Quadratic Programming Models:

Mean–Variance Optimization

Sirong Luo

Faculty of Statistics and Data Sciences
Shanghai University of Finance and Economics

Portfolio Return

Consider an investment environment where there is a universe of n risky assets. In the next few chapters we will be concerned with a one-period model of the problem of investing in these n risky assets. Assume a portfolio must be selected at some initial time t_0 and held until time t. Let $\mathbf{v}_0 = \begin{bmatrix} v_{1,0} & \cdots & v_{n,0} \end{bmatrix}^\top$ and $\mathbf{v} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^\top$ denote the vectors of asset prices at times t_0 and t respectively. The vector $\mathbf{v}_0$ is known whereas $\mathbf{v}$ is a vector of random variables. A vector $\mathbf{h} \in \mathbb{R}^n$ of share holdings in each of the assets defines a portfolio whose values at time t_0 and t are $W_0 := \mathbf{v}_0^\top \mathbf{h}$ and $W := \mathbf{v}^\top \mathbf{h}$ respectively. The value W_0 is known at time t_0 whereas W is a random variable. The gist of portfolio construction is to choose $\mathbf{h}$ to optimize some measure of satisfaction on the random variable W.

It is customary to use the initial portfolio value W_0 as a reference and to write the above problem in terms of the portfolio return

$$r_P = \frac{W - W_0}{W_0}.$$

The return of asset i, which is the same as that of a portfolio entirely invested in asset i, is similarly defined as

$$r_i = \frac{v_i - v_{i,0}}{v_{i,0}}$$

Instead of the vector of holdings $\mathbf{h} \in \mathbb{R}^n$, the portfolio construction problem is often stated in terms of percentage holdings $\mathbf{x} \in \mathbb{R}^n$ where

$$x_i = \frac{h_i v_{i,0}}{W_0} = \frac{h_i v_{i,0}}{\sum_{j=1}^n h_j v_{j,0}}.$$

Observe that $W = \mathbf{v}^\top \mathbf{h}$ can be equivalently written as

$$r_P = \sum_{i=1}^n r_i x_i = \mathbf{r}^\top \mathbf{x}$$

In spite of its wide popularity, this convention runs into difficulties in some cases. For example, the above quantity r_P does not make sense for a long-short portfolio associated with a pairs trading strategy. More broadly, the quantity r_P does not make sense for a situation where the initial value of a portfolio W_0 is zero as when one enters a futures contract or constructs a long-short portfolio with equal long and short cash positions.

As Meucci $(2005, 2010)$ nicely puts it, this difficulty can be amended by assuming that returns are measured relative to some predefined basis value b as opposed to the initial portfolio value W_0. In some cases, it is natural to choose $b = W_0$ but it is more proper to think of b as a general reference point.

To make this idea more precise, we associate with each asset and portfolio a basis b that satisfies the following four properties:

- The basis b for a long position of an asset is positive.
- The basis b is measured in the same unit as the asset values.
- The basis is homogeneous: the basis of k shares of an asset is k times the basis of one share.
- The basis is known at time t_0.

Equipped with this concept, we get a formal and unambiguous definition of asset and portfolio returns:

$$r_i = \frac{v_i - v_{i,0}}{b_i}, \quad r_P = \frac{W - W_0}{b_P}.$$

Likewise, we obtain a formal and unambiguous definition of percentage holdings:

$$x_i = \frac{h_i b_i}{b_P}.$$

Once again, the identity $W = \mathbf{v}^\top \mathbf{h}$ can be equivalently written as

$$r_P = \mathbf{r}^\top \mathbf{x}$$

Throughout this chapter $\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ will denote the vector of percentage holdings of a portfolio in a universe of n risky assets. When it is applicable and evident from the context, we shall assume the usual basis values $b_i = v_{i,0}$ and $b_P = W_0$ respectively.

Two Assets

Suppose we are combining two assets whose random returns are r_1 and r_2. Let

$$\mu_1 := \mathbb{E}(r_1), \quad \mu_2 := \mathbb{E}(r_2),$$

and

$$\sigma_1^2 := \text{var}(r_1), \quad \sigma_2^2 = \text{var}(r_2), \quad \sigma_{12} = \text{cov}(r_1, r_2) = \rho \cdot \sigma_1 \cdot \sigma_2.$$

In this case a portfolio of these two assets is determined by the proportion invested in one of the two assets. Let x denote the proportion in asset 1. Thus the portfolio return is

$$r_P = x \cdot r_1 + (1 - x) \cdot r_2,$$

the portfolio expected return is

$$\begin{aligned} \mu_P := \mathbb{E}(r_P) &= x \cdot \mathbb{E}(r_1) + (1 - x) \cdot \mathbb{E}(r_2) \\ &= x \cdot \mu_1 + (1 - x) \cdot \mu_2 \end{aligned}$$

Two Assets

The portfolio variance is

$$\sigma_P^2 = x^2\sigma_1^2 + (1-x)^2\sigma_2^2 + 2 \cdot x(1-x) \cdot \rho \cdot \sigma_1 \cdot \sigma_2.$$

In the special case when one of the assets, say asset 2 , is the asset with risk-free return r_f we get

$$\mu_P = x \cdot \mu_1 + (1-x) \cdot r_f = r_f + (\mu_1 - r_f)\, x, \quad \sigma_P^2 = x^2\sigma_1^2.$$

In this case the portfolio selection is particularly simple: a target level of expected return μ_P corresponds to one particular portfolio obtained by choosing $x = (\mu_P - r_f) / (\mu_1 - r_f)$. The situation with three assets leads to a more interesting situation.

Three Risky Assets

Suppose now that there are three assets with random returns r_1, r_2, and r_3. As before, let

$$\mu_j = \mathbb{E}(r_j), \quad \sigma_j^2 := \text{var}(r_j) \qquad \text{for } j = 1, 2, 3,$$

and

$$\sigma_{ij} := \text{cov}(r_i, r_j) = \rho_{ij} \cdot \sigma_i \cdot \sigma_j \qquad \text{for } i, j = 1, 2, 3$$

Now a portfolio determines the holdings in the three assets. Let x_j denote the proportion (weight) invested in asset j, for $j = 1, 2, 3$. Notice that these proportions should add up to one if the portfolio is fully invested in the three assets:

$$x_1 + x_2 + x_3 = 1$$

Similar to what we did before, the portfolio return is

$$r_P = r_1 x_1 + r_2 x_2 + r_3 x_3.$$

Three Risky Assets

So the portfolio expected return is

$$\mu_P = \mu_1 x_1 + \mu_2 x_2 + \mu_3 x_3,$$

and the portfolio variance is

$$\sigma_P^2 = \sigma_1^2 x_1^2 + \sigma_2^2 x_2^2 + \sigma_3^2 x_3^2 + 2\left(\sigma_{12} x_1 x_2 + \sigma_{23} x_2 x_3 + \sigma_{13} x_1 x_3\right)$$

Observe that now there are multiple portfolios that can achieve a target expected level of return. A portfolio is efficient if it has minimum risk for a given target return, or equivalently, if it has the maximum expected return for a given target risk. This naturally leads to the following quadratic programming formulation.

Three Risky Assets

To find a portfolio of minimum risk (variance) with expected return at least $\bar{\mu}$ solve the following mean-variance optimization model:

$$\begin{array}{ll} \min_{\mathbf{x}} & \sum_{i=1}^{3} \sigma_{ii} x_i^2 + 2 \sum_{i=1}^{3} \sum_{j=i+1}^{3} \sigma_{ij} x_i x_j \\ \text{s.t.} & \mu_1 x_1 + \mu_2 x_2 + \mu_3 x_3 \geq \bar{\mu} \\ & x_1 + x_2 + x_3 = 1 \end{array}$$

The efficient frontier is the set of efficient portfolios. The efficient frontier is often "visualized" by plotting the expected return against the standard deviation of the efficient portfolios. To generate portfolios on the efficient frontier, we can minimize variance, for varying target return $\bar{\mu}$:

$$\begin{array}{ll} \min_{\mathbf{x}} & \sum_{i=1}^{3} \sigma_{ii} x_i^2 + 2 \sum_{i=1}^{3} \sum_{j=i+1}^{3} \sigma_{ij} x_i x_j \\ \text{s.t.} & \mu_1 x_1 + \mu_2 x_2 + \mu_3 x_3 \geq \bar{\mu} \\ & x_1 + x_2 + x_3 = 1 \end{array}$$

Three Risky Assets

We can also maximize return, for varying target variance $\bar{\sigma}^2 > 0$:

$$\begin{array}{ll} \max_{\mathbf{x}} & \mu_1 x_1 + \mu_2 x_2 + \mu_3 x_3 \\ \text{s.t.} & \sum_{i=1}^{3} \sigma_{ii} x_i^2 + 2 \sum_{i=1}^{3} \sum_{j=i+1}^{3} \sigma_{ij} x_i x_j \leq \bar{\sigma}^2 \\ & x_1 + x_2 + x_3 = 1 \end{array}$$

Or we can maximize quadratic utility, for varying risk aversion $\gamma > 0$:

$$\begin{array}{ll} \max_{\mathbf{x}} & \mu_1 x_1 + \mu_2 x_2 + \mu_3 x_3 - \frac{\gamma}{2} \left(\sum_{i=1}^{3} \sigma_{ii} x_i^2 + 2 \sum_{i=1}^{3} \sum_{j=i+1}^{3} \sigma_{ij} x_i x_j \right) \\ \text{s.t.} & x_1 + x_2 + x_3 = 1 \end{array}$$

Any Number of Risky Assets

Let us now take a leap to the most general case. Assume we have n risky assets. Let $\mathbf{r} \in \mathbb{R}^n$ be the n-dimensional random vector of returns, i.e., r_i denotes the return of asset i between times t_0 and t. Let $\boldsymbol{\mu} \in \mathbb{R}^n$ denote the vector of expected returns, and $\mathbf{V} \in \mathbb{R}^{n \times n}$ denote the return covariance matrix. More precisely,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix}$$

where $\mu_i := \mathbb{E}(r_i), \sigma_{ij} := \operatorname{cov}(r_i, r_j), i, j = 1, \ldots, n.$

From the linearity properties of expectation, it follows that the expected return and variance of a given portfolio $\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ of the risky assets are respectively

$$\boldsymbol{\mu}^\top \mathbf{x} = \sum_{j=1}^{n} \mu_j x_j$$

and

$$\mathbf{x}^\top \mathbf{V} \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} x_i x_j = \sum_{i=1}^{n} \sigma_{ii} x_i^2 + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sigma_{ij} x_i x_j$$

The problem of selecting a portfolio can be formally stated as a tradeoff between these two components. A fully invested portfolio is efficient if it has minimum risk for a given level of return, or equivalently if it has maximum expected return for a given level of risk.

A fully invested efficient portfolio can then be characterized as the solution to the following quadratic program:

$$\begin{aligned} \max_{\mathbf{x}} \quad & \boldsymbol{\mu}^\top \mathbf{x} - \frac{1}{2}\gamma \cdot \mathbf{x}^\top \mathbf{V}\mathbf{x} \\ & \mathbf{1}^\top \mathbf{x} = 1 \end{aligned} \tag{6.1}$$

for some risk-aversion coefficient $\gamma > 0$. The set of efficient portfolios can also be obtained as the set of solutions to the quadratic program:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{x}^\top \mathbf{V}\mathbf{x} \\ \text{s.t.} \quad & \boldsymbol{\mu}^\top \mathbf{x} \geq \bar{\mu} \\ & \mathbf{1}^\top \mathbf{x} = 1, \end{aligned} \tag{6.2}$$

and also as the set of solutions to

$$\begin{aligned} \max_{\mathbf{x}} \quad & \boldsymbol{\mu}^\top \mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}^\top \mathbf{V}\mathbf{x} \leq \bar{\sigma}^2 \\ & \mathbf{1}^\top \mathbf{x} = 1 \end{aligned} \tag{6.3}$$

by varying $\bar{\mu}$ and $\bar{\sigma}$ respectively.

We shall refer to the equivalent mean-variance models (6.1), (6.2), and 6.3 as the basic mean-variance models as they include only the following three essential components: mean and variance of return, and the full investment constraint. Observe that these three optimization models are convex because the quadratic function $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{V}\mathbf{x}$ is convex as the covariance matrix $\mathbf{V}$ is positive semidefinite.

Asset Allocation and Security Selection

There are two distinct levels of portfolio analysis that are amenable to meanvariance models. The conventional top-down investment approach to portfolio construction consists of two main steps, namely asset allocation and security selection.
On the one hand, the asset allocation decision is concerned with portfolio choices among broad asset classes. At the coarsest level, these asset classes could be stocks, bonds, and cash. At a more refined level, some of these broad asset classes could be subdivided. For instance, stocks can be divided according to geography or market capitalization. The asset allocation decision involves only a small number of assets, typically ranging from a handful to a dozen or so. It generally involves simple constraints such as budget constraints and upper and lower bounds on individual positions.

On the other hand, the security selection decision is concerned with the specific securities within each particular asset class. For instance, if the relevant asset class is equities in the S&P 500 market index, then the security selection problem is concerned with the specific portfolio holdings at the individual stock level. The security selection problem typically involves a large number of securities, ranging from a few hundred to potentially thousands.

Minimum Risk and Characteristic Portfolios

Consider the simplified version of (6.1) that is obtained in the limit when $\gamma \to \infty$:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{x}^\top \mathbf{V}\mathbf{x} \\ & \mathbf{1}^\top \mathbf{x} = 1 \end{aligned} \tag{6.4}$$

The model (6.4) corresponds to the problem of finding the minimum-risk fully invested portfolio. We discussed this problem in Example 5.7 where the optimal solution was shown to be

$$\mathbf{x}^* = \frac{1}{\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1}} \mathbf{V}^{-1}\mathbf{1}$$

A related problem that is often of interest is to find the minimum-risk portfolio with unit exposure to a vector of attributes a associated with the assets. As we will see later, some interesting attributes could be the betas of the assets relative to a benchmark, the asset volatilities, or the asset expected returns. The characteristic portfolio of a vector of attributes a is the solution to the problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{x}^\top \mathbf{V} \mathbf{x} \\ & \mathbf{a}^\top \mathbf{x} = 1 \end{aligned} \tag{6.5}$$

Using the solution of (5.9) obtained in Chapter 5, it follows that the solution to (6.5) is

$$\mathbf{x}^* = \frac{1}{\mathbf{a}^\top \mathbf{V}^{-1} \mathbf{a}} \mathbf{V}^{-1} \mathbf{a}$$

Observe that a characteristic portfolio $\mathbf{x}^* = \left(1/\mathbf{a}^\top \mathbf{V}^{-1} \mathbf{a}\right) \mathbf{V}^{-1} \mathbf{a}$ is not necessarily fully invested as its components may not necessarily add up to one. Observe that the variance of the characteristic porfolio $\mathbf{x}^* = \left(1/\mathbf{a}^\top \mathbf{V}^{-1} \mathbf{a}\right) \mathbf{V}^{-1} \mathbf{a}$ is

$$\left(\mathbf{x}^*\right)^\top \mathbf{V} \mathbf{x}^* = \frac{1}{\mathbf{a}^\top \mathbf{V}^{-1} \mathbf{a}}$$

Two-Fund Separation Theorem

Consider the basic mean-variance model

$$\begin{aligned} \max_{\mathbf{x}} & \boldsymbol{\mu}^\top \mathbf{x} - \frac{1}{2}\gamma \cdot \mathbf{x}^\top \mathbf{V}\mathbf{x} \\ & \mathbf{1}^\top \mathbf{x} = 1 \end{aligned} \tag{6.6}$$

for some risk-aversion coefficient $\gamma > 0$. We next derive an interesting result often called the two-fund separation theorem. The theorem states that every fully invested efficient portfolio is a combination of two particular efficient portfolios. Applying the optimality conditions (5.10) from Theorem 5.6 to problem (6.6) we obtain the solution

$$\mathbf{x}^* = \lambda \cdot \frac{1}{\mathbf{1}^\top \mathbf{V}^{-1}\boldsymbol{\mu}} \mathbf{V}^{-1}\boldsymbol{\mu} + (1-\lambda) \cdot \frac{1}{\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1}} \mathbf{V}^{-1}\mathbf{1}$$

where $\lambda = \mathbf{1}^\top \mathbf{V}^{-1}\boldsymbol{\mu}/\gamma$. The following two-fund theorem readily follows.

Theorem (6.1 Two-fund theorem)

Consider model (6.6) for some $\gamma > 0$. There exist two efficient portfolios (funds), namely

$$\frac{1}{\mathbf{1}^\top \mathbf{V}^{-1}\boldsymbol{\mu}}\mathbf{V}^{-1}\boldsymbol{\mu} \text{ and } \frac{1}{\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1}}\mathbf{V}^{-1}\mathbf{1},$$

such that every efficient portfolio, that is, every solution to (6.6), is a combination of these two portfolios.

Observe that one of the two portfolios in the two-fund theorem is the minimum risk portfolio $\left(1/\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1}\right)\mathbf{V}^{-1}\mathbf{1}$ and the other one is a multiple of the characteristic portfolio $\left(1/\boldsymbol{\mu}^\top \mathbf{V}^{-1}\boldsymbol{\mu}\right)\mathbf{V}^{-1}\boldsymbol{\mu}$ of the vector of attributes $\boldsymbol{\mu}$.

One-Fund Separation Theorem

We next derive the one-fund or mutual fund separation theorem. This result is similar in spirit to the two-fund separation theorem. It states that if there is a risk-free asset, then every efficient portfolio is a combination of the risk-free asset and a particular fund.

Consider the case when, in addition to the universe of n risky assets, there is an additional asset $n+1$ with risk-free return r_f. In this case, problem (6.1) extends as follows

$$\begin{aligned} \max_{\mathbf{x}, x_{n+1}} \quad & \boldsymbol{\mu}^\top \mathbf{x} + r_f \cdot x_{n+1} - \frac{1}{2}\gamma \cdot \mathbf{x}^\top \mathbf{V}\mathbf{x} \\ & \mathbf{1}^\top \mathbf{x} + x_{n+1} = 1 \end{aligned} \tag{6.7}$$

By substituting $x_{n+1} = 1 - \mathbf{1}^\top \mathbf{x}$ in the objective and dropping the constraint, problem (6.7) can be rewritten as the following unconstrained optimization problem:

$$\max_{\mathbf{x}} \left(\boldsymbol{\mu} - r_f \mathbf{1}\right)^\top \mathbf{x} - \frac{1}{2}\gamma \cdot \mathbf{x}^\top \mathbf{V}\mathbf{x}$$

Applying the optimality conditions (5.4) from Theorem 5.2, we obtain the following solution to (6.7):

$$\mathbf{x}^* = \frac{1}{\gamma} \cdot \mathbf{V}^{-1}\left(\boldsymbol{\mu} - r_f \mathbf{1}\right) = \lambda \cdot \frac{1}{\mathbf{1}^\top \mathbf{V}^{-1}\left(\boldsymbol{\mu} - r_f \mathbf{1}\right)} \mathbf{V}^{-1}\left(\boldsymbol{\mu} - r_f \mathbf{1}\right), x_{n+1}^* = 1 - \mathbf{1}^\top \mathbf{x}^*,$$

where $\lambda = \mathbf{1}^\top \mathbf{V}^{-1}\left(\boldsymbol{\mu} - r_f \mathbf{1}\right)/\gamma$. The following one-fund theorem readily follows.

Theorem (6.2 One-fund theorem)

Suppose the investment universe includes n risky assets and a risk-free asset. Then there exists a fully invested efficient portfolio (fund) namely

$$\frac{1}{\mathbf{1}^\top \mathbf{V}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})} \mathbf{V}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})$$

such that every efficient portfolio - that is, every solution to (6.7) for some $\gamma > 0$ - is a combination of this portfolio and the risk-free asset.

The portfolio $\left[1/\mathbf{1}^\top \mathbf{V}^{-1}(\boldsymbol{\mu} - r_f\mathbf{1})\right]\mathbf{V}^{-1}(\boldsymbol{\mu} - r_f\mathbf{1})$ is called the tangency portfolio. This name is motivated by the geometric interpretation illustrated in Figure 1. Consider the plot of expected return versus standard deviation for the efficient frontier portfolios. The portfolio $\left[1/\mathbf{1}^\top \mathbf{V}^{-1}(\boldsymbol{\mu} - r_f\mathbf{1})\right]\mathbf{V}^{-1}(\boldsymbol{\mu} - r_f\mathbf{1})$ lies exactly at the tangency point on this frontier defined by the straight line emerging from the point $(0, r_f)$. The point $(0, r_f)$ corresponds to the expected return versus standard deviation of the risk-free asset. The tangency line is also known as the capital allocation line (CAL) as it corresponds to portfolios with different allocations of capital between the tangency portfolio and the risk-free asset.

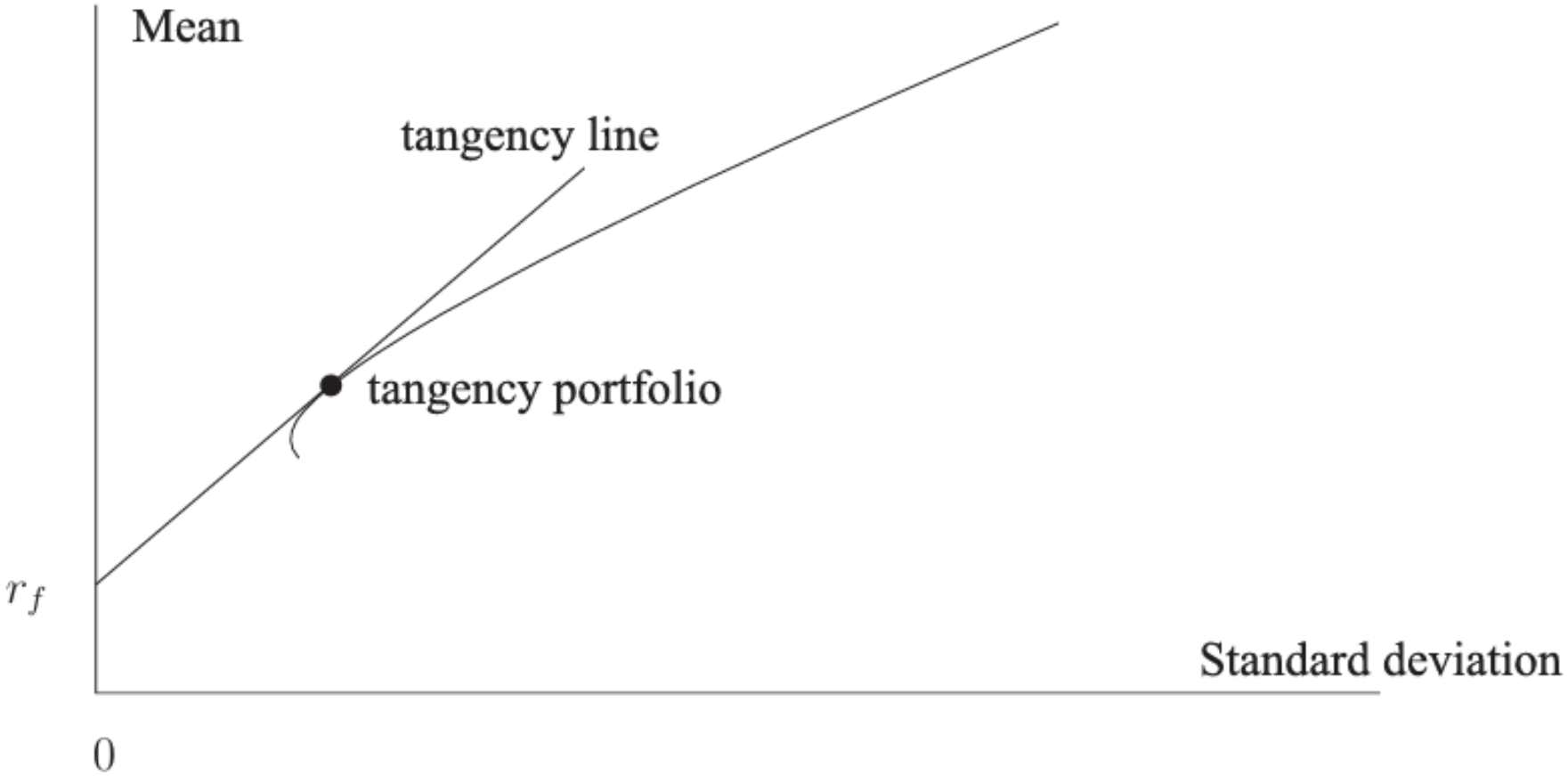

Figure 1: Tangency portfolio

Capital Asset Pricing Model (CAPM)

Under suitable equilibrium assumptions the tangency portfolio discussed above yields the main mathematical foundation for the capital asset pricing model (CAPM), a fundamental asset pricing model in financial economics. The key step in this derivation is that, in equilibrium, the tangency portfolio is precisely the market portfolio $\mathbf{x}_M$. That is,

$$\mathbf{x}_M = \frac{1}{\mathbf{1}^\top \mathbf{V}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})} \mathbf{V}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}). \tag{6.8}$$

From (6.8) we readily obtain

$$\mathbf{V}\mathbf{x}_M = \frac{1}{\mathbf{1}^\top \mathbf{V}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})} (\boldsymbol{\mu} - r_f \mathbf{1}), \tag{6.9}$$

and

$$\mathbf{x}_M^\top \mathbf{V}\mathbf{x}_M = \frac{(\boldsymbol{\mu} - r_f \mathbf{1})^\top \mathbf{x}_M}{\mathbf{1}^\top \mathbf{V}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})} = \frac{\mu_M - r_f}{\mathbf{1}^\top \mathbf{V}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})}, \tag{6.10}$$

where $\mu_M = \boldsymbol{\mu}^\top \mathbf{x}_M$ is the expected value of the market portfolio return. Combining (6.9) and (6.10) we get

$$\begin{aligned} \boldsymbol{\mu} - r_f \mathbf{1} &= \mathbf{1}^\top \mathbf{V}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}) \mathbf{V} \mathbf{x}_M \\ &= \left(\frac{1}{\mathbf{x}_M^\top \mathbf{V} \mathbf{x}_M} \mathbf{V} \mathbf{x}_M \right) (\mu_M - r_f) = \boldsymbol{\beta} \cdot (\mu_M - r_f) \end{aligned} \tag{6.11}$$

where $\boldsymbol{\beta} = \left(1/\mathbf{x}_M^\top \mathbf{V} \mathbf{x}_M\right) \mathbf{V} \mathbf{x}_M$. The above can be equivalently stated as

$$\mu_j - r_f = \beta_j (\mu_M - r_f), \text{ where } \beta_j = \frac{\sigma_{j,M}}{\sigma_M^2} \text{ for } j = 1, \ldots, n. \tag{6.12}$$

Equation (6.11) or its equivalent (6.12) is the formal statement of the capital asset pricing model (CAPM). The CAPM postulates that the excess return of asset j is determined entirely by its beta coefficient times the excess return of the market.

In the expression (6.12), $\sigma_{j,M}$ denotes the covariance between the return of asset j and the return of the market portfolio, and σ_M^2 denotes the variance of the market portfolio return. The last two quantities in turn have the following expressions in terms of the covariance matrix $\mathbf{V}$:

$$\sigma_{j,M} = \text{cov}\left(r_j, r_M\right) = \left(\mathbf{V}\mathbf{x}_M\right)_j, \quad \sigma_M^2 = \text{var}\left(r_M\right) = \mathbf{x}_M^\top \mathbf{V}\mathbf{x}_M$$

Common Constraints

Aside from a target expected return or a target variance, the only portfolio constraint in the basic mean-variance model is the full investment constraint

$$\mathbf{1}^\top \mathbf{x} = 1.$$

Furthermore, this constraint disappears if the portfolio is allowed to include holdings in a risk-free asset. In both cases the individual portfolio holdings could in principle take arbitrary positive and negative values as there is no explicit restriction on them. This motivates the following types of constraints that are often included in a mean-variance model:

- Budget constraints, such as fully invested portfolios.
- Upper and/or lower bounds on the size of individual positions.
- Upper and/or lower bounds on exposure to industries or sectors.
- Leverage constraints such as long-only, or 130/30 constraints.
- Turnover constraints.

The above types of constraints replace the single portfolio constraint

$$\mathbf{1}^\top \mathbf{x} = 1$$

by a more elaborate set of constraints of the form

$$\begin{aligned} \mathbf{A}\mathbf{x} &= \mathbf{b} \\ \mathbf{D}\mathbf{x} &\geq \mathbf{d}. \end{aligned}$$

Consequently, we get the following general version of the basic mean-variance model (6.1):

$$\begin{aligned} \max_{\mathbf{x}} \boldsymbol{\mu}^\top \mathbf{x} - \frac{1}{2}\gamma \cdot \mathbf{x}^\top \mathbf{V}\mathbf{x} \\ \mathbf{A}\mathbf{x} = \mathbf{b} \\ \mathbf{D}\mathbf{x} \geq \mathbf{d} \end{aligned} \tag{6.13}$$

The set of portfolios obtained via the model (6.13) can also be obtained via the following two equivalent models. The first one enforces a target expected return:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{x}^\top \mathbf{V}\mathbf{x} \\ \text{s.t.} \quad & \boldsymbol{\mu}^\top \mathbf{x} \geq \bar{\mu} \\ & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{D}\mathbf{x} \geq \mathbf{d}. \end{aligned} \tag{6.14}$$

The second one enforces a target variance of return:

$$\begin{aligned} \max_{\mathbf{x}} \quad & \boldsymbol{\mu}^\top \mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}^\top \mathbf{V}\mathbf{x} \leq \bar{\sigma}^2 \\ & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{D}\mathbf{x} \geq \mathbf{d} \end{aligned} \tag{6.15}$$

The models (6.13),(6.14), and (6.15) are still convex quadratic optimization models. Unlike the basic mean-variance model, they generally do not have an analytical closed-form solution due to the additional inequality constraints. However, they can be solved numerically very efficiently via optimization solvers. We next discuss how some of the above five types of constraints can be incorporated into a mean-variance model. The first three types of constraints have straightforward formulations. We concentrate on the last two, namely, leverage constraints and turnover constraints. A long-only constraint can readily be enforced via $\mathbf{x} \geq 0$. A relaxed version of this constraint, popular in certain contexts, is not to rule out leverage altogether but to limit it. For instance, a "130/30" leverage constraint means that the total value of the holdings in short positions must be at most 30% of the portfolio value.

In general, suppose that we want the value of the total short positions to be at most L. This means that we want to enforce the following restriction:

$$\sum_{j=1}^{n} \min(x_j, 0) \geq -L \quad \Leftrightarrow \quad \sum_{j=1}^{n} \max(-x_j, 0) \leq L$$

Although this is a correct mathematical formulation of the constraint, it is not ideal for computational purposes because of the non-smooth terms $\max(-x_j, 0)$.

In particular, if a constraint were written in this form the resulting meanvariance model would not be a quadratic program. To formulate this constraint efficiently in the quadratic optimization model, we trade terms of the form $\max(-x_j, 0)$ for new terms involving possibly new variables and linear inequalities. To that end, add the new vector of variables $\mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top$ and constraints

$$\mathbf{x} \geq -\mathbf{y}$$

$$\sum_{j=1}^{n} y_j \leq L$$

$$\mathbf{y} \geq \mathbf{0}$$

A turnover constraint is a constraint on the total change in the portfolio positions. This constraint is generally included as a way to limit certain kinds of costs such as taxes and transaction costs. Suppose that we have an initial portfolio $\mathbf{x}^0 = \begin{bmatrix} x_1^0 & \cdots & x_n^0 \end{bmatrix}^\top$ and we want to ensure that the new portfolio incurs a total turnover no larger than h. This means that we want to enforce the restriction

$$\sum_{j=1}^{n} \left| x_j^0 - x_j \right| \leq h$$

To formulate this constraint efficiently in the quadratic optimization model, add the new vector of variables $\mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top$ and constraints

$$x_j - x_j^0 \leq y_j$$
$$x_j^0 - x_j \leq y_j$$
$$\sum_{j=1}^{n} y_j \leq h$$

(see Exercise 6.3). The total turnover $\sum_{j=1}^{n} \left| x_j^0 - x_j \right|$ is also sometimes called the two-sided turnover.

Maximizing the Sharpe Ratio

The three equivalent mean-variance models (6.13),(6.14), and (6.15) define a frontier of efficient portfolios. These portfolios are determined by some optimal tradeoff of expected return and variance, or equivalently, standard deviation of return. The ratio of expected return to standard deviation, called Sharpe ratio or reward-to-risk ratio, singles out the efficient portfolio that offers the highest reward per measure of risk.

Definition (6.3 Sharpe ratio)

The Sharpe ratio of a given portfolio $\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ is the ratio of its expected return to its volatility (standard deviation) of return:

$$\text{Sharpe ratio} := \frac{\boldsymbol{\mu}^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{V} \mathbf{x}}}.$$

As we further elaborate in the next sections, sometimes $\boldsymbol{\mu}$ may not necessarily stand for the vector of expected absolute returns but instead it may make sense for $\boldsymbol{\mu}$ to stand for the vector of expected relative returns. In particular, if there is a risk-free asset, in the above definition of the Sharpe ratio it is usual to assume that $\boldsymbol{\mu}$ stands for the vector of expected excess returns. The excess return of an asset is simply the difference of its return and the risk-free return.

As an alternative or a complement to the equivalent mean-variance models (6.13),(6.14), and (6.15), consider the problem of finding the efficient portfolio with maximum Sharpe ratio. The natural formulation for this problem is the following:

$$\begin{aligned} \max_{\mathbf{x}} \ & \frac{\boldsymbol{\mu}^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{V}_\mathbf{x}}} \\ \text{s.t.} \ & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{D}\mathbf{x} \geq \mathbf{d} \end{aligned} \tag{6.16}$$

This natural formulation is evidently not a quadratic optimization model. Furthermore, the formulation is not convex as the objective function is not convex. We next show that this problem can be recast as a quadratic convex optimization problem via a suitable homogenization. To this end, make the following mild assumptions:

- There is a feasible portfolio $\mathbf{x}$ such that $\boldsymbol{\mu}^\top \mathbf{x} > 0$.
- The matrices $\mathbf{A}, \mathbf{D}$ and vector $\boldsymbol{\mu}$ satisfy the following technical condition:

$$\mathbf{Az} = \mathbf{0}, \mathbf{Dz} \geq \mathbf{0} \quad \Rightarrow \quad \boldsymbol{\mu}^\top \mathbf{z} \leq 0$$

The latter condition readily holds when the following stronger but easier to verify condition holds:

$$\mathbf{Az} = \mathbf{0}, \mathbf{Dz} \geq \mathbf{0} \quad \Rightarrow \quad \mathbf{z} = \mathbf{0}$$

The above assumptions ensure the soundness of the approach described next. To see what goes wrong when these assumptions do not hold, see the exercises at the end of the chapter.

The gist of the reformulation of (6.16) as a quadratic optimization problem is the following homogenization. Consider the change of variables obtained by putting $\mathbf{z} := \kappa \mathbf{x}$, where $\kappa > 0$ is a new scalar variable. The problem (6.16) can be rewritten as

$$\begin{aligned} \max_{\mathbf{z},\kappa} \ & \frac{\boldsymbol{\mu}^\top \mathbf{z}}{\sqrt{\mathbf{z}^\top \mathbf{V} \mathbf{z}}} \\ \text{s.t.} \ & \mathbf{A}\frac{\mathbf{z}}{\kappa} = \mathbf{b} \\ & \mathbf{D}\frac{\mathbf{z}}{\kappa} \geq \mathbf{d} \\ & \kappa > 0. \end{aligned} \tag{6.17}$$

The assumption $\boldsymbol{\mu}^\top \mathbf{x} > 0$ for some feasible $\mathbf{x}$ implies that we can choose $\kappa > 0$ such that $\boldsymbol{\mu}^\top \mathbf{z} = 1$. Using this together with the second assumption, it follows that the problem (6.17) is equivalent to

$$\begin{aligned} \min_{\mathbf{z},\kappa} \quad & \mathbf{z}^\top \mathbf{V} \mathbf{z} \\ \text{s.t.} \quad & \boldsymbol{\mu}^\top \mathbf{z} = 1 \\ & \mathbf{A}\mathbf{z} - \mathbf{b}\kappa = \mathbf{0} \\ & \mathbf{D}\mathbf{z} - \mathbf{d}\kappa \geq \mathbf{0} \\ & \kappa \geq 0. \end{aligned} \tag{6.18}$$

As the exercises at the end of the chapter detail, this approach also yields the following characterization of the portfolio with maximum Sharpe ratio in the case when we only include the full investment constraint $\mathbf{1}^\top \mathbf{x} = 1$.

Proposition (6.4)

Suppose the minimum-risk portfolio $\left(1/\mathbf{1}^\top \mathbf{V}^{-1}\mathbf{1}\right)\mathbf{V}^{-1}\mathbf{1}$ *has positive expected return; that is,* $\boldsymbol{\mu}^\top \mathbf{V}^{-1}\mathbf{1} > 0$. *Then the solution to the following maximum Sharpe ratio problem*

$$\begin{aligned} &\max_{\mathbf{x}} \frac{\boldsymbol{\mu}^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{V}}} \\ &s.t.\ \mathbf{1}^\top \mathbf{x} = 1 \end{aligned} \tag{6.19}$$

is the tangency portfolio

$$\mathbf{x}^* = \frac{1}{\mathbf{1}^\top \mathbf{V}^{-1}\boldsymbol{\mu}} \mathbf{V}^{-1}\boldsymbol{\mu}$$

Portfolio Management Relative to a Benchmark

In an investment portfolio, the security selection problem is concerned with determining the holdings of specific securities within a given asset class. It is customary to manage and evaluate the portfolio of securities relative to some predefined benchmark portfolio that represents a particular asset class. The benchmark portfolio provides a reference point. It serves the role of the market portfolio if the investment universe is restricted to the particular asset class that the benchmark represents. The management of a portfolio of securities relative to a benchmark could be passive or active. The goal of the former is to replicate the benchmark whereas the goal of the latter is to beat the benchmark.

Systematic (Beta) and Individual (Alpha) Returns

Both passive and active management rely on a fundamental decomposition of individual securities return into systematic and individual (or residual) components. The former is the component of return that can be explained by the security exposure to the benchmark. The latter is the component of return that is idiosyncratic to the individual security.

To make the above decomposition more precise, assume the investment universe determined by a particular asset class includes n individual securities. Let r_i denote the excess return of security i for $i = 1, \ldots, n$. Let r_B denote the excess return of the benchmark.

The return of security i can be decomposed via the following linear regression model:

$$r_i = \beta_i r_B + \theta_i,$$

where θ_i is the component of return uncorrelated to r_B; that is, $\mathrm{cov}(r_B, \theta_i) = 0$.

The coefficient β_i is the beta of security i relative to the benchmark B and is given by

$$\beta_i := \frac{\operatorname{cov}(r_i, r_B)}{\operatorname{var}(r_B)}.$$

The term $\beta_i r_B$ is the systematic component of return of security i. The term θ_i is the residual component of return of security i. The alpha of security i is the expected value of the residual return θ_i :

$$\alpha_i = \mathbb{E}(\theta_i)$$

Consider a portfolio of securities with percentage holdings $\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$. The above type of decomposition also applies to the portfolio return

$$r_P := \mathbf{r}^\top \mathbf{x} = r_1 x_1 + \cdots + r_n x_n$$

That is, we can decompose the portfolio return r_P as

$$r_P = \beta_P r_B + \theta_P,$$

where the systematic and residual components of the portfolio return are respectively

$$\beta_P r_B = \left(\boldsymbol{\beta}^\top \mathbf{x}\right) r_B = \left(\beta_1 x_1 + \cdots + \beta_n x_n\right) r_B$$

and

$$\theta_P = \boldsymbol{\theta}^\top \mathbf{x} = \theta_1 x_1 + \cdots + \theta_n x_n$$

Furthermore, it is easy to see that the beta and alpha of the portfolio are respectively

$$\beta_P = \boldsymbol{\beta}^\top \mathbf{x} = \beta_1 x_1 + \cdots + \beta_n x_n$$

and

$$\alpha_P = \mathbb{E}\left(\theta_P\right) = \boldsymbol{\alpha}^\top \mathbf{x} = \alpha_1 x_1 + \cdots + \alpha_n x_n$$

Active Return, Tracking Error, Information Ratio

Consider a portfolio with percentage holdings $\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$. The active return of the portfolio is the difference between the portfolio return and the benchmark return:

$$\mathbf{r}^\top \mathbf{x} - r_B$$

If the portfolio of benchmark holdings is $\mathbf{x}^B = \begin{bmatrix} x_1^B & \cdots & x_n^B \end{bmatrix}$, then $r_B = \mathbf{r}^\top \mathbf{x}^B$ and thus the active return can also be written as

$$\mathbf{r}^\top \mathbf{x} - r_B = \mathbf{r}^\top \left(\mathbf{x} - \mathbf{x}^B \right).$$

The vector $\mathbf{x} - \mathbf{x}^B$ is the vector of active holdings of the portfolio.

The active risk or tracking error ψ^2 of a portfolio is the standard deviation of the portfolio active return. In other words,

$$\psi^2 := \text{var}\left(\mathbf{r}^\top \left(\mathbf{x} - \mathbf{x}^B\right)\right)$$

Some straightforward matrix calculations show that if $\mathbf{V}$ is the covariance matrix of securities returns, then

$$\psi^2 = \text{var}\left(\mathbf{r}^\top \left(\mathbf{x} - \mathbf{x}^B\right)\right) = \left(\mathbf{x} - \mathbf{x}^B\right)^\top \mathbf{V} \left(\mathbf{x} - \mathbf{x}^B\right).$$

A straightforward calculation also shows that the active risk can be decomposed as

$$\psi^2 = \left(\beta_P - 1\right)^2 \sigma_B^2 + \omega_P^2,$$

where $\sigma_B^2 = \text{var}\left(r_B\right)$ and $\omega_P^2 = \text{var}\left(\theta_P\right)$. The first term $\left(\beta_P - 1\right)^2 \sigma_B^2$ is the component of active risk due to the active beta $\beta_P - 1$ of the portfolio. The second term ω_P^2 is the portfolio residual risk. Observe that the active risk and residual risk are the same when $\beta_P = 1$.

Definition (6.5 Information ratio)

The information ratio (IR) of a portfolio P is the ratio of expected residual return to volatility (standard deviation) of residual return:

$$IR_P := \frac{\alpha_P}{\omega_P}.$$

Portfolio Optimization with Benchmark Considerations

The consideration of a benchmark in portfolio construction typically leads to mean-variance models that include some adjustments and constraints induced by the benchmark.
The following are some of the most common adjustments and constraints when a mean-variance model is used for portfolio construction relative to a benchmark:

- Use expected residual returns $\boldsymbol{\alpha}^\top \mathbf{x}$ instead of expected total return $\boldsymbol{\mu}^\top \mathbf{x}$.
- Use active risk $\psi^2 = \left(\mathbf{x} - \mathbf{x}^B\right)^\top \mathbf{V} \left(\mathbf{x} - \mathbf{x}^B\right)$ instead of total risk $\mathbf{x}^\top \mathbf{V} \mathbf{x}$.
- Bounds on the size of active positions. These adjustments and constraints are typically of the form

$$L_i \leq x_i - x_i^B \leq U_i, i = 1, \ldots, n,$$

that restrict the deviations between the portfolio holdings and the benchmark holdings.

- Bounds on the beta of the portfolio. Again this type of constraint is typically of the form

$$L \leq \boldsymbol{\beta}^{\top}\mathbf{x} - 1 \leq U$$

As an example, the optimization problem might be

$$\begin{aligned} \max_{\mathbf{x}} \quad & \boldsymbol{\alpha}^{\top}\mathbf{x} \\ \text{s.t.} \quad & \left(\mathbf{x} - \mathbf{x}^{B}\right)^{\top} \mathbf{V} \left(\mathbf{x} - \mathbf{x}^{B}\right) \leq \bar{\psi}^{2} \\ & \mathbf{1}^{\top}\mathbf{x} = 1 \\ & L \leq \boldsymbol{\beta}^{\top}\mathbf{x} - 1 \leq U. \end{aligned} \tag{6.20}$$

Estimation of Inputs to Mean–Variance Models

The estimation of input parameters, namely the covariance matrix of returns $\mathbf{V}$ and the vector of total expected returns $\boldsymbol{\mu}$ or residual expected returns $\boldsymbol{\alpha}$, is one of the most critical and challenging steps in the use of mean-variance models. We next describe some of the central ideas that underlie most popular approaches to this fundamental problem.

Throughout this section assume the investment universe has n assets and let r_i denote the excess return of asset i for $i = 1, \ldots, n$. Let $\mathbf{r} \in \mathbb{R}^n$ denote the vector of excess returns. A rudimentary approach to estimate $\boldsymbol{\mu}$ and $\mathbf{V}$ via sample means and sample covariances is based on historical data. More precisely, given a time series of realized excess returns $\mathbf{r}(1), \mathbf{r}(2), \ldots, \mathbf{r}(T)$, the vectors of sample means and sample covariance are respectively

$$\hat{\boldsymbol{\mu}} := \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}(t), \quad \hat{\mathbf{V}} := \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{r}(t)-\hat{\boldsymbol{\mu}})(\mathbf{r}(t)-\hat{\boldsymbol{\mu}})^{\top}.$$

The vector $\hat{\mu}$ and matrix $\hat{\mathbf{V}}$ provide estimates of μ and $\mathbf{V}$. However, these estimators have three major shortcomings:

- The sample mean and sample covariance do not incorporate other data that could contain useful forecasting information.
- For an investment universe with n assets, there are a total of $n + \frac{1}{2}n(n+1)$ $= \frac{1}{2}n(n+3)$ different parameters to estimate. Although this could be manageable for a small asset allocation model, it is not viable for an equity portfolio management model, as the number of securities n in a stock universe could easily range in the hundreds or thousands.
- The sample mean and sample covariance inevitably contain a fair amount of estimation errors, which, as we further explain in the next chapter, are magnified by the mean-variance optimizer.

The first two shortcomings above can be largely mitigated by assuming some kind of structure in the portfolio returns $\mathbf{r}$, as the following subsections detail. The next chapter is devoted entirely to the third shortcoming.

Single-Factor Model

The task of estimating a risk model can be drastically simplified by assuming that each asset has two components of risk: market risk and residual risk. This is a single-factor risk model. Historically this model was introduced by Sharpe as an intellectual precursor of the capital asset pricing model (CAPM). The model assumes that excess returns are decomposed as in the following regression model:

$$r_i = \beta_i r_M + \theta_i.$$

Here β_i is the beta of asset i, and θ_i is its residual return, uncorrelated with r_M. The model also assumes that the residual returns θ_i are uncorrelated with each other. The rationale for the model is that a single common factor r_M, typically the return of the market portfolio, accounts for all of the common shocks between pairs of assets. The parameter β_i is also called the factor loading or factor exposure of asset i. The component θ_i is also called the residual or specific return of asset i, as it is the portion of r_i not accounted for by the common factor r_M.

A bit of algebra shows that in this model the expected return of asset i is

$$\mathbb{E}(r_i) = \beta_i \mathbb{E}(r_M) + \mathbb{E}(\theta_i),$$

the covariance between two different assets i and j is

$$\text{cov}(r_i, r_j) = \beta_i \beta_j \sigma_M^2,$$

and the variance of asset i is

$$\text{var}(r_i) = \beta_i^2 \sigma_M^2 + \omega_i^2,$$

where $\sigma_M^2 = \text{var}(r_M), \omega_i^2 = \text{var}(\theta_i)$. Using matrix-vector notation, the single-factor risk model assumption can be succinctly written as

$$\mathbf{r} = \boldsymbol{\beta} r_M + \boldsymbol{\theta}$$

and the vector of expected returns and covariance matrix can be written as

$$\mathbb{E}(\mathbf{r}) = \boldsymbol{\beta}\mathbb{E}\left(r_M\right) + \mathbb{E}(\boldsymbol{\theta}), \mathbf{V} = \sigma_M^2 \boldsymbol{\beta}\boldsymbol{\beta}^\top + \mathbf{D},$$

where $\mathbf{D}$ is the diagonal matrix $\mathbf{D} = \text{diag}\left(\omega_1^2, \ldots, \omega_n^2\right) = \text{cov}(\boldsymbol{\theta})$.
We observe that under the single-factor model, the estimation of the covariance matrix only requires the estimation of $\boldsymbol{\beta}, \sigma_M^2$, and $\mathbf{D}$. That is a total of $n + 1 + n = 2n + 1$ parameters in contrast to the $\frac{1}{2}n(n+1)$ parameters for a non-structured covariance matrix. The particular structure of the covariance matrix for a singlefactor risk model also enables the derivation of some interesting properties of minimum-risk portfolios.

A basic estimation of the parameters of a single-factor model can be performed as follows. Assume we have some historical data of realized returns $\mathbf{r}(1), \ldots, \mathbf{r}(T)$ as well as the corresponding returns for the factor $\mathbf{r}_M(1), \ldots, r_M(T)$. Use these data to run n simple linear regressions

$$r_i = \alpha_i + \beta_i r_M + \epsilon_i, \quad i = 1, \ldots, n$$

Each of these linear regressions yields estimates $\hat{\beta}_i$ of β_i, $\hat{\alpha}_i$ of $\mathbb{E}(\theta_i)$, and $\hat{\omega}_i$ of $\operatorname{var}(\epsilon_i) = \operatorname{var}(\theta_i)$. Using the historical data $r_M(1), \ldots, r_M(T)$ for the factor, we can also obtain an estimate $\hat{\sigma}_M^2$ of $\operatorname{var}(r_M)$.
The above basic regression method can be enhanced to produce more accurate estimates. In particular, it is known that the quality of the estimates of $\boldsymbol{\beta}$ can be improved via a shrinkage procedure as explained by Blume (1975). The basic idea, which can be traced back to the classical work of Stein (1956), is that improved estimates on $\boldsymbol{\beta}$ can be obtained by taking a convex combination of the raw estimates $\hat{\boldsymbol{\beta}}$ and $\mathbf{1}$:

$$(1 - \tau)\hat{\boldsymbol{\beta}} + \tau \mathbf{1}$$

for some shrinkage factor τ. The articles of Ledoit and Wolf (2003, 2004) elaborate further on using shrinkage for improved estimates of the covariance matrix. Efron and Morris (1977) present a related and entertaining discussion of shrinkage estimation applied to baseball statistics.
The estimates of σ_M and of ω_i can also be improved by using techniques such as exponential smoothing and generalized autoregressive conditional heteroskedasticity (GARCH) (Campbell et al., 1997; Engle, 1982).
The CAPM is related to, although not the same as, a single-factor risk model. In the context of a single-factor model where the factor is the market portfolio r_M, the CAPM postulates

$$\mathbb{E}(r_i) = \beta_i \mathbb{E}(r_M)$$

In other words, the expected value of the asset-specific return is zero. The CAPM thus gives a straightforward estimation procedure for the vector of expected returns $\boldsymbol{\mu} = \mathbb{E}(\mathbf{r})$, namely $\hat{\boldsymbol{\mu}} := \hat{\boldsymbol{\beta}}\hat{\mu}_M$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\mu}_M$ are estimates of $\boldsymbol{\beta}$ and $\mathbb{E}(r_M)$ respectively. As we discuss in Section 6.6 below, other alternatives for estimating expected returns are often used in equity portfolio management.

Constant Correlation Models

A second way of imposing structure on the asset returns is to assume that the correlation between any two different assets in the investment universe is the same. Under this assumption, the estimation of the covariance matrix only requires an estimate of each individual asset volatility σ_i and the average correlation ρ between different pairs of assets. This yields a "quick and dirty" estimate of the covariance matrix given by

$$\text{cov}\,(r_i, r_j) = \rho\sigma_i\sigma_j, \quad i \neq j$$

In this model the estimation of the covariance matrix only requires estimates of σ and ρ. That is a total of $n + 1$ parameters.

Under the reasonable assumption that $\rho > 0$, the constant correlation model can be seen as the following kind of single-factor model with predetermined factor loadings. Assume the following single-factor model for volatility scaled excess returns:

$$\frac{r_i}{\sigma_i} = f + \theta_i$$

where f is a common factor to all scaled returns and θ_i is a specific scaled return on asset i. It is easy to see that this particular single-factor model yields a constant correlation model with ρ being the variance of the single factor f. Using matrix notation, the constant correlation covariance matrix can be written as

$$\mathbf{V} = \rho \boldsymbol{\sigma} \boldsymbol{\sigma}^\top + (1 - \rho)\,\mathrm{diag}(\boldsymbol{\sigma})^2$$

A basic estimation procedure for this model is straightforward: first, using historical data, compute estimates $\hat{\sigma}_i$ of σ_i and estimates $\hat{\rho}_{ij}$ of each correlation ρ_{ij} for all $i \neq j$. Finally, take the average

$$\hat{\rho} := \frac{1}{n(n-1)} \sum_{i \neq j} \hat{\rho}_{ij}$$

as an estimate of ρ.

Multiple-Factor Models

Multiple-factor models are a generalization of the single-factor model discussed above. These models are based on the assumption that the return of each asset can be explained by a small collection of common factors in addition to some other specific return. Aside from simplifying the estimation task, multiple-factor models provide a useful breakdown of risk, incorporate some economic logic, and are fairly flexible.

A multi-factor model assumes that excess returns are as follows:

$$r_i = \sum_{k=1}^{K} B_{ik} f_k + u_i,$$

where

- r_i : excess return of asset i
- B_{ik} : exposure of asset i to factor k
- f_k : rate of return of factor k
- u_i : specific (or residual) return of asset i.

It is convenient to rewrite the relation above in matrix form as

$$\mathbf{r} = \mathbf{B}\mathbf{f} + \mathbf{u}$$

A bit of matrix algebra shows that the expected value and covariance of $\mathbf{r}$ are respectively

$$\mathbb{E}[\mathbf{r}] = \mathbf{B}\mathbb{E}[\mathbf{f}] + \mathbb{E}[\mathbf{u}], \quad \mathbf{V} = \mathbf{B}\mathbf{F}\mathbf{B}^\top + \Delta,$$

where $\mathbf{F} = \mathrm{cov}(\mathbf{f})$ and $\Delta = \mathrm{cov}(\mathbf{u})$. Observe that Δ is diagonal since the u_i are assumed to be uncorrelated with each other.

The construction and estimation of a multi-factor model hinges on the choice of factors. For an equity universe, the following three main classes of factors are commonly used:

- Macroeconomic factors: inflation, economic growth, etc.
- Fundamental factors: earning/price, dividend yield, market cap, etc.
- Statistical factors: principal component analysis, hidden factors.

Empirical evidence suggests that the second type of fundamental factors works better than the other two (Connor, 1995). This is also the prevalent class of factors used by most risk model providers. In this approach we have

$$\mathbf{r} = \mathbf{B}\mathbf{f} + \mathbf{u}$$

where the matrix of factor loadings $\mathbf{B}$ is predetermined. The estimation of the corresponding covariance matrix is as follows. Using historical data for the asset returns, infer the corresponding historical data for factor returns by solving each of the weighted least-squares problems

$$\min(\mathbf{r}(t) - \mathbf{B}\mathbf{f}(t))^\top \mathbf{D}^{-1}(\mathbf{r}(t) - \mathbf{B}\mathbf{f}(t)).$$

The matrix $\mathbf{D}$ is a diagonal matrix whose entries are estimates of the asset variances. A common proxy is to use instead the reciprocal of the market capitalizations of the assets. The solution to this weighted least-squares problem is

$$\mathbf{f}(t) = \left(\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top\right)^{-1} \mathbf{B}^\top \mathbf{D}^{-1}\mathbf{r}(t).$$

Each row of the matrix $\left(\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^{\top}\right)^{-1}\mathbf{B}^{\top}\mathbf{D}^{-1}$ can be interpreted as a factor mimicking portfolio.
Equipped with this historical data of factor returns, we can estimate the factor covariance matrix. The residuals $\mathbf{u}(t) := \mathbf{r}(t) - \mathbf{B}\mathbf{f}(t)$ can then be used to estimate the covariance matrix Δ of asset-specific returns.
The connection between the CAPM and single-factor models has an analogous counterpart in the context of multi-factor models, namely the arbitrage pricing theory (APT). A combination of an arbitrage argument and the assumption that the set of factors $\mathbf{f}$ account for all of the common shocks to the returns of all assets in the investment universe implies that

$$\mathbb{E}(\mathbf{r}) = \mathbf{B}\mathbb{E}(\mathbf{f})$$

Like the CAPM, the APT model also yields a straightforward estimation procedure for $\boldsymbol{\mu} = \mathbb{E}(\mathbf{r})$.

Estimation of Alpha

In a benchmark-relative context, an estimate of expected residual returns α is typically the relevant estimate instead of an estimate of expected total return μ. According to the CAPM or the more general APT model, the expected residual returns are zero. However, numerous articles have documented certain anomalies that are systematically associated with the over- and underperformance of the return of securities after controlling for their systematic component of return. Some of these anomalies include the SMB (small minus big market capitalization) and HML (high minus low book-to-price) factors introduced in the classical article by Fama and French (1992).
A generic approach for generating alpha is to rely on signals unveiled via a judicious type of analysis. A signal could be an empirical observation such as momentum that suggests that the recent performance (good or bad) of individual securities will persist in the near term. A signal could also be a financial principle such as "firms with low book-to-price ratio will outperform" or "firms with higher earnings per share will outperform".

The following is a reasonable and popular rule of thumb for transforming a signal into a forecast of alpha (for a detailed discussion see Grinold and Kahn (1999)):

$$\text{alpha} = (\text{ residual volatility }) \cdot \mathrm{IC} \cdot \text{ score } .$$

Here the residual volatility is the standard deviation of residual return. The score is a numerical score associated with the signal. The score is assumed to be scaled so that its cross-sectional mean and standard deviation are respectively 0 and 1. Finally, the information coefficient IC is a measure of the forecasting quality of the signal; that is, the correlation between the raw signal score and the residual return. In addition to proper scaling, the signal score should be neutralized so that the alphas do not include biases or undesirable bets on the benchmark or on risk factors. As we illustrate in the exercises at the end of the chapter, neutralization can be achieved in various ways, as multiple portfolios hedge out a bet on the benchmark or on other risk factors.

Performance Analysis

How can the performance of a portfolio manager be evaluated? Are the ex post results due to skill or luck? The goal of performance analysis is to answer these questions. The efficient market hypothesis suggests that skillful active management is impossible. However, there is considerable evidence against the efficient market hypothesis (Shleifer, 2000).
Empirical results also suggest that an average active fund manager underperforms their benchmark on a risk-adjusted basis. Furthermore, empirical evidence also shows that good performance does not persist: The winners this year are almost as likely to be winners or losers next year. These are bleak conclusions about asset management. So how could we tell which asset managers are the good ones?

The fundamental goal of performance analysis is to separate skill from luck. The simplest type of performance analysis is a cross-sectional comparison of returns over some time period. This would distinguish winners from losers. However, these kinds of comparisons have several drawbacks. First, they typically do not represent the complete universe of investment managers but only those in existence during a specific time period. They generally contain survivorship bias. Perhaps worst of all, cross-sectional comparisons do not adjust for risk. By contrast, time-series analysis of returns can do a better job at separating skill from luck by measuring both return and risk. An even more complete picture can be obtained via time-series analysis of returns and portfolio holdings.

Return-Based Performance Analysis (Basic)

The development of the CAPM and the notion of market efficiency in the 1960s encouraged academics to tackle the problem of performance analysis. According to the CAPM, consistent exceptional returns are unlikely. Academics devised tests to check if the theory was correct. As a byproduct the first performance analysis techniques emerged. One approach, proposed by Jensen, consists of regressing the time series of realized portfolio excess returns against benchmark excess return:

$$r_P(t) = \alpha_P + \beta_P r_B(t) + \epsilon_P(t).$$

Jensen's alpha is simply the intercept α_P of this regression. According to the CAPM, this intercept is zero. The regression yields not only alpha and beta, but t-statistics that give information about their statistical significance. The t-statistic for α_P is

$$t\text{-stat} = \frac{\alpha_P}{\mathrm{SE}(\alpha_P)}$$

As a rule of thumb, a t-statistic of 2 or more indicates that the performance of the portfolio is due to skill rather than luck. Assuming normality, the probability of observing such a large t-statistic purely by chance is smaller than 5%.

The t-statistic and the information ratio are closely related. The main difference between them is that the information ratio is annualized. By contrast, the t-statistic scales with the number of years of data. If we observe returns over a period of T years, the information ratio is approximately the t-statistic divided by the square root of the number of years of observation:

$$IR \approx \frac{t\text{-stat}}{\sqrt{T}}.$$

The standard error of the information ratio is approximately

$$\mathrm{SE}(IR) \approx \frac{1}{\sqrt{T}}$$

A simple alternative to Jensen's approach is to compare Sharpe ratios for the portfolio and the benchmark. A portfolio with

$$\frac{\bar{r}_P}{\sigma_P} > \frac{\bar{r}_B}{\sigma_B},$$

where $\bar{r}$ denotes mean excess return over the period, has demonstrated positive performance. Once again, the statistical significance of this relationship is relevant for distinguishing luck from skill. If we assume that the standard errors of the portfolio and benchmark volatilities are fairly small compared to $\bar{r}$ standard errors, then the standard error of the Sharpe ratio is approximately $1/\sqrt{N}$, where N is the number of observations. Hence a statistically significant demonstration of skill occurs when

$$\frac{\bar{r}_P}{\sigma_P} - \frac{\bar{r}_B}{\sigma_B} > 2\sqrt{\frac{2}{N}}.$$

Return-Based Style Analysis

Style analysis was developed by Nobel laureate William Sharpe (1992). The popularity of this concept was aided by a study (Brinson et al., 1991) concluding that 91.5% of the variation in returns of 82 mutual funds could be explained by the allocation to bills, stocks, and bonds. Later studies considering asset allocation across a broader range of asset classes have shown that as much as 97% of fund returns can be explained by asset allocation alone.
Style analysis attempts to determine the effective asset mix of a fund using only the time series of returns for the fund and for a number of carefully chosen asset classes. Like a factor model approach, style analysis assumes that portfolio returns have the form

$$r_P(t) = \sum_{j=1}^{m} w_j f_j(t) + u_P(t)$$

where the $f_j(t)$ are the returns of m benchmark asset classes. The holdings $w_j, j = 1, \ldots, m$, represent the style of the portfolio.

That is, the effective allocation to the m asset classes that could be replicated via a passive portfolio. The term $u_P(t)$ represents the selection return; that is, the portion of the portfolio return that style cannot explain. The effective holdings can be estimated via the quadratic program

$$\begin{aligned} &\min_{\mathbf{w}} \operatorname{var}\left(u_P(t)\right) \\ &\text{s.t.} \sum_{j=1}^{m} w_j = 1 \\ &\quad w_j \geq 0, j = 1, \ldots, m \end{aligned} \tag{6.21}$$

Notice that there are two key differences between this model and conventional multiple regression. First, the weights are constrained to be non-negative and to add up to 1. Second, instead of minimizing the sum of squared errors $\sum_{t=1}^{T} u_P(t)^2$, we minimize the variance of these quantities. The reason for the first restriction is that the w_j are to be interpreted as an effective asset allocation representing the style of the fund. In essence, they create a fund-specific benchmark. The reason for the second restriction is that we want to allow for a non-zero selection effect by the fund manager. The model finds the style that minimizes the variance of this effect. Once the optimal weights are determined, the average value of $u_P(t)$ gives the value added by the manager's selection skills, which can be negative or positive.

Assume the data available for style analysis are the return time series $r_P(t)$, $f_1(t), \ldots, f_m(t)$ for $t = 1, \ldots, T$. For ease of notation, put

$$\mathbf{r} := \begin{bmatrix} r_P(1) \\ \vdots \\ r_P(T) \end{bmatrix}, \quad \mathbf{F} := \begin{bmatrix} f_1(1) & \cdots & f_m(1) \\ \vdots & \ddots & \vdots \\ f_1(T) & \cdots & f_m(T) \end{bmatrix}, \quad \mathbf{1} := \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Then the objective function in (6.21) can be written as

$$\begin{aligned} \operatorname{var}(\mathbf{r} - \mathbf{F}\mathbf{w}) =& \frac{1}{T}\|\mathbf{r} - \mathbf{F}\mathbf{w}\|^2 - \frac{1}{T^2}\left(\mathbf{1}^\top(\mathbf{r} - \mathbf{F}\mathbf{w})\right)^2 \\ =& \left(\frac{\|\mathbf{r}\|^2}{T} - \frac{\left(\mathbf{1}^\top\mathbf{r}\right)^2}{T^2}\right) - 2\left(\frac{\mathbf{r}^\top\mathbf{F}}{T} - \frac{\mathbf{1}^\top\mathbf{r}}{T^2}\mathbf{1}^\top\mathbf{F}\right)\mathbf{w} \\ &+ \mathbf{w}^\top\left(\frac{1}{T}\mathbf{F}^\top\left(I - \frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)\mathbf{F}\right)\mathbf{w} \end{aligned}$$

- Style analysis provides an improvement tool for measuring performance. The constructed style usually tracks the performance of the fund more accurately than a predefined benchmark. Style analysis has also some limitations. For instance, the weights may not necessarily match the style disclosed by the fund manager. However, as Sharpe puts it: "If it acts like a duck, it is ok to assume it is a duck."
- Style analysis also makes the simplifying assumptions that the weights are constant. This is clearly not the case in actively managed funds, even without active trading. There exist some variations of style analysis that allow for weights to change. The model gets a bit more technical because it needs to incorporate some "regularization" term that prevents the weights from changing too much too often.