsymbol{x} \neq 0$.

Let $\boldsymbol{W}$ be a nonrandom $d \times q$ matrix so that $\boldsymbol{W}^\top \boldsymbol{Y}$ is a random vector of q linear combinations of $\boldsymbol{Y}$. Then we can generalize and write

$$\text{COV}(\boldsymbol{W}^\top \boldsymbol{Y}) = \boldsymbol{W}^\top\text{COV}(\boldsymbol{Y})\boldsymbol{W}$$

6.3 Multivariate t-Distribution

The random vector $\boldsymbol{Y}$ has a multivariate $t_\nu(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ distribution if

$$\boldsymbol{Y} = \boldsymbol{\mu} + \sqrt{\frac{\nu}{W}}\boldsymbol{Z}$$

where W is chi-squared distributed with ν degrees of freedom, $\boldsymbol{Z}$ is $N_d(0, \boldsymbol{\Lambda})$ distributed, and W and Z are independent. Thus, the multivariate t-distribution is a continuous scale mixture of multivariate normal distributions.

For $\nu > 1$, $\boldsymbol{\mu}$ is the mean vector of $\boldsymbol{Y}$. For $0 < \nu \leq 1$, the expectation does not exists, but $\boldsymbol{\mu}$ can still be regarded as the "center" of the distribution because for any value of ν, the vector $\boldsymbol{\mu}$ contains the medians of the components of $\boldsymbol{Y}$ and the contours of the density of $\boldsymbol{Y}$ are ellipses centered at $\boldsymbol{\mu}$. For $\nu > 2$, the covariance matrix of $\boldsymbol{Y}$ is

$$\boldsymbol{\Sigma} = \frac{\nu}{\nu - 2}\boldsymbol{\Lambda}$$

where $\boldsymbol{\Lambda}$ is known as the **scale matrix**.

6.4 Exercises with Solutions

6.4.1 Exercise 1

Suppose that $E(X) = 1$, $E(Y) = 1.5$, $\mathrm{Var}(X) = 2$, $\mathrm{Var}(Y) = 2.7$, and $\mathrm{Cov}(X,Y) = 0.8$.

a.) Compute $E(0.2X + 0.8Y)$ and $\mathrm{Var}(0.2X + 0.8Y)$.

$$E(0.2X + 0.8Y) = 0.2E(X) + 0.8E(Y) = 0.2(1) + 0.8(1.5) = 1.4$$

$$\mathrm{Var}(0.2X+0.8Y) = 0.2^2\mathrm{Var}(X)+0.8^2\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y) = (0.2^2)(2)+(0.8^2)(2.7)+2(0.8) = 3.408$$

b.) For what value of w is $\mathrm{Var}\{wX + (1-w)Y\}$ minimized?

$$\begin{aligned}
\frac{d}{dw}\mathrm{Var}\{wX + (1-w)Y\} &= 2w\mathrm{Var}(X) - 2(1-w)\mathrm{Var}(Y) + (2-4w)\mathrm{Cov}(X,Y) = 0\\
w(2\mathrm{Var}(X) + w\mathrm{Var}(Y) &- 4\mathrm{Cov}(X,Y)) - 2\mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y) = 0\\
w^* = \frac{2\mathrm{Var}(Y) - 2\mathrm{Cov}(X,Y)}{2\mathrm{Var}(X) + 2\mathrm{Var}(Y) - 4\mathrm{Cov}(X,Y)} &= \frac{2}{2}\frac{2.7 - 2(0.8)}{2 + 2.7 - 2(0.8)} \approx 0.3548
\end{aligned}$$

6.4.2 Exercise 2

Let X_1, X_2, Y_1, and Y_2 be random variables.

a.) Show $\mathrm{Cov}(X_1 + X_2, Y_1 + Y_2) = \mathrm{Cov}(X_1, Y_1) + \mathrm{Cov}(X_1, Y_2) + \mathrm{Cov}(X_2, Y_1) + \mathrm{Cov}(X_2, Y_2)$.

$$\begin{aligned}
\mathrm{Cov}\left(\sum_{i=1}^{2} X_i, \sum_{i=1}^{2} Y_i\right) &= E\{(X_1 + X_2)(Y_1 + Y_2)\} - E(X_1 + X_2)E(Y_1 + Y_2)\\
&= E(X_1Y_1 + X_1Y_2 + X_2Y_1 + X_2Y_2)\\
&\quad - \{E(X_1)E(Y_1) + E(X_1)E(Y_2) + E(X_2)E(Y_1) + E(X_2)E(Y_2)\}\\
&= \{E(X_1Y_1) - E(X_1)E(Y_1)\} + \cdots + \{E(X_2Y_2) - E(X_2)E(Y_2)\}\\
&= \mathrm{Cov}(X_1, Y_1) + \mathrm{Cov}(X_1, Y_2) + \mathrm{Cov}(X_2, Y_1) + \mathrm{Cov}(X_2, Y_2)
\end{aligned}$$

b.) Generalize to n of both variables.

$$\begin{aligned}
\mathrm{Cov}\left(\sum_{j=1}^{n} X_j, \sum_{k=1}^{n} Y_k\right) &= E\{(X_1 + \cdots + X_n)(Y_1 + \cdots + Y_n)\} - E(X_1 + \cdots + X_n)E(Y_1 + \cdots + Y_n)\\
&= E(X_1Y_1 + \cdots + X_jY_k + \cdots + X_kY_j \cdots X_nY_n)\\
&\quad - \{E(X_1)E(Y_1) + \cdots + E(X_j)E(Y_k) + \cdots + E(X_k)E(Y_j) + \cdots + E(X_n)E(Y_n)\}\\
&= \{E(X_1Y_1) - E(X_1)E(Y_1)\} + \cdots + \{E(X_jY_k) - E(X_j)E(Y_k)\} + \cdots\\
&\quad + \{E(X_kY_j) - E(X_k)E(Y_j)\} + \{E(X_nY_n) - E(X_n)E(Y_n)\}\\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{Cov}(X_j, Y_k)
\end{aligned}$$

6.4.3 Exercise 5

Show that if X is uniformly distributed on $[-a, a]$ for any $a > 0$ and if $Y = X^2$, then X and Y are uncorrelated but they are not independent.

For Y and X to be uncorrelated, $Corr(X, Y) = Corr(X, X^2) = 0 \rightarrow Cov(X, X^2) = 0 \rightarrow E(X^3) - E(X)E(X^2) = 0$. Obviously, the first moment of this distribution is simply the center of the distribution and thus 0, and since this is a symmetric distribution, any odd moment on the domain $(-a, a)$ will evaluate to 0 as any odd power integrates to an even power which must have the same value for $-a$ and a. Therefore, X and Y are uncorrelated but are obviously not independent as they are defined in terms of each other, $Y = X^2$.

7 Copulas

Copulas are a popular framework for both defining multivariate distributions and modeling multivariate data. A copula characterizes the dependence - and only the dependence - between the components of a multivariate distribution. The primary financial application of copula models is risk assessment and management of portfolios that contain assets which exhibit co-movements in extreme behavior.

A **copula** is a multivariate CDF whose univariate marginal distributions are all Uniform(0,1), denoted C_Y for a random vector $\boldsymbol{Y}$. Since C_Y is the CDF Of $\{F_{Y_1}(Y_1), ..., F_{Y_d}(Y_d)\}$, by the definition of a CDF we have

$$\begin{aligned} C_Y(u_1, ..., u_d) &= P\{F_{Y_1}(Y_1) \leq u_1, ..., F_{Y_d}(Y_d) \leq u_d\} && (52) \\ &= P\{Y_1 \leq F_{Y_1}^{-1}(u_1), ..., Y_d \leq F_{Y_d}^{-1}(u_d)\} && (53) \\ &= F_Y\{F_{Y_1}^{-1}(u_1), ..., F_{Y_d}^{-1}(u_d)\} && (54) \end{aligned}$$

Now, let $u_j = F_{Y_j}(y_j)$, for $j = 1, ..., d$ so that we see

$$F_Y(y_1, ..., y_d) = C_Y\{F_{Y_1}(y_1), ..., F_{Y_d}(y_d)\}$$

The above equation states that the joint CDF F_Y can be decomposed into the copula C_Y, which contains all information about the dependencies among $(Y_1, ..Y_d)$, and the univariate marginal CDFs $F_{Y_1}, ..., F_{Y_d}$, which contain all information about the univariate marginal distributions.

Let

$$c_Y(u_1, ..., u_d) = \frac{\partial^d}{\partial u_1 \cdots \partial u_d} C_Y(u_1, ..., u_d)$$

be the density associated with C_Y.

7.1 Special Copulas

All d-dimensional copula functions C have domain $[0, 1]^d$ and range $[0, 1]$. There are three copulas of special interest because they represent independence and two extremes of dependence.

7.1.1 Independence Copula

The d-dimensional **independence copula** C_0 is the CDF of d mutually independent Uniform(0,1) random variables. It equals

$$C_0(u_1, ..., u_d) = u_1 \cdots u_d$$

and the associated density is uniform on $[0,1]^d$; that is, $c_0(u_1, ..., u_d) = 1$ on $[0,1]^d$, and zero elsewhere.

7.1.2 Co-monotonicity Copula

The d-dimensional **co-monotonicity copula** C_+ characterizes perfect positive dependence, which means that the components of a given random vector can be represented as increasing functions of a single random variable. Let U be Uniform(0,1). Then, the co-monotonicity copula is the CDF of $\boldsymbol{U} = (U, ..., U)$; that is, $\boldsymbol{U}$ contains d copies of U so that all of the components of $\boldsymbol{U}$ are equal. Thus,

$$C_+(u_1, .., u_d) = P(U \leq u_1, ..., U \leq u_d) = P\{U \leq \min(u_1, ..., u_d)\} = \min(u_1, ..., u_d)$$

The co-monotonicity copula is also an **upper bound for all copula functions**:

$$C(u_1, ..., u_d) \leq C_+(u_1, ..., u_d) \ \forall \ (u_1, ..., u_d) \in [0,1]^d$$

7.1.3 Counter-monotonicity Copula

The 2-dimensional **counter-monotonicity copula** C_- is defined as the CDF of $(U, 1-U)$, which has perfect negative dependence, which means that the components of a given random vector can be represented as decreasing functions of a single random variable.. Therefore,

$$C_-(u_1, u_2) = P(U \leq u_1, 1 - U \leq u_2) = P(1 - u_2 \leq U \leq u_1) = \max(u_1 + u_2 - 1, 0)$$

All 2-dimensional copula functions are bounded below by the above equation. It is not possible to have a counter-monotonicity copula with $d > 2$. For example, suppose U_1 is counter-monotonic to U_2 and U_2 is counter-monotonic to U_3, then U_1 and U_3 will be co-monotonic instead of counter-monotonic. Hence we define a different **lower bound for all copula functions**:

$$\max(u_1 + ... + u_d + 1 - d, 0) \leq C(u_1, ..., u_d) \ \forall \ (u_1, ..., u_d) \in [0,1]^d$$

7.2 Gaussian and t-Copulas

There is a 1:1 correspondence between correlation matrices and Gaussian copulas. The Gaussian copula with correlation matrix $\boldsymbol{\Omega}$ will be denoted $C_{Gauss}(u_1, ..., u_d|\boldsymbol{\Omega})$. If a random vector $\boldsymbol{Y}$ has a Gaussian copula, then $\boldsymbol{Y}$ is said to have a **meta-Gaussian distribution**. This, does not, of course, mean that $\boldsymbol{Y}$ has a multivariate Gaussian distribution, since the univariate marginal distributions of $\boldsymbol{Y}$ could be any distributions at all. Similarly, let $C_t(u_1, ..., u_d|\boldsymbol{\Omega}, \nu)$ denote the copula of a random vector that has a multivariate t-distribution with tail index ν.

7.3 Archimedean Copulas

An **Archimedean copula** with a strict generator has the form

$$C(u_1, ..., u_d) = \phi^{-1}\{\phi(u_1) + ... + \phi(u_d)\}$$

where the generator function ϕ satisfies the following conditions:

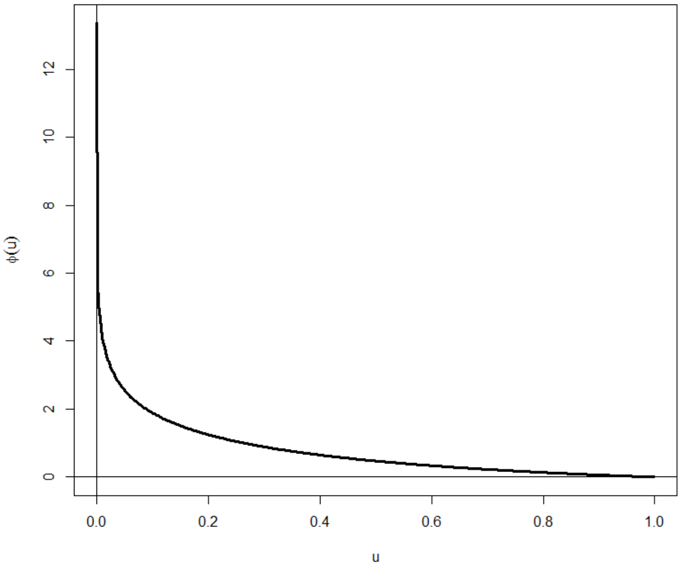
1. ϕ is a continuous, strictly decreasing, and convex function mapping $[0,1]$ onto $[0,\infty]$
2. $\phi(0) = \infty$
3. $\phi(1) = 0$

The independence copula C_0 is an Archimedean copula with generator function $\phi(u) = -\log(u)$.

7.3.1 The Frank Copula

The Frank copula has generator

$$\phi_{Fr}(u|\theta) = -\log\left(\frac{e^{-\theta u}-1}{e^{-\theta}-1}\right), \quad -\infty < \theta < \infty$$



The inverse generator is

$$\phi_{Fr}^{-1}(y|\theta) = -\frac{1}{\theta}\log\{e^{-y}(e^{-\theta}-1)+1\}$$

Therefore, the bivariate Frank copula is

$$C_{Fr}(u_1, u_2|\theta) = -\frac{1}{\theta}\log\left\{1 + \frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right\}$$

Using the approximations $e^x - 1 \approx x$ and $\log(1+x) \approx x$ as $x \to 0$, we can show that as $\theta \to 0$, $C_{Fr}(u_1, u_2) \to u_1 u_2$, the bivariate independence copula C_0, so we define the Frank copula to be the independence copula at $\theta = 0$.

As $\theta \to -\infty$, the bivariate Frank copula converges to the counter-monotonicity copula C_-. To see this, first note that as $\theta \to -\infty$

$$C_{Fr}(u_1, u_2|\theta) \sim -\frac{1}{\theta}\log\left\{1 + e^{-\theta(u_1+u_2-1)}\right\}$$

If $u_1 + u_2 - 1 > 0$, then as $\theta \to -\infty$, the exponent $-\theta(u_1 + u_2 - 1)$ converges to ∞ and

$$\log\left\{1 + e^{-\theta(u_1+u_2-1)}\right\} \sim -\theta(u_1 + u_2 - 1)$$

so that $C_{Fr}(u_1, u_2|\theta) \to u_1 + u_2 - 1$. Similarly, if $u_1 + u_2 - 1 \to 0$, then $u_1 + u_2 - 1 \to -\infty$, and $C_{Fr}(u_1, u_2|\theta) \to 0$. Putting these results together, we see that $C_{Fr}(u_1, u_2|\theta)$ converges to

$\max(0, u_1 + u_2 - 1)$, the counter-monotonicity copula C_- as $\theta \to -\infty$.

As $\theta \to \infty$, we note that our denominator $e^{-\theta} - 1 \to -1$ and after rearranging our expression in the logarithm, we are left with

$$e^{-\theta(u_1+u_2)} + e^{-\theta u_1} + e^{-\theta u_2}$$

Assuming $u_1, u_2 > 0$ this expression will converge to the term with the smallest coefficient on θ as the smallest coefficient indicates the slowest approach to 0, leaving only the term in the exponent after the logarithm is taken, that is, either $-\theta u_1$ or $-\theta u_2$. The $-1/\theta$ out front will send this to either u_1 or u_2, whichever is smaller. Therefore, $C_{Fr}(u_1, u_2|\theta) \to \min(u_1, u_2)$, the co-monotonicity copula C_+.

7.3.2 The Clayton Copula

The **Clayton copula**, with generator function $\phi_{Cl}(u|\theta) = \frac{1}{\theta}(u^{-\theta} - 1),\ \theta > 0$, is

$$C_{Cl}(u_1, ..., u_d|\theta) = (u_1^{-\theta} + \cdots + u_d^{-\theta} + 1 - d)^{-1/\theta}$$

We define the Clayton copula for $\theta = 0$ as

$$\lim_{\theta \downarrow 0} C_{Cl}(u_1, ..., u_d|\theta) = u_1 \cdots u_d$$

which is the independence copula C_0.

It is possible to extend the range of θ to include $-1 \leq \theta < 0$, bu then the generator $u^{-\theta}01/\theta$ is finite at $u = 0$ in violation of assumption 2 of strict generators. Thus, the generator is not strict if $\theta < 0$. As a result, it is necessary to define $C_{Cl}(u_1, ..., u_d|\theta)$ to equal 0 for small values of u_i in this case.

As $\theta \to -1$, the bivariate Clayton copula converges to the counter-monotonicity copula C_-, and as $\theta \to \infty$ the Clayton copula converges to the co-monotonicity copula C_+.

7.3.3 The Gumbel Copula

The Gumbel copula has the generator $\phi_{Gu}(u|\theta) = (-\log u)^\theta, \theta \geq 1$, and consequently is equal to

$$C_{Gu}(u_1, ..., u_d|\theta) = \exp\big[-\{(-\log u_1)^\theta + \cdots + (-\log u_d)^\theta\}^{1/\theta}\big]$$

The Gumbel copula is the independence copula C_0 when $\theta = 1$, and converges to the co-monotonicity copula C_+ as $\theta \to \infty$, but the Gumbel copula cannot have negative dependence.

7.3.4 The Joe Copula

The Joe copula is similar to the Gumbel copula; it cannot have negative dependence, but it allows even stronger upper tail dependence and is closer to being the reverse of the Clayton copula in the positive dependence case. The Joe copula has the generator $\phi_{Joe}(u|\theta) = -\log\{1 - (1-u)^\theta\}, \theta \geq 1$. In the bivariate case, the Joe copula is equal to

$$C_{Joe}(u_1, u_2|\theta) = 1 - [(1-u_1)^\theta + (1-u_2)^\theta - (1-u_1)^\theta(1-u_2)^\theta]^{1/\theta}$$

The Joe copula is the independence copula C_0 when $\theta = 1$, and converges to the co-monotonicity copula C_+ as $\theta \to \infty$.

In applications, it is useful that the different copula families have different properties, since this increases our ability to find a copula that fits the data.

7.4 Rank Correlation

The Pearson correlation coefficient defined as

$$\text{Corr}(X, Y) = \rho_{XY} = \sigma_{XY}/\sigma_X\sigma_Y$$

is not convenient for fitting copulas to data, since it depends on the univariate marginal distributions as well as the copula. Rank correlation coefficients remedy this problem, since they depend only on the copula. Define the rank of Y_i to be

$$\text{rank}(Y_i) = \sum_{j=1}^{n} I(Y_j \leq Y_i)$$

A key property of ranks is that they are unchanged by strictly monotonic transformations of variables. In particular, through transform by CDF, so the distribution of any rank statistic depends only on the copula of the observations, not the univariate marginals.

7.4.1 Kendall's Tau

Let (Y_1, Y_2) be a bivariate random vector and let (Y_1^*, Y_2^*) be an independent copy of (Y_1, Y_2). Then (Y_1, Y_2) and (Y_1^*, Y_2^*) are called a **concordant pair** if the relative rankings of each Y_i and Y_i^* agree, and the pair is called a **discordant pair** if the relative rankings disagree. **Kendall's tau** is the probability of a concordant pair minus the probability of a discordant pair. Therefore, Kendall's tau for (Y_1, Y_2) is

$$\begin{aligned}\rho_\tau(Y_1, Y_2) &= P\{(Y_1 - Y_1^*)(Y_2 - Y_2^*) > 0\} - P\{(Y_1 - Y_1^*)(Y_2 - Y_2^*) < 0\} \\ &= E[\text{sign}\{(Y_1 - Y_1^*)(Y_2 - Y_2^*)]\end{aligned}$$

where the **sign function** is

$$\text{sign}(x) = \begin{cases} 1, & x > 0, \\ -1, & x < 0, \\ 0, & x = 0 \end{cases}$$

It is clear that ρ_τ is symmetric in its arguments and takes values in $[-1, 1]$. It is easy to check that if g and h are increasing functions, then

$$\rho_\tau\{g(Y_1), h(Y_2)\} = \rho_\tau(Y_1, Y_2)$$

If g and h are the marginal CDFs of Y_1 and Y_2, then $\rho_\tau(Y_1, Y_2)$ is the Kendall's tau for a pair of random variables distributed according the copula of (Y_1, Y_2).

For a random vector $\boldsymbol{Y}$, we define the **Kendall's' tau correlation matrix $\boldsymbol{\Omega}_\tau$** to be the matrix whose (j, k) entry is Kendall's tau for the j^{th} and k^{th} components of $\boldsymbol{Y}$, that is

$$[\boldsymbol{\Omega}(\boldsymbol{Y})]_{jk} = \rho_\tau(Y_j, Y_k)$$

If we have a bivariate sample, then the sample Kendall's tau is

$$\widehat{\rho}_\tau(\boldsymbol{Y}_{1:n}) = \binom{n}{2}^{-1} \sum_{1 \leq i \leq j \leq n} \text{sign}\{(Y_{i,1} - Y_{j,1})(Y_{i,2} - Y_{j,2})\}$$

Note that $\binom{n}{2}$ is the number of summands, so $\widehat{\rho}_\tau$ is the average of $\text{sign}\{(Y_{i,1} - Y_{j,1})(Y_{i,2} - Y_{j,2})\}$ across all distinct pairs of observations. The sample Kendall's tau correlation matrix is defined analogously.

7.4.2 Spearman's Rho

Spearman's rho, for a bivariate random vector (Y_1, Y_2), is denoted as $\rho_S(Y_1, Y_2)$ and is defined to be the Pearson correlation coefficient of $\{F_{Y_1}(Y_1), F_{Y_2}(Y_2)\}$:

$$\rho_S(Y_1, Y_2) = \mathrm{Corr}\{F_{Y_1}(Y_1), F_{Y_2}(Y_2)\}$$

For a bivariate sample, spearman's rho is

$$\widehat{\rho_S}(\boldsymbol{Y}_{1:n}) = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} \left\{ \mathrm{rank}(Y_{i,1}) - \frac{n+1}{2} \right\} \left\{ \mathrm{rank}(Y_{i,2}) - \frac{n+1}{2} \right\}$$

The set of ranks for any variable is, of course, the integers 1 to n, and hence $(n+1)/2$ is the mean of its ranks.

If $\boldsymbol{Y} = (Y_1, ..., Y_d)$ is a random vector, then the **Spearman's correlation matrix $\boldsymbol{\Omega}_S$** of $\boldsymbol{Y}$ is the correlation matrix of $\{F_{Y_1}(Y_1), ..., F_{Y_d}(Y_d)\}$. The sample matrix is defined analogously.

7.5 Tail Dependence

Tail dependence measures association between the extreme values of two random variables and depends only on their copula. The **coefficient of lower tail dependence** is denoted by λ_l and defined as

$$\begin{aligned}
\lambda_l &:= \lim_{q \downarrow 0} P\{Y_2 \leq F_{Y_2}^{-1}(q) | Y_1 \leq F_{Y-1}^{-1}(q)\} \\
&= \lim_{q \downarrow 0} \frac{P\{Y_2 \leq F_{Y_2}^{-1}(q), Y_1 \leq F_{Y-1}^{-1}(q)\}}{P\{Y_1 \leq F_{Y_1}^{-1}(q)\}} \\
&= \lim_{q \downarrow 0} \frac{P\{F_{Y_2}(Y_2) \leq (q), F_{Y_1}(Y_1) \leq (q)\}}{P\{F_{Y_1}(Y_1) \leq q\}} \\
&= \lim_{q \downarrow 0} \frac{C_Y(q, q)}{q}
\end{aligned}$$

The **coefficient of upper tail dependence** λ_u is

$$\begin{aligned}
\lambda_u &:= \lim_{q \uparrow 1} P\{Y_2 \geq F_{Y_2}^{-1}(q) | Y_1 \geq F_{Y_1}^{-1}(q)\} \\
&= 2 - \lim_{q \uparrow q} \frac{1 - C_Y(q, q)}{1 - q}
\end{aligned}$$

Knowing whether or not there is tail dependence is important for risk management. If there are no tail dependencies among the returns on the assets in a portfolio, then there is little risk and simultaneous very negative returns, and the risk of an extreme negative return on the portfolio is low. Conversely, if there are tail dependencies, then the likelihood of extreme negative returns occurring simultaneously on several assets in the portfolio can be high.

7.6 Exercises with Solutions

7.6.1 Exercise 1

Kendall's tau rank correlation between X and Y is 0.55. Both X and Y are positive. What is Kendall's tau between X and $1/Y$?

Since Kendall's tau measures the difference in probabilities of concordance and discordance between two points, we know that transforming one of the variables with a monotonically decreasing function $1/Y$ will convert all of the inequalities and consequentially the concordant probability region to the discordant probability region, and vice versa. Thus $\rho_\tau(X, 1/Y) = 1 - 0.55 = 0.45$.

What is the Kendall's Tau between $1/X$ and $1/Y$?

Transforming both variables by the same monotonically decreasing function $1/x$ will preserve Kendall's tau obviously; thus $\rho_\tau(1/X, 1/Y) = 0.55$

7.6.2 Exercise 2

Suppose that X is Uniform$(0,1)$ and $Y = X^2$. Then the spearman rank correaltino and the Kendall's tau between X and Y will both equal 1, but the Pearson correlation between X and Y will be less than 1. Why?

Since Kendall's Tau is invariant for monotonic increasing transformations $(\rho_\tau\{(g(X), h(Y))\} = \rho_\tau(X, Y))$, $Y = X^2$ is strictly increasing on $[0,1)$, and thus Kendall's Tau clearly simplifies as follows:

$$\rho_\tau(X, Y) = \rho_\tau(X, X) = E[sign\{(X - X^*)(X - X^*)\} = E[sign\{(X - X^*)^2\}$$

$(X - X^*)^2$ is clearly positive for all values of X and X^*, thus $\rho_\tau(X, Y) = 1$.

The reason that Spearman's rho evaluates to 1 stems from the following relation:

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = F_X(\sqrt{y}) = \sqrt{y} = x$$

Thus values in X and Y take on the same value even when their CDF is applied. Thus we see that Spearman's rho evaluates as follows:

$$\rho_\tau(X, Y) = Corr\{F_X(X), F_Y(Y)\} = Corr\{X, X\} = \frac{Cov(X, X)}{Var(X)} = 1$$

Spearman's (standard) correlation coefficient however, is $\neq 1$ as we see below.

$$\begin{aligned}
Corr(X, X^2) &= \frac{Cov(X, X^2)}{\sqrt{Var(X)Var(X^2)}} && (55)\\
&= \frac{E(X^3) - E(X)(Var(X) + E(X)^2)}{\sqrt{Var(X)(E(X^4) - E(X^2)^2)}} && (56)\\
&= \frac{1/4 - 1/2(1/12 + 1/4)}{\sqrt{(1/12)(1/5 - (1/12 + 1/4)^2)}} && (57)\\
&\approx 0.9682 < 1 && (58)
\end{aligned}$$

7.6.3 Exercise 3

Show that an Archimedean copula with generator function $\phi(u) = -\log(u)$ is equal to the independence copula C_0. Does the same hold with $\phi(u) = -\log_{10}(u)$?

Generator function $\phi(u) = -\log(u)$ has inverse generator function $\phi^{-1}(y) = e^{-y}$. Hence the copula is

$$C(\boldsymbol{u}) = \phi^{-1}\{\phi(u_1) + \cdots \phi(u_d)\} = \exp\{-(-\log(u_1) - \cdots - \log(u_d)\} = u_1 \cdots u_d$$

which is the independence copula C_0. The same holds for any logarithm function and is not dependent on the base of the logarithm seen obviously through evaluation of a copula with a logarithmic generator.

7.6.4 Exercise 5

Show that the generator of a Frank copula satisfies the 3 assumptions of a strict generator.

Assumption 1 is not shown here (ϕ is continuous, strictly decreasing, and a convex mapping of $[0,1]$ onto $[0,\infty]$).

$\phi_{Fr}(0|\theta) = -\log\left\{\frac{e^{-\theta(0)}-1}{e^{-\theta}-1}\right\} = -\log(0) = \infty$ and $\phi_{Fr}(1|\theta) = -\log\left\{\frac{e^{-\theta(1)}-1}{e^{-\theta}-1}\right\} = -\log(1) = 0$

7.6.5 Exercise 8

7.6.6 Exercise 10

8 Time Series Basics

A process is said to be **strictly stationary** if for every m and n, $(Y_1, ..., Y_n)$ and $(Y_{1+m}, ..., Y_{n+m})$ have the same distribution; i.e. it does not depend on the origin.

A process is said to be **weakly stationary** if

- $E(Y_t) = \mu$ for all t
- $Var(Y_t) = \sigma^2$ for all t
- $Cov(Y_t, Y_s) = \gamma(|t-s|)$ for all t and s

The sequence $Y_1, Y_2, ...$ is **weak white noise** with mean μ and variance σ^2 if

- $E(Y_t) = \mu$ for all t
- $Var(Y_t) = \sigma^2$ for all t
- $Cov(Y_t, Y_s) = 0$ for all $t \neq s$

Because of the lack of dependence, past values of a white noise process contain no information that can be used to predict future values. More precisely, suppose that $Y_1, Y_2, ...$ is an i.i.d. WN(μ, σ^2) process. Then,

$$E(Y_{t+h}|Y_1, ..., Y_t) = \mu \text{ for all } h \geq 1$$

The above equation says that one cannot predict the future deviations of a white noise process from its mean.

8.1 Estimating Parameters

Suppose we observe $Y_1, ..., Y_n$ from a weakly stationary process. To estimate the mean μ and the variance σ^2, we use the sample mean $\bar{Y}$ and the sample variance s^2; however, to estimate the autocovariance function, we use the **sample autocovariance function**

$$\widehat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(Y_{t+h} - \bar{Y})(Y_t - \bar{Y}) = \frac{1}{n}\sum_{t=h+1}^{n}(Y_t - \bar{Y})(Y_{t-h} - \bar{Y})$$

From here, we estimate $\rho(.)$ using the **sample autocorrelation function** defined as

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}$$

8.1.1 Ljung-Box Test

The null hypothesis of the Ljung-Box test is $H_0 : \rho(1) = \rho(2) = ... = \rho(K) = 0$ for some K. If the test rejects, then we conclude that one or more of the ρ's is nonzero. This is implemented in R using the function Box.test().

8.2 ARIMA Processes

ARIMA models are needed to make stationary processes sufficiently parsimonious with a finite, small, number of parameters.

8.2.1 AR(p)

Let us first examine the AR process with $p = 1$. Let $\epsilon_1, \epsilon_2, ...$ be weak $WN(\mu, \sigma^2)$. We say that $Y_1, Y_2, ...$ is an **AR(1) process** if for some constant parameters μ and ϕ,

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \epsilon_t$$

for all t. We interpret the term $\phi(Y_{t-1} - \mu)$ as representing "memory" or "feedback" of the past into the present value of the process.

If $Y_1, Y_2, ...$ is a weakly stationary process, then $|\phi| < 1$. If $|\phi| < 1$, then repeated substitution of shows that

$$Y_t = \mu + \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + ... = \mu + \sum_{h=0}^{\infty}\phi^h\epsilon_{t-h}$$

The above equation is known as the **infinite moving average** $[MA(\infty)]$ representation of the process. The properties of the $AR(1)$ process are listed below.

$$E(Y_t) = \mu \ \forall t \tag{59}$$

$$Var(Y_t) = \gamma(0) = \sigma_Y^2 = \frac{\sigma_\epsilon^2}{1 - \phi^2} \ \forall t \tag{60}$$

$$Cov(Y_t, Y_{t+h}) = \gamma(h) = \phi^{|h|}\frac{\sigma_\epsilon^2}{1 - \phi^2} \ \forall t, h \tag{61}$$

$$Corr(Y_t, Y_{t+h}) = \rho(h) = \phi^{|h|} \ \forall t, h \tag{62}$$

For AR(1) processes:

- $|\phi| < 1$: stationary process

- $|\phi| = 1$: random walk
- $|\phi| > 1$: explosive

Residuals and Model Checking
Once μ and ϕ have been estimated, one can estimate the white noise process $\epsilon_1, ..., \epsilon_n$. Rearranging the model, we can write

$$\epsilon_t = (Y_t - \mu) - (Y_{t-1} - \mu)$$

and thus $\widehat{\epsilon}_2, \widehat{\epsilon}_3, ..., \widehat{\epsilon}_n$ are defined as

$$\widehat{\epsilon}_t = (Y_t - \widehat{\mu}) - \widehat{\phi}(Y_{t-1} - \widehat{\mu})$$

We have seen that the ACF of an $AR(1)$ decays geometrically to zero if $|\phi| < 1$ and also alternates sign if $\phi < 0$; however, this is a limited range of behavior and not representative of most time series. To extend the model so that it is **AR(p)**, we let the last p values of the the process feed into the current value of Y_t so that it takes the following form

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + ... + \phi_p(Y_{t-p} - \mu) + \epsilon_t$$

where $\epsilon_1, \epsilon_2, ...$ is weak WN$(0, \sigma_\epsilon^2)$.

If the $AR(p)$ model fits the time series well, then the residuals should look like white noise. Residual autocorrelation can be detected by examining the sample ACF of the residuals and using the Ljung-Box test.

8.2.2 MA(q)

The moving average process remedies the problem of using large values of p in AR processes. Let us first consider the pure moving average process with $q = 1$.

A process Y_t is a **moving average process** if Y_t can be expressed as a weighted average (moving average) of the past values of the white noise process $\{\epsilon_t\}$. The **MA(1)** process is

$$Y_t - \mu = \epsilon_t + \theta\epsilon_{t-1}$$

Properties of the MA(1) process:

1. $E(Y_t) = \mu$
2. $Var(Y_t) = \sigma_\epsilon^2(1 + \theta^2)$
3. $\gamma(1) = \theta\sigma_\epsilon^2$
4. $\gamma(h) = 0$ if $|h| > 1$
5. $\rho(1) = \frac{\theta}{1+\theta^2}$
6. $\rho(h) = 0$ if $|h| > 1$

The **MA(q)** process is

$$Y_t - \mu = \epsilon_t + \theta_1\epsilon_{t-1} + ... + \theta_1\epsilon_{t-q}$$

One can show that $\gamma(h) = 0$ and $\rho(h) = 0$ if $|h| > q$.

8.2.3 ARMA(p,q)

Stationary time series with complex autocorrelation behavior often are more parsimoniously modeled by mixed autoregressive and moving average processes than by either by themselves. Let us first introduce the **backwards operator** B

$$BY_t = Y_{t-1} \rightarrow B^h Y_t = Y_{t-h}$$

The **ARMA(p,q)** model is defined as

$$(1 - \phi_1 B - ... - \phi_p B^p)(Y_t - \mu) = (1 + \theta_1 B + ... + \theta_q B^q)\epsilon_t$$

Note that w hite noise process is $ARMA(0, 0)$ since if $p = q = 0$, then the model reduces to

$$(Y_t - \mu) = \epsilon_t$$

Consider the $ARMA(1, 1)$ model

$$Y_t = \phi Y_{t-1} + \theta\epsilon_{t-1} + \epsilon_t$$

Properties of the $ARMA(1, 1)$ model

1. $\gamma(0) = \frac{(1+\theta^2+2\phi\theta)\sigma_\epsilon^2}{1-\phi^2}$
2. $\rho(1) = \frac{(1+\phi\theta)(\phi+\theta)}{1+\theta^2+2\phi\theta}$
3. $\rho(h) = \phi\rho(h-1),\ h \geq 2$

Let us now introduce the **differencing operator**

$$\Delta Y_t = Y_t - BY_t = Y_t - Y_{t-1}$$

Δ^k is called the k^{th}**-order differencing operator** and can be defined using binomial expansion:

$$\Delta^k Y_t = (1 - B)^k Y_t = \sum_{l=0}^{k} \binom{k}{l} (-1)^l Y_{t-l}$$

8.2.4 ARIMA(p,d,q)

A time series Y_t is said to be an **ARIMA(p,d,q)** process if $\Delta^d Y_t$ is $ARMA(p, q)$. An $ARIMA(p, d, q)$ is stationary only if $d = 0$.

The **integral** of a process Y_t is the process w_t, where

$$w_t = w_{t_0} + Y_{t_0+1} + ... + Y_t,$$

Here t_0 is an arbitrary starting time pointa nd w_{t_0} is the starting value of the w_t process.

We will say that a process is **I(d)** if it is stationary after being differenced d times.

If $E(Y_t)$ has an m-th degree polynomial trend, then the mean of $E(\Delta^d Y_t)$ has an $(m - d)$th-degree trend for $d \leq m$. For $d > m, E(\Delta^d Y_t) = 0$ (proved in exercises).

8.3 Unit Root Tests

We have seen that it can be difficult to tell whether a time series is best modeled as stationary or nonstationary, so in order to help decide we invoke the power of hypothesis testing. What is meant by a unit root? Recall that an $ARMA(p,q)$ process can be written as

$$(Y_t - \mu) = \phi_1(Y_{t-1} - \mu) + ... + \phi_p(Y_{t-p} - \mu) + \epsilon_t + \theta_1\epsilon_{t-1} + ... + \theta_q\epsilon_{t-q}$$

The condition for $\{Y_t\}$ to be stationary is that all roots of the polynomial

$$1 - \phi_1 x - ... - \phi_p x^p$$

have absolute values greater than 1. We can generalize the above function to the following form

$$p(x) = b_0 + b_1 x + ... + b_p x^p$$

The fundamental theorem of algebra states that $p(x)$ can be factored as

$$b_p(x - r_1)(x - r_2)...(x - r_p)$$

where $r_1, ..., r_p$ are the roots of $p(x)$. In R, the roots of a polynomial can be found using the function polyroot().

Dickey-Fuller Test
The DF test is based on the $AR(1)$ model

$$Y_t = \phi Y_{t-1} + \epsilon_t$$

H_0 is that there exists a unit root ($\phi = 1$), and the alternative H_1 is stationary ($\phi < 1$). This model is equivalent to

$$\Delta Y_t = \pi Y_{t-1} + \epsilon_t$$

where $\pi = \phi - 1$. The ADF expands on this form by adding a time trend so that it is of the form

$$\Delta Y_t = \beta_0\beta_1 t + \pi Y_{t-1} + \sum_{j=1}^{p}\gamma_j \Delta Y_{t-j} + \epsilon_t$$

KPSS Test
The KPSS test is based on the following model

$$Y_t = \xi t + r_t + \epsilon_t, \quad r_t = r_{t-1} + u_t, \quad u_t \sim N(0, \sigma_u^2)$$

From the above information, we note that

$$r_t = r_0 + u_1 + ... + u_t \rightarrow Y_t = r_0 + \xi t + u_1 + ... + u_t + \epsilon_r$$

The KPSS test tests $H_0 : \sigma_u^2 = 0 \leftrightarrow u_1 = ... = u_t = 0$ against $H_1 : \sigma_u^2 > 0 \leftrightarrow u_i = 0$ for some i.

Automatic Selection of an ARIMA Model
auto.arima() in R will automatically select an $ARIMA(p,d,q)$ model for you by first applying the KPSS for $d = 0, 1, 2, ...$ until H_0 is accepted and then p, q are selected via AIC/BIC.

8.4 Forecasting

Forecasting means predicting future values of a time series using the current **information set**, which is the set of present and past values of the time series. For a general $AR(1)$ process, the k-step ahead forecast is given by

$$\widehat{Y}_{n+k} = \widehat{\mu} + \widehat{\phi}^k (Y_n - \widehat{\mu})$$

Note that we needn't include any future values of $_t$ since out best prediction of their value is 0. If $|\widehat{\phi}| < 1$, as is true for a stationary series, then as k increases, the forecasts will converge geometrically fast to $\widehat{\mu}$.

Forecasting $AR(p)$ processes is similar. For a k-step ahead forecast,

$$\widehat{Y}_{n+k} = \widehat{\mu} + \sum_{i=1}^{p} \widehat{\phi}_i (Y_{n+k-i} - \widehat{\mu}), \ k \geq 2$$

Forecasting ARMA and ARIMA is similar. To forecast one step ahead in an $ARMA(p,q)$ model,

$$\widehat{Y}_{n+1} = \widehat{\mu} + \widehat{\phi}(Y_n - \widehat{\mu}) + \widehat{\theta}\widehat{\epsilon}_n$$

To forecast k-steps ahead in an $ARMA(p,q)$ model,

$$\widehat{Y}_{n+k} = \widehat{\mu} + \sum_{i=1}^{p} \widehat{\phi}_i (Y_{n+k-i} - \widehat{\mu}) + \sum_{j=1}^{q} \widehat{\theta}_j \widehat{\epsilon}_{n+k-j}, \ k \geq 2$$

Notice that as long as $k > q$, the MA component drops to 0 in our forecast. To examine forecasting $ARIMA(p,d,q)$ models, let us consider the following example. Suppose that Y_t is $ARIMA(1,1,0)$ so that ΔY_t is $AR(1)$. To forecast Y_{n+k}, $k \geq 1$, one first fits an $AR(1)$ model to the ΔY_t process and forecasts ΔY_{n+k}, $k \geq 1$. Let the forecasts be denoted by $\widehat{\Delta Y}_{n+k}$, $k \geq 1$. Then,

$$Y_{n+1} = Y_n + \Delta Y_{n+1}$$

$$\widehat{Y}_{n+1} = Y_n + \widehat{\Delta Y}_{n+1}$$

$$\widehat{Y}_{n+2} = \widehat{Y}_{n+1} + \widehat{\Delta Y}_{n+2} = Y_n + \widehat{\Delta Y}_{n+1} + \widehat{\Delta Y}_{n+2}$$

and so on.

Forecast Errors
The variance of forecast errors in an $AR(1)$ process converges to $\gamma(0)$, the marginal covariance of the $AR(1)$ process. This is an example of the general principle that for any stationary ARMA process, the variance of the forecast error converges to the marginal variance.

Forecasting with Simulation
There are two important advantages to using simluation for forecasting:

1. Simulation can be used in situations where standard software does not compute forecast limits

2. Simulation does not require that the noise series be Gaussian

9 Time Series Further Topics

9.1 Seasonal and Multiplicative ARIMA Models

To remove seasonal nonstationarity, one uses seasonal differencing. Let s be the period. Defined $\Delta_s = 1 - B^s$ so that $\Delta_s Y_t = Y_t - Y_{t-s}$. Be careful to distinguish between Δ_s and $\Delta^s = (1-B)^s$. The series $\Delta_s Y_t$ is called the **seasonally differenced series**. These operators commute so that

$$\Delta(\Delta_s Y_t) = \Delta_s(\Delta Y_t)$$

A multiplicative model **ARIMA{**$(p, d, q) \times (p_s, d_s, q_s)_s$**}** is defined as

$$(1-\phi_1 B-\cdots-\phi_p B^p)\{1-\phi_1^* B^s-\cdots-\phi_{p_s}^*(B^s)^{p_s}\}\{\Delta^d(\Delta_s^{d_s} Y_t)-\mu\} = (1+\theta_1 B+\cdots+\theta_q B^q)\{1+\theta_q^* B^s+\cdots+\theta_{q_s}*(B^s)^{q_s}\}\epsilon_t$$

Box-Cox Transformation for Time Series
Although a transformation can be selected by trial-and-error, another possibility is automatic selection by maximum likelihood estimation using the model

$$(\Delta^d Y_t^{(\alpha)} - \mu) = \phi_1(\Delta^d Y_{t-1}^{(\alpha)} - \mu) + \cdots + \phi_p(\Delta^d Y_{t-p}^{(\alpha)} - \mu) + \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q}$$

This model states that Y_t follows an ARIMA model with Gaussian noise that has a constant variance after transformation.

9.2 Time Series and Regression

In a multiple linear regression $Y_i = \beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p} + \epsilon_i$, the errors ϵ_i are assumed to be mutually independent. However, if the data $\{(X_i, Y_i), i = 1, ..n\}$ are time series, then it is likely that the errors are correlated, a problem we will call **residual correlation**. Residual correlation causes standard errors and confidence intervals (which incorrectly assume uncorrelated noise) to be incorrect. This is later solved by correcting/adjusting the estimated covariance matrix of the coefficient estimates.

In the extreme case where the residuals are an integrated process, the least-squares estimator is inconsistent, meaning that it will not converge to the true parameter as the sample size converges to ∞. If an $I(1)$ process is regressed on another $I(1)$ process and the two processes are independent (so that the regression coefficient is 0), it is quite possible to obtain a highly significant result, that is, to strongly reject the true null that the regression coefficient is 0. This is called **spurious regression**.

The problem with spurious regression is that the test is based on the incorrect assumption of independent error. The residuals of the problem of correlated noise can be detected by looking at the sample ACF of the residuals and using a **Durbin-Watson Test**. The DW test has a null that the first p autocorrelation coefficients are all 0, where p can be selected by the user in the durbinWatsonTest() in the car package (p is called max.lag)

9.2.1 Linear Regression with ARMA Errors

When residual analysis shows that the residuals are correlated, then one of the key assumptions of the linear model does not hold, and test and confidence intervals based on this assumption are invalid and cannot be trusted. Solution: replace the assumption of independent noise by the weaker assumption that the noise process is stationary but possibly correlated. Assuming that the noise is an ARMA process is an approach referred to as the **ARMAX Model** where the X indicates the inclusion of exogenous regression variables. It is defined as follows

$$Y_t = \beta_0 + \beta_1 X_{t,1} + ... + \beta_p X_{t,p} + \epsilon_t$$

where

$$(1-\phi_1 B - ... - \phi_p B^p)\epsilon_t = (1+\theta_1 B + ... + \theta_q B^q)u_t$$

and $\{u_t\}$ is white noise.

9.3 Multivariate Time Series

Define a d-dimensional random vector $Y_t = (Y_{1,t}, ..., Y_{d,t})^T$ representing quantities that were measured at time t. This is called a **multivariate time series**.

9.3.1 The Cross-Correlation Function (CCF)

Suppose that Y_j and Y_i are the two component series of a stationary multivariate time series. The **cross correlation function** between Y_j and Y_i is defined as

$$\rho_{Y_j,Y_i}(h) = Corr\{Y_j(t), Y_i(t-h)\}$$

This represents the correlation between Y_j at time t and Y_i at h steps earlier. Unlike the ACF, the CCF is not symmetric in lag h, that is, $\rho_{Y_j,Y_i}(h) \neq \rho_{Y_j,Y_i}(-h)$. Instead, $\rho_{Y_j,Y_i}(h) \neq \rho_{Y_i,Y_j}(-h)$. Cross-correlations can suggest how the component series might be influencing each other or might be influenced by a common factor. Like all correlations, cross-correlations only show statistical association, not casuation, but a causal relationship might be deduced from other knowledge.

9.3.2 Multivariate White Noise

A d-dimensional multivariate time series $Y_1, Y_2, ...$ is weak WN($\boldsymbol{\mu}, \boldsymbol{\Sigma}$) process if

1. $E(\boldsymbol{Y}_t) = \boldsymbol{\mu}\ \forall t$
2. $COV(\boldsymbol{Y}_t) = \boldsymbol{\Sigma}\ \forall t$
3. $\forall\ t \neq s$, all components of $\boldsymbol{Y}_t$ are uncorrelated with all components of $\boldsymbol{Y}_s$

Notice that if $\boldsymbol{\Sigma}$ is not diagonal, then there is cross-correlation between the components of $\boldsymbol{Y}_t$ because $Corr(\boldsymbol{Y}_{j,t}, \boldsymbol{Y}_{i,s}) = \boldsymbol{\Sigma}_{j,i}$; in other words, there may be nonzero **contemporaneous** correlations.

9.3.3 Multivariate ACF Plots and Multivariate Ljung-Box Test

The ACF for multivariate time series includes the d marginal ACFs for each univariate series $\{\rho_{Y_i}(h) | i = 1, ..., d\}$ and the $d(d-1)/2$ CCFs for all unordered pairs of the univariate series $\{\rho_{Y_j,Y_i}(h) | 1 \leq j < i \leq d\}$. In R, $n \times 1$ vectors can be combined using the cbind() function which then makes them compatible for acf().

Let $\rho(h)$ denote the $d \times d$ lag-h cross-correlation matrix for a d-dimensional multivariate times series. The null hypothesis of the multivariate Ljung-Box test is $H_0 : \rho(1) = \rho(2) = ... = \rho(K) = 0$ for some K. If the multivariate Ljung-Box test rejects, then we conclude that one or more correlation is non-zero.

9.3.4 Multivariate ARMA

A d-dimensional multivariate times series $\boldsymbol{Y_1, Y_2, ...}$ is a multivariate ARMA(p, q) process with mean $\boldsymbol{\mu}$ if for $d \times d$ matrices $\boldsymbol{\Phi_1, ..., \Phi_p}$ and $\boldsymbol{\Theta_1, ..., \Theta_q}$,

$$\boldsymbol{Y}_t - \boldsymbol{\mu} = \boldsymbol{\Phi_1}(\boldsymbol{Y_{t-1}} - \boldsymbol{\mu}) + ... + \boldsymbol{\Phi_p}(\boldsymbol{Y_{t-p}} - \boldsymbol{\mu}) + \boldsymbol{\epsilon}_t + \boldsymbol{\Theta_1}\boldsymbol{\epsilon}_{t-1} + ... + \boldsymbol{\Theta_q}\boldsymbol{\epsilon_{t-q}}$$

where $\boldsymbol{\epsilon_1, ..., \epsilon_n}$ is a multivariate weak WN$(\mathbf{0}, \mathbf{\Sigma})$ process.

For a d-dimensional AR(1) process,

$$E(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-k}) = \mathbf{\Phi}^k \boldsymbol{Y}_{t-k} \ \forall \ k > 0$$

It can be shown that eigenvectors of $\mathbf{\Phi}$ which have magnitude > 1 cause the mean to explode.

Prediction Using VARMA

1. AR(p): $\widehat{\boldsymbol{Y}}_{n+h} = \widehat{\boldsymbol{\mu}} + \widehat{\mathbf{\Phi}}_1(\widehat{\boldsymbol{Y}}_{n+h-1} - \widehat{\boldsymbol{\mu}}) + ... + \widehat{\mathbf{\Phi}}_p(\widehat{\boldsymbol{Y}}_{n+h-p} - \widehat{\boldsymbol{\mu}})$
2. AR(1): $\widehat{\boldsymbol{Y}}_{n+h} = \widehat{\boldsymbol{\mu}} + \widehat{\mathbf{\Phi}}_1^h(\boldsymbol{Y}_n - \widehat{\boldsymbol{\mu}})$

9.4 Long-Memory Processes

Stationary ARMA processes have only short memories in that their autocorrelation functions decay to zero exponentially fast. That is, $\exists \ D > 0$ and $r < 1$ such that

$$\rho(k) < D|r|^k \quad \forall \, k$$

In contrast, many financial time series appear to have long memory since their ACFs decay at a (slow) polynomial rate rather than a (fast) geometric rate. That is,

$$\rho(k) \sim Dk^{-\alpha}$$

for some D and $\alpha > 0$. A polynomial rate of decay is sometimes called a hyperbolic rate.

Fractional Differencing
For integer values of d, we have

$$\Delta^d = (1-B)^d = \sum_{k=0}^{d} \binom{d}{k} (-B)^k$$

We can extend this to non-integer d however. We define

$$\binom{d}{k} = \frac{d(d-1)\cdots(d-k+1)}{k!} \quad \forall \ d > -1, k \geq 0$$

If $d > -1$, $\Delta^d = (1-B)^d$ is always defined such that

$$\Delta^d Y_t = \sum_{k=0}^{\infty} \binom{d}{k} (-1)^k Y_{t-k}$$

9.4.1 FARIMA Processes

A process Y_t is a **FARIMA**(p, d, q) process if $\Delta^d Y_t$ is an ARMA(p, q) process. We say that Y_t is a fractionally integrated process of order d, or more simply, an $I(d)$ process. In these processes, $d = 0$ is the ordinary ARMA case, but d can be negative now. If $-1/2 < d < 1/2$, then the process is stationary; furthermore, if $0 < d < 1/2$, then it is a long-memory stationary process. If $d > 1/2$, then Y_t can be differenced an integer numbe rof times to become a stationary process, though perhaps with long-memory. Suppose $1/2 < d < 3/2$. This means that ΔY_t is fractionally integrated of order $d-1 \in (-1/2, 1/2)$ and ΔY_t has long-memory if $1 < d < 3/2$ so that $d-1 \in (0, 1/2)$.

In R, the function fracdiff() from the fracdiff package will fit a FARIMA(p, d, q) process where p, d, q must be input.

9.5 Exercises with Solutions

10 Cointegration

Cointegration analysis is a technique that is frequently applied in econometrics. In finance it can be used to find trading strategies based on mean-reversion. The **cointegrating vector**is the vector of coefficients which is applied to two or more assets that achieves stationarity.

A vector $\boldsymbol{Y}_t = (Y_{1,t}, ..., Y_{d,t})^T$ is cointegrated if

1. $Y_{d,t}$ is $I(1), j = 1, ..., d$
2. $\exists\, \lambda = (\lambda_1, ..., \lambda_d)^T \neq \mathbf{0}$ such that $\lambda^T \boldsymbol{Y}_t = \lambda_1 Y_{1,t} + ... + \lambda_d Y_{d,t}$ is $I(0)$

Phillips-Ouliaris Test
The Phillips-Ouliaris test regresses one integrated series on others and aplpies the Phillips-Perron unit root test to the residuals. H_0 is unit root nonstationary, which implies that the series are **not** cointegrated. Consequently, a small p-value implies that the series **are** cointegrated and therefore sutiable for regresson analysis.

In R, the Phillips-Ouliaris test is po.test() in the teseries package.

10.1 Vector Error Correction Models (VECM)

The idea behind error correction is simplest when there are only two series, $Y_{1,t}$ and $Y_{2,t}$. In this case, the error correction model is

$$\Delta Y_{1,t} = \phi_1(Y_{1,t-1} - \lambda Y_{2,t-1}) + \epsilon_{1,t}$$

$$\Delta Y_{2,t} = \phi_2(Y_{1,t-1} - \lambda Y_{2,t-1}) + \epsilon_{2,t}$$

where the ϵ's are white noise.Subtracting λ times the second equation from the first equation gives

$$\Delta(Y_{1,t} - \lambda Y_{2,t}) = (\phi_1 - \lambda\phi_2)(Y_{1,t-1} - \lambda Y_{2,t-1}) + (\epsilon_{1,t} - \lambda\epsilon_{2,t})$$

Let $\mathcal{F}_t$ denote the information set at time t. If $\phi_1 - \lambda\phi_2 < 0$, then $E[\Delta(Y_{1,t} - \lambda Y_{2,t})|\mathcal{F}_t]$ is opposite sign to $Y_{1,t-1} - \lambda Y_{2,t-1}$. This causes error correction because whenever $Y_{1,t-1} - \lambda Y_{2,t-1}$ is positive, it's expected change is negative and vice versa.

Rearranging that last equation, you can show that $Y_{1,t-1} - \lambda Y_{2,t-1}$ follows an $AR(1)$ process with coefficient $1 + \phi_1 - \lambda\phi_2$.

- $\phi_1 - \lambda\phi_2 > 0 \rightarrow 1 + \phi_1 - \lambda\phi_2 > 1 \rightarrow Y_{1,t-1} - \lambda Y_{2,t-1}$ is explosive
- $\phi_1 - \lambda\phi_2 = 0 \rightarrow 1 + \phi_1 - \lambda\phi_2 = 1 \rightarrow Y_{1,t-1} - \lambda Y_{2,t-1}$ is a random walk
- $-2 < \phi_1 - \lambda\phi_2 < 0 \rightarrow 1 + \phi_1 - \lambda\phi_2 < 1 \rightarrow Y_{1,t-1} - \lambda Y_{2,t-1}$ is stationary

The lower bound of -2 for the last scenario represents the point at which you are "over-correcting", which leads to diversion.

To generalize to more than two series, Let $\boldsymbol{Y}_t = (Y_{1,t}, Y_{2,t})^T$ and define the residuals analogously. Then

$$\Delta \boldsymbol{Y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}^T \boldsymbol{Y}_{t-1} + \boldsymbol{\epsilon}_t$$

where $\boldsymbol{\alpha} = (\phi_1, \phi_2)^T$ and the cointegrating vector $\boldsymbol{\beta} = (1, -\lambda)^T$. Here, $\boldsymbol{\alpha}$ specifies the speed of mean-reversion and is called the **loading matrix**.

VAR(p) Model

$$\Delta \boldsymbol{Y}_t = \boldsymbol{\Gamma}_1 \Delta \boldsymbol{Y}_{t-1} + ... + \boldsymbol{\Gamma}_{p-1} \Delta \boldsymbol{Y}_{t-p+1} + \boldsymbol{\Pi} \boldsymbol{Y}_{t-1} + \boldsymbol{\mu} + \boldsymbol{\Phi} \boldsymbol{D}_t + \boldsymbol{\epsilon}_t$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{D}_t$ is a vector of nonstochastic regressors, and

$$\boldsymbol{\Pi} = \boldsymbol{\alpha} \boldsymbol{\beta^T}$$

It is easy to show that the columns of $\boldsymbol{\beta}$ are the cointegration vectors. Since $\boldsymbol{Y}_t$ is $I(1)$, $\Delta \boldsymbol{Y}_t$ on the left hand side is $I(0)$ and therefore $\boldsymbol{\Pi} \boldsymbol{Y}_{t-1} = \boldsymbol{\alpha} \boldsymbol{\beta^T} \boldsymbol{Y}_{t-1}$ is also $I(0)$. It follows that each of the r components of $\boldsymbol{\beta} \boldsymbol{Y}_{t-1}$ is $I(0)$.

10.2 Exercises with Solutions

11 GARCH Models

Financial market data often exhibits volatility clustering, where time series show periods of high volatility and periods of low volatility. ARMA models are used to model the conditional expectation of a process given the past, but in an ARMA model the conditional variance given the past is constant. What does this mean for, say, modeling stock returns? Suppose we have noticed that recent daily returns have been unusually volatile. We might expect that tomorrow's return is also more variable than usual; however, an ARMA model cannot capture this type of behavior because its conditional variance is constant. So we need better time series models if we want to model the non-constant volatility which leads us to GARCH models.

GARCH is an acronym meaning Generalized Auto-Regressive Conditional Heteroskedasticity. In these models the conditional variance has a structure very similar to the structure of the conditional expectation in an AR model.

11.1 Estimating Conditional Means and Variances

Consider regression modeling with a constant conditional variance, $Var(Y_t|X_{1,t}, ..., X_{p,t}) = \sigma^2$. Then the general form for the regression of Y_t on $X_{1,t}, ..., X_{p,t}$ is

$$Y_t = f(X_{1,t}, ..., X_{p,t}) + \epsilon_t$$

where ϵ_t is independent of $X_{1,t}, ..., X_{p,t}$ and has expectation equal to 0 and a constant conditional variance σ_ϵ^2. The function $f(\cdot)$ is the conditional expectation of Y_t given $X_{1,t}, ..., X_{p,t}$. Moreover, the conditional variance of Y_t is σ_ϵ^2. The above equation can be modified to allow conditional heteroskedasticity. Let $\sigma^2(X_{1,t}, ..., X_{p,t})$ be the conditional variance of Y_t given $X_{1,t}, ..., X_{p,t}$. Then the model

$$Y_t = f(X_{1,t}, .., X_{p,t}) + \epsilon_t \sigma(X_{1,t}, ..., X_{p,t})$$

where ϵ_t has conditional mean equal to 0 and conditional variance equal to 1, gives the correct conditional variance of Y_t.

11.2 ARCH(1) Processes

Suppose for now that $\epsilon_1, \epsilon_2, ...$ is Gaussian white noise with unit variance. Then

$$E(\epsilon_t|\epsilon_{t-1}, ...) = 0 \text{ and } \mathrm{Var}(\epsilon_t|\epsilon_{t-1}, ...) = 1$$

The process a_t is an ARCH(1) process under the model

$$a_t = \epsilon_t \sqrt{\omega + \alpha a_{t-1}^2}$$

Note that this model is a special case of generalized conditional heteroskedasticity models with $f = 0$ and $\sigma = \sqrt{\omega + \alpha a_{t-1}^2}$. It is required that $\omega > 0$ and $\alpha \geq 0$ so that $\omega + \alpha a_{t-1}^2 > 0$ for all t. It is also required that $\alpha < 1$ in order for $\{a_t\}$ to be stationary with a finite variance. Define

$$\sigma_t^2 = \text{Var}(a_t | a_{t-1}, ...)$$

to be the conditional varaince of a_t given past values. Since ϵ_t is independent of a_{t-1} and $E(\epsilon_t^2) = \text{Var}(\epsilon_t) = 1$, we have

$$E(a_t | a_{t-1}, ...) = 0$$

and

$$\begin{aligned} \sigma_t^2 &= E\{(\omega + \alpha a_{t-1}^2)\epsilon_t^2 | a_{t-1}, a_{t-2}, ...\} \\ &= (\omega + \alpha a_{t-1}^2) E\{\epsilon_t^2 | a_{t-1}, a_{t-2}, ...\} \\ &= \omega + \alpha a_{t-1}^2 \end{aligned}$$

The above result is crucial to the mechanics of GARCH processes. If a_{t-1} has an unusually large absolute value, then σ_t is larger than usual and so a_t is also expected to have an unusually large magnitude. This volatility propagates since when a_t has a large magnitude that makes σ_{t+1}^2 large, then a_{t+1} tends to be large in magnitude, and so on. This logic also holds when the variability is unexpectedly small. Thus, unusual volatility in a_t tends to persist, though not forever. The conditional variance tends to revert to the unconditional variance provided that $\alpha < 1$, so that the process is stationary with a finite variance.

The unconditional, that is, marginal, variance of a_t denoted by $\gamma_a(0)$ is obtained by taking expectations of $(\omega + \alpha a_{t-1}^2)$, which gives us

$$\gamma_a(0) = \omega + \alpha \gamma_a(0)$$

for a stationary model. This has a positive solution $\gamma_a(0) = \omega/(1 - \alpha)$ provided $\alpha < 1$. At $\alpha = 1$ the unconditional variance is infinite and you end up with what is known as the integrated GARCH or I-GARCH model.

Straightforward calculations using $E(a_t | a_{t-1}, ...) = 0$ show that the ACF of a_t is

$$\rho_a(h) = 0 \text{ if } h \neq 0$$

Although a_t is an uncorrelated process, the process a_t^2 has a more interesting ACF. If $\alpha < 1$, then

$$\rho_{a^2}(h) = \alpha^{|h|} \text{ for all } h$$

If $\alpha \geq 1$, then a_t^2 is either non-stationary or has an infinite variance, so it does not have an ACF. This geometric decay in the ACF of a_t^2 for an ARCH(1) process is analogous to the geometric decay in the ACF of an AR(1) process. To complete the analogy, define

$$\eta_t = a_t^2 - \sigma_t^2$$

and note that $\{\eta_t\}$ is a mean zero weak white noise process, but not an i.i.d. white noise process. Adding η_t to $(\sigma_t^2 = \omega + \alpha a_{t-1}^2)$ we have

$$\sigma_t^2 + \eta_t = a_t^2 = \omega + \alpha a_{t-1}^2 + \eta_t$$

which is a direct representation of $\{a_t^2\}$ as an AR(1) process with $\mu = \omega$, $\phi = \alpha$, and $\epsilon_t = \eta_t$.

11.3 The AR(1) + ARCH(1) Model

Let a_t be an ARCH(1) process so that $a_t = \epsilon_t \sqrt{\omega + \alpha a_{t-1}^2}$, where ϵ_t is i.i.d. $N(0, 1)$, and suppose that

$$y_t - \mu = \phi(y_{t-1} - \mu) + a_t$$

Noe that y_t is an AR(1) process, except the noise term is not i.i.d. white noise, but rather an ARCH(1) process which is only weak white noise.

Because a_t is an uncorrelated process, it has the same ACF as independent white noise, and therefore, y_t has the same ACF as an AR(1) process with independent white noise

$$\rho_y(h) = \phi^{|h|} \text{ for all } h$$

in the stationary case. Moreover, a_t^2 has the ARCH(1) ACF:

$$\rho_{a^2}(h) = \alpha^{|h|} \text{ for all } h$$

11.4 ARCH(p) Models

As before, let ϵ_t be Gaussian white noise with unit variance. Then a_t is an ARCH(p) process if

$$a_t = \sigma_t \epsilon_t, \text{ where } \sigma_t = \sqrt{\omega + \sum_{i=1}^{p} \alpha_i a_{t-i}^2}$$

is the conditional standard deviation of a_t given the past values $a_{t-1}, a_{t-2}, ...$ of this process. The ARCH(p) process is uncorrelated and has a constant mean (both conditional and unconditional) and a constant unconditional variance.

11.5 ARIMA(p_M, d, q_M) + GARCH(p_V, q_V) Models

The GARCH(p, q) model is

$$a_t = \sigma_t \epsilon_t, \text{ where } \sigma_t = \sqrt{\omega + \sum_{i=1}^{p} \alpha_i a_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2}$$

Because past values of the σ_t process are fed back into the present value, the conditional standard deviation can exhibit more persistent periods of high or low volatility than seen in an ARCH process. In the stationary case, the process a_t is uncorrelated with a constant unconditional mean and variance and a_t^2 has an ACF like that of an ARMA process.

A very general time series model lets a_t be GARCH(p_V, q_V) and uses a_t as the noise term in an ARIMA(p_M, d, q_M) model. The subscripts distinguish between the conditional variance (V) or GARCH parameters and the conditional mean (M) or ARIMA parameters.

11.5.1 Two Types of Residuals

When one fits a model of this type to a time series Y_t, there is an ordinary residual, denoted $\widehat{a}_t$, which is the difference between Y_t and its conditional expectation, and there is also a standardized residual, denoted $\widehat{\epsilon}_t$, which is $\widehat{a}_t$ divided by its estimated conditional standard deviation $\widehat{\sigma}_t$. The standardized residuals are used for model checking, meaning that $\widehat{\epsilon}_t$ nor its square $\widehat{\epsilon}_t^2$ show serial correlation.

11.6 GARCH Models as ARMA Models

Assume that

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i a_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

To simplify notation, if $q > p$, then define $\alpha_i = 0$ for $i = p+1, ..., q$. Similarly, if $p > q$, then define $\beta_j = 0$ for $j = q+1, ..., p$. Define

$$\nu = \frac{\omega}{1 - \sum_{i=1}^{max(p,q)} (\alpha_i + \beta_i)}$$

Straightforward algebra shows that

$$a_t^2 - \nu = \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i)(a_{t-i}^2 - \nu) - \sum_{j=1}^{q} \beta_j \eta_{t-j} + \eta_t$$

so that a_t^2 is an ARMA$(\max(p,q), q)$ process with mean $\mu = \nu$, AR coefficients $\phi_i = \alpha_i + \beta_i$ and MA coefficients $\theta_j = -\beta_j$. As a byproduct of these calculations, we obtain a necessary condition for a_t to be stationary:

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

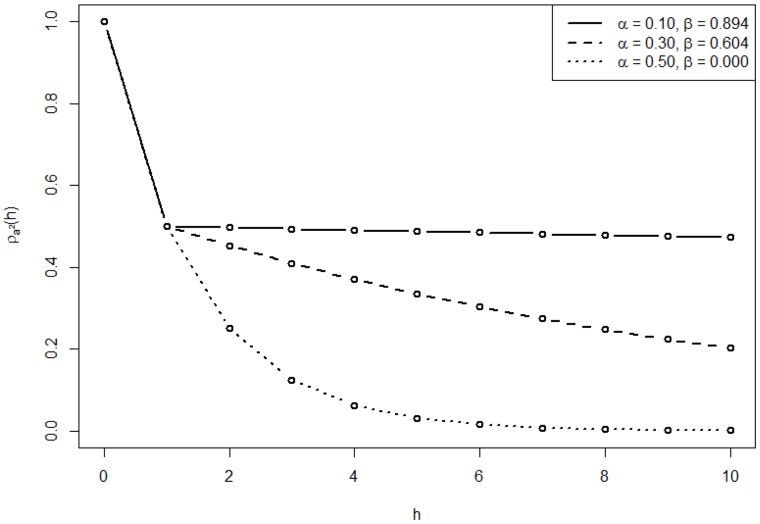
11.7 GARCH(1,1) Processes

If a_t is GARCH$(1,1)$, then as we have just seen, a_t^2 is ARMA$(1,1)$. Therefore, the ACF of a_t^2 can be written with GARCH parameters as follows:

$$\rho_{a^2}(h) = \frac{\alpha(1 - \alpha\beta - \beta^2)}{1 - 2\alpha\beta - \beta^2},\ h = 1$$

$$\rho_{a^2}(h) = (\alpha + \beta)^{h-1} \rho_{a^2}(1),\ h \geq 2$$

These formulas also hold in an AR(1) + GARCH(1,1) model, and the ACF of y_t^2 also decays with $h \geq 2$ at a geometric rate in the stationary case, provided some additional assumptions hold.



The capability of the GARCH$(1,1)$ model to fit the lag-1 autocorrelation and the subsequent rate of decay separately is important in practice. It appears to be the main reason that the GARCH$(1,1)$ model fits so many financial time series.

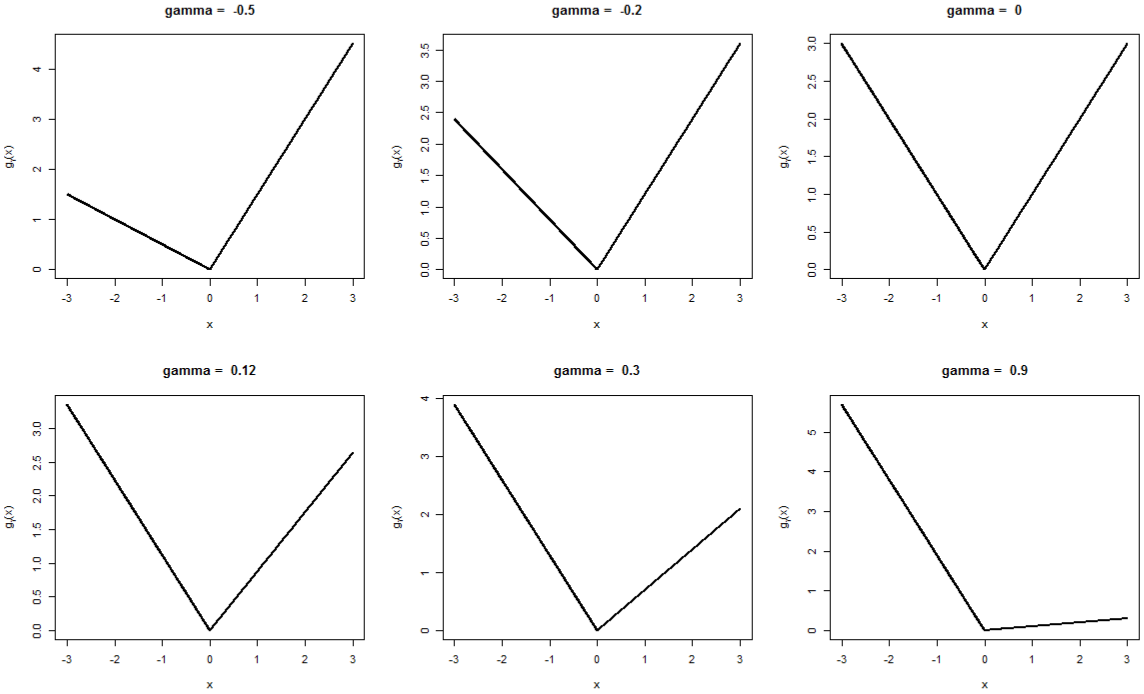
11.8 APARCH Models

In some financial time series, large negative returns appear to increase volatility more than do positive returns of the same magnitude. This is called the **leverage effect**. Standard GARCH models cannot model the leverage effect because they model σ_t as a function of past values of a_t^2–whether the past values of a_t are positive or negative is not taken into account. The problem here is that the square function x^2 is symmetric in x. The solution is to replace the square function with a flexible class of non-negative functions that include asymmetric functions. Asymmetric Power ARCH (APARCH) models do this by modeling σ_t^δ, where $\delta > 0$ is a new parameter.

The APARCH(p,q) model for the conditional standard deviation is

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

where $\delta > 0$ and $-1 < \gamma_i < 1, i = 1, ..., p$. The effect of a_{t-i} upon σ_t is through the function g_{γ_t}, where $g_\gamma(x) = |x| - \gamma x$. When $\gamma > 0$, $g_\gamma(-x) > g_\gamma(x)$ for any $x > 0$, so there is a leverage effect. If $\gamma < 0$, then there is a leverage effect in the opposite direction to what is expected–positive past values of a_t increase volatility more than negative past values of the same magnitude.



11.9 Linear Regression with ARMA+GARCH Errors

Consider the following model with ARMA disturbances:

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \cdots + \beta_p X_{t,p} + e_t$$

where

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(e_t - \mu) = (1 + \theta_1 B + \cdots + \theta_q B^q)a_t$$

and $\{a_t\}$ is i.i.d. white noise. This model is sufficient for serially correlated errors, but it does not accommodate volatility clustering ,which is often found in the residuals. Thus, we model the noise as an ARMA+GARCH process. Therefore, we will now assume that, instead of being i.i.d. white noise, $\{a_t\}$ is a GARCH process so that

$$a_t = \sigma_t \epsilon_t$$

where

$$\sigma_t = \sqrt{\omega + \sum_{i=1}^{p} \alpha_i a_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2}$$

and $\{\epsilon_t\}$ is i.i.d. white noise.

11.10 Forecasting ARMA+GARCH Processes

Let us first compare the prediction of a Gaussian white noise process and the prediction of a GARCH$(1,1)$ process with Gaussian innovations. Both have an ARMA$(0,0)$ model for the conditional mean so their forecasts are equal to the marginal mean, which will be called μ. For Gaussian white noise, the prediction limits are $\mu \pm z_{\alpha/2}$, where σ is the marginal standard deviation. For a GARCH$(1,1)$ process $\{Y_t\}$, the prediction limits at time origin n for h-steps ahead forecasting are $\mu \pm z_{\alpha/2}\sigma_{n+h|n}$ where $\sigma_{n+h|n}$ is the conditional standard deviation of Y_{n+h} given the information available at time n. As h increases, $\sigma_{n+h|n}$ converges to σ, so for long lead times the prediction intervals for the two models are similar.

11.11 Multivariate GARCH Processes

Financial asset returns tend to move together over time, as do their respective volatilities, across both assets and markets. Multivariate modeling of the volatility matrix proves difficult, however.

Analogous to positivity constraints in univariate GARCH models, a well-defined multivariate volatility matrix process must be positive-definite at each point in time, and model-based forecasts should as well. From a practical perspective, a well-defined inverse of a volatility matrix is frequently needed in applications. Additionally, a positive conditional variance estimate for a portfolio's return, which are a linear combination of asset returns, is essential; fortunately, this is guaranteed by positive definiteness.

11.11.1 Basic Setting

Let $\boldsymbol{Y}_t = (Y_{1,t}, ..., Y_{d,t})'$ denote a d-dimensional vector process and let $\mathcal{F}_t$ denote the information set at time index t, generated by $\boldsymbol{Y}_t$, $\boldsymbol{Y}_{t-1}$,... . We may partition the process as

$$\boldsymbol{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{a}_t$$

in which $\boldsymbol{\mu}_t = E(\boldsymbol{Y}_t|\mathcal{F}_{t-1})$ is the conditional mean vector at time index t, and $\{\boldsymbol{a}_t\}$ is the mean zero weak white noise innovation vector process with unconditional covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{a}} = \mathrm{Cov}(\boldsymbol{Y}_t|\mathcal{F}_{t-1})$. Let

$$\boldsymbol{\Sigma}_t = \mathrm{Cov}(\boldsymbol{a}_t|\mathcal{F}_{t-1}) = \mathrm{Cov}(\boldsymbol{Y}_t|\mathcal{F}_{t-1})$$

denote the conditional covariance matrix or volatility matrix at time index t. Multivariate time series modeling is primarily concerned with the time evolutions of $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$, the conditional mean and conditional covariance matrix.

We assume that $\boldsymbol{\mu}_t$ follows a stationary VAR(p) model with

$$\boldsymbol{\mu}_t = \boldsymbol{\mu} + \sum_{l=1}^{p} \boldsymbol{\Phi}_l (\boldsymbol{Y}_{t-l} - \boldsymbol{\mu})$$

where p is a non-negative integer, $\boldsymbol{\mu}$ is the $d \times 1$ unconditional mean vector, and the $\boldsymbol{\Phi}_l$ are $d \times d$ coefficient matrices, respectively.

The relationship between the innovation process and the volatility process is defined by

$$\boldsymbol{a}_t = \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim F(\boldsymbol{0}, \boldsymbol{I}_d)$$

in which $\boldsymbol{\Sigma}_t^{1/2}$ is a symmetric **matrix square root** of $\boldsymbol{\Sigma}_t$, such that $\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\Sigma}_t^{1/2} = \boldsymbol{\Sigma}_t$. The iid white noise $\boldsymbol{\epsilon}_t$ are **standardized** innovations from a multivariate distribution F with mean zero and a covariance matrix equal to the identity.

12 Factor Models and Principal Component Analysis

High-dimensional data can be challenging to analyze. They are difficult to visualize, need extensive computer resources, and often require special statistical methodology. Fortunately, in many practical applications, high-dimensional data have most of their variation in a lower-dimensional space that cab be found using **dimension reduction techniques**.

12.1 Principal Component Analysis (PCA)

PCA finds structure in the covariance or correlation matrix and uses this structure to locate low-dimensional subspaces containing most of the variation in the data. PCA starts with a sample $\boldsymbol{Y}_i = (Y_{i,1}, ..., Y_{i,d}), i = 1, ..., n$ of d-dimensional random vectors with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Consider the following model $\boldsymbol{Y}_i = \boldsymbol{\mu} + W_i \boldsymbol{o} + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i$ is a random vector uncorrelated with W_i having a "small" covariance matrix. Then most of the variation among the $\boldsymbol{Y}_i - \boldsymbol{\mu}$ vectors is in the space spanned by $\boldsymbol{o}$, but there is small variation in other directions due to $\boldsymbol{\epsilon}_i$. For simplicity of notation, we assume that the mean $\bar{\boldsymbol{Y}}$ has been subtracted from each $\boldsymbol{Y}_i$.

$$\boldsymbol{\Sigma} = \boldsymbol{O} diag(\lambda_1, ..., \lambda_d) \boldsymbol{O}^\top$$

where $\boldsymbol{O}$ is an orthogonal matrix whose columns $\boldsymbol{o}_1, ..., \boldsymbol{o}_d$ are the eigenvectors of $\boldsymbol{\Sigma}$ and $\lambda_1, ..., \lambda_d$ are the corresponding eigenvalues. The columns of $\boldsymbol{O}$ have been arranged so that the eigenvalues are ordered from largest to smallest. This is not essential, but it is convenient. We also assume no ties among the eigenvalues, which almost certainly will be true in actual applications.

A **normal linear combination** of $\boldsymbol{Y}_i$ is of the form $\boldsymbol{\alpha}^\top \boldsymbol{Y}_i = \sum_{j=1}^{p} \alpha_j Y_{i,j}$, where $||\boldsymbol{\alpha}|| = \sqrt{\sum_{j=1}^{p} \alpha_i^2} = 1$. The first principal component is the normed linear combination with the greatest variance. The variance in the direction $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is any fixed vector with norm 1, is

$$Var(\boldsymbol{\alpha}^\top \boldsymbol{Y}_i) = \boldsymbol{\alpha}^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}$$

The first PC maximizes the variance over α; the maximizer is $\boldsymbol{\alpha} = \boldsymbol{o}_1$, corresponding to the largest eigenvalue, and is called the **first principal axis**. The projections $\boldsymbol{o}_1^\top \boldsymbol{Y}_i, i = 1, ..., n$ onto this vector are called the **principal component scores**. Requiring that the norm of $\boldsymbol{\alpha}$ be fixed is essential, because otherwise the variance is unbounded and the maximizer does not exist.

After the first PC has been determined, you search for the maximum variation that is perpendicular to the first axis; this means maximizing the variance subject to $||\boldsymbol{\alpha}|| = 1$ and $\boldsymbol{\alpha}^\top \boldsymbol{o}_1 = 0$. In general, $\mathbf{0}_1, ..., \mathbf{0}_d$ are the principal axes and the set of projections $\boldsymbol{o}_j^\top \boldsymbol{Y}_i, i = 1, ..., n$ onto the j^{th} eigenvector is the j^{th} PC. Moreover, $\lambda_i = \boldsymbol{o}_i^\top \boldsymbol{\Sigma} \boldsymbol{o}_i$ is the variance of the i^{th} PC, $\lambda_i/(\lambda_1 + \cdots + \lambda_d)$ is the proportion of variance due to this PC, and $(\lambda_1 + \cdots + \lambda_i)/(\lambda_1 + \cdots + \lambda_d)$ is the proportion of variance explained by the first i PCs. The principal components are mutually uncorrelated since for $j \neq k$ we have

$$Cov(\boldsymbol{o}_j^\top \boldsymbol{Y}_i, \boldsymbol{o}_k^\top \boldsymbol{Y}_i) = \boldsymbol{o}_j^\top \boldsymbol{\Sigma} \boldsymbol{o}_k = 0$$

Let $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_n^\top)^\top$ be the original data and let

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{o}_1^\top \boldsymbol{Y}_1 & \cdots & o_d^\top \boldsymbol{Y}_1 \\ \vdots & \ddots & \vdots \\ o_1^\top \boldsymbol{Y}_n & \cdots & o_d^\top \boldsymbol{Y}_n \end{bmatrix}$$

be the matrix of the PCs. Then $\boldsymbol{S} = \boldsymbol{Y}\boldsymbol{O}$.

12.2 Factor Models

A **factor model** for excess equity returns is

$$R_{j,t} = \beta_{0,j} + \beta_{1,j} F_{1,t} + \cdots + \beta_{p,j} F_{p,t} + \epsilon_{j,t}$$

where $R_{j,t}$ is either the return or the excess return on the j^{th} asset at time t, $F_{1,t}, ..., F_{p,t}$ are variables, called **risk factors**, that represent the "state of the financial markets and world economy" at time t, and $\epsilon_{1,t}, ..., \epsilon_{n,t}$ are uncorrelated, mean-zero random variables called the **unique risks** of the individual stocks. The assumption that unique risks are uncorrelated means that all cross-correlation between the returns is due to the factors. Notice that the factors do not depend on j since they are common to all returns. The parameter $\beta_{i,j}$ is called a **factor loading** and specifies the sensitivity of the j^{th} return to the i^{th} factor.

The **CAPM** is a factor model where $p = 1$ and $F_{1,t}$ is the excess return on the market portfolio. A **factor** can be any variable thought to affect asset returns (e.g. market portfolio returns, GDP growth rate, interest rates, etc.).

12.3 Fama and French Three-Factor Model

Fama and French developed a model that has 3 factors: excess return of the market portfolio, small minus big (SMB), high minus low (HML). **SMB** represents the difference in returns on a portfolio of large stocks (small and big refer to the size of the market value (the share price times the number of outstanding shares)). **HML** represents the difference in returns on a portfolio of high book-to-market value (BE/ME) stocks and a portfolio of low BE/ME stocks. **Book value** is the net worth of the firm according to its accounting balance sheet. Their model of the return on the j^{th} asset for the t^{th} holding period is

$$R_{j,t} - \mu_{f,t} = \beta_{0,j} + \beta_{1,j}(R_{M,t} - \mu_{f,t}) + \beta_{2,j} SMB_t + \beta_{3,j} HML_t + \epsilon_{j,t}$$

where $\mu_{f,t}$ is the risk-free rate for the t^{th} holding period. Returns on portfolios have little autocorrelation, so the returns themselves, rather than residuals from a time series model, can be used.

12.3.1 Estimating Expectations and Covariance of Asset Returns

Let us start with 2 factors for simplicity. With $p = 2$, we have

$$R_{j,t} = \beta_{0,j} + \beta_{1,j} F_{1,t} + \beta_{2,j} F_{2,t} + \epsilon_{j,t}$$

It follows that

$$E(R_{j,t}) = \beta_{0,j} + \beta_{1,j} E(F_{1,t}) + \beta_{2,j} E(F_{w,t})$$
$$Var(R_{j,t}) = \beta_{1,j}^2 Var(F_1) + \beta_{2,j}^2 Var(F_2) + 2\beta_{1,j}\beta_{2,j} Cov(F_1, F_2) + \sigma_{\epsilon,j}^2$$

More generally, let $\boldsymbol{F}_t^\top = (F_{1,t}, ..., F_{p,t})$ be the vector of the p factors at time t and suppose that $\boldsymbol{\Sigma}_F$ is the $p \times p$ covariance matrix of $\boldsymbol{F}_t$. Define the vector of intercepts and matrix of loadings as follows.

$$\boldsymbol{\beta}_0^\top = (\beta_{0,1}, ..., \beta_{0,n})$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,j} & \cdots & \beta_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \cdots & \beta_{p,j} & \cdots & \beta_{p,n} \end{bmatrix}$$

Additionally, define $\boldsymbol{\epsilon}^\top = (\epsilon_{1,t}, ..., \epsilon_{n,t})$ and let $\boldsymbol{\Sigma}_\epsilon$ be the $n \times n$ diagonal covariance matrix of $\boldsymbol{\epsilon}$:

$$\boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} \sigma_{\epsilon,1}^2 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{\epsilon,j}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \sigma_{\epsilon,n}^2 \end{bmatrix}$$

Finally, let $\boldsymbol{R}_t^\top = (R_{1,t}, ..., R_{n,t})$ be the vector of all returns at time t. We can now reexpress the model using matrix notation as follows.

$$\boldsymbol{R}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}^\top \boldsymbol{F}_t + \boldsymbol{\epsilon}_t$$

Therefore, the $n \times n$ covariance matrix of $\boldsymbol{R}_t$ is

$$\boldsymbol{\Sigma}_R = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_F \boldsymbol{\beta} + \boldsymbol{\Sigma}_\epsilon$$

Variance/Covariance:

$$Var(R_j) = \boldsymbol{\beta}_j^\top \boldsymbol{\Sigma}_F \boldsymbol{\beta}_j + \sigma_{\epsilon_j}^2 \quad Cov(R_j, R_j') = \boldsymbol{\beta}_j^\top \boldsymbol{\Sigma}_F \boldsymbol{\beta}_{j'}$$

To make use of these above relations, we estimate $\boldsymbol{\beta}$ with regression coefficients, $\boldsymbol{\Sigma}_F$ with the sample covariance of the factors, and $\boldsymbol{\Sigma}_\epsilon$ with the diagonal matrix of the mean residual sum of squared errors from the regressions.

12.4 Cross-Sectional Factor Models

A **CSF Model** is a regression model using data from many assets but from only a single holding period. For example, suppose that R_j, $(B/M)_j$, and D_j are the return, book-to-market value, and dividend yeild for the j^{th} asset for some fixed time t. Since t is fixed, it will not be made xplicit in the notation. Then a possible CSF model is

$$R_j = \beta_0 + \beta_1 (B/M)_j + \beta_2 D_j + \epsilon_j$$

The two fundamental differences between TSF and CSF models is that with a TSF model one estimates parameters, one asset at a time, using multiple holding periods, while in a CSF model one estimates parameters, one single holding period at a time, using multiple assets. The other major difference is that in a time series factor model, the factors are directly measured and the loadings are the unknown parameters to be estimated by regression. In a CSF model, the opposit is true; the loadings are directly measured and the factor values are estimated by regression.

12.5 Statistical Factor Models

In a statistical factor model, neither the factor values nor the loadings are directly observable. All that is available is the sample $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$, or only the sample covariance matrix. Let us start with the multifactor model notation and the return covariance matrix which for convenience will be repeated as

$$\boldsymbol{R}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}^\top \boldsymbol{F}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\Sigma}_R = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_F \boldsymbol{\beta} + \boldsymbol{\Sigma}_\epsilon$$

Here $\boldsymbol{\beta}^\top$ is $d \times p$ where d is the dimension of R_t and p is the number of factors. We now need a set of constraints which prohibit factor correlation and standardize the factors such that $\boldsymbol{\Sigma}_F = \boldsymbol{I}$. Under these constraints, the model simplifies to

$$\boldsymbol{\Sigma}_R = \boldsymbol{\beta}^\top \boldsymbol{\beta} + \boldsymbol{\Sigma}_\epsilon$$

The last constraint that enables us to determine $\boldsymbol{\beta}$ is that $\boldsymbol{\beta}\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{\beta}^\top$ is diagonal (though not the only possible constraint). To estimate $\boldsymbol{\Sigma}_R$,

$$\widehat{\boldsymbol{\Sigma}}_R = \boldsymbol{O}^\top \boldsymbol{O}$$

where $\boldsymbol{O}^\top$ is the $d \times p$ matrix whose columns are the first d principal axes (eigenvectors) and the rank of $\widehat{\boldsymbol{\Sigma}}_R$ is only p so less than full rank.

13 Portfolio Selection

The objectives of a portfolio are to maximize expected return and minimize the risk of the portfolio. We start with a simple example of one risky asset, which could be a portfolio, for example, a mutual fund. Assume that the expected return is 0.15 and the standard deviation of the return is 0.25. Assume that there is a **risk-free asset**, such as, a 90-day T-bill, and the risk-free rate is 6%, so the return on the risk-free asset is 6%, or .06. The standard deviation of the return on the risk-free asset is 0 by definition of "risk-free."

We are faced with the problem of constructing an investment portfolio that we will hold for one time period, which is called the **holding period** and which could be a day, a month, a quarter, a year, 10 years, and so forth. For now, we are only looking at returns over one time period. Suppose that a fraction w of our wealth is invested in the risky asset and the remaining fraction $1 - w$ is invested in the risk-free asset. Then the return is

$$E(R) = w(0.15) + (1 - w)(0.06) = 0.06 + 0.09w$$

and the variance of the return is

$$\sigma_R^2 = w^2(.25)^2 + (1 - w)^2(0)^2 = w^2(.25)^2$$

To decide what proportion w of one's wealth to invest in the risky asset, one chooses either the expected return $E(R)$ one wants or the amount of risk σ_R with which one is willing to live.

13.1 Two Risky Assets

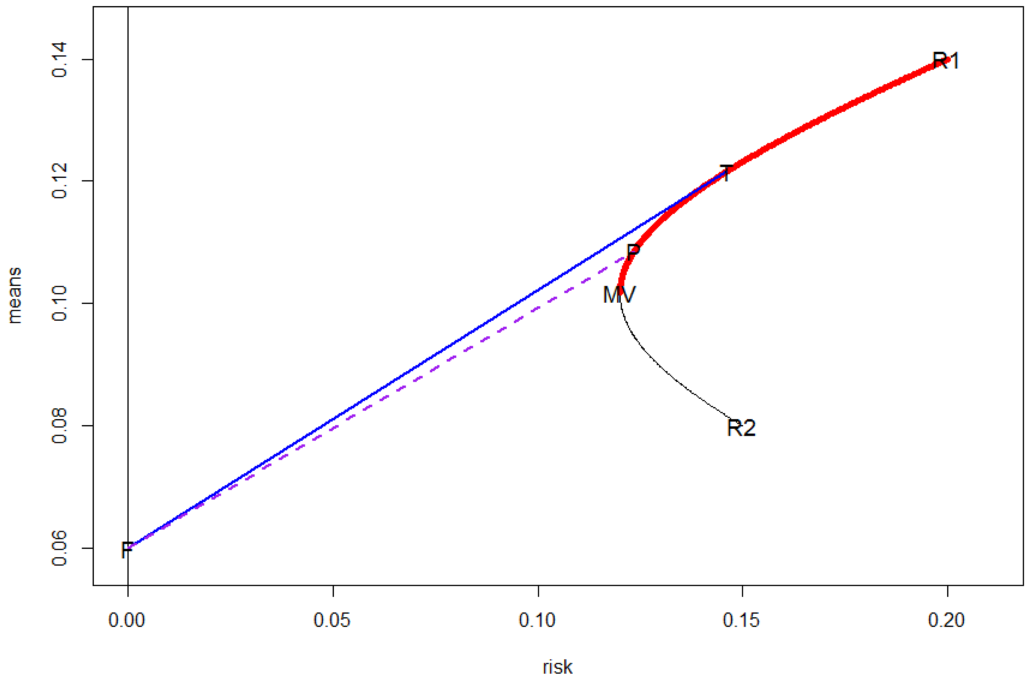
Suppose we have two risky assets with returns R_1 and R_2 that we mix so that our expected return is

$$E(R_p) = w\mu_1 + (1-w)\mu_2$$

Thus the variance of the portfolio is

$$\sigma_R^2 = w^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w(1-w)\rho_{12}\sigma_1\sigma_2$$

The curve shown following is the locus of values of $(\sigma_R, E(R))$ when $0 \leq w \leq 1$. The leftmost point on this locus achieves the minimum value of the risk and is called the **minimum variance portfolio**. The points on this locus that have an expected return at least as large as the minimum variance portfolio are called the **efficient frontier**. Portfolios on the efficient frontier are known as **mean-variance efficient portfolios**.



In the graph above, F = risk-free asset, T = tangency portfolio, and $R1$ and $R2$ are the first and second risky assets. All points on the curve connecting $R1$ and $R2$ is attainable with $0 \leq w \leq 1$.

13.2 Combining Two Risky Assets with a Risk-Free Asset

If $E(R_P)$ and σ_{R_P} are the expected return and volatility on a portfolio and μ_f is the risk-free rate, then

$$\frac{E(R_P) - \mu_f}{\sigma_{R_P}}$$

is the **Sharpe's ratio** of the portfolio. Sharpe's can be thought of as a "reward-to-risk" ratio. It is the ratio of the reward quantified by the excess expected return to the risk as measured by the standard deviation. The point T on the efficient frontier is the portfolio with the highest Sharpe's ratio and this portfolio is known as the **tangency portfolio**.

The optimal or efficient portfolios mix the tangency portfolio with the risk-free asset. Each efficient portfolio has two properties:

- it has a higher expected return than any other portfolio with the same or smaller risk, and
- it has a smaller risk than any other portfolio with the same or higher expected return

Thus we can only improve (reduce) the risk of an efficient portfolio by accepting a worse (smaller) expected return, and we can only improve (increase) the expected return of an efficient portfolio by accepting worse (higher) risk.

Define $V_1 = \mu_1 - \mu_f$ and $V_2 = \mu_2 - \mu_f$, the excess expected returns. Then the tangency portfolio uses the weight

$$w_T = \frac{V_1\sigma_2^2 - V_2\rho_{12}\sigma_1\sigma_2}{V_1\sigma_2^2 + V_2\sigma_1^2 - (V_1 + V_2)\rho_{12}\sigma_1\sigma_2}$$

for the first risky asset and weight $(1 - w_T)$ for the second.

Let $R_T, E(R_T)$, and σ_T be the return, expected return, and standard deviation of the return on the tangency portfolio. Then $E(R_T)$ and σ_T can be found by first finding w_T and then using the above weight and then using the formulas

$$E(R_T) = w_T\mu_1 + (1 - w_T)\mu_2$$

and

$$\sigma_T = \sqrt{w_T^2\sigma_1^2 + (1 - w_T)^2\sigma_2^2 + 2w_T(1 - w_T)\rho_{12}\sigma_1\sigma_2}$$

13.2.1 Combining the Tangency Portfolio with the Risk-Free Asset

Let R_p be the return on the portfolio that allocates a fraction w of the investment to the tangency portfolio and $1 - w$ to the risk-free asset. Then $R_p = wR_T + (1 - w)\mu_f = \mu_f + \omega(R_T - R_f)$, so that

$$E(R_p) = \mu_f + w\{E(R_T) - \mu_f\} \text{ and } \sigma_{R_P} = w\sigma_T$$

13.3 Selling Short

Often some of the weights in an efficient portfolio are negative. A negative weight on an asset means that this asset is sold short. **Selling short** is a way to profit if a stock price goes down. To sell a stock short, one sells the stock without owning it. The stock must be borrowed from a broker or another customer of the broker. At a later point in time, one buys the stock and gives it back to the lender. This closes the short position. Owning a stock is called a **long position**.

13.4 Risk-Efficient Portfolios with N Risky Assets

In this section, we use quadratic programming to find efficient portfolios with an arbitrary number of assets. Assume that we have N risky assets and that the return on the i^{th} risky asset is R_i and has expected value μ_i. Define

$$\boldsymbol{R} = \begin{bmatrix} R_1 \\ \vdots \\ R_N \end{bmatrix}$$

to be the random vector of returns,

$$E(\boldsymbol{R}) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}$$

and $\boldsymbol{\Sigma}$ to be the covariance matrix of $\boldsymbol{R}$. Let

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix}$$

be a vector of portfolio weights so that $w_1 + \cdots + w_N = \boldsymbol{1}^\top \boldsymbol{w} = 1$. The expected return on the portfolio is

$$\sum_{i=1}^{N} \boldsymbol{w}_i \mu_i = \boldsymbol{w}^\top \boldsymbol{\mu}$$

and the variance of the return is

$$\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}.$$

Given a target μ_P, the efficient portfolio minimizes the above variance subject to

$$\boldsymbol{w}^\top \boldsymbol{\mu} = \mu_P \text{ and } \boldsymbol{w}^\top \boldsymbol{1} = 1$$

Quadratic programming is used to minimize a quadratic objective function subject to linear constraints. In applications to portfolio optimization, the objective function is the variance of the portfolio return. The objective function is a function of N variables, such as the weights of the N assets, that are denoted by an $N \times 1$ vector $\boldsymbol{x}$. Suppose that the quadratic objective function to be minimized is

$$\frac{1}{2} \boldsymbol{x}^\top \boldsymbol{D} \boldsymbol{x} - \boldsymbol{d}^\top \boldsymbol{x}$$

where $\boldsymbol{D}$ is an $N \times N$ matrix and $\boldsymbol{d}$ is an $N \times 1$ vector. There are two types of linear constraints on $\boldsymbol{x}$, inequality and equality constraints. The linear inequality constraints are

$$\boldsymbol{A}_{neq}^\top \boldsymbol{x} \geq \boldsymbol{b}_{neq},$$

where $\boldsymbol{A}_{neq}$ is an $m \times N$ matrix, $\boldsymbol{b}_{neq}$ is an $m \times 1$ vector, and m is the number of inequality constraints. The equality constraints are

$$\boldsymbol{A}_{eq}^\top \boldsymbol{x} = \boldsymbol{b}_{eq}$$

where $\boldsymbol{A}_{eq}$ is an $n \times N$ matrix, $\boldsymbol{b}_{eq}$ is an $n \times 1$ vector, and n is the number of equality constraints.

To apply quadratic programming to find an efficient portfolio, we use $\boldsymbol{x} = \boldsymbol{w}$, $\boldsymbol{D} = \boldsymbol{2\Sigma}$, and $\boldsymbol{d}$ equal to an $N \times 1$ vector of zeros so that $\frac{1}{2} \boldsymbol{x}^\top \boldsymbol{D} \boldsymbol{x} - \boldsymbol{d}^\top \boldsymbol{x} = \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}$, the return variance of the portfolio. There are two equality constraints: one such that the weights sum to 1 and the other such that the portfolio return is a specified target μ_P. Therefore, we define

$$\boldsymbol{A}_{eq}^\top = \begin{bmatrix} \boldsymbol{1}^\top \\ \boldsymbol{\mu}^\top \end{bmatrix}$$

and

$$\boldsymbol{b}_{eq} = \begin{bmatrix} 1 \\ \mu_P \end{bmatrix},$$

so that $\boldsymbol{A}_{eq}^{\top}\boldsymbol{x} = \boldsymbol{b}_{eq}$ becomes

$$\begin{bmatrix} \mathbf{1}^{\top}\boldsymbol{w} \\ \boldsymbol{\mu}^{\top}\boldsymbol{w} \end{bmatrix} = \begin{bmatrix} 1 \\ \mu_P \end{bmatrix}$$

Investors often wish to impose additional inequality constraints. If an investor cannot or does not wish to sell short, then the constraint

$$\boldsymbol{w} \geq \mathbf{0}$$

can be used. Here $\mathbf{0}$ is a vector of N zeros. In this case $\boldsymbol{A}_{neq}$ is the $N \times N$ identical matrix and $\boldsymbol{b}_{neq} = \mathbf{0}$.

To avoid concentrating the portfolio in just one or a few stocks, an investor may wish to constrain the portfolio so that no w_i exceeds a bound λ, which enforces a certain degree of diversification to the portfolio. In this case, $\boldsymbol{w} \leq \lambda\mathbf{1}$, so that $\boldsymbol{A}_{neq}$ is minus the $N \times N$ identity matrix and $\boldsymbol{b}_{neq} = -\lambda\mathbf{1}$.

To find the efficient frontier, one uses a grid of values of μ_P and finds the corresponding efficient portfolios. For each portfolio, σ_P^2, which is the minimized value of the objective function, can be calculated. Then one can find the minimum variance portfolio by finding the portfolio with the smallest value of the σ_P^2. The efficient frontier is the set of efficient portfolios with expected return above the expected return of the minimum variance portfolio.

13.5 Resampling and Efficient Portfolios

The theory of portfolio optimization assumes that the expected returns and the covariance matrix of the returns is known. In practice, one must replace these quantities with estimates. However, the effects of estimation error, especially with smaller values of N, can result in portfolios that only appear efficient. This is remedied by the bootstrap.

14 The Capital Asset Pricing Model (CAPM)

The CAPM provides a theoretical justification for the widespread practice of passive investing by holding **index funds**. The CAPM can provide estimates of expected rates of return on individual investments and can establish "fair" rates of return on invested capital in regulated firms or in firms working on a cost-plus basis.

The CAPM starts with the question, what would be the risk premiums on securities if the following assumptions were true?

1. The market prices are "in equilibrium." In particular, for each asset, supply equals demand.

2. Everyone has the same forecasts of expected returns and risks.

3. All investors choose portfolios optimally according to the principles of efficient diversification. This implies that everyone holds a tangency portfolio of risky assets as well as the risk-free asset.

4. The market rewards people for assuming unavoidable risk, but there is no reward for needless risks due to inefficient portfolio selection. Therefore, the risk premium on a single security is not due to its "standalone" risk, but rather to its contribution to the risk of the tangency portfolio.

14.1 The Capital Market Line (CML)

The **capital market line** related the excess expected return on an efficient portfolio to its risk. **Excess expected return** is the expected return minus the risk-free rate and is also called the risk premium. The CML is

$$\mu_R = \mu_f + \frac{\mu_M - \mu_f}{\sigma_M}\sigma_R$$

where R is the return on a given efficient portfolio (mixture of the market portfolio and the risk-free asset), $\mu_R = E(R)$, μ_f is the risk-free rate, R_M is the return on the market portfolio, $\mu_M = E(R_M)$, σ_M is the standard deviation of R_M, and σ_R is the standard deviation of R. The risk premium of R is $\mu_R - \mu_f$ and the risk premium of the market portfolio is $\mu_M - \mu_f$.

Think of the CML as showing how μ_R depends on σ_R, for the portfolio R is the only thing which varies. The slope of the CML is, of course,

$$\frac{\mu_M - \mu_f}{\sigma_M},$$

which can be interpreted as the ratio of the risk premium to the standard deviation of the market portfolio. This is Sharpe's famous "reward-to-risk ratio," which is widely used in finance. We can rewrite the CML as

$$\frac{\mu_R - \mu_f}{\sigma_R} = \frac{\mu_M - \mu_f}{\sigma_M},$$

which says that the reward-to-risk ratio for any efficient portfolio equals that ratio for the market portfolio - all efficient portfolios have the same Sharpe's ratio as the market portfolio.

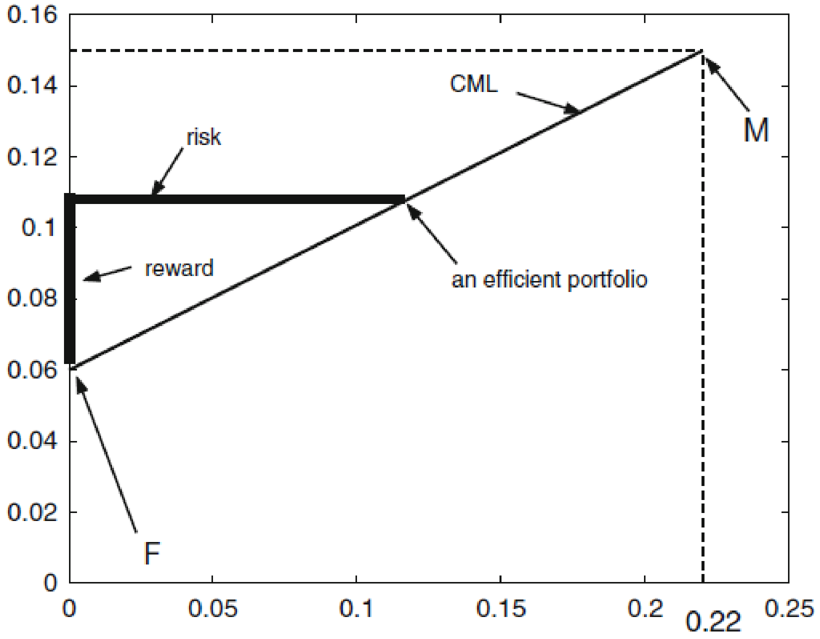


Fig. 17.1. *CML when $\mu_f = 0.06$, $\mu_M = 0.15$, and $\sigma_M = 0.22$. All efficient portfolios are on the line connecting the risk-free asset (F) and the market portfolio (M). Therefore, the reward-to-risk ratio is the same for all efficient portfolios, including the market portfolio. This fact is illustrated by the thick lines, whose lengths are the risk and reward for a typical efficient portfolio.*

The CAPM says that the optimal way to invest is to

1. Decide on the risk σ_R that you can tolerate, $0 \leq \sigma_R \leq \sigma_M$;
2. Compute $w = \sigma_R/\sigma_M$;
3. Invest w proportion of your investment in a market index fund, that is, a fund that tracks the market as a whole;
4. Invest $1-w$ proportion of your investment in risk-free Treasury bills, or a money-market fund.

Alternatively,

1. Choose the reward $\mu_R - \mu_f$ that you want; the only constraint is that $\mu_f \leq \mu_R \leq \mu_M$ so that $0 \leq w \leq 1$;
2. Compute $w = (\mu_R - \mu_f)/(\mu_M - \mu_f)$;
3. Steps 3 and 4 are the same

14.2 Betas and the Security Market Line

The **security market line** (SML) relates the excess return on an asset to the slope of its regression on the market portfolio. The SML differs from the CML in that the SML applies to all assets while the CML applies only to efficient portfolios.

Suppose that there are many securities indexed by j. Define

σ_{jM} = covariance between the returns on the j^{th} security and the market portfolio, and

$$\beta_j = \frac{\sigma_{jM}}{\sigma_M^2}$$

The best linear predictor of R_j based on R_M is

$$\widehat{R}_j = \beta_{0,j} + \beta_J R_M,$$

where β_j is defined as above.

Suppose we have a bivariate time series $(R_{j,t}, R_{M,t})_{t=1}^n$ of returns on the j^{th} asset and the market portfolio. Then, the estimate slope of the linear regression of $R_{j,t}$ on $R_{M,t}$ is

$$\widehat{\beta}_j = \frac{\sum_{t=1}^n (R_{j,t} - \bar{R}_j)(R_{M,t} - \bar{R}_M)}{\sum_{t=1}^n (R_{M,t} - \bar{R}_M)^2}$$

which, after multiplying the numerator and denominator by the same factor n^{-1}, becomes an estimate of σ_{jM} divided by an estimate of σ_M^2 and therefore an estimate of β_j.

Let μ_j be the expected return on the j^{th} security. Then $\mu_j - \mu_f$ is the **risk premium** for that security. Using CAPM, it can be shown that

$$\mu_j - \mu_f = \beta_j(\mu_M - \mu_f).$$

This equation represents the security market line (SML). The SML says that the risk premium of the j^{th} asset is the product of its beta (β_j) and the risk premium of the market portfolio ($\mu_M - \mu_f$). Therefore, β_j measures both riskiness of the j^{th} asset and the reward for assuming that riskiness.

Consequently, β_j is a measure of how "aggressive" the j^{th} asset is. By definition, the beta for the market portfolio is 1; i.e., $\beta_M = 1$. This suggest the rules-of-thumb

$$\begin{aligned} \beta_j > 1 &\rightarrow \text{"aggressive"} \\ \beta_j = 1 &\rightarrow \text{"average risk"} \\ \beta_j < 1 &\rightarrow \text{"not aggressive"} \end{aligned}$$

Consider what would happen if an asset like J did exist. Investors would not want to buy it because, since it is below the SML, its risk premium is too low for the risk given by its beta. They would invest less in J and more in other securities. Thereof,re the price of J would decline and **after** this decline its expected return would increase. After that increase, the asset J would be on the SML, or so the theory predicts.

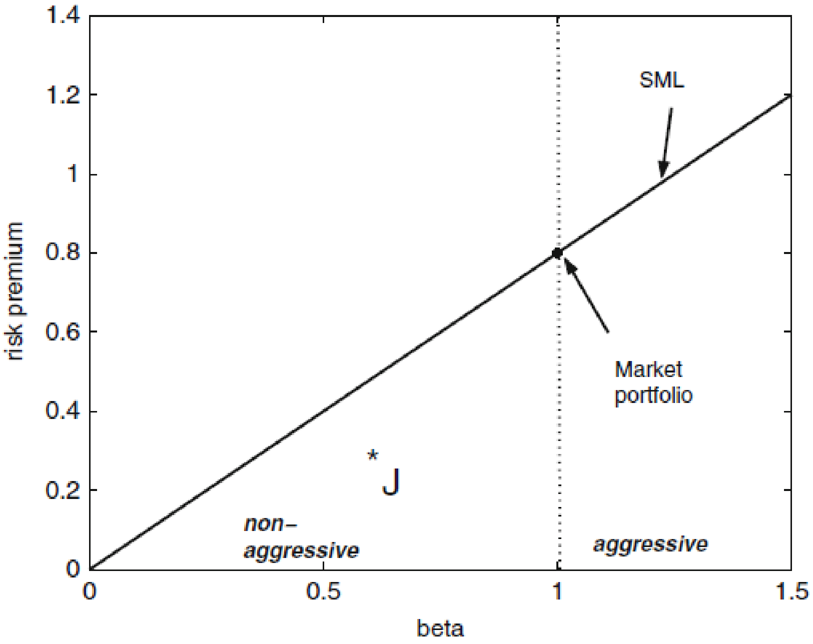


Fig. 17.2. *Security market line (SML) showing that the risk premium of an asset is a linear function of the asset's beta. J is a security not on the line and a contradiction to the CAPM. Theory predicts that the price of J decreases until J is on the SML. The vertical dotted line separates the nonaggressive and aggressive regions.*

14.2.1 Comparison of CML and SML

The CML applies only to the return R of an efficient portfolio. It can be arranged so as to relate the excess expected return of that portfolio to the excess expected return of the market portfolio:

$$\mu_R - \mu_f = \left(\frac{\sigma_R}{\sigma_M}\right)(\mu_M - \mu_f)$$

The SML applies to **any** asset and like the CML relates its excess expected return to the excess expected return of the market portfolio:

$$\mu_j - \mu_f = \beta_j(\mu_M - \mu_f)$$

If we take an efficient portfolio and consider it as an asset, then μ_R and μ_j both denote the expected return on that portfolio/asset.

14.3 The Security Characteristic Line

Let $R_{j,t}$ be the return at time t on the j^{th} asset. Similarly, let $R_{M,t}$ and $\mu_{f,t}$ be the return on the market portfolio and the risk-free turn at time t. The **security characteristic line** is a regression model:

$$R_{j,t} = \mu_{f,t} + \beta_j(R_{M,t} - \mu_{f,t}) + \epsilon_{j,t}$$

where $\epsilon_{j,t} \sim N(0, \sigma^2_{\epsilon,j})$. It is often assumed that the $\epsilon_{j,t}$s are uncorrelated across assets, that is, that $\epsilon_{j,t}$ is uncorrelated with $\epsilon_{j',t}$ for $j \neq j'$.

Let $\mu_{j,t} = E(R_{j,t})$ and $\mu_{M,t} = E(R_{M,t})$. Taking expectations in the above equation gives us

$$\mu_{j,t} = \mu_{f,t} + \beta_j(\mu_{M,t} - \mu_{f,t}),$$

which is the SML; the SML gives us information about expected returns, but not about the variance of the returns. For the latter we need the characteristic line. The characteristic line is said to be a **return-generating process** since it gives us a probability model of the returns, not just a model of their expected values.

The characteristic line implies that

$$\sigma^2_j = \beta^2_j \sigma^2_M + \sigma^2_{\epsilon,j},$$

that

$$\sigma_{jj'} = \beta_j \beta_{j'} \sigma^2_M$$

for $j \neq j'$, and that

$$\sigma_{Mj} = \beta_j \sigma^2_M$$

Hence the total risk of the j^{th} asset is

$$\sigma_j = \sqrt{\beta^2_j \sigma^2_M + \sigma^2_{\epsilon,j}}.$$

The squared risk has two components: $\beta^2_j \sigma^2_M$ is called the **market** or **systematic component of risk** and $\sigma^2_{\epsilon,j}$ is called the **unique, non-market,** or **unsystematic component of risk**.

14.3.1 Reducing Unique Risk by Diversification

The market component of risk cannot be reduced by diversification, but the unique component can be reduced or even eliminated by sufficient diversification.

Suppose that there are N assets with returns on a holding period t. The return on a weighted portfolio is thus

$$R_{P,t} = w_1 R_{1,t} + \cdots + w_N R_{N,t}.$$

Let $R_{M,t}$ be the return on the market portfolio. According to the characteristic line model $R_{j,t} = \mu_{f,t} + \beta_j(R_{M,t} - \mu_{f,t}) + \epsilon_{j,t}$, so that

$$R_{P,t} = \mu_{f,t} + \left(\sum_{j=1}^{N} \beta_j w_j\right)(R_{M,t} - \mu_{f,t}) + \sum_{j=1}^{N} w_j \epsilon_{j,t}.$$

Therefore, the portfolio beta is

$$\beta_P = \sum_{j=1}^{N} w_j \beta_j,$$

and the "epsilon" for the portfolio is

$$\epsilon_{P,t} = \sum_{j=1}^{N} w_j \epsilon_{j,t}.$$

We now assume that the epsilons are uncorrelated. Therefore,

$$\sigma^2_{\epsilon,P} = \sum_{j=1}^{N} w_j^2 \sigma^2_{\epsilon,j}.$$

Suppose the assets in the portfolio are equally weighted; that is, $w_j = 1/N$ for all j. Then

$$\beta_P = \frac{1}{N} \sum_{j=1}^{N} \beta_j,$$

and

$$\sigma^2_{\epsilon,P} = \frac{N^{-1} \sum_{j=1}^{N} \sigma^2_{\epsilon,j}}{N} = \frac{\bar{\sigma}^2_{\epsilon}}{N}.$$

15 Risk Management

The financial world has always been risky, and financial innovations such as the development of derivatives markets and the packaging of mortgages have now made risk management more important than ever, but also more difficult.

There are many different types of risk.

- **Market risk** is due to changes in prices.
- **Credit risk** is the danger that a counterparty does not meet contractual obligations, for example, that interest or principal on a bond is not paid.
- **Liquidity risk** is the potential extra cost of liquidating a position because buyers are difficult to locate.
- **Operational risk** is due to fraud, mismanagement, human errors, and similar problems.

The two most important metrics are value-at-risk (VaR) and expected shortfall (ES). VaR uses two parameters, the time horizon and the confidence level, which are denoted by T and $1 - \alpha$, respectively. Given these, the VaR is a bound such that the loss over the horizon is less than this bound with probability equal to the confidence coefficient.

If $\mathcal{L}$ is the loss over the holding period T, then $\mathrm{VaR}(\alpha)$ is the α-th upper quantile of $\mathcal{L}$. Equivalently, if $\mathcal{R} = -\mathcal{L}$ is the revenue, then $\mathrm{VaR}(\alpha)$ is minus the α-th quantile for $\mathcal{R}$. For continuous loss distributions, $\mathrm{VaR}(\alpha)$ solves

$$P\{\mathcal{L} > \mathrm{VaR}(\alpha)\} = P\{\mathcal{L} \geq \mathrm{VaR}(\alpha)\} = \alpha$$

and for any loss distribution, continuous or not,

$$\mathrm{VaR}(\alpha) = \inf\{x : P(\mathcal{L} > x) \leq \alpha\}.$$

The problem with VaR is that it discourages diversification, which leads us to measuring the expected loss given that the loss exceed VaR, or **expected shortfall**.

For any loss distribution, continuous or not,

$$\mathrm{ES}(\alpha) = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}(u)du,$$

which is the average of $\mathrm{VaR}(u)$ over all u that are less than or equal to α. If $\mathcal{L}$ has a continuous distribution,

$$ES(\alpha) = E\{\mathcal{L}|\mathcal{L} > \mathrm{VaR}(\alpha)\} = E\{\mathcal{L}|\mathcal{L} \geq VaR(\alpha)\}.$$

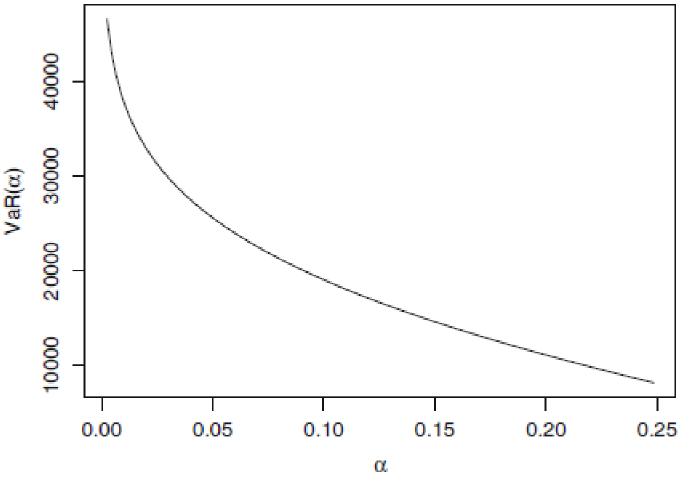


Fig. 19.1. $\mathrm{VaR}(\alpha)$ *for* $0.025 < \alpha < 0.25$ *when the loss distribution is normally distributed with mean* -4000 *and standard deviation* 18,000.

15.1 Estimating VaR and ES with One Asset

We start with **non-parametric** estimates of VaR and ES, meaning that the loss distribution is not assumed to be in a parametric family such as the normal or t-distributions. Suppose that we want a confidence coefficient of $1-\alpha$ for the risk measures, and consequently the α-upper quantile of the loss distribution. This quantile is estimated as the α-quantile of a sample of historic returns, called $\widehat{q}(\alpha)$. If S is the size of the current position, then the non-parametric estimate of VaR is

$$\widehat{\mathrm{VaR}}^{np}(\alpha) = -S \times \widehat{q}(\alpha),$$

with the minus sign converting revenue to a loss.

To estimate ES, let $R_1, ..., R_n$ be the historic returns and define $\mathcal{L}_i = -S \times R_i$. Then

$$\widehat{\mathrm{ES}}^{np}(\alpha) = \frac{\sum_{i=1}^n \mathcal{L}_i I\{\mathcal{L}_i > \widehat{\mathrm{VaR}}(\alpha)\}}{\sum_{i=1}^n I\{\mathcal{L}_i > \widehat{\mathrm{VaR}}(\alpha)\}} = -S \times \frac{\sum_{i=1}^n R_i I\{R_i < \widehat{q}(\alpha)\}}{\sum_{i=1}^n I\{R_i < \widehat{q}(\alpha)\}},$$

which is the average of all $\mathcal{L}_i$ exceeding $\widehat{\mathrm{VaR}}^{np}(\alpha)$. Here $I\{\mathcal{L}_i > \widehat{\mathrm{VaR}}^{np}(\alpha)\}$ is the indicator that $\mathcal{L}_i$ exceeds $\widehat{\mathrm{VaR}}^{np}(\alpha)$, and similarly for $I\{R_i < \widehat{q}(\alpha)\}$.

15.1.1 Parametric Estimation

Parametric estimation allows us to use GARCH models to adapt the risk measures to the current estimate of volatility. Let $F(y|\boldsymbol{\theta})$ be a parametric family of distributions used to model the return distribution and suppose that $\widehat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$, such as, the MLE computed from historic returns. Then $F^{-1}(\alpha|\boldsymbol{\theta})$ is an estimate of the α-quantile of the return distribution and

$$\widehat{\text{VaR}}^{par}(\alpha) = -S \times F^{-1}(\alpha|\widehat{\boldsymbol{\theta}})$$

is a parametric estimate of $\text{VaR}(\alpha)$. As before, S is the size of the current position.

Let $f(y|\boldsymbol{\theta})$ be the density of $F(y|\boldsymbol{\theta})$. Then the estimate of ES is

$$\widehat{\text{ES}}^{par}(\alpha) = -\frac{S}{\alpha} \times \int_{-\infty}^{F^{-1}(\alpha|\widehat{\boldsymbol{\theta}})} x f(x|\widehat{\boldsymbol{\theta}}) dx.$$

15.2 Estimating VaR and ES Using ARMA+GARCH Models

Daily equity returns typically have a small amount of autocorrelation and a greater amount of volatility clustering. When calculating risk measures, the autocorrelation can be ignored if it is small enough, however the volatility clustering remains a problem. In this section, we use ARMA+GARCH models so that $\text{VaR}(\alpha)$ and $\text{ES}(\alpha)$ can adjust to periods of high or low volatility.

Assume that we have n returns, $R_1, ..., R_n$ and we need to estimate VaR and ES for the next return R_{n+1}. Let $\widehat{\mu}_{n+1|n}$ and $\widehat{\sigma}_{n+1|n}$ be the estimated conditional mean and varaince of tomorrow's return R_{n+1}, conditional on the current information set, which in this context is simply $\{R_1, ..., R_n\}$. We will also assume that R_{n+1} has a conditional t-distribution with tail index ν. After fitting an ARMA+GARCH model, we have estimates of $\widehat{\nu}$, $\widehat{\mu}_{n+1|n}$, and $\widehat{\sigma}_{n+1|n}$. The estimated conditional scale parameter is

$$\widehat{\lambda}_{n+1|n} = \sqrt{(\widehat{\nu}-2)/\widehat{\nu}}\widehat{\sigma}_{n+1|n}.$$

16 Bayesian Data Analysis and Monte Carlo Markov Chains

In Bayesian statistics all unknowns, and in particular unknown parameters, are considered to be random variables and their probability distributions specify our beliefs about their likely values. Here, we let $\boldsymbol{\theta}$ denote the vector of all unknowns and call it the parameter vector, which could include both the unobserved white noise and the future values of the series being forecast.

16.1 Prior and Posterior Distributions

We now assume that $\boldsymbol{\theta}$ is a continuously distributed parameter vector. The **prior distribution** with density $\pi(\boldsymbol{\theta})$ expresses our beliefs about $\boldsymbol{\theta}$ prior to observing data. The likelihood function is interpreted as the conditional density of the data $\boldsymbol{Y}$ gives $\boldsymbol{\theta}$ and written as $f(\boldsymbol{y}|\boldsymbol{\theta})$. The joint density of $\boldsymbol{\theta}$ and $\boldsymbol{Y}$ is the product of the prior and the likelihood; that is,

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta}).$$

The marginal density of $\boldsymbol{Y}$ is found by integrating $\boldsymbol{\theta}$ out of the joint density so that

$$f(\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta}),$$

and the conditional density of $\boldsymbol{\theta}$ given $\boldsymbol{Y}$ is

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{\pi(\boldsymbol{\theta}) f(\boldsymbol{Y}|\boldsymbol{\theta})}{f(\boldsymbol{Y})} = \frac{\pi(\boldsymbol{\theta}) f(\boldsymbol{Y}|\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta}) f(\boldsymbol{Y}|\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

and is known as the **posterior density**. Notice that π always denotes densities of $\boldsymbol{\theta}$, while f is used to denote densities of our data.

16.2 Conjugate Priors

A family of distributions is called a **conjugate prior family** for statistical model if the posterior is in this family whenever the prior is in the family. Suppose that the prior for θ is Beta(α, β) so that the prior density is

$$\pi(\theta) = K_1 \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where K_1 is a constant ($K_1 = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$). The parameters in a prior density mus the known, so here α and β are chosen by the data analyst in accordance with the prior knowledge about the value of θ.

Suppose now that the stock price is observed on n days and increases on Y days (and does not increase on $n-Y$ day). Then the likelihood is

$$f(Y|\theta) = K_2 \theta^{Y}(1-\theta)^{n-Y},$$

where $K_2 = \binom{n}{Y}$ is another constant. The joint density of θ and Y is thus

$$\pi(\theta)f(Y|\theta) = K_3 \theta^{\alpha+Y-1}(1-\theta)^{\beta+n-Y-1},$$

where $K_3 = K_1 K_2$. Then, the posterior density is

$$\pi(\theta|Y) = \frac{\pi(\theta)f(Y|\theta)}{\int_0^1 \pi(\theta)f(Y|\theta)d\theta} = \theta^{\alpha+Y-1}(1-\theta)^{\beta+n-Y-1}$$

where

$$K_4 = \frac{1}{\int_0^1 \theta^{\alpha+Y-1}(1-\theta)^{\beta+n-Y-1}d\theta}$$

The posterior distribution is Beta$(\alpha+Y, \beta+n-Y)$.

We did not need to keep track of the values of $K_1, ..., K_4$. Since $\pi(\theta|Y)$ is proportional to a Beta$(\alpha+Y, \beta+n-Y)$ density and since all densities integrate to 1, we can deduce that the constant of proportionality is 1 and the posterior is Beta$(\alpha+Y, \beta+n-Y)$. It follows that

$$K_4 = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+Y)\Gamma(\beta+n-Y)}.$$

The mean of the posterior is

$$E(\theta|Y) = \frac{\alpha+Y}{\alpha+\beta+n}$$

and the posterior variance is

$$\text{Var}(\theta|Y) = \frac{(\alpha+Y)(\beta+n-Y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} = \frac{E(\theta|Y)\{1-E(\theta|Y)\}}{(\alpha+\beta+n+1)}.$$