Portfolio Optimization
Part 1 – Unconstrained Portfolios

John Norstad

j.norstad@mac.com
http://homepage.mac.com/j.norstad

September 11, 2002
Updated: November 24, 2005

Abstract

We recapitulate the single-period results of Markowitz [2] and Sharpe [10][1] in the context of the lognormal random walk model, iso-elastic utility, and continuous portfolio rebalancing. We formally derive the solution to the unconstrained optimization problem and examine the mathematical properties of the resulting efficient frontier and efficient portfolios. We derive the two-fund separation theorem both in the presence of a risk-free asset and in its more general form. We derive and briefly discuss the Capital Asset Pricing Model (CAPM). We present several examples parameterized using market return data for US stocks, bonds, cash, and inflation.

We assume that the reader is familiar with the material presented in references [4, 5, 6, 7, 8, 9].

[1] Our exposition closely follows Sharpe's mathematical supplement in [10]. Indeed, this homework paper is little more than an attempt to fill in the details in Sharpe's presentation, make them rigorous, rework them in the lognormal model, and provide some examples.

Contents

List of Figures

List of Tables

1 Introduction

The portfolio optimization problem is the following: Given a finite set of assets, find the portfolio that combines the assets using an optimal fixed percentage asset allocation given the risk preferences of an individual investor.

We make the following strong assumptions:

- The prices of the assets follow jointly distributed lognormal random walks.
- We invest for a single period of time.
- We decide on our asset allocation at the beginning of the time period and do not change it during the time period, except for continuous rebalancing to keep the asset allocation constant throughout the time period.
- Our goal is to maximize the expected utility of our end-of-period wealth.
- We have a risk-averse iso-elastic (CRRA) utility function.
- Asset allocation is unconstrained. In particular, both shorting (negative allocations) and leverage (allocations greater than 100%) are permitted.

The assets can be individual securities or groups of securities such as sectors, styles, or even entire markets.

We use two main results from random walk theory and basic portfolio theory.

1.1 The Random Walk Theorem

The first result is from random walk theory. It is Theorem 4.1 in reference [9]. We restate it here as our Theorem 1.1:

Theorem 1.1 *For a random variable s,*

$$\frac{ds}{s} = \alpha dt + \sigma dX \quad \textit{iff} \quad \frac{ds}{s} = e^{\mu dt + \sigma dX} - 1$$

where dX is $N[0, dt]$ and $\alpha = \mu + \frac{1}{2}\sigma^2$.

In this case, s follows a lognormal random walk. The logarithm of $s(1)/s(0)$ is normally distributed with mean μ and standard deviation σ. α is the yearly instantaneous expected return, μ is the yearly continuously compounded expected return and σ is the standard deviation of those returns. Over any time horizon t, $s(t)$ is lognormally distributed with:

$$s(t) = s(0)e^{\mu t + \sigma X} \quad \textit{where } X \textit{ is } N[0, t]$$

As we discussed in section 5 of [9], when we talk about rates of return on investments, it is very important to distinguish between the different ways to measure the returns.

α = instantaneous return
μ = continuously compounded return
$r1$ = simply compounded arithmetic mean (average) return
$r2$ = simply compounded geometric mean (annualized or median) return

These ways to measure returns are related by the following equations:

$$\alpha = \mu + \frac{1}{2}\sigma^2 \tag{1}$$
$$r1 = e^{\alpha} - 1 \tag{2}$$
$$r2 = e^{\mu} - 1 \tag{3}$$

In this paper we use instantaneous returns.

1.2 The Portfolio Choice Theorem

The second result is from basic portfolio theory. It is Theorem 4.1 in reference [4]. We restate it here as our Theorem 1.2:

Theorem 1.2 *For time horizon t, a utility-maximizing investor with initial wealth w_0 and an iso-elastic utility function with coefficient of relative risk aversion A, when faced with a decision among a set of competing feasible investment alternatives F all of whose elements have lognormally distributed returns under the random walk model, acts to select an investment $I \in F$ which maximizes*

$$\alpha_I - \frac{1}{2}A\sigma_I^2$$

where α_I and σ_I are the expected instantaneous yearly return and standard deviation for investment I respectively.

Note that the optimal investment is independent of both the time horizon t and the investor's initial wealth w_0. This is because we are using iso-elastic utility, which has the property of constant relative risk aversion with respect to both wealth and time. See references [4] and [5] for details.

2 Linear Combinations of Assets

We have a finite number of assets in which we can invest. We wish to consider fixed linear combinations of these assets, with the goal being to compute optimal asset allocations.

Markowitz, Sharpe, and other theorists often use simply compounded yearly returns for the assets and assume they are normally distributed. Linear combinations of normally distributed random variables are also normally distributed.[2] This greatly simplifies the reasoning.

There are several problems with the assumption that asset returns are normally distributed.

1. The assumption of normal returns violates the principle of limited liability. The most an investor stands to lose with an investment in normal risky assets like stocks is 100%. With normally distributed returns the maximum loss is unlimited. With lognormally distributed returns the maximum loss is 100% ($e^{-\infty} - 1 = 0 - 1 = -1 = -100\%$).

2. The assumption also violates the Central Limit Theorem, which says that the sum of independent and identically distributed random variables with finite variance approaches a normal distribution in the limit as the number of summands goes to infinity. But with returns on investments, over time the returns are compounded, which means that they are multiplied instead of being added. It is the logarithms of the returns that are added. So the Central Limit Theorem implies that returns are lognormally distributed, not normally distributed.

3. Normally distributed returns do not match the historical data, especially over long time horizons. Lognormally distributed returns are a much closer match.

These are the reasons we use the lognormal model in this paper. Our goal is to reproduce the classical results in this model. We assume that all the asset returns are lognormally distributed.

Thus our portfolios are linear combinations of lognormal assets. Unfortunately, linear combinations of lognormally distributed random variables are not lognormally distributed! (Their products are, but not their sums.) So we have a problem that must be addressed before we can proceed further.

[2]For a proof, see reference [6].

Consider a collection of n assets each of which follows a lognormal random walk and a portfolio P formed by combining the assets in fixed proportions. Let:

w_i = proportion of portfolio P invested in asset i, $1 \leq i \leq n$.
w = column vector of proportions w_i.
α_i = expected instantaneous return of asset i, $1 \leq i \leq n$.
x = column vector of expected returns α_i.
$\rho_{i,j}$ = covariance of asset i with asset j, $1 \leq i \leq n$ and $1 \leq j \leq n$.
$V = n \times n$ matrix of covariances $\rho_{i,j}$.
σ_i = instantaneous standard deviation of asset $i = \sqrt{\rho_{i,i}}$.
α_P = expected instantaneous return of portfolio P.
σ_P = instantaneous standard deviation of portfolio P.

For each asset, the price of the asset s_i follows the Ito process:

$$\frac{ds_i}{s_i} = \alpha_i dt + \sigma_i dX_i \quad \text{where } dX_i \text{ is } N[0, dt]$$

Note that over the short time period dt the random variable dX_i is normally distributed, not lognormally distributed. Thus over the short time period the percent change in the asset value, ds_i/s_i, is normally distributed with mean $\alpha_i dt$ and variance $\sigma_i^2 dt$.

This is the key observation. In an Ito process, asset returns are normally distributed over short time periods. (Actually, this is strictly speaking true only in the limit, over infinitesimal time periods, but we will ignore this mystery here.[3])

Over the same short time period dt, therefore, the percent change in the value of our portfolio, which is a linear combination of normally distributed component assets, is also normally distributed, with expected return $\alpha_P dt$ and variance $\sigma_P^2 dt$ where:

$$\alpha_P = w'x = \sum_{i=1}^{n} w_i \alpha_i \tag{4}$$

$$\sigma_P^2 = w'Vw = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \rho_{i,j} \tag{5}$$

(Equation (5) is Proposition 8 in reference [8].)

Consider our portfolio at the end of the short time period dt. The component asset prices have gone up and down randomly over the period, so our portfolio is no longer in balance. We now make the critical assumption that we *rebalance* the portfolio by buying and selling the proper small amounts of the component assets to restore the asset allocation specified by the proportions w_i.

[3]For a full and rigorous treatment of the continuous-time issues discussed in this section, see Merton's exquisitely difficult book [3].

Assuming we do this rebalancing, over the next short time period dt our portfolio starts over with the original asset proportions, and we have the same equations as above for the expected return and standard deviation of the portfolio over the new time period.

Thus, if we rebalance at the end of each short time period dt, over time our portfolio's value p follows the Ito process:

$$\frac{dp}{p} = \alpha_P dt + \sigma_p dX \quad \text{where } dX \text{ is } N[0, dt]$$

By The Random Walk Theorem 1.1, p follows a lognormal random walk.

In the limit as $dt \to 0$, this implies that if we do continuous rebalancing of our portfolio, our linear combination of lognormal random walk assets also follows a lognormal random walk, with the instantaneous expected return α_P and instantaneous standard deviation σ_P given by equations (4) and (5).

This notion of continuous rebalancing is of course unrealistic, indeed impossible. So our model and our results have no hope of exactly matching what happens with real investors and their real portfolios. It is close enough so that it is still quite a good approximation, however, provided the investor rebalances reasonably often (e.g., at least yearly over a long time horizon). This assumption of continuous rebalancing is certainly much less onerous than those we are forced to make if we use the normally distributed returns model.[4]

[4] In one simulation we ran of investing over a 40 year time horizon with a 50/50 stock/bond portfolio, the difference between the mean and median ending values with yearly rebalancing and continuous rebalancing was less that 2% in each case. This is a close enough approximation for practical use. The yearly rebalancing case has a bit higher expected return and is a bit riskier because of the tendency for the asset allocation to drift towards a heavier weighting in stocks in between the yearly rebalancings.

3 Linear Combinations of Two Assets

We now know how to compute the instantaneous expected return and standard deviation of a portfolio which is formed from a continuously rebalanced linear combination of jointly distributed lognormal random walk assets. The equations (4) and (5) from the previous section are:

$$\alpha_P = w'x = \sum_{i=1}^{n} w_i \alpha_i$$

$$\sigma_P^2 = w'Vw = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \rho_{i,j}$$

In this section we look at linear combinations of two assets to illustrate the risk reduction benefits of diversification that are a consequence of the second formula above.

With two assets let $w_1 = w$, $w_2 = 1 - w$, and $c =$ the correlation coefficient of the two assets.

Note that $\rho_{1,1} = \sigma_1^2$, $\rho_{2,2} = \sigma_2^2$, and $\rho_{1,2} = \rho_{2,1} = c\sigma_1\sigma_2$. Our equations above become:

$$\alpha_P = w\alpha_1 + (1 - w)\alpha_2 \tag{6}$$

$$\sigma_P^2 = w^2\sigma_1^2 + 2w(1 - w)c\sigma_1\sigma_2 + (1 - w)^2\sigma_2^2 \tag{7}$$

$$\sigma_P = \sqrt{w^2\sigma_1^2 + 2w(1 - w)c\sigma_1\sigma_2 + (1 - w)^2\sigma_2^2} \tag{8}$$

The equation (6) for the expected return of the portfolio is simply a linear interpolation of the expected returns of the two assets. E.g., if $w = 0.5$, the portfolio expected return is half way between the expected returns of the two assets. The equations for the variance and standard deviation are more complicated, and those are the ones we will investigate.

As an example, consider the following pair of assets:

Asset 1: $\alpha_1 = 5\%$ and $\sigma_1 = 8\%$
Asset 2: $\alpha_2 = 11\%$ and $\sigma_2 = 19\%$

In this example we constrain w to the range $0 \le w \le 1$.

In Figure 1, we graph our equations to show the risk and return characteristics of the linear combinations of these two assets for various values of the correlation coefficient c. The graph shows how combining assets with imperfect correlation significantly reduces risk, with lower correlations providing greater benefits.

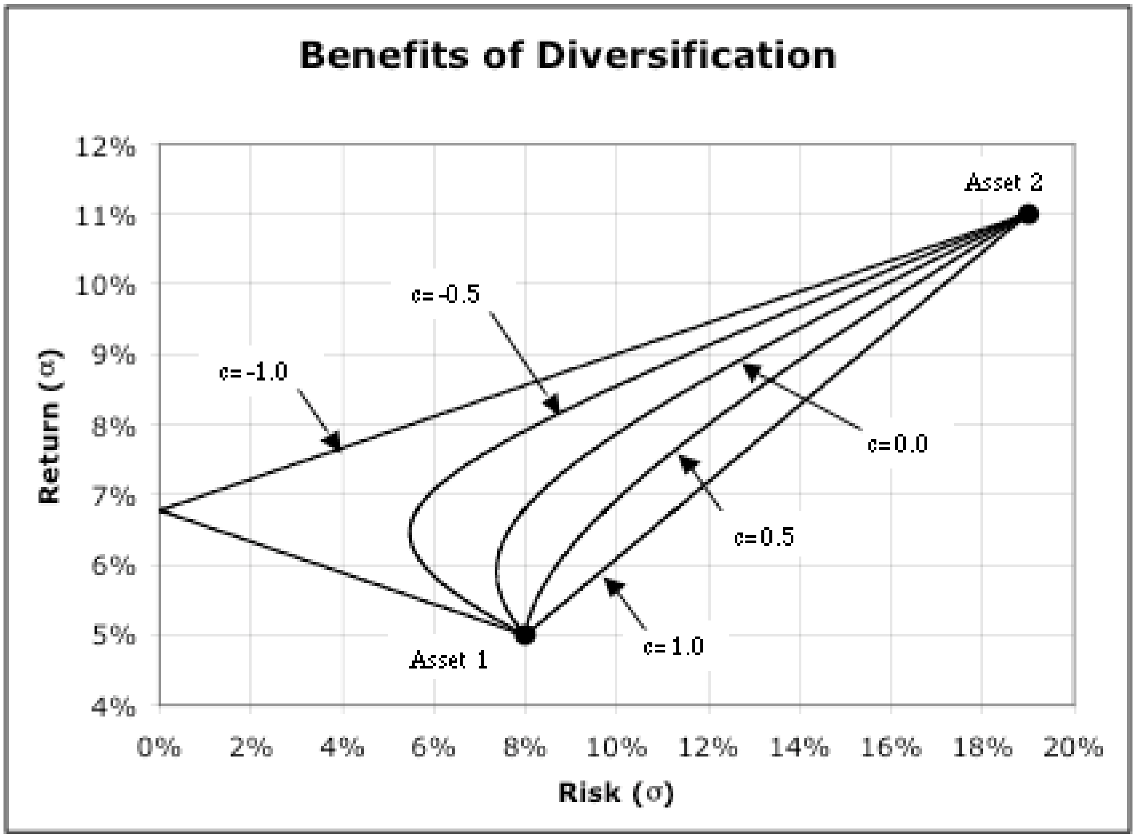


Figure 1: The Benefits of Diversification

When $c = 1.0$ we have perfectly correlated assets and the risk/return graph of the linear combinations of the two assets is a straight line joining the two assets. It's not difficult to see why this happens. Equation (7) for σ_P^2 becomes:

$$\begin{aligned}
\sigma_P^2 &= w^2\sigma_1^2 + 2w(1-w)c\sigma_1\sigma_2 + (1-w)^2\sigma_2^2 \\
&= w^2\sigma_1^2 + 2w(1-w)\sigma_1\sigma_2 + (1-w)^2\sigma_2^2 \\
&= [w\sigma_1 + (1-w)\sigma_2]^2 \\
\sigma_P &= w\sigma_1 + (1-w)\sigma_2
\end{aligned}$$

When $c < 1.0$ we have assets that are not perfectly correlated. In equation (7) above for σ_P^2, notice that if we keep w fixed, as c decreases the value of σ_P^2 also decreases. So as c decreases the risk/return curves move further to the left. We can see this in the graph, where the curve for $c = 0.5$ is to the left of the one for $c = 1.0$, the curve for $c = 0.0$ is to the left of the one for $c = 0.5$, and so on.

When $c = -1.0$ we have perfectly negatively correlated assets. In Figure 1 the risk/return curve is a pair of straight lines. The first line starts at asset 2 and goes left all the way to the Y axis ($\sigma = 0$), then jogs back to asset 1. It's easy to see why this happens. Our equation becomes:

$$\begin{aligned}\sigma_P^2 &= w^2\sigma_1^2 + 2w(1-w)c\sigma_1\sigma_2 + (1-w)^2\sigma_2^2 \\ &= w^2\sigma_1^2 - 2w(1-w)\sigma_1\sigma_2 + (1-w)^2\sigma_2^2 \\ &= [w\sigma_1 - (1-w)\sigma_2]^2\end{aligned}$$

The expression inside the square brackets has the value zero when $w = \dfrac{\sigma_2}{\sigma_1+\sigma_2}$ (about $w = 0.7$ in our example). For smaller values of w the value is negative, and for larger values of w it is positive. Thus:

$$\begin{aligned}&\text{For } w < \frac{\sigma_2}{\sigma_1+\sigma_2}: && \sigma_P = -[w\sigma_1 - (1-w)\sigma_2] = \sigma_2 - w(\sigma_1+\sigma_2) \\ &\text{For } w = \frac{\sigma_2}{\sigma_1+\sigma_2}: && \sigma_P = 0 \\ &\text{For } w > \frac{\sigma_2}{\sigma_1+\sigma_2}: && \sigma_P = +[w\sigma_1 - (1-w)\sigma_2] = \sigma_1 - (1-w)(\sigma_1+\sigma_2)\end{aligned}$$

The first equation is the straight line from asset 2 to the Y axis. The last equation is the straight line from the Y axis to asset 1.

4 Parameter Estimation for the Examples

We give several examples later based on historical market return data for US stocks, bonds, cash, and inflation.[5] We have the following time series of data for 1926 through 1994, as simply compounded yearly rates of return:[6]

C = cash (30 day US Treasury bills)
B = bonds (20 year US Treasury bonds)[7]
S = stocks (S&P 500 large US stocks)
I = inflation (CPI = consumer price index)

We used Microsoft Excel to compute parameters for the lognormal random walk model based on this raw data.[8]

First, we convert each return to continuous compounding by taking the natural logarithm of 1 plus the simply compounded return. This gives four new time series for the continuously compounded returns:

$$\begin{aligned} \hat{C} &= \log(1+C) \\ \hat{B} &= \log(1+B) \\ \hat{S} &= \log(1+S) \\ \hat{I} &= \log(1+I) \end{aligned}$$

These series are used to compute the instantaneous returns, variances, standard deviations, and covariances as follows, where X and $Y \in \{\hat{C}, \hat{B}, \hat{S}, \hat{I}\}$:

$\rho_{X,Y}$ = instantaneous covariance = $\mathrm{Cov}(X, Y)$
σ_X = instantaneous standard deviation = $\sqrt{\rho_{X,X}}$
μ_X = continuously compounded expected return = $E(X)$
α_X = instantaneous expected return = $\mu_X + \frac{1}{2}\sigma_X^2$
$r1_X$ = arithmetic mean (average) simply compounded expected return = $e^{\alpha_X} - 1$
$r2_X$ = geometric mean (annualized) simply compounded expected return = $e^{\mu_X} - 1$

Table 1 presents the results of these estimates.

---

[5] Using historical data to estimate parameters is common but not required. Analysts can and do form their own estimates for all of the parameters.
[6] The data is from Table 2.4 in reference [1].
[7] Note that the bonds are of long maturity (20 years).
[8] The Excel spreadsheet used to do the parameter estimation and all the other calculations, examples, and graphs is available at the author's web site.

Nominal Summary Statistics

	cash	bonds	stocks	cpi
σ	3.11%	7.87%	19.33%	4.44%
μ	3.62%	4.71%	9.71%	3.08%
α	3.67%	5.02%	11.58%	3.17%
$r1$	3.74%	5.15%	12.27%	3.22%
$r2$	3.69%	4.82%	10.19%	3.12%

Nominal Covariances

	cash	bonds	stocks	cpi
cash	0.00096	0.00054	-0.00012	0.00057
bonds	0.00054	0.00620	0.00246	-0.00052
stocks	-0.00012	0.00246	0.03738	0.00037
cpi	0.00057	-0.00052	0.00037	0.00197

Nominal Correlation Coefficients

	cash	bonds	stocks	cpi
cash	1.00	0.22	-0.02	0.42
bonds	0.22	1.00	0.16	-0.15
stocks	-0.02	0.16	1.00	0.04
cpi	0.42	-0.15	0.04	1.00

Table 1: Nominal Historical Return Data

The example in Section 7.4 is inflation-adjusted and uses real returns instead of nominal returns. To estimate the parameters for this example, we convert the nominal simply compounded returns into real simply compounded returns, then convert those returns to continuous compounding:

$$
\begin{aligned}
C_{real} &= (C - I)/(1 + I) \qquad & \hat{C}_{real} &= \log(1 + C_{real}) \\
B_{real} &= (B - I)/(1 + I) \qquad & \hat{B}_{real} &= \log(1 + B_{real}) \\
S_{real} &= (S - I)/(1 + I) \qquad & \hat{S}_{real} &= \log(1 + S_{real})
\end{aligned}
$$

Finally, we compute the covariances, correlation coefficients, and summary statistics as before. Table 2 presents the results.

Real Summary Statistics

	cash	bonds	stocks
σ	4.22%	9.60%	19.65%
μ	0.54%	1.64%	6.63%
α	0.63%	2.10%	8.56%
$r1$	0.63%	2.12%	8.94%
$r2$	0.54%	1.65%	6.86%

Real Covariances

	cash	bonds	stocks
cash	0.00178	0.00245	0.00091
bonds	0.00245	0.00922	0.00459
stocks	0.00091	0.00459	0.03862

Real Correlation Coefficients

	cash	bonds	stocks
cash	1.00	0.60	0.11
bonds	0.60	1.00	0.24
stocks	0.11	0.24	1.00

Table 2: Real Historical Return Data

Note how the three assets are significantly more correlated when we adjust for inflation. Compare the correlation coefficients here with those in Table 1 for nominal returns.

Note that the returns decrease significantly when we adjust for inflation, but the standard deviations do not. In fact, the standard deviations go up a bit.

5 The Simplest Case – A Single Risky Asset

In this section we solve the simple problem of optimizing portfolio asset allocation when the only choices available are a single risky asset and a risk-free asset. In this paper we only consider the *unconstrained* problem. That is, we permit asset allocation proportions outside the range $[0, 1]$. Asset allocation proportions greater than one represent leverage. Asset allocation proportions less than zero represent short selling or borrowing.[9]

Let:

w = proportion of portfolio invested in risky asset.
α = expected instantaneous return of risky asset.
σ = instantaneous standard deviation of risky asset.
r = instantaneous risk-free rate of return.
α_P = expected instantaneous return of portfolio.
σ_P = instantaneous standard deviation of portfolio.
A = iso-elastic coefficient of relative risk aversion.

With continuous rebalancing, our portfolio follows a lognormal random walk with:

$$\alpha_P = w\alpha + (1 - w)r \tag{9}$$

$$\sigma_P = w\sigma \tag{10}$$

By The Portfolio Choice Theorem 1.2, the problem is to maximize $f(w)$ where:

$$\begin{aligned} f(w) &= \alpha_P - \frac{1}{2}A\sigma_p^2 \\ &= w\alpha + (1 - w)r - \frac{1}{2}Aw^2\sigma^2 \end{aligned}$$

Take the derivative of this quadratic equation, set it equal to 0, and solve for w:

$$f'(w) = \alpha - r - A\sigma^2 w = 0$$

$$w = \frac{\alpha - r}{A\sigma^2} \tag{11}$$

The numerator in this equation is the difference between the expected return of the risky asset and the risk-free rate. This is called the asset's *risk premium.* When the risky asset is stocks it is called the *equity risk premium.* Notice that as the premium increases, so does the proportion w allocated to the risky asset, as long as the denominator stays the same.

[9] We assume that it is possible to both borrow and lend at the same risk-free interest rate. The literature deals with the problem when borrowing and lending rates are different, but we do not address that problem here.

The denominator is the product of two factors, the coefficient of risk aversion and the variance of the risky asset. It either of these factors increases, the proportion w allocated to the risky asset decreases, as long as the other factors remain the same.

Thus, as expected, risk-averse investors prefer large risk premia and small variances (volatility), and more risk-averse investors with larger values of A have smaller allocations to the risky asset.

5.1 Example 1 – US Stocks Plus a Risk-Free Asset

In our first example we consider portfolios constructed from mixtures of large US stocks and a risk-free asset. We use the following parameters for stocks from Table 1.

$\alpha = 11.58\% =$ expected instantaneous return for large US stocks.
$\sigma = 19.33\% =$ instantaneous standard deviation of large US stocks.

For the risk-free rate we use:

$r = 4.0\% =$ instantaneous risk-free interest rate.

Given an iso-elastic coefficient of relative risk aversion A, our equation (11) for the optimal allocation to stocks becomes:

$$w = \frac{\alpha - r}{A\sigma^2} = \frac{0.1158 - 0.04}{A \times 0.1933^2} = \frac{2.03}{A}$$

Figure 2 graphs the efficient frontier for this set of feasible investments (all linear combinations of large US stocks and the risk-free asset). The values for the risk and return for each portfolio are calculated from our equations (9) and (10) and from equations (2) and (3) on page 4.

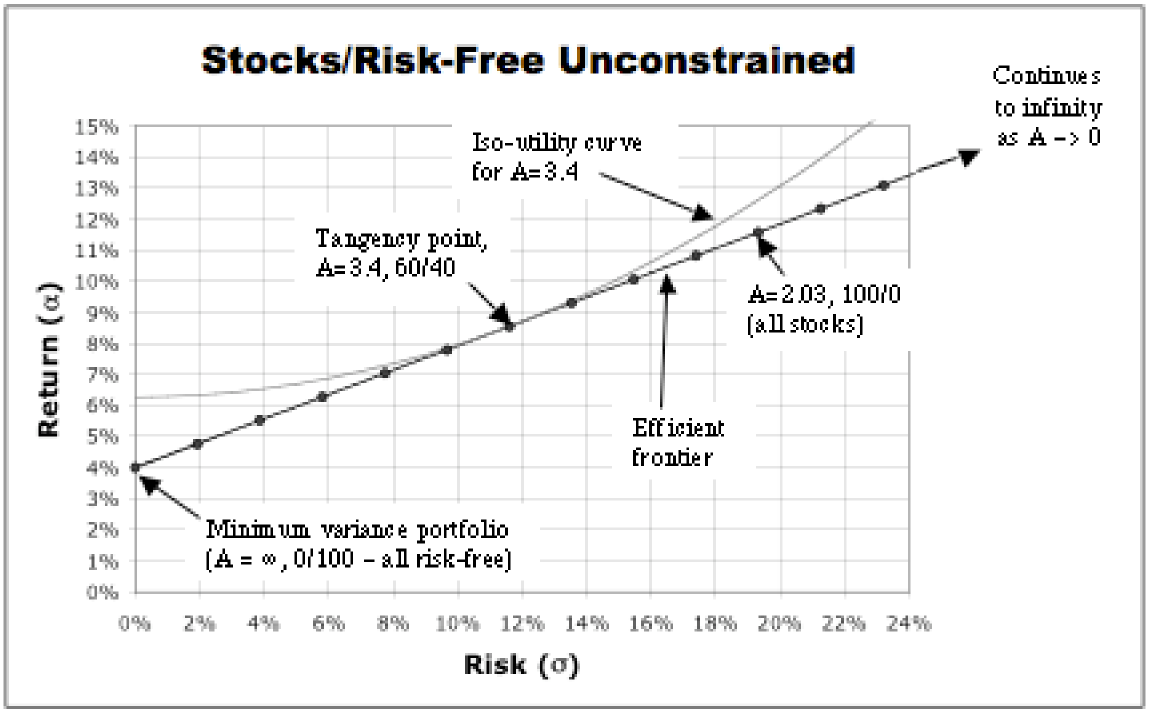
The graph also shows the iso-utility curve tangent to the efficient frontier for an investor with $A = 3.4$. This investor has an optimal moderate 60/40 stock/risk-free portfolio.[10]

Notice that an aggressive 100% stock portfolio is held by an investor with $A = 2.03$. More risk-tolerant investors with $A < 2.03$ actually have optimal portfolios with more than 100% stocks! These investors borrow money at the risk-free rate to leverage their stock holdings.

At the other end of the spectrum, an investor with infinite risk aversion ($A = \infty$) has no stocks in his portfolio, only the risk-free asset.

Note that the efficient frontier is a straight line. As we will see, this is always the case when one of the available assets is risk-free.

[10] We will use our "moderate" or "average" investor with $A = 3.4$ as a running example throughout the rest of this paper.



A	w	$1-w$	σ	α	$r1$	$r2$
∞	0%	100%	0.00%	4.00%	4.08%	4.08%
20.27	10%	90%	1.93%	4.76%	4.87%	4.85%
10.13	20%	80%	3.87%	5.52%	5.67%	5.59%
6.76	30%	70%	5.80%	6.27%	6.47%	6.29%
5.07	40%	60%	7.73%	7.03%	7.28%	6.96%
4.05	50%	50%	9.67%	7.79%	8.10%	7.60%
3.38	60%	40%	11.60%	8.55%	8.92%	8.19%
2.90	70%	30%	13.53%	9.30%	9.75%	8.75%
2.53	80%	20%	15.47%	10.06%	10.58%	9.27%
2.25	90%	10%	17.40%	10.82%	11.43%	9.75%
2.03	100%	0%	19.33%	11.58%	12.27%	10.19%
1.84	110%	−10%	21.27%	12.33%	13.13%	10.60%
1.69	120%	−20%	23.20%	13.09%	13.99%	10.96%

A = iso-elastic coefficient of relative risk aversion.
w = portfolio stock percentage.
$1-w$ = portfolio risk-free percentage.
σ = instantaneous standard deviation of portfolio.
α = expected instantaneous return of portfolio.
$r1$ = arithmetic mean (average) simply compounded expected return.
$r2$ = geometric mean (annualized) simply compounded expected return.

Figure 2: Example 1 – US Stocks Plus a Risk-Free Asset

6 Lagrange Multipliers

In the next section we will derive the general solution to the unconstrained optimization problem. Our solution will involve maximizing a quadratic objective function of the portfolio weight variables subject to the *budget constraint* which says that the sum of the weights must equal 1. We use a standard technique called a *Lagrange multiplier* to deal with the budget constraint. In this section we introduce and discuss this technique.

We give a simple example to illustrate the technique. Suppose we have the following quadratic function f of two variables x and y which we wish to maximize:

$$f(x, y) = 50x + 116y - 12x^2 - 10xy - 76y^2$$

We solve this problem by taking the two partial derivatives, setting them equal to 0, and solving the resulting pair of simultaneous linear equations:

$$\frac{\partial f}{\partial x} = 50 - 24x - 10y = 0$$

$$\frac{\partial f}{\partial y} = 116 - 10x - 152y = 0$$

$$\begin{aligned} 24x + 10y &= 50 \\ 10x + 152y &= 116 \end{aligned}$$

These equations are easy to solve. We can do it by hand, with a calculator, or using a computer program like Excel or Matlab. The approximate (rounded) solution is:

$$\begin{aligned} x &= 1.8 \\ y &= 0.64 \end{aligned}$$

This problem is just simple calculus of multiple variables up to this point. It extends easily to situations involving more than two variables, although solving the linear equations becomes more tedious and using a computer program that can do matrix algebra is a big help with more variables.

Note that in our example we have $x + y = 2.44$. What happens if we modify the problem to insist that $x + y$ have some other value? For example, consider the following constrained version of our example:

Maximize:

$$f(x, y) = 50x + 116y - 12x^2 - 10xy - 76y^2$$

subject to the budget constraint:

$$x + y = 1$$

In this problem we are finding the apex of the intersection of the surface defined by f and the vertical plane defined by the budget constraint, which is a parabola.

We introduce a new third variable λ called the *Lagrange multiplier* and we define the following modified objective function of three variables which has no constraints:

$$\begin{aligned} \hat{f}(x,y,\lambda) &= f(x,y) + \lambda(1-x-y) \\ &= 50x + 116y - 12x^2 - 10xy - 76y^2 + \lambda(1-x-y) \end{aligned}$$

We now take the three partial derivatives of our modified objective function and set them equal to 0:

$$\begin{aligned} \frac{\partial \hat{f}}{\partial x} &= 50 - 24x - 10y - \lambda = 0 \\ \frac{\partial \hat{f}}{\partial y} &= 116 - 10x - 152y - \lambda = 0 \\ \frac{\partial \hat{f}}{\partial \lambda} &= 1 - x - y = 0 \end{aligned}$$

$$\begin{aligned} 24x + 10y + \lambda &= 50 \\ 10x + 152y + \lambda &= 116 \\ x + y &= 1 \end{aligned}$$

Note that the last equation is our budget constraint.

In matrix algebra notation this set of three simultaneous linear equations in three unknowns can be expressed as:

$$\hat{V}\hat{x} = \hat{c}$$

where:

$$\hat{V} = \begin{pmatrix} 24 & 10 & 1 \\ 10 & 152 & 1 \\ 1 & 1 & 0 \end{pmatrix} \qquad \hat{x} = \begin{pmatrix} x \\ y \\ \lambda \end{pmatrix} \qquad \hat{c} = \begin{pmatrix} 50 \\ 116 \\ 1 \end{pmatrix}$$

The solution is:

$$\begin{pmatrix} x \\ y \\ \lambda \end{pmatrix} = \hat{x} = \hat{V}^{-1}\hat{c} = \begin{pmatrix} 0.487 \\ 0.513 \\ 33.2 \end{pmatrix}$$

Notice that we have $x + y = 1$ as required.

We claim that these values of x and y are also a solution to the original problem with the constraint.

It's easy to see why the modified solution must satisfy the constraint condition. This is because any solution to the modified problem satisfies $\frac{\partial \hat{f}}{\partial \lambda} = 0$, which by design is the constraint condition.

Let $\lambda_0 = 33.2 =$ the solution we derived for λ above, and define $g(x, y) = \hat{f}(x, y, \lambda_0)$. Note that the rest of our solution $x_0 = 0.487$ and $y_0 = 0.513$ maximizes the function $g(x, y)$.[11] Note also that along the line $x + y = 1$, $g(x, y) = f(x, y)$. So because (x_0, y_0) maximizes g along the line, it also maximizes f along the line.

What role does the Lagrange multiplier λ play in our solution? To answer this question, notice that:

$$\hat{f}(x, y, \lambda) = f(x, y) + \lambda(1 - x - y)$$

$$\frac{\partial \hat{f}}{\partial x} = \frac{\partial f}{\partial x} - \lambda$$

$$\frac{\partial \hat{f}}{\partial y} = \frac{\partial f}{\partial y} - \lambda$$

In our solution we set the partial derivatives of $\hat{f}$ equal to 0 and solve the resulting equations for x, y and λ. Thus at the solution we have:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = \lambda$$

This says that at the solution the rate of change of our objective function f with respect to x is the same as with respect to y, and λ is the rate of change.

It's not difficult to see why $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ must be equal at the solution. Suppose for example that $\frac{\partial f}{\partial x} > \frac{\partial f}{\partial y}$. Then for some small value δ we could replace x by $x' = x + \delta$ and replace y by $y' = y - \delta$ and we would have $f(x', y') > f(x, y)$ with $x' + y' = 1$. But this would imply that f does not in fact attain its maximum value at (x, y) under the budget constraint.

Another way to say this is that because of our budget constraint $x + y = 1$, the two partial derivatives of our original objective function f may not be 0 at the solution, but they still must have the same value, and the multiplier λ is that value.[12]

[11] This part of the argument is a bit tricky. Our full solution (x_0, y_0, λ_0) does not in fact maximize $\hat{f}(x, y, \lambda)$. It turns out to be a saddle point, and the $\hat{f}$ function has no global or local maxima or minima. But the function $g(x, y)$ which sets $\lambda = \lambda_0$ in $\hat{f}$ does have a maximum value at the point (x_0, y_0). The "slice" of $\hat{f}$ defined by $\lambda = \lambda_0$ is a quadratic surface with its apex at (x_0, y_0).

[12] Note that the fact that our two partial derivatives are equal is a consequence of the fact that the coefficients of x and y in our linear constraint equation are equal. In a more general kind of problem with a different constraint equation this might not be the case.

Our function f is in "equilibrium" at the solution in the sense that its rate of change with respect to x and y is the same at the solution.

This property of the Lagrange multiplier will play a critical role in our derivation of the CAPM equations in chapters 11 and 12, where we will discuss the issue again in that context.

7 General Solution to the Unconstrained Problem

We now solve the general unconstrained problem of optimizing portfolio asset allocation when the choices available are all linear combinations of a finite number of assets.

We only treat the unconstrained problem in this paper. All values for the asset proportions are permitted, even those outside the range $[0, 1]$, which represent short-selling and leverage.

The solution for the simplest case of a single risky asset plus a risk-free asset was trivial. The solution for the general problem is a bit more complicated but we follow the same general idea. First, we derive the equations for the risk and return of the portfolio as functions of the asset proportions in the portfolio, the expected returns of the assets, and the pairwise covariances of the assets. We then state the problem in terms of maximizing a function of the asset proportions parameterized by the coefficient of relative risk aversion. We use a Lagrange multiplier to deal with the budget constraint which says that the sum of the asset proportions must equal 1. We solve the resulting transformed maximization problem by taking partial derivatives, setting them equal to zero, and solving the resulting set of simultaneous linear equations.

Let:

$n =$ number of assets.
$w_i =$ proportion of portfolio invested in asset $i, 1 \leq i \leq n$.
$w =$ column vector of proportions w_i.
$\alpha_i =$ expected instantaneous return of asset $i, 1 \leq i \leq n$.
$x =$ column vector of expected returns α_i.
$\rho_{i,j} =$ covariance of asset i with asset $j, 1 \leq i \leq n$ and $1 \leq j \leq n$.
$V = n \times n$ matrix of covariances $\rho_{i,j}$.
$\alpha_P =$ expected instantaneous return of portfolio.
$\sigma_P =$ instantaneous standard deviation of portfolio.
$A =$ iso-elastic coefficient of relative risk aversion.

Assuming continuous rebalancing, by equations (4) and (5) on page 6, our portfolio follows a lognormal random walk with:

$$\begin{aligned} \alpha_P &= w'x \\ \sigma_P^2 &= w'Vw \end{aligned}$$

By The Portfolio Choice Theorem 1.2, the problem is to maximize $f(w)$ where:

$$\begin{aligned} f(w) &= \alpha_P - \frac{1}{2}A\sigma_P^2 \\ &= w'x - \frac{1}{2}Aw'Vw \\ &= \sum_{i=1}^{n} w_i\alpha_i - \frac{1}{2}A\sum_{i=1}^{n}\sum_{j=1}^{n} w_iw_j\rho_{i,j} \end{aligned}$$

subject to the budget constraint:

$$\sum_{i=1}^{n} w_i = 1$$

To deal with the budget constraint, we introduce a Lagrange multiplier λ and a new objective function $\hat{f}$ with no constraints:

$$\begin{aligned} \hat{f}(w,\lambda) &= f(w) + \lambda\left(1 - \sum_{i=1}^{n} w_i\right) && (12) \\ &= w'x - \frac{1}{2}Aw'Vw + \lambda\left(1 - \sum_{i=1}^{n} w_i\right) && (13) \\ &= \sum_{i=1}^{n} w_i\alpha_i - \frac{1}{2}A\sum_{i=1}^{n}\sum_{j=1}^{n} w_iw_j\rho_{i,j} + \lambda\left(1 - \sum_{i=1}^{n} w_i\right) && (14) \end{aligned}$$

To solve the problem, we want to take the $n+1$ partial derivatives of $\hat{f}$ and set them equal to 0.

Taking the partial derivative with respect to w_i is a bit tricky because of the double summation in the middle of the expression above. It's easiest to see how to do this by writing out the summation in full:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_iw_j\rho_{i,j} =$$

$$\begin{array}{ccccccccc} w_1w_1\rho_{1,1} & + & \cdots & + & w_1w_i\rho_{1,i} & + & \cdots & + & w_1w_n\rho_{1,n} & + \\ w_2w_1\rho_{2,1} & + & \cdots & + & w_2w_i\rho_{2,i} & + & \cdots & + & w_2w_n\rho_{2,n} & + \\ & & & & \cdots & & & & & \\ w_iw_1\rho_{i,1} & + & \cdots & + & w_iw_i\rho_{i,i} & + & \cdots & + & w_iw_n\rho_{i,n} & + \\ & & & & \cdots & & & & & \\ w_nw_1\rho_{n,1} & + & \cdots & + & w_nw_i\rho_{n,i} & + & \cdots & + & w_nw_n\rho_{n,n} & \end{array}$$

The only terms involving w_i are those in column i and row i, so the partial derivatives of all the other terms are 0. The term at the intersection of column i and row i involves w_i^2 while all the others are linear in w_i. Also note that

$\rho_{i,j} = \rho_{j,i}$ by Proposition 7 in reference [8]. These observations reveal that the partial derivative of the double summation with respect to w_i is:

$$2\rho_{i,i}w_i + \sum_{\substack{j=1 \\ j\neq i}}^{n} \rho_{i,j}w_j + \sum_{\substack{j=1 \\ j\neq i}}^{n} \rho_{j,i}w_j = 2\sum_{j=1}^{n} \rho_{i,j}w_j$$

We can now take the $n+1$ partial derivatives of $\hat{f}$ and set them equal to 0:

$$\frac{\partial \hat{f}}{\partial w_i} = \alpha_i - A\sum_{j=1}^{n} \rho_{i,j}w_j - \lambda = 0 \quad (\text{for } 1 \leq i \leq n) \tag{15}$$

$$\frac{\partial \hat{f}}{\partial \lambda} = 1 - \sum_{i=1}^{n} w_i = 0 \tag{16}$$

Rewrite these equations as:

$$\sum_{j=1}^{n} \rho_{i,j}w_j + \frac{\lambda}{A} = \frac{\alpha_i}{A} \quad (\text{for } 1 \leq i \leq n) \tag{17}$$

$$\sum_{i=1}^{n} w_i = 1 \quad (\text{note that this is the budget constraint}) \tag{18}$$

This is a set of $n+1$ linear equations in $n+1$ unknowns which we can solve using linear algebra. Define vectors and matrices as follows:

$$\hat{V} = \begin{pmatrix} \rho_{1,1} & \cdots & \rho_{1,n} & 1 \\ \vdots & & \vdots & \vdots \\ \rho_{n,1} & \cdots & \rho_{n,n} & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}$$

$$\hat{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \lambda/A \end{pmatrix} \qquad \hat{x} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \\ 0 \end{pmatrix} \qquad \hat{y} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

Then equations (17) and (18) become:

$$\hat{V}\hat{w} = \frac{1}{A}\hat{x} + \hat{y} \tag{19}$$

Notice that the matrix $\hat{V}$ is the covariance matrix V for the asset returns enhanced by adding an extra row and column to accommodate the Lagrange multiplier. $\hat{V}$ is called the *enhanced covariance matrix*.

We assume for the moment that the matrix $\hat{V}$ is non-singular and hence has an inverse.[13]

Let:

$$\hat{c} = \hat{V}^{-1}\hat{x} \quad (20)$$

$$\hat{d} = \hat{V}^{-1}\hat{y} \quad (21)$$

Our solution is:

$$\hat{w} = \frac{1}{A}\hat{c} + \hat{d} \quad (22)$$

Note that the column vectors $\hat{c}$ and $\hat{d}$ depend only on the expected returns and covariances of the assets and are independent of the coefficient of relative risk-aversion A. Each optimal asset proportion is:

$$w_i = \frac{1}{A}c_i + d_i \quad (23)$$

For an infinitely risk-averse investor with $A = \infty$, the solution becomes simply $w_i = d_i$, and the resulting portfolio has minimum variance.

The sign of c_i determines whether other investors (with $A < \infty$) have more or less than d_i invested in asset i, and whether as investors become more risk-averse the proportion of asset i increases ($c_i < 0$) or decreases ($c_i > 0$).

The computation of the solution vectors $\hat{c}$ and $\hat{d}$ is easily done using Microsoft Excel, Matlab, or any other computer program which can do basic matrix algebra.

[13] We will return to address the issue of singularity in section 8.

7.1 Math Details

For the sake of exposition, we deliberately omitted some of the mathematical details in our derivation of the solution above. In this section we supply those missing details.

The solution vector $\hat{w}$ we derived in equation (22) is a *critical point* of the enhanced objective function $\hat{f}$ (all of the partial derivatives of $\hat{f}$ are 0 at the solution). We claim that the first n elements of this vector $w_1 \ldots w_n$ are also a solution to the original problem, which was to maximize the function f subject to the budget constraint. We now formally prove this claim.

Let $w =$ the first n elements of the solution vector $\hat{w}$ in equation (22):

$$w_i = \hat{w}_i = \frac{1}{A}c_i + d_i$$

The last element of the solution vector $\hat{w}$ in equation (22) gives the solution for the Lagrange multiplier λ:

$$w_{n+1} = \frac{\lambda}{A} = \frac{1}{A}c_{n+1} + d_{n+1}$$

$$\lambda = c_{n+1} + Ad_{n+1}$$

For any vector of asset weights v, define the function $g(v)$ as follows:

$$g(v) = \hat{f}(v, \lambda) = f(v) + \lambda \left(1 - \sum_{i=1}^{n} v_i \right) \tag{24}$$

Note that g is a function of the n variables $v_1 \ldots v_n$, and λ is a constant. By holding λ constant we have reduced the dimension of the vector space in which we are working from $n+1$ back down to n.

Our first goal is to prove that w globally maximizes g.

We begin by computing the first derivatives of g. By equation (15):

$$\frac{\partial g}{\partial v_i} = \frac{\partial \hat{f}}{\partial v_i} = \alpha_i - A \sum_{j=1}^{n} \rho_{i,j} v_j - \lambda \tag{25}$$

Our solution vector w is a critical point of g where all the first partial derivatives are 0:

$$\frac{\partial g}{\partial v_i}(w) = 0 \quad \text{for all } i$$

The problem we face is that the critical point w is not necessarily a global maximum. It might be a local maximum, or a minimum, or a saddle point. We need to prove that we have found a global maximum and not one of these other possibilities.

We begin by using equation (25) to compute the second partial derivatives of g:

$$\frac{\partial^2 g}{\partial v_i \partial v_j} = -A\rho_{i,j} \tag{26}$$

The *Hessian matrix* of g is defined to be $H(g) =$ the $n \times n$ matrix of all the second partial derivatives of g. Equation (26) shows that:

$$H(g) = -AV \tag{27}$$

where V is the covariance matrix.

It is interesting to note that $H(g)$ is a constant matrix. It is independent of the vector v. This is a consequence of the fact that the objective function is quadratic in the decision variables.

The key observation we need at this point is that the covariance matrix V is *positive semidefinite*. This means that for any vector $v \neq 0$ we have:

$$v'Vv \geq 0$$

Covariance matrices are *always* positive semidefinite. We give the simple proof later in Lemma 10.2, and we will not reproduce that proof here.

Because V is positive semidefinite, and because the coefficient of relative risk aversion $A > 0$, by equation (27), the Hessian matrix $H(g)$ must be *negative semidefinite*, which means that for any vector $v \neq 0$ we have:

$$v'H(g)v \leq 0$$

We are now ready to prove that the solution vector w maximizes the function g. Let $v \neq w$ be any other vector of asset weights. We must show that $g(v) \leq g(w)$.

Let $p(t)$ be the straight line path connecting the two points w and v defined parametrically by the following equation:

$$p(t) = w + t(v - w)$$

Note that $p(0) = w$ and $p(1) = v$.

We want to examine the behavior of the function g along this path from w to v. Our goal is to show that g is non-increasing along this path. To this end we define the following function $h(t)$ for the value of g along the path:

$$h(t) = g(p(t))$$

We have $h(0) = g(w)$ and $h(1) = g(v)$, so the goal is to prove that $h(1) \leq h(0)$.

By the chain rule, the first derivative of h is:

$$h'(t) = \sum_{j=1}^{n} \frac{\partial g}{\partial v_j}(p(t))(v_j - w_j)$$

Note that because $p(0) = w$ and w is a critical point of g, we have $h'(0) = 0$.

A second application of the chain rule computes the second derivative of h:

$$\begin{aligned} h''(t) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 g}{\partial v_i \partial v_j}(p(t))(v_i - w_i)(v_j - w_j) \\ &= (v-w)' H(g)(v-w) \\ &\leq 0 \quad \text{because } H(g) \text{ is negative semidefinite and } v \neq w \end{aligned}$$

$h'(0) = 0$ and $h''(t) \leq 0$ for all t, so we must have $h'(t) \leq 0$ for all $t \geq 0$. This in turn implies that $h(t) \leq h(0)$ for all $t \geq 0$. In particular, $g(v) = h(1) \leq h(0) = g(w)$, and we have our result – the solution vector w does indeed globally maximize the function g.

We are now (finally) ready to prove the main result, that the solution vector w is a solution to the original problem – it maximizes the objective function f subject to the budget constraint.

Recall the equation (24) we used to define the function g above:

$$g(v) = f(v) + \lambda\left(1 - \sum_{i=1}^{n} v_i\right) \tag{28}$$

Consider the line L defined by the budget constraint:

$$\sum_{i=1}^{n} v_i = 1$$

Our solution vector w is on this line: $w \in L$. We have shown that w globally maximizes the function g. Along the line L defined by the budget constraint, equation (28) shows that f and g are equal:

$$f(v) = g(v) \quad \text{for all } v \in L$$

$g(w)$ is the maximum value attained by the function g along the line L, so $f(w) = g(w)$ must be the maximum value attained by the function f along the line L. In other words, w maximizes f subject to the budget constraint.

It is interesting to note that it was necessary to project $\hat{f}$ onto g by holding λ constant in our argument. The full solution vector $\hat{w}$ is a saddle point of $\hat{f}$, not a global maximum, and in fact $\hat{f}$ is unbounded over its entire domain, with no maxima or minima of any kind (global or local). It turns out that the Hessian matrix of $\hat{f}$ is not negative semidefinite, but rather indefinite. This is a typical complexity that arises in problems using Lagrange multipliers.

For the purposes of computation, note how important it is that the covariance matrix V is positive semidefinite. We cannot feed in any random old symmetric matrix for the covariances and get back a meaningful result from the equations. The matrix must represent a valid collection of covariances. It must be positive semidefinite. This is an important consideration in any computer program that uses the equations to optimize portfolios. Any such program must verify the asset correlations and/or covariances entered by the user to check that they are consistent by verifying that the covariance matrix is positive semidefinite.[14]

There is another proof that w maximizes f subject to the budget constraint that is simpler because it avoids calculus and uses only matrix algebra. It is not, however, as intuitive as the first proof we gave above. We give the second proof here for the sake of completeness. We begin with a lemma.

Lemma 7.1 *For any square matrix V of dimension n with elements $\rho_{i,j}$:*

$$(v-w)'V(v-w) = v'Vv - w'Vw - 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_j(v_i - w_i)\rho_{i,j}$$

Proof:

$$\begin{aligned}
(v-w)'V(v-w) &= \sum_{i=1}^{n}\sum_{j=1}^{n}(v_i - w_i)(v_j - w_j)\rho_{i,j} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \rho_{i,j} + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \rho_{i,j} - 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_j v_i \rho_{i,j} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \rho_{i,j} - \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \rho_{i,j} + \\
&\quad 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \rho_{i,j} - 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_j v_i \rho_{i,j} \\
&= v'Vv - w'Vw - 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_j(v_i - w_i)\rho_{i,j}
\end{aligned}$$

We can now prove that w maximizes f subject to the budget constraint. Let v be any other vector of asset weights which also satisfies the budget constraint, and let (w, λ) be the solution we derived above. First note that:

$$\lambda\sum_{i=1}^{n}(v_i - w_i) = \lambda\left(\sum_{i=1}^{n} v_i - \sum_{i=1}^{n} w_i\right) = \lambda(1-1) = \lambda \times 0 = 0$$

[14]This can be done by using the theorem that a symmetric real matrix is positive semidefinite if and only if all of its eigenvalues are non-negative, using any of the many available matrix algebra libraries that can compute eigenvalues.

Then we have:

$$
\begin{aligned}
f(v) - f(w) &= (v-w)'x - \frac{1}{2}A(v'Vv - w'Vw) \\
&= \text{(by the Lemma above)} \\
&\quad (v-w)'x - \frac{1}{2}A\left[(v-w)'V(v-w) + 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_j(v_i - w_i)\rho_{i,j}\right] \\
&= \sum_{i-1}^{n}(v_i - w_i)\alpha_i - A\sum_{i=1}^{n}\sum_{j=1}^{n} w_j(v_i - w_i)\rho_{i,j} - \frac{1}{2}A(v-w)'V(v-w) \\
&= \sum_{i-1}^{n}(v_i - w_i)\alpha_i - A\sum_{i=1}^{n}\sum_{j=1}^{n} w_j(v_i - w_i)\rho_{i,j} - \\
&\quad \lambda\sum_{i=1}^{n}(v_i - w_i) - \frac{1}{2}A(v-w)'V(v-w) \\
&= \sum_{i=1}^{n}\left(\alpha_i - A\sum_{j=1}^{n} w_j\rho_{i,j} - \lambda\right)(v_i - w_i) - \frac{1}{2}A(v-w)'V(v-w) \\
&= \sum_{i=1}^{n}(0)\,(v_i - w_i) - \frac{1}{2}A(v-w)'V(v-w) \\
&= -\frac{1}{2}A(v-w)'V(v-w) \\
&\leq 0 \quad \text{(because } V \text{ is positive semidefinite and } A > 0\text{)}
\end{aligned}
$$

We have shown that $f(v) <= f(w)$ for all v which satisfy the budget constraint. Thus w maximizes f subject to the budget constraint, and our second proof is complete.

7.2 Example 2 – Bonds and Stocks

In this example the feasible set is all unconstrained linear combinations of long US bonds (asset 1) and large US stocks (asset 2). Both assets are risky, and we do not include a risk-free asset in this example. The parameters from Table 1 are:

$$V = \begin{pmatrix} 0.00620 & 0.00246 \\ 0.00246 & 0.03738 \end{pmatrix} = \text{covariance matrix}$$

$$x = \begin{pmatrix} 5.02\% \\ 11.58\% \end{pmatrix} = \text{expected instantaneous returns}$$

The solution vectors are:

$$\hat{c} = \begin{pmatrix} -1.6958 \\ 1.6958 \\ 0.0565 \end{pmatrix} \qquad \hat{d} = \begin{pmatrix} 0.9033 \\ 0.0967 \\ -0.0058 \end{pmatrix}$$

The bond and stock proportions are:

$$\begin{aligned} w_1 &= 0.9033 - 1.6958/A \quad \text{(bonds)} \\ w_2 &= 0.0967 + 1.6958/A \quad \text{(stocks)} \end{aligned}$$

Figure 3 shows the efficient frontier.

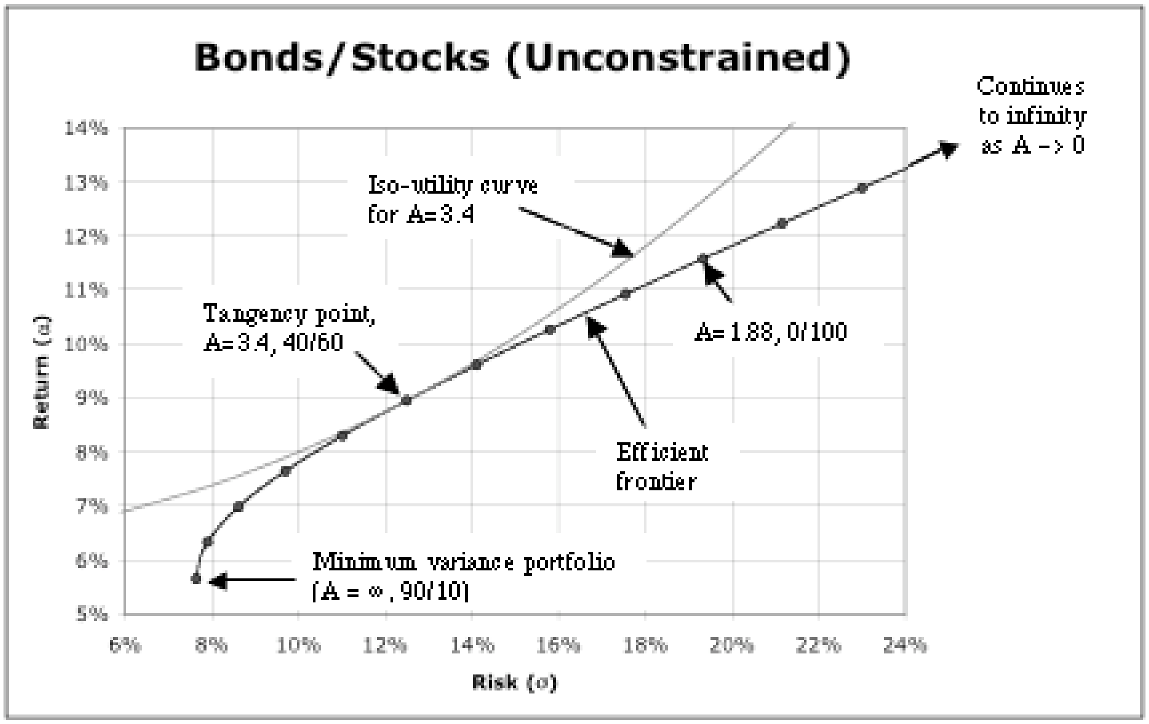
The minimum variance portfolio for $A = \infty$ contains 90.33% bonds and 9.67% stocks. Thus even very risk-averse investors allocate nearly 10% of their portfolios to stocks!

Our average investor with $A = 3.4$ has a moderate 60/40 stock/bond portfolio.

The 100% stock portfolio is held by an aggressive investor with $A = 1.88$.

More aggressive investors with $A < 1.88$ have negative bond allocations and stock allocations greater than 100%. These investors short bonds and use the proceeds to buy extra stocks.

Notice that as A increases (the investor becomes more risk-averse), the proportion of bonds increases and the proportion of stocks decreases, because of the signs (negative and positive respectively) of the first two values in the solution vector $\hat{c}$.



A	w_1	w_2	σ	α	$r1$	$r2$
∞	90%	10%	7.64%	5.65%	5.82%	5.51%
16.41	80%	20%	7.91%	6.33%	6.54%	6.20%
8.34	70%	30%	8.62%	6.99%	7.24%	6.84%
5.59	60%	40%	9.69%	7.64%	7.94%	7.44%
4.20	50%	50%	11.01%	8.30%	8.65%	8.00%
3.37	40%	60%	12.50%	8.95%	9.37%	8.52%
2.81	30%	70%	14.11%	9.61%	10.09%	9.00%
2.41	20%	80%	15.80%	10.27%	10.81%	9.44%
2.11	10%	90%	17.55%	10.92%	11.54%	9.84%
1.88	0%	100%	19.33%	11.58%	12.27%	10.19%
1.69	−10%	110%	21.15%	12.23%	13.01%	10.51%
1.54	−20%	120%	23.00%	12.89%	13.75%	10.79%

A = iso-elastic coefficient of relative risk aversion.
w_1 = portfolio bond percentage.
w_2 = portfolio stock percentage.
σ = instantaneous standard deviation of portfolio.
α = expected instantaneous return of portfolio.
$r1$ = arithmetic mean (average) simply compounded expected return.
$r2$ = geometric mean (annualized) simply compounded expected return.

Figure 3: Example 2 – Bonds and Stocks

7.3 Example 3 – Cash, Bonds and Stocks

In this example the feasible set is all unconstrained linear combinations of cash (30 day US T-Bills) (asset 1), long US bonds (asset 2), and large US stocks (asset 3). All three assets are risky. This example extends Example 2 by adding the new "cash" asset. The parameters from Table 1 are:

$$V = \begin{pmatrix} 0.00096 & 0.00054 & -0.00012 \\ 0.00054 & 0.00620 & 0.00246 \\ -0.00012 & 0.00246 & 0.03738 \end{pmatrix} = \text{covariance matrix}$$

$$x = \begin{pmatrix} 3.67\% \\ 5.02\% \\ 11.58\% \end{pmatrix} = \text{expected instantaneous returns}$$

The solution vectors are:

$$\hat{c} = \begin{pmatrix} -3.2088 \\ 1.2573 \\ 1.9515 \\ 0.0393 \end{pmatrix} \qquad \hat{d} = \begin{pmatrix} 0.9177 \\ 0.0587 \\ 0.0235 \\ -0.0009 \end{pmatrix}$$

The optimal asset proportions are:

$$\begin{aligned} w_1 &= 0.9177 - 3.2088/A \quad \text{(cash)} \\ w_2 &= 0.0587 + 1.2573/A \quad \text{(bonds)} \\ w_3 &= 0.0235 + 1.9515/A \quad \text{(stocks)} \end{aligned}$$

Figure 4 shows the efficient frontier.

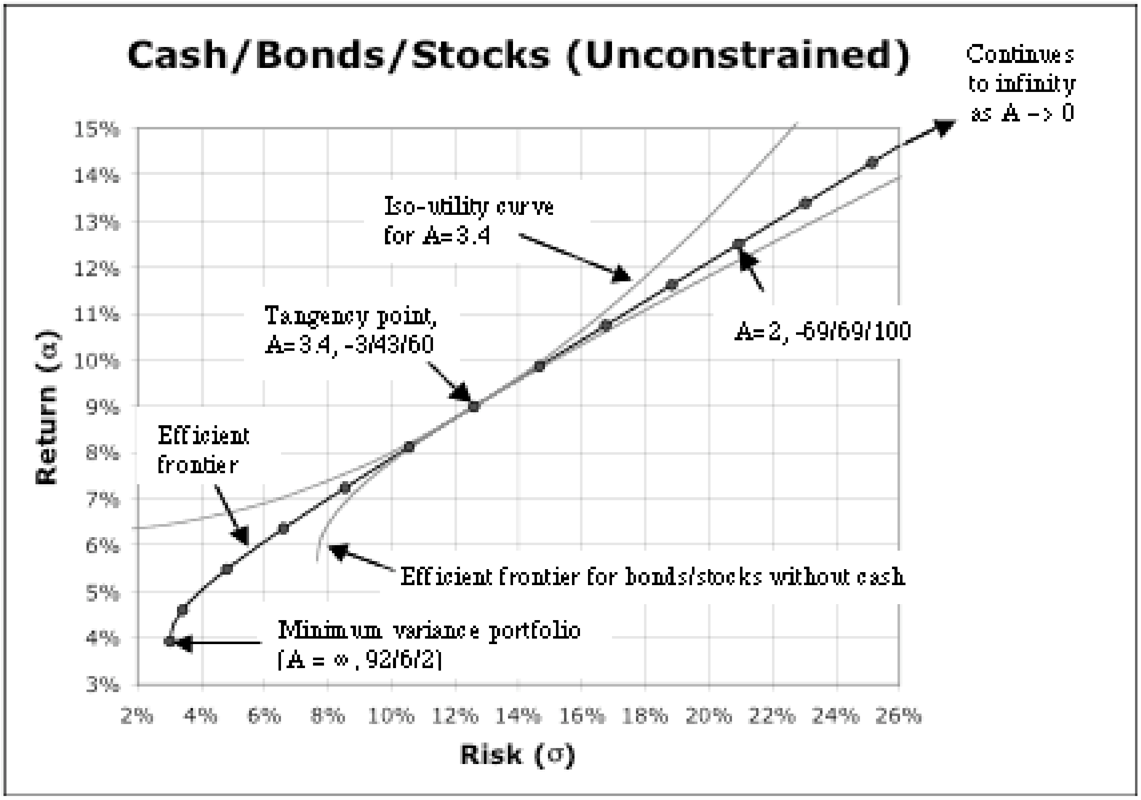
The minimum variance portfolio for $A = \infty$ contains 91.77% cash, 5.87% bonds and 2.35% stocks.

Our average investor with $A = 3.4$ has a 60/43/-3 stock/bond/cash portfolio. This is almost the same 60/40 stock/bond portfolio we saw in example 2, except in this case a small 3% of the portfolio is used to short cash and buy extra bonds.

The 100% stock portfolio is held by an aggressive investor with $A = 2$.

Notice that as A increases (the investor becomes more risk-averse), the proportions of bonds and stocks decrease and the proportion of cash increases.

The graph also shows the bond/stock efficient frontier from Example 2. The addition of the cash asset significantly improves the frontier by moving it to the northwest (in the direction of greater return and less risk), especially at the southwest (less risky) end of the curve. With the extra cash asset, more conservative investors use mixtures of cash and bonds instead of just bonds to reduce their risk, with the ratio of cash to bonds increasing as risk aversion increases.



A	w_1	w_2	w_3	σ	α	$r1$	$r2$
∞	92%	6%	2%	3.02%	3.93%	4.01%	3.96%
25.52	79%	11%	10%	3.43%	4.60%	4.71%	4.65%
11.06	63%	17%	20%	4.81%	5.48%	5.64%	5.51%
7.06	46%	24%	30%	6.60%	6.36%	6.57%	6.34%
5.18	30%	30%	40%	8.54%	7.24%	7.51%	7.12%
4.10	13%	37%	50%	10.55%	8.12%	8.46%	7.85%
3.39	−3%	43%	60%	12.60%	8.99%	9.41%	8.55%
2.88	−19%	49%	70%	14.66%	9.87%	10.38%	9.20%
2.51	−36%	56%	80%	16.75%	10.75%	11.35%	9.80%
2.23	−52%	62%	90%	18.84%	11.63%	12.33%	10.36%
2.00	−69%	69%	100%	20.93%	12.51%	13.32%	10.87%
1.81	−85%	75%	110%	23.03%	13.39%	14.32%	11.33%
1.66	−102%	82%	120%	25.14%	14.26%	15.33%	11.74%

A = iso-elastic coefficient of relative risk aversion.
w_1 = portfolio cash percentage.
w_2 = portfolio bond percentage.
w_3 = portfolio stock percentage.
σ = instantaneous standard deviation of portfolio.
α = expected instantaneous return of portfolio.
$r1$ = arithmetic mean (average) simply compounded expected return.
$r2$ = geometric mean (annualized) simply compounded expected return.

Figure 4: Example 3 – Cash, Bonds and Stocks

7.4 Example 4 – Cash, Bonds and Stocks (Real)

If the purpose of investing is to finance future consumption (and it usually is), then inflation is important. Indeed, it is critical over long time horizons (e.g., when investing for retirement.)

The easiest way to deal with inflation is to use inflation-adjusted returns and other parameters.

In this section we redo Example 3 above only this time we use real returns (inflation-adjusted returns) instead of nominal returns. We use the real parameters from Table 2.

$$V = \begin{pmatrix} 0.00178 & 0.00245 & 0.00091 \\ 0.00245 & 0.00922 & 0.00459 \\ 0.00091 & 0.00459 & 0.03862 \end{pmatrix} = \text{covariance matrix}$$

$$x = \begin{pmatrix} 0.63\% \\ 2.10\% \\ 8.56\% \end{pmatrix} = \text{expected instantaneous returns}$$

The solution vectors are:

$$\hat{c} = \begin{pmatrix} -3.3843 \\ 1.4419 \\ 1.9425 \\ 0.0071 \end{pmatrix} \qquad \hat{d} = \begin{pmatrix} 1.0933 \\ -0.1258 \\ 0.0325 \\ -0.0017 \end{pmatrix}$$

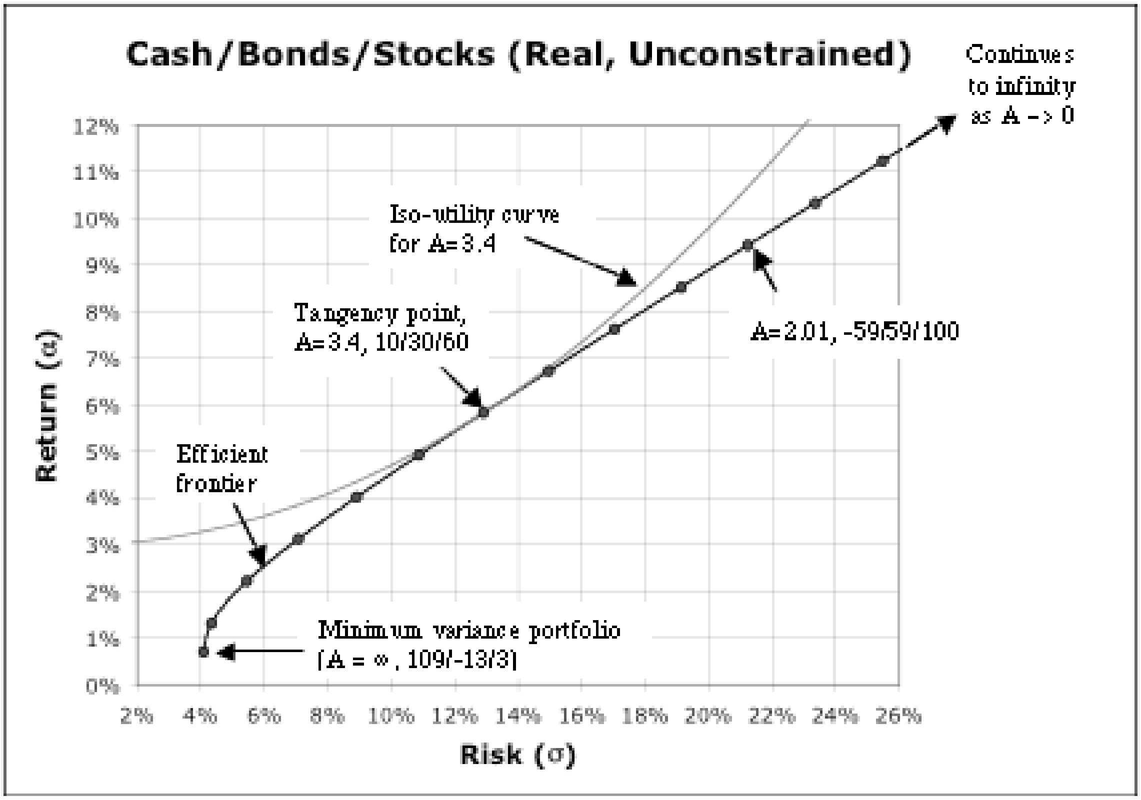
The optimal asset proportions are:

$$\begin{aligned} w_1 &= 1.0933 - 3.3843/A \quad \text{(cash)} \\ w_2 &= -0.1258 + 1.4419/A \quad \text{(bonds)} \\ w_3 &= 0.0325 + 1.9425/A \quad \text{(stocks)} \end{aligned}$$

Figure 5 shows the efficient frontier.

Our average investor with $A = 3.4$ holds 10% cash, 30% bonds, and 60% stocks.

When comparing this example to the nominal Example 3, the most interesting thing to note is that cash plays a more important role in this real example.



A	w_1	w_2	w_3	σ	α	$r1$	$r2$
∞	109%	−13%	3%	4.09%	0.71%	0.71%	0.62%
28.80	98%	−8%	10%	4.34%	1.31%	1.32%	1.23%
11.60	80%	0%	20%	5.45%	2.22%	2.24%	2.09%
7.26	63%	7%	30%	7.07%	3.12%	3.17%	2.91%
5.29	45%	15%	40%	8.91%	4.02%	4.10%	3.69%
4.16	28%	22%	50%	10.87%	4.92%	5.04%	4.43%
3.42	10%	30%	60%	12.89%	5.82%	6.00%	5.12%
2.91	−7%	37%	70%	14.95%	6.72%	6.96%	5.77%
2.53	−24%	44%	80%	17.03%	7.63%	7.92%	6.37%
2.24	−42%	52%	90%	19.13%	8.53%	8.90%	6.93%
2.01	−59%	59%	100%	21.24%	9.43%	9.89%	7.44%
1.82	−77%	67%	110%	23.36%	10.33%	10.88%	7.90%
1.66	−94%	74%	120%	25.48%	11.23%	11.89%	8.31%

A = iso-elastic coefficient of relative risk aversion.
w_1 = portfolio cash percentage.
w_2 = portfolio bond percentage.
w_3 = portfolio stock percentage.
σ = instantaneous standard deviation of portfolio.
α = expected instantaneous return of portfolio.
$r1$ = arithmetic mean (average) simply compounded expected return.
$r2$ = geometric mean (annualized) simply compounded expected return.

Figure 5: Example 4 – Cash, Bonds and Stocks (Real)

8 Singularities in the Unconstrained Problem

In our general solution to the unconstrained problem we made the assumption that the enhanced matrix $\hat{V}$ was non-singular and hence had an inverse. We now fulfill our promise to return to consider the situation where $\hat{V}$ is singular.

Suppose $\hat{V}$ is singular. By Definition 1 in reference [7]:

$$\hat{V}\hat{w} = 0 \quad \text{for some } \hat{w} \neq 0$$

$$\begin{pmatrix} \rho_{1,1} & \cdots & \rho_{1,n} & 1 \\ \vdots & & \vdots & \vdots \\ \rho_{n,1} & \cdots & \rho_{n,n} & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \\ w_{n+1} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

Multiplying out each row gives:

$$\begin{aligned} \sum_{j=1}^{n} \rho_{i,j} w_j + w_{n+1} &= 0 \quad \text{for } 1 \leq i \leq n \\ \sum_{j=1}^{n} w_j &= 0 \end{aligned}$$

We first notice that we must have $w_i \neq 0$ for some $i \leq n$. If this were not the case, the equation for the first row would imply that $w_{n+1} = 0$, which in turn would imply that $\hat{w} = 0$.

Multiply each of the first equations above by w_i and sum them to get:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} w_i \left(\sum_{j=1}^{n} \rho_{i,j} w_j + w_{n+1} \right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \rho_{i,j} + \sum_{i=1}^{n} w_i w_{n+1} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \rho_{i,j} + w_{n+1} \sum_{i=1}^{n} w_i \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \rho_{i,j} + w_{n+1} \times 0 \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \rho_{i,j} \end{aligned}$$

Consider the portfolio P which is a continuously rebalanced linear combination of amount w_i in asset X_i for each $i = 1 \ldots n$. Then:

$$P = \sum_{j=1}^{n} w_j X_j$$

$$\sigma_P^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \rho_{i,j} = 0$$
$$\alpha_P = \sum_{i=1}^{n} w_i \alpha_i$$

P has variance 0 and is therefore risk-free with a constant instantaneous rate of return α_P:

$$P = \sum_{j=1}^{n} w_j X_j = \alpha_P$$

We consider the three cases $\alpha_P < 0$, $\alpha_P > 0$, and $\alpha_P = 0$.

First suppose $\alpha_P < 0$. Reverse the signs of all the w_i. The variance of the portfolio remains 0 and the sign of α_P reverses, so we get the case $\alpha_P > 0$.

Now suppose $\alpha_P > 0$. Our portfolio is risk-free with a positive rate of return. Recall that the sum of the proportions w_i is 0. This is called a "zero-budget" portfolio. We can invest in this portfolio without putting up any money of our own as follows: For each w_i for which $w_i < 0$, we sell amount w_i of asset i short. We use the proceeds to invest amount w_i in all of the assets i for which $w_i > 0$. The resulting portfolio earns a positive rate of return α_P with no risk! This is called an "arbitrage opportunity." By multiplying the values w_i by an arbitrarily large constant we can make an arbitrarily large risk-free profit.

Arbitrage opportunities are difficult to find in the real world. In our optimization problem, if an arbitrage opportunity exists, there is no efficient frontier because the rates of returns of the feasible portfolios are unbounded.

Now consider the last case where we have $\alpha_P = 0$. We know that $w_a \neq 0$ for some $a \leq n$.

$$P = \sum_{j=1}^{n} w_j X_j = 0$$
$$X_a = \sum_{\substack{j=1 \\ j \neq a}}^{n} \frac{w_j}{w_a} X_j$$

In this case the assets are not linearly independent, and asset X_a is a linear combination of the other assets.

Clearly the set of all feasible portfolios which is formed from continuously rebalanced linear combinations of the full set of assets is the same feasible set as the one formed from continuously rebalanced linear combinations of the assets minus asset X_a. Thus we can simply remove X_a from the set of assets and solve the problem for the remaining set of assets.

In general, if the assets are not linearly independent, and if there are no arbitrage opportunities, we can remove dependent ones until we get a set that is independent. The matrix $\hat{V}$ for the subset is non-singular and we can solve the optimization problem for the smaller set of assets. The feasible set and the efficient frontier for the smaller problem are the same as those for the original problem.

Suppose the enhanced covariance matrix $\hat{V}$ is singular. Then one of the eigenvalues of the matrix must be 0, and the corresponding eigenvector x is a non-zero solution to $\hat{V}x = 0$. The first n elements of this eigenvector are the asset weights of a zero-budget portfolio with variance zero. This can be useful in computer programs for finding the linear dependency or arbitrage opportunity, assuming the availability of a linear algebra library that can do eigen decompositions.

9 The Two-Fund Separation Theorem

We now turn our attention to the general unconstrained problem when one of the assets is risk-free. Without loss of generality, we'll assume that the first asset is the risk-free one. We'll also assume that the enhanced matrix $\hat{V}$ is non-singular so that we have a solution.

By definition, a risk-free asset has a standard deviation of 0, a variance of 0, and a covariance of 0 with all other assets. Thus:

$$\rho_{i,1} = \rho_{1,i} = 0 \quad \text{for all } 1 \le i \le n$$

Recall our solution (22) to the general unconstrained problem on page 25:

$$\hat{w} = \frac{1}{A}\hat{c} + \hat{d}$$

We will focus on the solution vector $\hat{d}$. Recall that Equation (21) on page 25 defines this vector as follows:

$$\hat{d} = \hat{V}^{-1}\hat{y}$$

So:

$$\hat{V}\hat{d} = \hat{y}$$

Writing out the matrix and vectors and substituting $\rho_{i,1} = \rho_{1,i} = 0$ gives:

$$\hat{V}\hat{d} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & \rho_{2,2} & \cdots & \rho_{2,n} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & \rho_{n,2} & \cdots & \rho_{n,n} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \\ d_{n+1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} = \hat{y}$$

The first row of $\hat{V}$ times the vector $\hat{d}$ is d_{n+1}, which must equal the first element of $\hat{y}$, which is 0. So:

$$d_{n+1} = 0$$

Now consider the outlined submatrix and subvectors above. Call them $\tilde{V}$, $\tilde{d}$, and $\tilde{y}$. Because $d_{n+1} = 0$, we have $\tilde{V}\tilde{d} = \tilde{y}$.

We first show that $\tilde{V}$ must be non-singular. Suppose it is singular. Then for some $\tilde{w} \neq 0$ we have $\tilde{V}\tilde{w} = 0$:

$$\begin{pmatrix} \rho_{2,2} & \cdots & \rho_{2,n} \\ \vdots & & \vdots \\ \rho_{n,2} & \cdots & \rho_{n,n} \end{pmatrix} \begin{pmatrix} w_2 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

Let $w_1 = -\sum_{i=2}^{n} w_i$ and $w_{n+1} = 0$. Then:

$$\hat{V}\hat{w} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & \rho_{2,2} & \cdots & \rho_{2,n} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & \rho_{n,2} & \cdots & \rho_{n,n} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ w_{n+1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} = 0$$

Thus $\hat{V}$ is singular, a contradiction.

$\tilde{V}$ is non-singular, and we have $\tilde{V}\tilde{d} = \tilde{y} = 0$, so we must have $\tilde{d} = 0$. Thus we have now shown that $d_i = 0$ for all $2 \leq i \leq n+1$.

Because all but the first element of the vector $\hat{d}$ are 0, the last row of $\hat{V}$ times the vector $\hat{d}$ is d_1, which must equal the last element of $\hat{y}$, which is 1. So $d_1 = 1$. We have now shown that when the first asset is risk-free, we have:

$$\hat{d} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Note the intuition behind this result. The vector $\hat{d}$ represents the asset proportions held by an infinitely risk-averse investor. Such an investor allocates all of his portfolio to the first asset, the risk-free one.

The solutions for the optimal asset proportions w_i become:

$$\begin{aligned} w_1 &= \frac{1}{A}c_1 + 1 \quad \text{(the risk-free asset)} \\ w_i &= \frac{1}{A}c_i \quad (2 \leq i \leq n, \text{ the risky assets}) \end{aligned}$$

Now consider the part of the optimal portfolio which contains just the risky assets. We call this the *optimal risky portfolio.* For $2 \leq i \leq n$, the proportion of risky asset i in the optimal risky portfolio is:

$$\frac{w_i}{\sum_{j=2}^{n} w_j} = \frac{\frac{1}{A}c_i}{\sum_{j=2}^{n} \frac{1}{A}c_j} = \frac{\frac{1}{A}c_i}{\frac{1}{A}\sum_{j=2}^{n} c_j} = \frac{c_i}{\sum_{j=2}^{n} c_j}$$

Note that these proportions are all independent of the coefficient of relative risk aversion A, which canceled out in the equation above. In other words, the optimal risky portfolio is the same for all investors. The total optimal portfolio is always a linear combination of the risk-free asset and the optimal risky portfolio.

In other words, we have reduced the general case to the simple case where the available assets are a risk-free asset and a single risky asset. The efficient frontier is a straight line that contains the risk-free asset and the optimal risky portfolio. This is the two-fund separation theorem for the case where one of the assets is risk-free.[15]

There is a good way to visualize this separation theorem. First, draw a graph of the efficient frontier for just the risky assets. Mark the point on the Y axis of this graph where the return is equal to the risk-free rate. The efficient frontier for the full set of assets including all the risky assets and the risk-free asset is the straight line starting at the risk-free point on the Y axis that is tangent to the curved frontier for just the risky assets. The portfolio located at the tangency point is the optimal risky portfolio. See Example 5 below for an example of such a graph.

It is important to think about the meaning of a "risk-free asset" in this context. We are solving the portfolio optimization problem for a single period investment over some time horizon. For this purpose, the risk-free asset must be an investment which pays off all of its guaranteed interest at the exact end of the time horizon in question – e.g., a zero-coupon US treasury bill, note, or bond (depending on the horizon). This notion of a "risk-free" asset specifically does not include money market funds, savings accounts, bonds with coupon payments before the end of the time horizon, or bonds with a maturity different from the time horizon.

[15] We will derive a more general version of this theorem in section 13.

9.1 Example 5 – Bonds, Stocks and a Risk-Free Asset

In this example the feasible set is all unconstrained linear combinations of a risk-free asset (asset 1), long US bonds (asset 2), and large US stocks (asset 3). We use 3.7% for the instantaneous risk-free rate, corresponding to a simply-compounded interest rate of 3.77%. The parameters from Table 1 are:

$$V = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.00620 & 0.00246 \\ 0 & 0.00246 & 0.03738 \end{pmatrix} = \text{covariance matrix}$$

$$x = \begin{pmatrix} 3.7\% \\ 5.02\% \\ 11.58\% \end{pmatrix} = \text{expected instantaneous returns}$$

The solution vectors are:

$$\hat{c} = \begin{pmatrix} -3.3481 \\ 1.3287 \\ 2.0194 \\ 0.0370 \end{pmatrix} \qquad \hat{d} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

The optimal risky portfolio proportions are:

$$\begin{aligned} \text{Bonds} &= \frac{c_2}{c_2 + c_3} = \frac{1.3287}{1.3287 + 2.0194} = 40\% \\ \text{Stocks} &= \frac{c_3}{c_2 + c_3} = \frac{2.0194}{1.3287 + 2.0194} = 60\% \end{aligned}$$

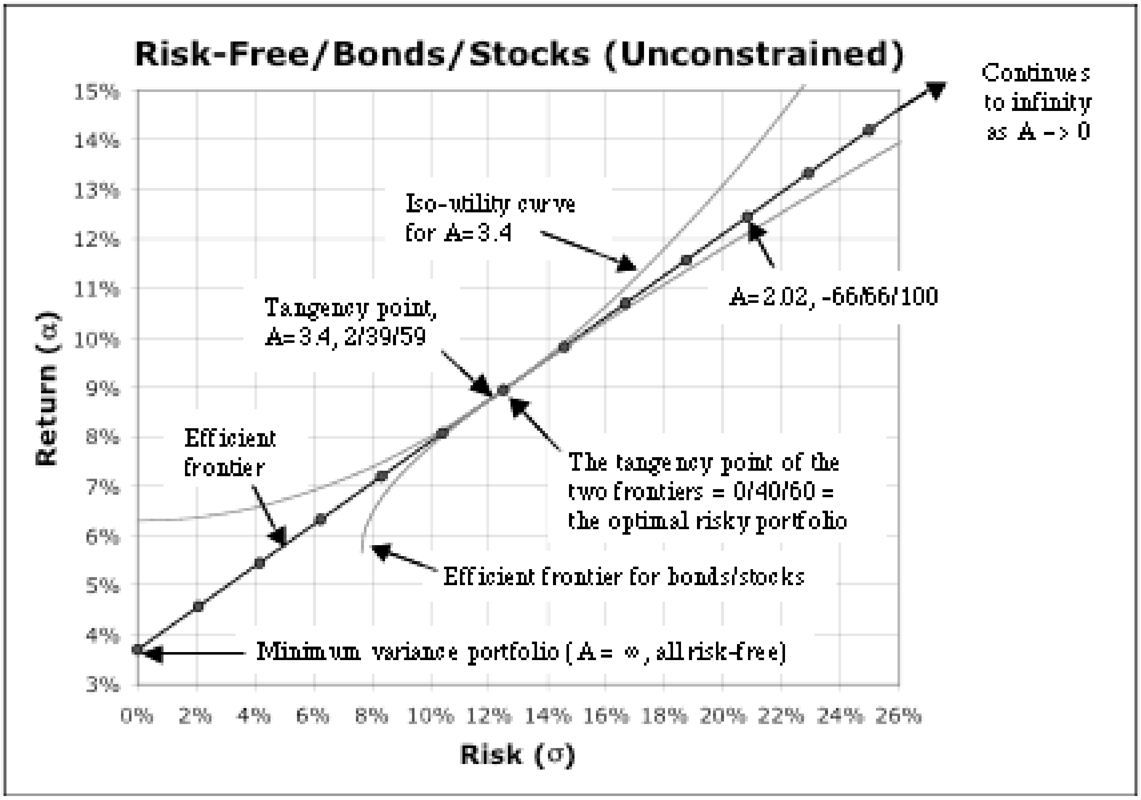
The optimal asset proportions are:

$$\begin{aligned} w_1 &= 1 - 3.3481/A \quad \text{(risk-free)} \\ w_2 &= 1.3287/A \quad \text{(bonds)} \\ w_3 &= 2.0194/A \quad \text{(stocks)} \end{aligned}$$

Or, stated another way:

$$\begin{aligned} w_1 &= 1 - 3.3481/A \quad \text{(risk-free)} \\ 1 - w_1 &= 3.3481/A \quad \text{(60/40 stocks/bonds)} \end{aligned}$$

Figure 6 shows the efficient frontier.



A	w_1	w_2	w_3	σ	α	$r1$	$r2$
∞	100%	0%	0%	0.00%	3.70%	3.77%	3.77%
20.19	83%	7%	10%	2.08%	4.57%	4.68%	4.66%
10.10	67%	13%	20%	4.16%	5.45%	5.60%	5.51%
6.73	50%	20%	30%	6.24%	6.32%	6.53%	6.32%
5.05	34%	26%	40%	8.32%	7.20%	7.46%	7.09%
4.04	17%	33%	50%	10.41%	8.07%	8.41%	7.82%
3.37	1%	39%	60%	12.49%	8.95%	9.36%	8.51%
2.88	−16%	46%	70%	14.57%	9.82%	10.32%	9.16%
2.52	−33%	53%	80%	16.65%	10.70%	11.29%	9.76%
2.24	−49%	59%	90%	18.73%	11.57%	12.27%	10.31%
2.02	−66%	66%	100%	20.81%	12.45%	13.25%	10.83%
1.84	−82%	72%	110%	22.89%	13.32%	14.25%	11.29%
1.68	−99%	79%	120%	24.97%	14.19%	15.25%	11.71%

A = iso-elastic coefficient of relative risk aversion.
w_1 = portfolio risk-free percentage.
w_2 = portfolio bond percentage.
w_3 = portfolio stock percentage.
σ = instantaneous standard deviation of portfolio.
α = expected instantaneous return of portfolio.
$r1$ = arithmetic mean (average) simply compounded expected return.
$r2$ = geometric mean (annualized) simply compounded expected return.

Figure 6: Example 5 – Bonds, Stocks and a Risk-Free Asset

10 The Unconstrained Efficient Frontier

We have learned how to find the optimal unconstrained asset allocation for a give iso-elastic coefficient of relative risk aversion A. The solution to the problem is a set of portfolio weights w_i. Given the portfolio weights we can compute the risk σ_P and expected return α_P of the optimal portfolio.

As we let A range over its domain $(0, \infty]$, the resulting pairs of values (σ_P, α_p) trace out the efficient frontier curve, which is conventionally graphed with σ_P on the X axis and α_P on the Y axis.[16] Our general solution to the unconstrained optimization problem defines this curve parametrically as functions of A.

In this section we investigate the mathematical properties of the efficient frontier curve. We begin with a sequence of four lemmas.

Lemma 10.1 $\sum_{i=1}^{n} c_i = 0$ *and* $\sum_{i=1}^{n} d_i = 1$.

Proof:

Recall from Equation 23 on page 25 that:

$$w_i = \frac{1}{A} c_i + d_i$$

Then:

$$\begin{aligned} 1 &= \sum_{i=1}^{n} w_i \quad \text{(by the budget constraint)} \\ &= \frac{1}{A} \sum_{i=1}^{n} c_i + \sum_{i=1}^{n} d_i \end{aligned}$$

This equation holds for all A, so it holds for $A = \infty$, which proves that $\sum_{i=1}^{n} d_i = 1$, which in turn implies that we must have $\sum_{i=1}^{n} c_i = 0$.

[16]Even though Markowitz originally did it the other way around.

Lemma 10.2 *Suppose V is an $n \times n$ matrix of covariances $\rho_{i,j}$ for a set of random variables $X_1 \ldots X_n$. Then V is positive semidefinite. That is, for any any column vector z of values $z_1 \ldots z_n$, we have:*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j \rho_{i,j} = z'Vz \geq 0$$

Proof:

By Proposition 8 and Definition 2 in reference [8]:

$$z'Vz = \mathrm{Var}\left(\sum_{i=1}^{n} z_i X_i\right) \geq 0$$

Lemma 10.3 $\sum_{i=1}^{n} c_i \alpha_i = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \rho_{i,j} \geq 0.$

Proof:

$$\begin{aligned}
\sum_{i=1}^{n} c_i \alpha_i &= \begin{pmatrix} c_1 & \cdots & c_n & c_{n+1} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \\ 0 \end{pmatrix} \\
&= \hat{c}'\hat{x} \\
&= \hat{c}'\hat{V}\hat{c} \quad \text{(because } \hat{c} = \hat{V}^{-1}\hat{x}\text{, so } \hat{V}\hat{c} = \hat{x}\text{)} \\
&= \begin{pmatrix} c_1 & \cdots & c_n & c_{n+1} \end{pmatrix} \begin{pmatrix} \rho_{1,1} & \cdots & \rho_{1,n} & 1 \\ \vdots & & \vdots & \vdots \\ \rho_{n,1} & \cdots & \rho_{n,n} & 1 \\ 1 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \\ c_{n+1} \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^{n} c_i \rho_{i,1} + c_{n+1} & \cdots & \sum_{i=1}^{n} c_i \rho_{i,n} + c_{n+1} & \sum_{i=1}^{n} c_i \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \\ c_{n+1} \end{pmatrix} \\
&= \sum_{j=1}^{n} c_j \left(\sum_{i=1}^{n} c_i \rho_{i,j} + c_{n+1}\right) + c_{n+1} \sum_{i=1}^{n} c_i \\
&= \sum_{j=1}^{n} c_j \sum_{i=1}^{n} c_i \rho_{i,j} + c_{n+1} \sum_{j=1}^{n} c_j + c_{n+1} \sum_{i=1}^{n} c_i \\
&= \sum_{j=1}^{n} c_j \sum_{i=1}^{n} c_i \rho_{i,j} \quad \text{(by Lemma 10.1)}
\end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \rho_{i,j} \\ &\geq 0 \quad \text{(by Lemma 10.2)} \end{aligned}$$

Lemma 10.4 $\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} c_i d_j \rho_{i,j} = 0.$

Proof:

First, we have:

$$\begin{aligned} \hat{c}'\hat{V}\hat{d} &= \hat{c}'\hat{V}\left(\hat{V}^{-1}\hat{y}\right) \qquad \text{(because } \hat{d} = \hat{V}^{-1}\hat{y}\text{)} \\ &= \hat{c}'\hat{y} \\ &= \begin{pmatrix} c_1 & \cdots & c_n & c_{n+1} \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \\ &= c_{n+1} \end{aligned}$$

Second, we have:

$$\begin{aligned} \hat{c}'\hat{V}\hat{d} &= \begin{pmatrix} c_1 & \cdots & c_n & c_{n+1} \end{pmatrix} \begin{pmatrix} \rho_{1,1} & \cdots & \rho_{1,n} & 1 \\ \vdots & & \vdots & \vdots \\ \rho_{n,1} & \cdots & \rho_{n,n} & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ \vdots \\ d_n \\ d_{n+1} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{n} c_i\rho_{i,1} + c_{n+1} & \cdots & \sum_{i=1}^{n} c_i\rho_{i,n} + c_{n+1} & \sum_{i=1}^{n} c_i \end{pmatrix} \begin{pmatrix} d_1 \\ \vdots \\ d_n \\ d_{n+1} \end{pmatrix} \\ &= \sum_{j=1}^{n} d_j \left(\sum_{i=1}^{n} c_i\rho_{i,j} + c_{n+1}\right) + d_{n+1}\sum_{i=1}^{n} c_i \\ &= \sum_{j=1}^{n} d_j \sum_{i=1}^{n} c_i\rho_{i,j} + c_{n+1}\sum_{j=1}^{n} d_j + d_{n+1}\sum_{i=1}^{n} c_i \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i d_j \rho_{i,j} + c_{n+1} \qquad \text{(by Lemma 10.1)} \end{aligned}$$

Setting these two equations equal yields the result.

This concludes the Lemmas. Our next goal is to calculate some equations and derivatives so that we can draw conclusions about the shape and properties of the efficient frontier.

First recall equation 23 on page 25:

$$w_i = \frac{1}{A} c_i + d_i$$

As a notational convenience, define:

$$B = \frac{1}{A}$$

A is the coefficient of relative risk aversion, so we can think of B as a coefficient of relative risk tolerance. Our equation now becomes:

$$w_i = Bc_i + d_i$$

We first derive equations for α_P and σ_P as functions of B:

$$\begin{aligned}
\alpha_P &= \sum_{i=1}^{n} w_i \alpha_i \\
&= \sum_{i=1}^{n} (Bc_i + d_i) a_i \\
&= B \sum_{i=1}^{n} c_i \alpha_i + \sum_{i=1}^{n} d_i \alpha_i \\
\sigma_P^2 &= \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \rho_{i,j} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} (Bc_i + d_i)(Bc_j + d_j) \rho_{i,j} \\
&= B^2 \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \rho_{i,j} + B \sum_{i=1}^{n} \sum_{j=1}^{n} c_i d_j \rho_{i,j} + B \sum_{i=1}^{n} \sum_{j=1}^{n} c_j d_i \rho_{i,j} + \sum_{i=1}^{n} \sum_{j=1}^{n} d_i d_j \rho_{i,j} \\
&= B^2 \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \rho_{i,j} + \sum_{i=1}^{n} \sum_{j=1}^{n} d_i d_j \rho_{i,j} \quad \text{(by Lemma 10.4)}
\end{aligned}$$

Let:

$$\begin{aligned}
k &= \sum_{i=1}^{n} c_i \alpha_i = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \rho_{i,j} \quad \text{(by Lemma 10.3)} \\
\alpha_{min} &= \sum_{i=1}^{n} d_i \alpha_i \\
\sigma_{min}^2 &= \sum_{i=1}^{n} \sum_{j=1}^{n} d_i d_j \rho_{i,j} \geq 0 \quad \text{(by Lemma 10.2)}
\end{aligned}$$

$$\sigma_{min} = \sqrt{\sigma_{min}^2} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} d_i d_j \rho_{i,j}}$$

Then our equations become:

$$\begin{aligned}
\alpha_P &= kB + \alpha_{min} \\
\sigma_P^2 &= kB^2 + \sigma_{min}^2 \\
\sigma_p &= \sqrt{kB^2 + \sigma_{min}^2} = (kB^2 + \sigma_{min}^2)^{\frac{1}{2}}
\end{aligned}$$

Note the interesting property that the rate of change of σ_P^2 with respect to B^2 is the same as the rate of change of α_P with respect to B. This is our constant k. Also note that if $k = 0$ we have a degenerate solution where the efficient frontier is the single point $(\sigma_{min}, \alpha_{min})$. This can only happen if all the asset expected returns α_i are the same.

If $\sigma_{min} = 0$ we have $\sigma_P = \sqrt{k}B$, both α_P and σ_P are linear in B, and the efficient frontier is a straight line. The minimum variance portfolio for $B = 0$ is risk-free.

The first derivatives are:

$$\begin{aligned}
\frac{d\alpha_P}{dB} &= k \geq 0 \\
\frac{d\sigma_P}{dB} &= kB(kB^2 + \sigma_{min}^2)^{-\frac{1}{2}} \geq 0
\end{aligned}$$

Assume we do not have the degenerate case of a single point, so $k > 0$, and the efficient frontier is a curve defined for all $\sigma_P \geq \sigma_{min}$. We can now compute the first derivative of α_P with respect to σ_P:

$$\begin{aligned}
\frac{d\alpha_P}{d\sigma_P} &= \frac{\frac{d\alpha_P}{dB}}{\frac{d\sigma_P}{dB}} \\
&= \frac{k}{kB(kB^2 + \sigma_{min}^2)^{-\frac{1}{2}}} \\
&= \frac{1}{B}(kB^2 + \sigma_{min}^2)^{\frac{1}{2}} \\
&= (k + \sigma_{min}^2 B^{-2})^{\frac{1}{2}} \\
&= (k + \sigma_{min}^2 A^2)^{\frac{1}{2}} > 0
\end{aligned}$$

This equation says that α_P is an increasing function of σ_P. As the risk of the optimal portfolio increases, so does its return. The is the “no free lunch” principle of investing. Once a portfolio has been optimally diversified, the only way to get a higher expected return is to undertake greater risk, and the only way to lower risk is to sacrifice expected return.[17]

Note that at the minimum variance portfolio corresponding to $A = \infty$ the derivative is ∞ and the line tangent to the efficient frontier at this point is vertical.

For large values of A near the minimum variance portfolio the slope is large. This means that conservative investors get relatively large increases in expected return for undertaking small amounts of extra risk. The most "bang for the buck" in becoming more aggressive is at the conservative end of the frontier.

On the other hand, for smaller values of A farther up the efficient frontier, the slope is smaller. The most "bang for the buck" in becoming more conservative is in the more aggressive ranges of the frontier. This effect is not unlimited or as pronounced as its partner at the conservative end of the frontier, however, since the slope is bounded below by $\sqrt{k}$.

Again assume non-degeneracy with $k > 0$. We now compute the second derivative of α_P with respect to σ_P:

$$\begin{aligned}
\frac{d}{dB}\frac{d\alpha_P}{d\sigma_P} &= -\sigma_{min}^2 B^{-3}(k+\sigma_{min}^2 B^{-2})^{-\frac{1}{2}} \\
\frac{d^2\alpha_P}{d\sigma_P^2} &= \frac{\frac{d}{dB}\frac{d\alpha_P}{d\sigma_P}}{\frac{d\sigma_P}{dB}} \\
&= \frac{-\sigma_{min}^2 B^{-3}(k+\sigma_{min}^2 B^{-2})^{-\frac{1}{2}}}{kB(kB^2+\sigma_{min}^2)^{-\frac{1}{2}}} \\
&= -\frac{\sigma_{min}^2 B^{-3}(kB^2+\sigma_{min}^2)^{\frac{1}{2}}}{kB(k+\sigma_{min}^2 B^{-2})^{\frac{1}{2}}} \\
&= -\frac{\sigma_{min}^2 B^{-3}(kB^2+\sigma_{min}^2)^{\frac{1}{2}}}{k(kB^2+\sigma_{min}^2)^{\frac{1}{2}}} \\
&= -\frac{\sigma_{min}^2}{k}\frac{1}{B^3} \\
&= -\frac{\sigma_{min}^2}{k}A^3 \leq 0
\end{aligned}$$

The second derivative is 0 if and only $\sigma_{min} = 0$, in which case there is a risk-free portfolio with variance 0 and the efficient frontier is a straight line.

In all other cases the second derivative is negative, and the efficient frontier is concave.

[17]This principle is of course also an immediate consequence of the definition of an "efficient portfolio." But verifying that our first derivative is positive is a good sanity check on the math if nothing else.

For large values of A near the minimum variance portfolio the second derivative has a large negative magnitude, which means the slope is rapidly decreasing. This is the typical sharp curve at the southwest end of the frontier.

As $A \to 0$ the slope decreases much more slowly, and the rate of change of the slope approaches 0. From the equation for the first derivative we see that the slope approaches $\sqrt{k}$ in the limit.

For a given point (σ_P, α_P) on the efficient frontier corresponding to a coefficient of risk tolerance B, let y_P = the Y intercept of the straight line that goes through the point and has slope $\sqrt{k}$. Then:

$$\begin{aligned} y_P &= \alpha_P - \sqrt{k}\sigma_P \\ &= kB + \alpha_{min} - \sqrt{k}\sqrt{kB^2 + \sigma_{min}^2} \\ &= \alpha_{min} + \left(kB - \sqrt{k^2B^2 + k\sigma_{min}^2}\right) \end{aligned}$$

Define y as follows:

$$\begin{aligned} y &= \lim_{B\to\infty} y_P \\ &= \alpha_{min} + \lim_{B\to\infty}\left(kB - \sqrt{k^2B^2 + k\sigma_{min}^2}\right) \\ &= \alpha_{min} + \lim_{B\to\infty}\left(\frac{-k\sigma_{min}^2}{kB + \sqrt{k^2B^2 + k\sigma_{min}^2}}\right) \\ &= \alpha_{min} + 0 \\ &= \alpha_{min} \end{aligned}$$

Thus as $A \to 0$ and $B \to \infty$ the efficient frontier asymptotically approaches from below the straight line with slope $\sqrt{k}$ and Y intercept α_{min}.

Consider our general solution to the unconstrained optimization problem. We're given the vector of expected returns x and the covariance matrix V as constants. Given a value for the coefficient of relative risk aversion A, we know how to compute the optimal asset allocation w and the risk σ_P and return α_P for the resulting optimal portfolio P.

It is also easy to work the equations in other directions. Given any of the variables we can compute the others.

If one of the asset proportions w_i is known, solve the following equation for A and then calculate all the other variables:

$$\begin{aligned} w_i &= \frac{1}{A}c_i + d_i \\ A &= \frac{c_i}{w_i - d_i} \end{aligned}$$

If the expected return α_P is known, solve the following equation for A and then calculate all the other variables:

$$\begin{aligned} \alpha_P &= \frac{1}{A}k + \alpha_{min} \\ A &= \frac{k}{\alpha_P - \alpha_{min}} \end{aligned}$$

If the risk σ_P is known, solve the following equation for A and then calculate all the other variables:

$$\begin{aligned} \sigma_P^2 &= \frac{1}{A^2}k + \sigma_{min}^2 \\ A &= \sqrt{\frac{k}{\sigma_P^2 - \sigma_{min}^2}} \end{aligned}$$

Note that we can also express α_P directly as a function of σ_P (and vice-versa) in case a non-parametric equation is desired:

$$\begin{aligned} \alpha_P &= \frac{1}{A}k + \alpha_{min} = k\sqrt{\frac{\sigma_P^2 - \sigma_{min}^2}{k}} + \alpha_{min} = \sqrt{k(\sigma_P^2 - \sigma_{min}^2)} + \alpha_{min} \\ \sigma_P &= = \sqrt{\frac{(\alpha_P - \alpha_{min})^2}{k} + \sigma_{min}^2} \end{aligned}$$

Rearranging these equations slightly gives:

$$\frac{\sigma_P^2}{\sigma_{min}^2} - \frac{(\alpha_P - \alpha_{min})^2}{k\sigma_{min}^2} = 1$$

This equation is the standard form for a hyperbola centered at $(0, \alpha_{min})$.

The efficient frontier curve is one quarter section of the full hyperbola defined by this equation – the quarter which satisfies $\sigma_P \geq \sigma_{min}$ and $\alpha_P \geq \alpha_{min}$.

We have derived the efficient frontier by considering investors with iso-elastic utility. What about other kinds of utility functions? Is it possible that a risk-averse investor with non-iso-elastic utility has an optimal portfolio that is not on the efficient frontier we have derived? The answer is no. In other words, restricting our attention to iso-elastic utility functions is sufficient to derive the entire efficient frontier.

This is not difficult to see. First, recall that σ_{min}^2 is our minimum variance, and our solution includes one and only one efficient portfolio for each variance $\sigma_P^2 \geq \sigma_{min}^2$. Each such solution maximizes the expected return α_P and is therefore the unique efficient portfolio for that level of risk. So the only possibility is that an investor with non-iso-elastic utility might have an efficient portfolio with variance less than σ_{min}^2.

Thus we need to show that σ_{min}^2 is the true minimum variance for all of the portfolios in the feasible set consisting of all unconstrained linear combinations of the given assets. Recall the our solution maximizes the following function for a given iso-elastic coefficient of relative risk aversion A:

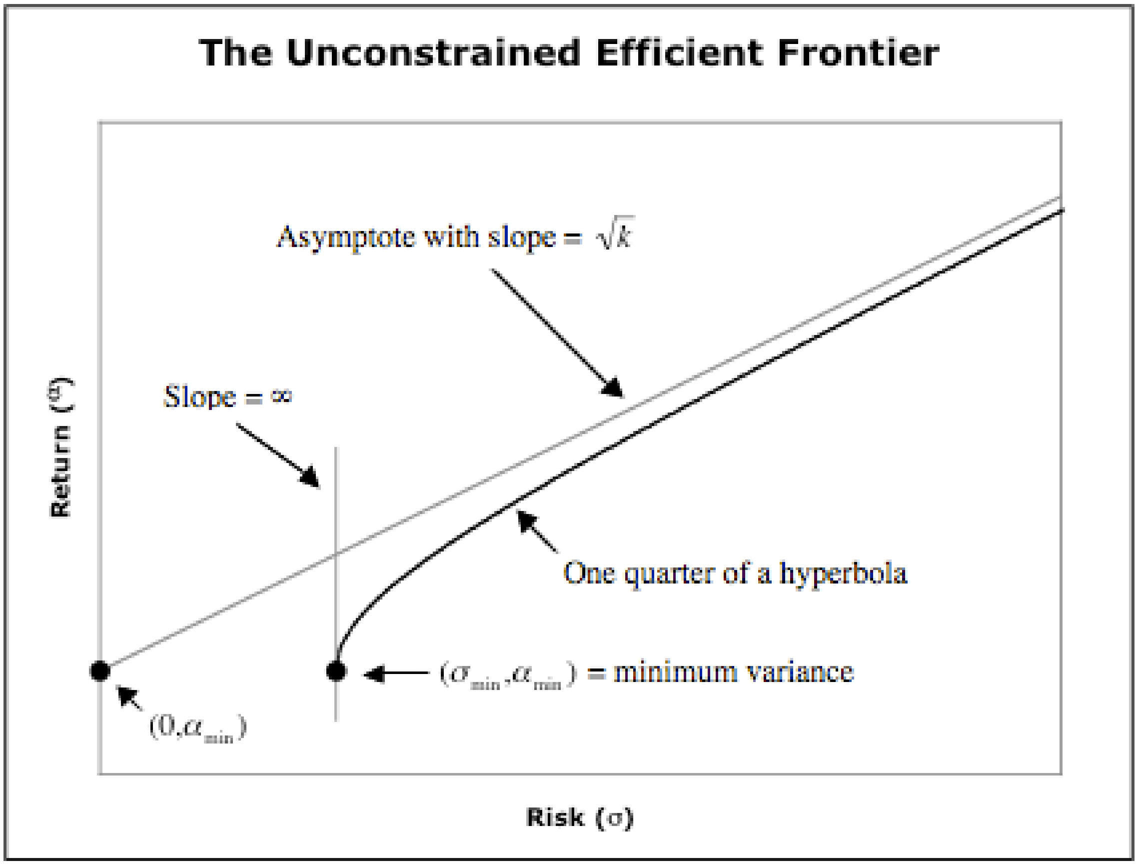
$$\begin{aligned} f(w) &= \alpha_P - \frac{1}{2}A\sigma_P^2 \\ &= \alpha_P - \frac{1}{2B}\sigma_P^2 \quad \text{(where } B = 1/A\text{)} \end{aligned}$$

If we multiply by $-2B$ this maximization problem turns into a minimization problem. Our solution minimizes:

$$g(w) = -2B\alpha_P + \sigma_P^2$$

When $B = 0$ our solution minimizes σ_P^2 over the entire feasible set, and its solution is the portfolio with $\sigma_P^2 = \sigma_{min}^2$.

Figure 7 summarizes and illustrates some of the material covered in this section. Note that the entire efficient frontier curve is completely defined by the three parameters σ_{min}, α_{min}, and k.



$$
\begin{aligned}
k &= \sum_{i=1}^{n} c_i \alpha_i = \text{rate of change of } \alpha_P \text{ with respect to } B \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \rho_{i,j} = \text{rate of change of } \sigma_P^2 \text{ with respect to } B^2 \\
\alpha_{min} &= \sum_{i=1}^{n} d_i \alpha_i = \text{expected return of minimum variance portfolio} \\
\sigma_{min}^2 &= \sum_{i=1}^{n} \sum_{j=1}^{n} d_i d_j \rho_{i,j} = \text{variance of minimum variance portfolio} \\
\alpha_P &= kB + \alpha_{min} = k\frac{1}{A} + \alpha_{min} \\
\sigma_P^2 &= kB^2 + \sigma_{min}^2 = k\frac{1}{A^2} + \sigma_{min}^2
\end{aligned}
$$

Figure 7: The Unconstrained Efficient Frontier

11 Unconstrained Efficient Portfolios

In this section we examine the mathematical properties of unconstrained efficient portfolios and their relationships to both the underlying assets and to other feasible portfolios.

We begin with some new notational conventions.

Let $X_1 \ldots X_n$ be the underlying individual assets in our optimization problem and let Q be a feasible portfolio $Q = \sum_{i=1}^{n} w_i X_i$ with $\sum_{i=1}^{n} w_i = 1$. (Note that Q may or may not be efficient.)

Define:

$$\begin{aligned}
\rho_{Q,i} &= \mathrm{Cov}(Q, X_i) \\
&= \mathrm{Cov}\left(\sum_{j=1}^{n} w_j X_j, X_i\right) \\
&= \sum_{j=1}^{n} w_j \mathrm{Cov}(X_j, X_i) \qquad \text{(by Proposition 9 in reference [8])} \\
&= \sum_{j=1}^{n} w_j \rho_{i,j}
\end{aligned}$$

For any two feasible portfolios Q and R (efficient or not) we use the notation:

$$\rho_{Q,R} = \mathrm{Cov}(Q, R)$$

Recall that the general solution to the unconstrained optimization problem maximizes the expected utility of end-of-period wealth. The expected utility function maximized is:[18]

$$\begin{aligned}
f(w) &= \alpha_P - \frac{1}{2} A \sigma_P^2 \\
&= \sum_{i=1}^{n} w_i \alpha_i - \frac{1}{2} A \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \rho_{i,j}
\end{aligned}$$

The partial derivatives of this function are:

$$\frac{\partial f}{\partial w_i} = \alpha_i - A \sum_{j=1}^{n} w_j \rho_{i,j} = \alpha_i - A \rho_{P,i}$$

[18] $f(w)$ is not the actual iso-elastic expected utility function. See reference [4] for details. But maximizing $f(w)$ is equivalent to maximizing the actual expected utility function, and that's all we need for our purposes here. So we will continue to loosely refer to f as an "expected utility" function. We will also take the liberty of just saying "utility" instead of "expected utility."

We introduced a Lagrange multiplier λ and an enhanced objective function $\hat{f}$. Our solution satisfied:

$$\frac{\partial \hat{f}}{\partial w_i} = \alpha_i - A \sum_{j=1}^{n} w_j \rho_{i,j} - \lambda = \frac{\partial f}{\partial w_i} - \lambda = 0$$

Thus for any efficient portfolio P for coefficient of relative risk aversion A_P and corresponding Lagrange multiplier λ_P we have:

$$\frac{\partial f}{\partial w_i} = \alpha_i - A_P \rho_{P,i} = \lambda_P$$

This is an important property of efficient portfolios. It says that an efficient portfolio is in equilibrium in the sense that its marginal utility with respect to each individual asset is the same. This is also the economic meaning of our multiplier λ.[19]

We now have the following equation for the expected return of asset i which holds for any efficient portfolio P:

$$\alpha_i = \lambda_P + A_P \rho_{P,i} \tag{29}$$

A similar equation for expected return holds for any feasible portfolio Q (efficient or not), as we will now demonstrate. Suppose:

$$Q = \sum_{i=1}^{n} w_i X_i \text{ with } \sum_{i=1}^{n} w_i = 1$$

Then for any efficient portfolio P:

$$\begin{aligned}
\alpha_Q &= \sum_{i=1}^{n} w_i \alpha_i \\
&= \sum_{i=1}^{n} w_i (\lambda_P + A_P \rho_{P,i}) \\
&= \lambda_P \sum_{i=1}^{n} w_i + A_P \sum_{i=1}^{n} w_i \rho_{P,i} \\
&= \lambda_P + A_P \sum_{i-1}^{n} w_i \text{Cov}(P, X_i) \\
&= \lambda_P + A_P \text{Cov}\left(P, \sum_{i=1}^{n} w_i X_i\right) \qquad \text{(by Proposition 9 in reference [8])} \\
&= \lambda_P + A_P \text{Cov}(P, Q) \\
\alpha_Q &= \lambda_P + A_P \rho_{P,Q} \qquad (30)
\end{aligned}$$

[19]See section 6 where we discussed this property of Lagrange multipliers in more detail.

12 The Capital Asset Pricing Model

We now revisit the case where one of the assets is risk-free. Again, without loss of generality, we assume that the first asset X_1 is the risk-free one. Let:

$$r = \alpha_1 = \text{the instantaneous risk-free rate of return}$$

In the previous section we derived the following equation (29) which holds for any efficient portfolio P:

$$\alpha_i = \lambda_P + A_P \rho_{P,i}$$

In particular, for $i = 1$ we get:

$$\alpha_1 = \lambda_P + A_P \rho_{P,1}$$

Asset 1 is risk-free, so its covariance with all the other assets is 0. Substituting $\rho_{P,1} = 0$ and $\alpha_1 = r$ gives:

$$r = \lambda_P$$

This equation holds for all efficient portfolios P. It says that the marginal utility with respect to each asset is the risk-free rate.

For each asset i we now have:

$$\alpha_i = r + A_P \rho_{P,i}$$

As we showed in equation (30), we have the same result for any feasible portfolio Q (efficient or not):

$$\alpha_Q = r + A_P \rho_{P,Q}$$

The risk-free version of the two-fund separation theorem we derived in section 9 tells us that there is a unique efficient portfolio M for some coefficient of risk aversion A_M which is the optimal risky portfolio combining the risky assets. Substitute $P = M$ in our equation above and rearrange to get:

$$\alpha_Q - r = A_M \rho_{M,Q}$$

This equation holds for any feasible portfolio Q, so in particular it holds for $Q = M$ which gives:

$$\alpha_M - r = A_M \rho_{M,M} = A_M \sigma_M^2$$

Dividing these two equations gives:

$$\frac{\alpha_Q - r}{\alpha_M - r} = \frac{A_M \rho_{M,Q}}{A_M \sigma_M^2} = \frac{\rho_{M,Q}}{\sigma_M^2} = \beta_Q$$

$$\alpha_Q - r = \beta_Q(\alpha_M - r)$$

Our new term β_Q is called the *beta* of the portfolio Q: its covariance with M divided by the variance of M. This number is a measure of how much the portfolio varies relative to M. Our equation says that the risk premium of any feasible portfolio is its beta times the risk premium of the optimal risky portfolio.

To illustrate the meaning of beta, let:

$$\begin{aligned}
c_{M,Q} &= \text{the correlation coefficient of } M \text{ and } Q \\
&= \frac{\rho_{M,Q}}{\sigma_M \sigma_Q} \\
&= \beta_Q \frac{\sigma_M}{\sigma_Q} \\
\beta_Q &= c_{M,Q} \frac{\sigma_Q}{\sigma_M} \\
\alpha_Q - r &= c_{M,Q} \frac{\sigma_Q}{\sigma_M} (\alpha_M - r)
\end{aligned}$$

Consider the case where M and Q are perfectly correlated ($c_{M,Q} = +1$):

$$\alpha_Q - r = \frac{\sigma_Q}{\sigma_M} (\alpha_M - r)$$

In this case the risk premium of Q is its volatility relative to M times the risk premium of M. For example, consider the portfolio Q which is a 50/50 mixture of M and the risk-free asset. σ_Q is $\sigma_M/2$ and the risk premium of Q is one half the risk premium of M.

Now consider the case where M and Q are uncorrelated ($c_{M,Q} = 0$). In this case $\beta_Q = 0$ and the expected return of Q is the risk-free rate r. An investor who holds portfolio Q suffers a risk of σ_Q but receives no risk premium as compensation. This is quite inefficient.

Finally consider the case where M and Q are perfectly negatively correlated ($c_{M,Q} = -1$):

$$\alpha_Q - r = -\frac{\sigma_Q}{\sigma_M} (\alpha_M - r)$$

In this case an investor holding portfolio Q actually receives a negative risk premium or a risk "penalty" for undertaking the risk σ_Q! This is terribly inefficient. As an example, consider the portfolio Q that sells M short. In this case we have $\sigma_Q = \sigma_M$ and $\alpha_Q - r = -(\alpha_M - r)$. The risk penalty of Q is equal to the risk premium of M.

In general, if $c_{M,Q} > 0$, the risk premium of Q is positive, with higher risk premia for larger values of $c_{M,Q}$. If $c_{M,Q} < 0$, there is a risk penalty (a negative risk premium), with higher risk penalties for smaller values of $c_{M,Q}$. If $c_{M,Q} = 0$, there is neither a risk premium nor a risk penalty.

The *Sharpe ratio* of a portfolio Q is defined to be:

$$S_Q = \frac{\alpha_Q - r}{\sigma_Q}$$

This number is the risk premium of the portfolio divided by its risk. It is another way to measure how well a portfolio is diversified.

In risk/return graphs, S_Q is the slope of the line joining the risk-free asset and portfolio Q, and S_M is the slope of the efficient frontier (which is also our value $\sqrt{k}$). Note that we always have $S_Q \leq S_M$.

We derived the following equation above:

$$\alpha_Q - r = c_{M,Q} \frac{\sigma_Q}{\sigma_M} (\alpha_M - r)$$

Restated in terms of Sharpe ratios we get:

$$S_Q = c_{M,Q} S_M$$

Once again we see that portfolios which are highly correlated with M offer the best diversification, this time as measured by their Sharpe ratio.

What is the composition of the optimal risky portfolio M? We can easily divine the answer to this question if we assume that all investors agree on the expected returns and covariances of the assets and they all act rationally to select efficient portfolios.

Under these assumptions, all investors hold some combination of the risk-free asset and the same optimal risky portfolio M. When we aggregate the risky holdings over all investors we must get the total market portfolio which consists of all of the assets in their capitalization-weighted proportions. Thus M is the cap-weighted market portfolio.

Note that the CAPM equations which we have derived in this section are all valid whether or not M is the market portfolio. The equations do not depend on the assumptions that investors agree on expected returns and covariances or that they all act rationally to select efficient portfolios. The equations are *normative* in the sense that they describe the necessary relationships among assets and portfolios that must hold if estimates of the asset expected returns and covariances are given. The assumptions which lead to the conclusion that M is the market portfolio and/or the conclusion itself are *positive* in the sense that they are statements about how investors and/or markets actually behave.

In the context of the US stock market, CAPM is often used to model individual stocks, sectors, and asset classes such as style and size subsets. The betas of these securities and market subsets are usually estimated by performing linear regressions of the asset risk premia against the risk premia of a proxy for the total market portfolio, typically the S&P 500 index or a total market index like the Russell 3000 or the Wilshire 5000.

On the typical risk/return graph, the straight line efficient frontier is called the "capital market line." It has slope $S_M = \sqrt{k}$ and Y-intercept the risk-free rate r.

For a given portfolio or asset we can also graph the market risk premium on the X axis vs. the portfolio risk premium on the Y axis. This straight line is called the "security characteristic line." It has slope β and Y-intercept 0. It is often graphed along with observed data points for the realized market and portfolio premia, in which case the security characteristic line is the regression "best fit" line to the data points and may have a non-zero Y-intercept called the portfolio's "alpha."

We can also graph beta on the X axis vs. the portfolio expected return on the Y axis. This straight line is called the "security market line." Its slope is the market risk premium and its Y-intercept is the risk-free rate r.

A full treatment of CAPM would require a long paper or book of its own, and many such papers and books have been written. Our treatment here is just a cursory introduction. See reference [10] for the classic discussion by William Sharpe, who is the creator of this theory.

13 The General Two-Fund Separation Theorem

In this section we derive the general form of the two-fund separation theorem which does not rely on the assumption of a risk-free asset. We use the notation and terminology of section 10.

Suppose that P is the efficient portfolio for $A = A_P$ and Q is the efficient portfolio for $A = A_Q$. Let $B_P = 1/A_P$ and $B_Q = 1/A_Q$. We can compute the covariance of P and Q:

$$\begin{aligned}
\rho_{P,Q} &= \text{Cov}\left(\sum_{i=1}^{n}[B_P c_i + d_i]X_i, \sum_{j=1}^{n}[B_Q c_j + d_j]X_j\right) \\
&= \text{(by Proposition 9 in reference [8])} \\
&\quad \sum_{i=1}^{n}\sum_{j=1}^{n}(B_P B_Q c_i c_j + B_P c_i d_j + B_Q c_j d_i + d_i d_j)\rho_{i,j} \\
&= B_P B_Q \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \rho_{i,j} + B_P \sum_{i=1}^{n}\sum_{j=1}^{n} c_i d_j \rho_{i,j} + \\
&\quad B_Q \sum_{i=1}^{n}\sum_{j=1}^{n} c_j d_i \rho_{i,j} + \sum_{i=1}^{n}\sum_{j=1}^{n} d_i d_j \rho_{i,j} \\
&= B_P B_Q k + 0 + 0 + \sigma^2_{min} \quad \text{(by Lemma 10.4)} \\
&= B_P B_Q k + \sigma^2_{min}
\end{aligned}$$

We now consider linear combinations of efficient portfolios. Let P and Q be as above and consider the portfolio R with is the linear combination of proportion w in P and $1 - w$ in Q. The expected return and variance of R are:

$$\begin{aligned}
\alpha_R &= w\alpha_P + (1-w)\alpha_Q \\
&= w(B_P k + \alpha_{min}) + (1-w)(B_Q k + \alpha_{min}) \\
&= (B_P w + B_Q(1-w))k + \alpha_{min} \\
\sigma^2_R &= w^2\sigma^2_P + 2w(1-w)\rho_{P,Q} + (1-w)^2\sigma^2_Q \\
&= w^2(B_P^2 k + \sigma^2_{min}) + 2w(1-w)(B_P B_Q k + \sigma^2_{min}) + \\
&\quad (1-w)^2(B_Q^2 k + \sigma^2_{min}) \\
&= (w^2 B_P^2 + 2w(1-w)B_P B_Q + (1-w)^2 B_Q^2)k + \\
&\quad (w^2 + 2w(1-w) + (1-w)^2)\sigma^2_{min} \\
&= (B_P w + B_Q(1-w))^2 k + \sigma^2_{min}
\end{aligned}$$

Define C as:

$$C = B_P w + B_Q(1-w)$$

Then we have:

$$\begin{aligned} \alpha_R &= Ck + \alpha_{min} \\ \sigma_R^2 &= C^2 k + \sigma_{min}^2 \end{aligned}$$

Our portfolio R is the efficient portfolio for $B = 1/A = C$.

We have shown that all linear combinations of efficient portfolios are efficient. If we take any two distinct efficient portfolios P and Q, we can therefore generate all of the other efficient portfolios as unconstrained linear combinations of P and Q. This is the general form of the two-fund separation theorem.

In particular, this is true if one of the portfolios is the minimum variance portfolio and the other one is any other efficient portfolio. For example, let P be the efficient portfolio for $A = A_P = 1$ ($B_P = 1/A_P = 1$) and let Q be the minimum variance portfolio for $A = A_Q = \infty$ ($B_Q = 1/A_Q = 0$). Our equations become:

$$\begin{aligned} C &= w \\ \alpha_R &= wk + \alpha_{min} \\ \sigma_R^2 &= w^2 k + \sigma_{min}^2 \end{aligned}$$

References

[1] Zvi Bodie and Robert C. Merton. *Finance*. Prentice-Hall, preliminary edition, 1998.

[2] Harry M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Blackwell, second edition, 1991. (Originally published in 1959.).

[3] Robert C. Merton. *Continuous-Time Finance*. Blackwell, 1990.

[4] John Norstad. An introduction to portfolio theory. http://homepage.mac.com/j.norstad/finance, Apr 1999.

[5] John Norstad. An introduction to utility theory. http://homepage.mac.com/j.norstad/finance, Mar 1999.

[6] John Norstad. The normal and lognormal distributions. http://homepage.mac.com/j.norstad/finance, Feb 1999.

[7] John Norstad. Linear algebra review. http://homepage.mac.com/j.norstad/finance, Sep 2002.

[8] John Norstad. Probability review. http://homepage.mac.com/j.norstad/finance, Sep 2002.

[9] John Norstad. Random walks. http://homepage.mac.com/j.norstad/finance, Jan 2005.

[10] William F. Sharpe. *Portfolio Theory and Capital Markets*. McGraw-Hill, 2000. (Originally published in 1970.).