1

The Classic Capital Asset Pricing Model

1.1 Introduction

An investor is considering creating a portfolio of securities. How should they choose the proportions of these securities based on their attitude toward risk? What are the asset pricing implications of all the investors' portfolio choices? Are these choices consistent with the absence of arbitrage opportunities?

This chapter deals with these issues in the context of the static markets idealized in the first contributions that would give birth to financial economics. Section 1.2 explains how portfolio selection works when our investor optimizes a mean-variance criterion as in the seminal approach of Markovitz (1952). According to this criterion, an investor seeks to maximize the portfolio expected return for any given variance constraint—the celebrated "mean-variance model." The ensuing asset allocation leads, in equilibrium (supply equals demand for all assets), to a notion of "market portfolio" as well as a first theory of asset prices—the celebrated "capital asset pricing model (CAPM)" of Sharpe (1964), Lintner (1965), and Mossin (1966) (see section 1.3).

The market portfolio is the first instance of a "preference-free" construct in financial economics: the asset composition of this portfolio is independent of the investors' risk aversion. Remarkably, any desired profile in the investors' risk-return trade-off is obtained as a combination of a purely safe asset and the market portfolio (hence the name). We shall explain that there are nuances to be made regarding these preference-free attributes, but the notion of a market portfolio is important, representing as it does a benchmark for each individual asset evaluation: the striking conclusion of the CAPM is that each asset's expected return is proportional to the expected return on the market portfolio. The volatility (i.e., riskiness) of each asset return then increases with the asset exposure to market movements. This asset exposure is known as beta.

These explanations are elegant albeit quite coarse descriptions of asset markets. Section 1.4 develops the "arbitrage pricing theory (APT)" model of Ross (1976). The APT provides refinements of the CAPM and predicts that each asset's expected return links to a number of factors and not just the market, as the original CAPM suggests. For example, there might be macroeconomic forces explaining stock returns because investors require compensation (i.e., a risk premium) for investing in risky assets that perform poorly when the economy

performs equally poorly, which is usually the norm. The APT formalizes these concepts very elegantly.

How successful were these initial explanations of asset price fluctuations? It is well known that the single-factor CAPM cannot explain the cross-section of stock returns. It is one of the first asset pricing puzzles that financial economists had to face while developing the field. Section 1.5. explains that the empirical evidence suggests that at least three factors explain the cross-section of stock returns. In fact, the literature has produced a huge amount of evidence regarding many additional factors. We shall succinctly discuss this evidence and the extent to which some of them are merely "lucky factors."

The important point of this empirical evidence is that it provides motivation and guidance to the formulation of new models. These models can be seen as dynamic extensions of the APT. For example, a recurrent theme in financial economics is how "cyclical forces" affect asset evaluation and, in turn, how capital market developments feed back the business cycle. The original APT framework was the initial reference for thinking about these themes as studied in detail in more advanced parts of the book (see, e.g., chapters 8 and 9). More generally, it was a milestone for all subsequent work aimed to address various "asset pricing puzzles."

The empirical literature abounds with additional instances of asset pricing "anomalies"; that is, difficulties the CAPM (and related static models) encounters while explaining market behavior. Some of these anomalies are now understood to link to frictions that are behind the very price formation process. Information problems, agency issues, limited rationality, capital immobility, and decentralized markets are all examples of hindrances to market efficiency. In fact, the notion of "market efficiency" is elusive; financial economists actually agree that some (rigorously defined) inefficiency is needed to have markets function in the first place—an equilibrium degree of disequilibrium, in the words of Grossman and Stiglitz (1980) (see chapters 5 and 10). For example, bid-ask spreads are needed to incentivize financial institutions to trade securities in their roles of intermediaries. While these topics are dealt with in many junctures of this book (e.g., in chapters 5 and 10), it is important to emphasize them in the first chapter of this book.

However, the CAPM has many undisputed potentially useful features. A striking feature of the basic portfolio allocation rules derived in this chapter is that they are simple to understand. Thus, mean-variance asset allocation roughly suggests to invest more in assets that have higher expected returns and less volatility. In fact, we have already explained that the CAPM has been the basic starting point for further developments in the history of thought. But how successful the CAPM allocation rules have been in the market practice? This chapter provides some perspective on these rules. For example, section 1.3 explains that implementing the mean-variance model may be very difficult; one reason is that one of its main ingredients, the assets' expected returns, can only be estimated with poor precision. We review alternative rules, based on a Bayesian framework, where portfolio decisions incorporate the user's views on the asset expected returns (the famous Black-Litterman model). We also review the basic principles underlying risk parity in asset allocation: because expected returns are difficult to estimate, a possibility may be to proceed while ignoring the assets' expected returns in the first place. In its basic version, risk parity would suggest creating portfolios that weigh more assets that have less volatility. In fact, section 1.3 provides some foundations for this portfolio rule, based on "Knightian

uncertainty," a situation arising when decision makers do not understand the statistical laws underlying their environment; chapter 9 (in part II) studies the asset price implications of Knightian uncertainty in more detail (see section 9.6). Section 1.3 shows that risk parity is a portfolio strategy that investors adopt, in equilibrium, once Knightian uncertainty is high. More advanced parts of this book (see, e.g., chapters 9 or 11) explain that a wide adoption of risk parity trading strategies may be destabilizing as assets would be sold when market volatility is high, a circumstance which typically occurs when markets fall. Finally, section 1.5 briefly reviews developments in the industry practice based on factor investing. Because additional factors appear to help explain the cross-section of expected returns, it seems natural that investors should seek exposure to factors that go beyond the market. Factor investing aims to seek such exposures ("smart beta").

1.2 Portfolio Selection

This section describes the process of wealth allocation for a representative investor who cares about "mean-variance efficiency." This notion of efficiency constraints our investor to choices of portfolios that produce the highest expected return for a given level of risk.

1.2.1 Wealth Constraints

Our investor can invest into m risky assets and one safe asset. Let $S=[S_1,\cdots,S_m]$ be the risky asset's price vector, and let S_0 be the price of the riskless asset. We wish to determine the value of a portfolio of all these assets. Let $\theta=[\theta_1,\cdots,\theta_m]$, where θ_i is the number of the ith risky asset, and let θ_0 be the number of the riskless asset in this portfolio. The initial wealth is $w=S_0\theta_0+S\cdot\theta$. Terminal wealth is $w'=x_0\theta_0+x\cdot\theta$, where x_0 is the payoff promised by the riskless asset and $x=[x_1,\cdots,x_m]$ is the vector of the payoffs pertaining to the risky assets—that is, x_i is the payoff promised by the ith risky asset.

Let $R\equiv\frac{x_0}{S_0}$ and $\tilde{R}_i\equiv\frac{x_i}{S_i}$ denote the gross returns from investing into the safe asset and the ith risky asset. Accordingly, define $r\equiv R-1$ as the safe interest rate, $\tilde{\mu}\equiv[\tilde{\mu}_1,\cdots,\tilde{\mu}_m]^\top$, where $\tilde{\mu}_i\equiv\tilde{R}_i-1$ is the return on the ith risky asset and $\mu\equiv E(\tilde{\mu})$, the vector of the expected returns on the risky assets. Finally, we let $\pi=[\pi_1,\cdots,\pi_m]^\top$, where $\pi_i\equiv\theta_i S_i$ is the wealth invested in the ith asset. We have

$$w'=x_0\theta_0+\sum_{i=1}^{m}x_i\theta_i\equiv R\pi_0+\sum_{i=1}^{m}\tilde{R}_i\pi_i \quad\text{and}\quad w=\pi_0+\sum_{i=1}^{m}\pi_i. \tag{1.1}$$

Combining the two expressions for w' and w leaves

$$w'=\pi^\top(\tilde{R}-\mathbf{1}_m R)+Rw=\pi^\top(\mu-\mathbf{1}_m r)+Rw+\pi^\top(\tilde{\mu}-\mu).$$

We use the decomposition, $\tilde{\mu}-\mu=B\cdot\tilde{u}$, where B is a $m\times d$ volatility matrix, with $m\leq d$, and $\tilde{u}$ is a random vector with expectation zero and variance-covariance matrix equal to the identity matrix. With this decomposition, we can rewrite the budget constraint in eq. (1.1) as follows:

$$w'=\pi^\top(\mu-\mathbf{1}_m r)+Rw+\pi^\top B\tilde{u}. \tag{1.2}$$

We can now use eq. (1.2) and determine the expected return and the variance of the portfolio value. We have

$$E\left[w'(\pi)\right] = \pi^\top(\mu - \mathbf{1}_m r) + Rw \quad \text{and} \quad var\left[w'(\pi)\right] = \pi^\top \Sigma \pi, \tag{1.3}$$

where $\Sigma \equiv BB^\top$. Let $\sigma_i^2 \equiv \Sigma_{ii}$. We assume that Σ has full rank and that for all i, j, $\sigma_i^2 > \sigma_j^2 \Longrightarrow \mu_i > \mu_j$. The last condition does actually hold in the equilibrium of markets in which investors are averse to risk, a topic dealt with several times in this chapter; further, note more trivially that this condition implies that $r < \min_j(\mu_j)$.

1.2.2 Portfolio Choice: The "Capital Market Line"

We assume that the investor maximizes the expected return on their portfolio conditionally on a given level of uncertainty. That is, let $w^2 \cdot v_p^2$ define the maximum level of dollar variance the investor is willing to accept. We consider the following program based on eq. (1.3) and the uncertainty constraint:

$$\hat{\pi}\left(v_p\right) = \arg\max_{\pi \in \mathbb{R}^m} E\left[w'(\pi)\right] \quad \text{s.t.} \quad var\left[w'(\pi)\right] = w^2 \cdot v_p^2. \tag{1.P1}$$

The first-order conditions for [1.P1] are

$$\hat{\pi}\left(v_p\right) = (2\nu)^{-1}\Sigma^{-1}(\mu - \mathbf{1}_m r) \quad \text{and} \quad \hat{\pi}^\top \Sigma \hat{\pi} = w^2 \cdot v_p^2,$$

where ν is a Lagrange multiplier for the variance constraint. Plugging the first condition into the second, we obtain $(2\nu)^{-1} = \mp \frac{w \cdot v_p}{\sqrt{\text{Sh}}}$, where

$$\text{Sh} \equiv (\mu - \mathbf{1}_m r)^\top \Sigma^{-1}(\mu - \mathbf{1}_m r) \tag{1.4}$$

is the *Sharpe market performance*. To ensure efficiency, we take the positive solution. Substituting the positive solution for $(2\nu)^{-1}$ into the first-order condition, we obtain the portfolio that solves [1.P1]:

$$\frac{\hat{\pi}\left(v_p\right)}{w} \equiv \frac{\Sigma^{-1}(\mu - \mathbf{1}_m r)}{\sqrt{\text{Sh}}} \cdot v_p. \tag{1.5}$$

We are ready to determine the value of [1.P1], $E\left[w'(\hat{\pi}\left(v_p\right))\right]$, and hence the expected portfolio return, defined as

$$\mu_p(v_p) \equiv \frac{E\left[w'(\hat{\pi}(v_p))\right] - w}{w} = r + \sqrt{\text{Sh}} \cdot v_p, \tag{1.6}$$

where the last equality follows by a simple calculation. Eq. (1.6) describes what is known as the *Capital Market Line* (CML).

1.2.3 Without the Safe Asset: The "Efficient Portfolio Frontier"

Next, assume that the investor's choice space does not include the riskless asset. Their current wealth is now $w = \sum_{i=1}^m \pi_i$ while their terminal wealth is $w' = \sum_{i=1}^m \tilde{R}_i \pi_i$, such that, and relying on the definition of $\tilde{\mu}_i$,

$$w' = \sum_{i=1}^m \tilde{\mu}_i \pi_i + \sum_{i=1}^m \pi_i = \pi^\top \mu + w + \pi^\top B\tilde{u}, \tag{1.7}$$

where B and $\tilde{u}$ are as defined as in eq. (1.2). We can use eq. (1.7) to determine the expected return and the variance of the portfolio value, which are

$$E\left[w'(\pi)\right] = \pi^\top \mu + w, \text{ where } w = \pi^\top \mathbf{1}_m \quad \text{and} \quad var\left[w'(\pi)\right] = \pi^\top \Sigma \pi. \tag{1.8}$$

The program our investor now solves is

$$\hat{\pi}\left(v_p\right) = \arg\max_{\pi \in \mathbb{R}} E\left[w'(\pi)\right] \quad \text{s.t. } var\left[w'(\pi)\right] = w^2 \cdot v_p^2 \text{ and } w = \pi^\top \mathbf{1}_m. \tag{1.P2}$$

In appendix 1.A, we show that, provided $AC - D^2 > 0$ (a second-order condition, as explained in appendix A.1), the solution to [1.P2] is

$$\frac{\hat{\pi}\left(v_p\right)}{w} = \frac{C\mu_p(v_p) - D}{AC - D^2}\Sigma^{-1}\mu + \frac{A - D\mu_p(v_p)}{AC - D^2}\Sigma^{-1}\mathbf{1}_m, \tag{1.9}$$

where $A \equiv \mu^\top \Sigma^{-1}\mu$, $D \equiv \mathbf{1}_m^\top \Sigma^{-1}\mu$ and $C \equiv \mathbf{1}_m^\top \Sigma^{-1}\mathbf{1}_m$ and $\mu_p(v_p)$ is the expected portfolio return, defined as in eq. (1.6). In appendix 1.A, we also show that

$$v_p^2 = \frac{1}{C}\left[1 + \frac{1}{AC - D^2}\left(\mu_p(v_p)C - D\right)^2\right]. \tag{1.10}$$

Based on eq. (1.10), we define the *global minimum variance (GMV) portfolio* as the portfolio that achieves a variance equal to $v_{\text{gmv}}^2 \equiv C^{-1}$ and an expected return equal to $\mu_{\text{gmv}} \equiv D/C$. We shall return to this portfolio below (see section 1.2.5).

Note that for each v_p, there are two values of $\mu_p(v_p)$ that solve eq. (1.10). The optimal choice for our investor is that with the highest μ_p. We define the *efficient portfolio frontier* as the set of values (v_p, μ_p) that solve eq. (1.10) with the highest μ_p. It has the following expression:

$$\mu_p(v_p) = \frac{D}{C} + \frac{1}{C}\sqrt{\left(Cv_p^2 - 1\right)\left(AC - D^2\right)}. \tag{1.11}$$

The efficient portfolio frontier is increasing and concave in risk, v_p. It can be interpreted as a "production function"—that is, a technology through which expected returns are obtained using varying "levels of risk" as inputs (see, e.g., the two-asset case depicted in figure 1.1). Which portfolio on this frontier is selected by an investor depends on the investor's attitudes toward risk. We shall return to this selection problem in section 1.3.

1.2.4 Risk-Return Trade-Offs in the Two-Asset Case

Consider a simple example in which there are only two risky assets, $m = 2$. This simplified version of the model illustrates some crucial points regarding diversification very clearly and is also at the basis of famous contributions going beyond financial economics, notably in macroeconomics (see section 1.3.5.4). In this two-asset case, the efficient portfolio frontier is the risk-return trade-off resulting from any feasible combination of the two assets and does not require any optimization, as above: the budget constraint, $\frac{\pi_1}{w} + \frac{\pi_2}{w} = 1$, pins down a unique relation between the portfolio expected return and variance. Precisely, we have: $\mu_p = \frac{E\left[w'(\pi)\right] - w}{w} = \frac{\pi_1}{w}\mu_1 + \frac{\pi_2}{w}\mu_2$, or

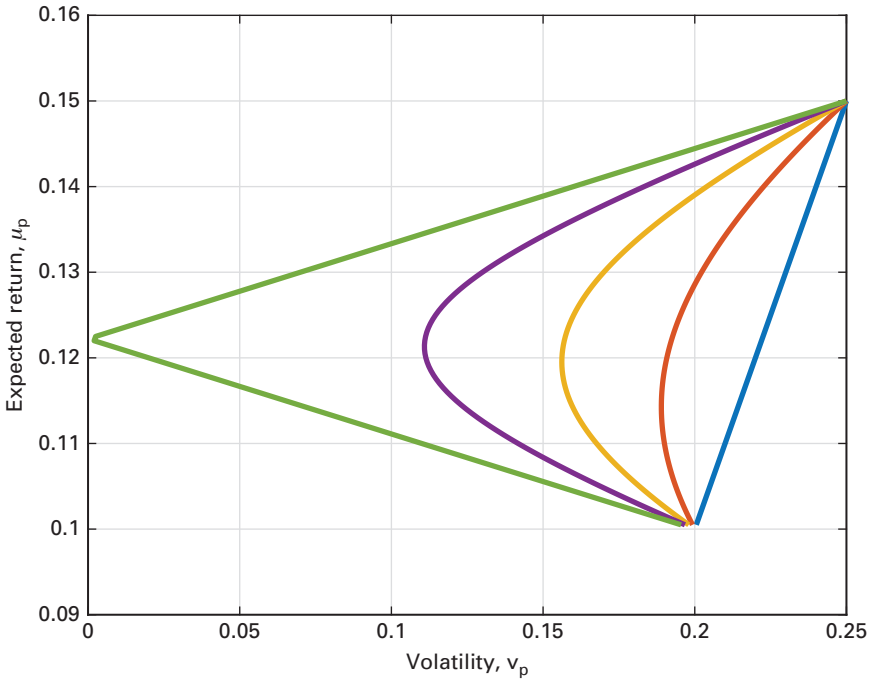

Figure 1.1
From top to bottom: portfolio frontiers corresponding to $\rho = -1, -0.5, 0, 0.5, 1$. Remaining parameter values are $\mu_1 = 0.10, \mu_2 = 0.15, \sigma_1 = 0.20, \sigma_2 = 0.25$. For each portfolio frontier, the efficient portfolio frontier includes those portfolios that yield the highest expected return for a given level of volatility.

$$\begin{cases} \mu_p = \mu_1 + (\mu_2 - \mu_1)\dfrac{\pi_2}{w} \\ v_p^2 = \left(1 - \dfrac{\pi_2}{w}\right)^2 \sigma_1^2 + 2\left(1 - \dfrac{\pi_2}{w}\right)\dfrac{\pi_2}{w}\sigma_{12} + \left(\dfrac{\pi_2}{w}\right)^2 \sigma_2^2 \end{cases}$$

whence

$$v_p = \frac{1}{\mu_2 - \mu_1}\sqrt{(\mu_2 - \mu_p)^2 \sigma_1^2 + 2(\mu_2 - \mu_p)(\mu_p - \mu_1)\rho\sigma_1\sigma_2 + (\mu_p - \mu_1)^2 \sigma_2^2}.$$

Note that when the returns on the two assets are perfectly correlated, $\rho = 1$, we have

$$\mu_p = \mu_1 + \frac{\mu_2 - \mu_1}{\sigma_2 - \sigma_1}(v_p - \sigma_1).$$

That is, the risk-return trade-off is linear. However, diversification pays as soon as the asset returns are not perfectly positively correlated: for any given expected return, the portfolio return volatility decreases as the asset correlation decreases. Figure 1.1 actually reveals that there are portfolios that are even less risky than the less risky asset and that risk can actually be zeroed in the extreme case where $\rho = -1$.

Naturally, in these examples, the portfolio frontier is given, in that the parameter values are given. However, these values should be determined in equilibrium. For example, the portfolio with zero volatility seems to offer investors too good opportunities: it yields

positive expected returns with no risk. Can this be sustainable? Many investors would spot this opportunity and implement a trade that would make this opportunity disappear. Let us illustrate. If $\rho = -1$, the portfolio that achieves zero volatility has $\frac{\pi_2}{w} = \frac{\sigma_1}{\sigma_1+\sigma_2}$, and the expected return is $\mu_{o,p} \equiv \frac{\sigma_2}{\sigma_1+\sigma_2}\mu_1 + \frac{\sigma_1}{\sigma_1+\sigma_2}\mu_2$. But as more investors go long to exploit this portfolio's nice opportunity, the two asset prices go up, and their returns decrease. This process eventually stops once μ_1 and μ_2 adjust in a way such that $\mu_{o,p} = r$, the risk-free rate.

Naturally, it is a simple example, which relies on hypothetical price adjustments. However, this example illustrates the type of standard restrictions arising in financial economics, which make prices economically viable or free from arbitrage opportunities, as we shall explain in many junctures of this book. For example, in the context of this chapter, eq. (1.15) given below is the restriction that generalizes the previous reasoning to the case in which assets returns are not perfectly and negatively correlated.

1.2.5 Risk Parity and the Global Minimum Variance Portfolio

Note that the portfolio in eq. (1.9) can be decomposed into two components as follows:

$$\frac{\hat{\pi}\left(v_p\right)}{w} = \ell\left(v_p\right)\frac{\pi_d}{w} + \left[1-\ell\left(v_p\right)\right]\frac{\pi_{\text{gmv}}}{w}, \quad \ell\left(v_p\right) \equiv \frac{D\left(\mu_p\left(v_p\right)C - D\right)}{AC - D^2}, \tag{1.12}$$

where

$$\frac{\pi_d}{w} \equiv \frac{\Sigma^{-1}\mu}{D}, \qquad \frac{\pi_{\text{gmv}}}{w} \equiv \frac{\Sigma^{-1}\mathbf{1}_m}{C}. \tag{1.13}$$

Therefore, any portfolio on the efficient portfolio frontier can be constructed by choosing a convex combination of two portfolios, π_d and π_{gmv}, with weights equal to $\ell\left(v_p\right)$ and $1-\ell\left(v_p\right)$. It is a *mutual fund (or separation) theorem*.[1] We shall use this representation of efficient portfolios to derive the zero-beta CAPM of section 1.3.2.

The two portfolios in (1.13) may be described as follows. The first, π_d, achieves the expected return that makes $\ell\left(v_p\right) = 1$, that is, $(v_p, \mu_p) = \left(\frac{\sqrt{A}}{D}, \frac{A}{D}\right)$. The second portfolio, π_{gmv}, is the GBV portfolio introduced in section 1.2.3, for we know from eq. (1.10) that the minimum variance occurs at $(v_p, \mu_p) = (\sqrt{1/C}, D/C)$, that is, when $\ell\left(v_p\right) = 0$.[2] This portfolio displays a very peculiar and interesting property: it requires an asset composition that is independent of the market expectations on asset returns, μ. It is a remarkable property: in many instances of this book, portfolio weights are instead proportional to the assets' expected excess returns (see, e.g., eq. (1.18) or the extensions to its dynamic counterparts in chapter 4 [section 4.6]). It is then quite peculiar that such a portfolio does not carry any information on the expected returns.

To facilitate the interpretation of this portfolio, assume that all asset returns are uncorrelated, in which case π_{gmv} in (1.13) collapses to a vector, where each component is

1. But then any efficient portfolio can be constructed through a convex combination of any two arbitrary but distinct efficient portfolios (see appendix 1.A).

2. The covariance of the global minimum variance portfolio with any other portfolio is always C^{-1} (see appendix 1.A).

given by

$$\frac{\pi_{\text{gmv},i}}{w} \equiv \frac{1/\sigma_i^2}{\sum_{j=1}^m 1/\sigma_j^2}, \quad i = 1, \cdots, m. \tag{1.14}$$

This portfolio expresses the rule to overweight assets that display lower volatility. It is the simplest example of a *risk parity* investment strategy, by which the risk contribution is the same for different assets: eq. (1.14) is an equally weighted portfolio in terms of risk and thus implicitly assumes that expected returns are the same for all assets. More advanced parts of this book (see, e.g., chapter 9 [section 9.11.5] or chapter 11 [section 11.10] explain that risk parity practice may feed endogenous risk: agents invested into such a strategy tend to sell when volatility increases, determining more volatility over a vicious circle. Roncalli (2014) provides an introduction to risk parity practice.

Note, finally, that if variances are all the same, $\sigma_i^2 = \bar{\sigma}^2$ for some $\bar{\sigma}^2$, then, the global minimum variance portfolio in (1.14) collapses to the famous one-over-N rule,

$$\frac{\pi_{\text{gmv},i}}{w} = \frac{1}{m},$$

where each asset is equally weighted and N stands for m in the notation of this chapter.

Implicit in these rules may be the acknowledgment of the idea that expected returns (and volatilities in the case of the one-over-N rule) are so difficult to calibrate that one may give up taking any assumptions on them in the first place. Indeed, it is well known that the actual performance of the mean-variance model is very sensitive to the estimates of the expected returns. Section 1.3.6 reviews a Bayesian approach to this problem, which enables a model user to incorporate their own views on market returns. Instead, section 1.3.7 proposes explanations for why investors would insist in using global minimum variance portfolios in a world of Knightian uncertainty where uncertainty reigns regarding the exact value taken by the assets' expected returns.

1.2.6 The Market Portfolio

The *market portfolio* is the portfolio at which the CML in eq. (1.6) and the efficient portfolio frontier in eq. (1.11) intersect. In fact, the market portfolio is the point at which the CML is *tangent* to the efficient portfolio frontier. For this reason, the market portfolio is also referred to as the "tangent portfolio." In figure 1.2, the market portfolio is at point M and has volatility equal to v_M and expected return equal to μ_M. At this point, the CML is tangent to the efficient portfolio frontier, AMC.[3] We now develop the reasons for referring to the tangent portfolio as the market portfolio.

1.2.6.1 Two-fund separation

Figure 1.2 illustrates that the CML dominates the efficient portfolio frontier AMC. Let us explain. On the one hand, the CML is the value of the investor's problem, [1.P1], obtained while investing in all the risky assets *and* the riskless asset. On the other hand, the efficient portfolio frontier is the value of the investor's problem, [1.P2], obtained while investing

3. The existence of the market portfolio requires a restriction on r, derived in eq. (1.15).

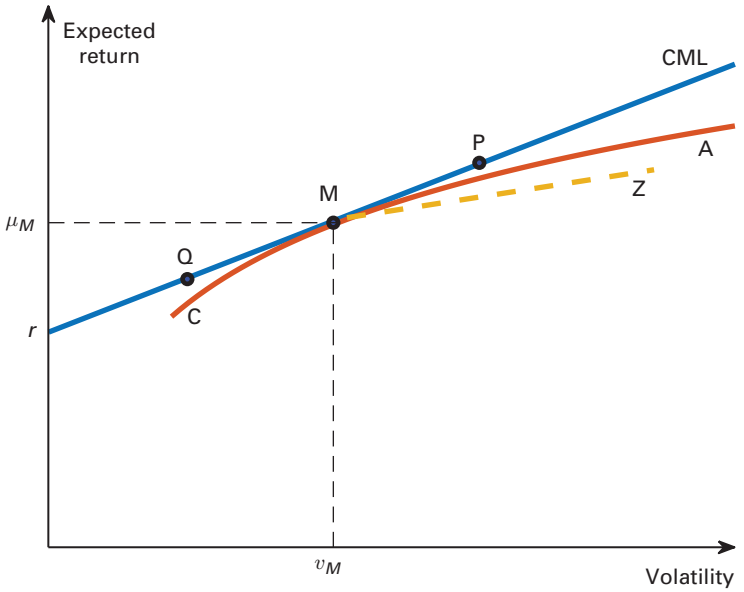

Figure 1.2
The capital market line and the market portfolio.

in the risky assets *only*.[4] Then, the CML and the efficient portfolio frontier can only be tangent to each other. If not, then there would exist a point on the efficient portfolio frontier that dominates some portfolio on the CML, and this is impossible. Likewise, the CML must have a portfolio in common with the efficient portfolio frontier—the portfolio that does not include the safe asset. We shall rely on these insights and provide an analytical characterization of the market portfolio.

So why do we term the market portfolio in this way? Figure 1.2 reveals that any portfolio on the CML can be obtained as a combination of the safe asset and the market portfolio M (a portfolio containing only the risky assets). Investors with high risk aversion would like to choose a portfolio such as Q, and investors with low risk aversion would perhaps pick a portfolio such as P. But no matter how risk averse an investor is, their optimal choice is a mix of the safe asset and the market portfolio M. Thus, the market portfolio allows an investor to calibrate their desired overall exposure to risk: they will invest those portions of their wealth in the market portfolio that makes their overall exposure to risk consistent with their risk appetite. It's a *two-fund separation theorem*. We shall make these statements more rigorous in the following sections.

This separation theorem has equilibrium implications that naturally lead us to refer the tangent portfolio to as "market portfolio." We know that any portfolio can be attained by investing in the portfolio M and by lending or borrowing funds in zero net supply. In

4. Figure 1.2 also depicts the dotted line MZ, which is the value of the investor's problem when they invest a proportion higher than 100% in the market portfolio, leveraged at an interest rate for borrowing higher than the interest rate for lending. In this case, the CML coincides with rM up to the point M. From M onward, the CML coincides with the highest between MZ and MA.

equilibrium, then, every investor must hold some proportions of M. But in aggregate, there is no net borrowing or lending, and all investors must then have portfolio holdings that sum up to the market portfolio, which is therefore the value-weighted portfolio of all the existing risky assets. This argument is formally developed in appendix 1.B.

Note, finally, that the market portfolio is a mere convex combination of all the existing risky assets. At first sight, it appears to be "preference-free" in that its construction is independent of the investors' risk attitudes. However, this proposition needs to be nuanced: below, we shall explain that the vector of expected returns μ still needs to be estimated and that models of risk and risk aversion might be needed to characterize μ.

1.2.6.2 Asset allocation puzzles

Do investors allocate wealth according to the previous predictions? Financial advisers are known to recommend that young investors hold more risky positions; they are also known to advise less conservative investors to increase their stock holdings compared with bonds. However, according to the theory in this section, the stocks-bonds mix in the market portfolio should be the same, regardless of the investors' attitude to risk (see the discussion around figure 1.2). It is an instance of asset allocation puzzles.

Campbell and Viceira (2002) provide an early synthesis of how these puzzles can be addressed through dedicated extensions of the framework in this section. These or additional analysts' recommendations could be understood while assuming that investors face more risks than those directly related to the price fluctuations of the assets in their portfolios. Some of these risks (not necessarily traded—e.g., labor income or random volatility) could indeed be hedged by investing more or less in risky positions than predicted by the simple theory in this section. Chapter 4 provides discussions of these cases while framing them into subsequently developed notions, according to which agents are assumed to have access to "stochastic opportunity sets."

1.2.6.3 Analytical characterization

We turn to characterize the market portfolio. We need to assume that the interest rate is sufficiently low to allow the CML to be tangent at the efficient portfolio frontier. The condition ensuring this outcome is that the return on the safe asset should be less than the expected return on the global minimum variance portfolio; namely,

$$r < \frac{D}{C} = \mu_{\text{gmv}}, \tag{1.15}$$

where μ_{gmv} is the global minimum variance portfolio identified through eq. (1.10). That is, we cannot find risky portfolios that yield less than the safe asset.

Next, let π_M be the market portfolio. To identify π_M, we note that it belongs to AMC if $\pi_M^\top \mathbf{1}_m = w$, where π_M also belongs to the CML and, therefore, by eq. (1.5), is such that

$$\frac{\pi_M}{w} = \frac{\Sigma^{-1}(\mu - \mathbf{1}_m r)}{\sqrt{\text{Sh}}} v_M. \tag{1.16}$$

Therefore, we search for the value v_M that solves

$$w = \mathbf{1}_m^\top \pi_M = \mathbf{1}_m^\top w \frac{\Sigma^{-1}(\mu - \mathbf{1}_m r)}{\sqrt{\text{Sh}}} v_M;$$

that is

$$v_M = \frac{\sqrt{\text{Sh}}}{D - Cr}. \tag{1.17}$$

Finally, we plug this value of v_M into the expression for π_M in eq. (1.16) and obtain

$$\frac{\pi_M}{w} = \frac{1}{D - Cr} \Sigma^{-1} (\mu - \mathbf{1}_m r). \tag{1.18}$$

Once again, the market portfolio belongs to the efficient portfolio frontier.[5] Indeed, on the one hand, the market portfolio cannot be above the efficient portfolio frontier as this would contradict the efficiency of the *AMC* curve (obtained by investing in risky assets only). On the other hand, by construction, the market portfolio belongs to the CML, and so it cannot be below the efficient portfolio frontier as the CML dominates the efficient portfolio frontier. Appendix 1.B makes this reasoning rigorous and shows that the market portfolio does indeed satisfy the tangency condition.

1.3 The Capital Asset Pricing Model

This section provides analysis that goes beyond the investors' efficient portfolio choices and derives the first asset evaluation formula appearing in the literature, based on the celebrated CAPM. First, we derive the CAPM, relying on arguments that have the same flavor as those in the original derivation of Sharpe (1964). Second, we derive the zero-beta CAPM of Black (1972), a model that does not require the existence of pure risk-free asset. Third, we explain how CAPM predictions relate to the behavior of risk averse investors. Finally, this section applies CAPM ideas and deals with such disparate topics as project evaluation or "speculative" demand of money.

1.3.1 Restrictions on Securities Expected Returns

Consider a portfolio comprising a proportion x of wealth invested in any asset i and the remaining proportion $1 - x$ invested in the market portfolio. That is, we consider a portfolio that is parametrized by x. The expected return and volatility of this portfolio are

$$\begin{cases} \tilde{\mu}_p \equiv x\mu_i + (1-x)\mu_M \\ \tilde{v}_p \equiv \sqrt{(1-x)^2\sigma_M^2 + 2(1-x)x\sigma_{iM} + x^2\sigma_i^2} \end{cases} \tag{1.19}$$

where we have defined $\sigma_M \equiv v_M$. Clearly, the market portfolio M is a special case of this portfolio. Relying on the examples of section 1.2.4, we know that the curve in (1.19) has the same shape as the curve *BMi* in figure 1.3. The curve *BMi* lies below the efficient portfolio frontier *AMC* because the latter results from optimizing a mean-variance criterion over all the existing assets, which then dominates any portfolio that only comprises the two assets i and M.

5. While the market portfolio depends on r, this portfolio does not obviously include any shares in the safe asset.

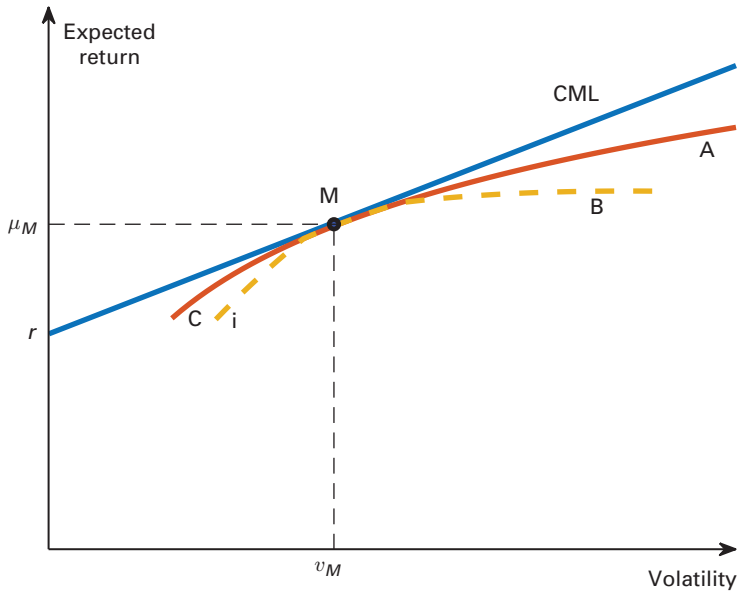

Figure 1.3
Construction of the capital asset pricing model.

For example, suppose that the *BMi* curve intersects the *AMC* curve. Then a feasible combination of assets in the *BMi* curve would dominate points on *AMC*, a contradiction, as *AMC* is the most efficient feasible combination of all existing assets. On the other hand, the *BMi* curve has a point in common with the *AMC* curve, which is *M* (the portfolio obtained with $x = 0$). Therefore, the curve *BMi* is tangent to the efficient portfolio frontier *AMC* at *M*, which is tangent to the CML at *M*, as we know.

That is, at *M*, the slopes of the *BMi* curve is the same as that of the efficient portfolio frontier *AMC*. This condition provides a restriction on the expected return μ_i on any asset *i* as we now show. Consider the *x*-parametrized curve (1.19). We determine its slope at *M* through the two slopes, $d\tilde{\mu}_p / dx$ and $d\tilde{v}_p / dx$, both evaluated at $x = 0$. We have

$$\frac{d\tilde{\mu}_p}{dx} = \mu_i - \mu_M, \quad \left.\frac{d\tilde{v}_p}{dx}\right|_{x=0} = -\frac{-(1-x)\sigma_M^2 + (1-2x)\sigma_{iM} + x\sigma_i^2|_{x=0}}{\tilde{v}_p|_{x=0}} = \frac{1}{\sigma_M}\left(\sigma_{iM} - \sigma_M^2\right).$$

Therefore,

$$\left.\frac{d\tilde{\mu}_p(x)}{d\tilde{v}_p(x)}\right|_{x=0} = \frac{\mu_i - \mu_M}{\frac{1}{\sigma_M}\left(\sigma_{iM} - \sigma_M^2\right)}. \tag{1.20}$$

On the other hand, the slope of the CML is $(\mu_M - r)/\sigma_M$, which, equated to the slope in eq. (1.20), yields

$$\mu_i - r = \beta_i (\mu_M - r), \quad \beta_i \equiv \frac{\sigma_{iM}}{v_M^2}, \quad i = 1, \cdots, m. \tag{1.21}$$

Eq. (1.21) is the celebrated *security market line* (SML) (see appendix 1.B for an alternative derivation). The assets for which $\beta_i > 1$ are usually referred to as "aggressive," and the

assets for which $\beta_i < 1$ are called "conservative."[6] The rationale is the following. Assuming a positive beta, the higher this beta, the more volatile (and risky) the asset and the higher its expected return. Assets where beta is negative return on average less than the safe interest because they provide diversification. Thus, according to the model, assets that command an expected return less than the safe interest rate can still be quite volatile, notably when their (negative) beta is sizable.

The SML can be interpreted as a projection of the excess return on asset i (i.e., $\tilde{\mu}_i - r$) on the excess returns on the market portfolio (i.e. $\tilde{\mu}_M - r$). In other words,

$$\tilde{\mu}_i - r = \alpha_i + \beta_i(\tilde{\mu}_M - r) + \varepsilon_i, \quad i = 1, \cdots, m, \tag{1.22}$$

where $\alpha_i = 0$ provided the model is correctly specified. Section 1.5 provides discussion regarding the empirical evidence on the CAPM.

Eq. (1.22) leads to the following decomposition of the volatility regarding the return in the ith asset:

$$\sigma_i^2 = \beta_i^2 v_M^2 + var\,(\varepsilon_i)\,, \quad i = 1, \cdots, m.$$

The quantity $\beta_i^2 v_M^2$ is known as *systematic* risk. The quantity $var\,(\varepsilon_i) \geq 0$ does, instead, capture the notion of *idiosyncratic* risk. In the next section, we shall show that idiosyncratic risk can be eliminated through a well-diversified portfolio—roughly, a portfolio that contains a large number of assets.

1.3.2 The Low-Beta Anomaly

Even though we discuss empirical properties of the CAPM in section 1.5, it is useful to discuss one of these properties now. Note that the SML line predicts that high-beta stocks should command higher returns than low beta (see eq. (1.21)). Yet historically, low-beta stocks have performed better on a risk-adjusted basis: stocks with lower beta tend to display a higher alpha (the intercept in eq. (1.22)) than stocks with higher beta. Why?

This anomaly has been known since at least Black, Jensen, and Scholes (1972). One explanation relies on market frictions. Consider an investor with a pronounced attitude toward risk-taking behavior. Theoretically, they might want to invest in a portfolio such as that indicated by point P in figure 1.2. This portfolio may be constructed through leverage: they should borrow money to invest more than 100% of their wealth into the market portfolio M. However, there may be constraints to leverage. To approximate the profit and loss (P&L) of a levered position, some investors may want to overweight their exposure to very risky assets; that is, those with high beta. High-beta stocks would display lower average returns as a result. Likewise, low-beta stocks may be underweighted and therefore might offer higher returns. There might also be investors able to borrow, and these investors

6. Interestingly, this notion of beta links to the following classical regression-based hedging context. Suppose we hold a perhaps highly illiquid asset that promises to deliver a return equal to $\tilde{z}$. To hedge against this stochastic return, go long a portfolio comprising a proportion π in the market portfolio and $1 - \pi$ in a safe asset. We minimize the variance of the overall position by choosing $\hat{\pi} = \arg\min_x var[\tilde{z} - ((1-\pi)\, r + \pi\, \tilde{\mu}_M)] = \beta_{\tilde{z}} \equiv cov(\tilde{z}, \tilde{\mu}_M)/v_M^2$. That is, the beta of the asset is the proportion of the market portfolio that we use to hedge the asset against systematic market movements.

would be less exposed to the "weighting biases" the constrained are subject to. However, investors able to borrow are likely to be subject to margin requirements. The tighter the constraints, the higher the low-high beta spread.

It seems natural to assess how robust these insights are when judged against market conditions. Frazzini and Pedersen (2014) construct a betting-against-beta (BAB) factor: a portfolio that is long low-beta stocks and short high-beta stocks and engineered in a way to be market-neutral—that is, to have zero beta. Theoretically, this portfolio has a positive expected return. Empirically, this portfolio produces substantial risk-adjusted returns. However, in times of distress, when liquidity is tight, this portfolio's performance deteriorates, just as the theory suggests: a market distress comes with higher margin requirements; therefore, more cash should be held by the investors who are able to borrow, making their asset holdings more similar to the investors unable to borrow. The price of high-beta (or resp., low-beta) stocks increases (resp., decreases) as a result.

The previous examples cover quite advanced material, but it is useful to point them out in this introductory chapter. The objective is not only to inform the reader about the importance of market frictions (the CAPM is based on abstract frictionless markets) but also to explain how apparently spectacular risk-adjusted returns might then need to be assessed against realistic market conditions, such as those resulting from a liquidity dry-up. Moreover, the BAB factor is an example of a possible determinant of the cross-section of expected returns. Section 1.5 explains that hundreds of additional successful factors have been discovered in the empirical literature. While we may be inclined to be more lenient on factors suggested by economic reasoning (as the BAB of this section), we may want to create stringent statistical criteria for the purpose of avoiding a proliferation of factors arising from pure statistical analysis. Section 1.5.5 reviews work aimed at addressing such data-mining concerns.

1.3.3 Zero-Beta CAPM

The CAPM predicts that the market portfolio is a benchmark for all the assets' required returns. We can actually use any benchmark on the efficient portfolio frontier and not just the market portfolio to gauge each asset's expected return. In particular, we pick an arbitrary portfolio on the efficient frontier and consider another portfolio that has returns uncorrelated with the initial arbitrary portfolio. The spread generated by these two portfolios becomes the evaluation benchmark. Because we only deal with portfolios that have no riskless assets, the procedure may be useful when the risk-free asset is not available for trading.

So, consider any portfolio on the efficient frontier, which can be generated by eq. (1.12). Given a fixed weight equal to l, define $v_{p,l} : \ell\left(v_{p,l}\right) = l$, and let $\hat{\pi}_l \equiv \hat{\pi}\left(v_{p,l}\right)$. The return generated by the portfolio $\frac{\hat{\pi}_l}{w}$ is obviously $\mu_l \equiv \frac{\hat{\pi}_l^\top \tilde{\mu}}{w}$. The vector of the covariances of all the asset returns with μ_l is

$$cov(\tilde{\mu}, \tilde{\mu}_l) = \Sigma \frac{\hat{\pi}_l}{w} = \frac{l}{D}\mu + \frac{1-l}{C}\mathbf{1}_m. \tag{1.23}$$

In particular, we have that, for any asset i,

$$cov(\tilde{\mu}_i, \tilde{\mu}_l) = \frac{l}{D}\mu_i + \frac{1-l}{C}, \tag{1.24}$$

and then

$$var(\tilde{\mu}_l) = cov(\tilde{\mu}_l, \tilde{\mu}_l) = \frac{l}{D}\mu_l + \frac{1-l}{C}. \tag{1.25}$$

Eqs. (1.24) and (1.25) can be solved for $\frac{l}{D}$ and $\frac{1-l}{C}$, and the solutions can be replaced back into eq. (1.23), leaving

$$\mu = \frac{\left(var(\tilde{\mu}_l)\mu_i - cov(\tilde{\mu}_i, \tilde{\mu}_l)\mu_l\right)\mathbf{1}_m + cov(\tilde{\mu}, \tilde{\mu}_l)\left(\mu_l - \mu_i\right)}{var(\tilde{\mu}_l) - cov(\tilde{\mu}_i, \tilde{\mu}_l)}.$$

The previous expression simplifies for assets that have returns uncorrelated with $\tilde{\mu}_l$, say, some asset $z_l : cov(\tilde{\mu}_{z_l}, \tilde{\mu}_l) = 0$. We now have

$$\mu = \mu_{z_l}\mathbf{1}_m + \frac{cov(\tilde{\mu}, \tilde{\mu}_l)}{var(\tilde{\mu}_l)}\left(\mu_l - \mu_{z_l}\right). \tag{1.26}$$

Eq. (1.26) is Black's (1972) zero-beta CAPM. We may express expected returns in terms of "benchmark" returns, just as with the SML of eq. (1.21), without however having to rely on the market portfolio and a riskless asset. The spread referred to as at the beginning of this section is $\mu_l - \mu_{z_l}$, where μ_{z_l} is the expected return on an asset that has zero correlation with the benchmark portfolio on the efficient frontier (whence, zero-beta). It goes without saying that eq. (1.26) does not rely on whether riskless assets are actually available for trading or not.

1.3.4 An Excursion into Risk Premiums and Certainty Equivalents

1.3.4.1 Project evaluation

The CAPM is a model that determines the required return for any asset and, thus, is the very first tool that we can use to evaluate any risky project, not necessarily a very liquid one. The main prediction of the model is that the market is a pricing factor as it is the only source of risk against which we assess the expected return on any other asset. Let $\tilde{C}$ denote the future, random cash flow promised by a given project. Its return is $\tilde{\mu}_C = \frac{\tilde{C}}{V} - 1$, where V is the current value of the project; and by the CAPM, the risk-adjusted discount rate for this project is

$$\mu_C \equiv E(\tilde{\mu}_C) = r + \beta_C\left(\mu_M - r\right) = r + \frac{1}{V}\frac{cov(\tilde{C}, \tilde{\mu}_M)}{\nu_M}\lambda, \tag{1.27}$$

where $\lambda \equiv \frac{\mu_M - r}{\nu_M}$ is the unit market risk premium and $\beta_C = cov(\tilde{\mu}_C, \tilde{\mu}_M)/\nu_M^2$ is the project beta. The value of the project is

$$V = \frac{E(\tilde{C})}{1 + \mu_C}. \tag{1.28}$$

Replacing the expression for μ_C in eq. (1.27) into eq. (1.28) and rearranging terms gives us

$$V = \frac{E(\tilde{C}) - \frac{\lambda}{\nu_M}cov(\tilde{C}, \tilde{\mu}_M)}{1 + r}. \tag{1.29}$$

The project value is the discounted expectation of the random cash flow, $\tilde{C}$, "corrected for risk," in the following sense. According to the CAPM, the market is the only source

of risk relevant for pricing, as explained. If the project cash flows are positively correlated with the market, investing in the project does not insulate the investments from this source of risk: the higher this correlation, the more aggressive the project is, just as we said in previous sections. But the more aggressive the project, the higher its expected return and, hence, the lower the project value. The risk-adjustment term, $\frac{\lambda}{v_M} cov(\tilde{C}, \tilde{x}_M)$, determines how low the project value has to be to induce an investor to undertake this project. This is the *risk premium* required to invest.

Were the project cash flows uncorrelated with the market (or were the market return the same as the riskless asset, $\lambda = 0$), eq. (1.29) would collapse to an actuarially fair evaluation formula, by which the project value is simply the discounted expectation of the future cash flows. Quite naturally, these statistical correlations occur under the probability laws generating returns and cash flows.

Is there any probability law such that the project value is the same as in eq. (1.29), but with the numerator replaced by the cash flow expectation under these hypothetical new laws and without any risk adjustments? The answer is in the affirmative. For obvious reasons, these laws are known as "risk neutral." They constitute indeed one fundamental pillar of asset evaluation, which this book heavily relies on.[7] Section 1.4 provides an introductory discussion to (and the next chapters provide a systematic treatment of) these foundational issues.

1.3.4.2 With utility functions

Consider an individual, who is considering entering into a gamble that yields a random payoff $\tilde{x}$ and costs V, with an initial wealth equal to w. We can think of this gamble as the investment decision in the previous section (*with* $\tilde{x} = \tilde{C}$) and set $r = 0$ to simplify the analysis. We assume the investor will enjoy a random utility $u(w - V + \tilde{x})$ from the project, where the function u is increasing and concave. The property by which the utility function is concave is tantamount to the assumption that our decision maker is risk averse. That is, the investor would prefer receiving the expected outcome of the project rather than exposing themselves to gambling for their utility. Formally, by Jensen's inequality, $E(u(w - V + \tilde{x})) \leq u(w - V + E(\tilde{x}))$.

We ask: What is the value V that induces the investor to enter into this project? Clearly, it cannot be higher than the solution V to

$$u(w) = E(u(w - V + \tilde{x})), \tag{1.30}$$

for otherwise the investor would be better off while staying away from the gamble (and experiencing utility equal to $u(w)$). We also assume that, should the project value be less than V, many additional investors with the same attitude for risk would step in and make this value increase to the extent that eq. (1.30) holds.

---

7. Informally, the project value can be expressed as $V = \frac{\mathbb{E}(\tilde{C})}{1+r}$, where $\mathbb{E}(\cdot)$ denotes the expectation under this risk neutral probability, say, Q. For the CAPM, the relation between Q and the original, natural probability, say P, is given by the so-called Radon-Nikodym derivative $\frac{dQ}{dP} = 1 - \frac{\lambda}{v_M}(\tilde{\mu}_M - \mu_M)$ (see section 1.4.2 for introductory details).

Next, define the demeaned payoff $\epsilon \equiv \tilde{x} - E(\tilde{x})$. We study the investor's attitude vis-à-vis risk in relation to the fair gamble, that is, the lottery that has payoff equal to ϵ. We have

$$u(w) = E(u(w + \Lambda + \epsilon)) \leq u(w + \Lambda), \tag{1.31}$$

where we have defined $\Lambda \equiv E(\tilde{x}) - V$. The first equality is simply eq. (1.30) (with the change in notation), and the inequality is Jensen's inequality again. Thus, Λ is positive. That is, due to risk aversion, the investor is willing to pay less than the expected cash flow to enter the project. For obvious reasons, we refer to Λ as the risk premium required to invest in the project.

Suppose, for example, that a representative investor has exponential utility, $u(w) = -e^{-\tau w}$, and that the cash flow is normally distributed, $\tilde{C} \sim N(E(\tilde{C}), \sigma_\epsilon^2)$, such that by eq. (1.30),

$$V = E(\tilde{C}) - \frac{1}{2}\tau\sigma_\epsilon^2. \tag{1.32}$$

The project value is at discount compared with the expected cash flow by a factor equal to $\Lambda = \frac{1}{2}\tau\sigma_\epsilon^2$, the risk premium. Note that this risk premium relates to the investor's risk aversion coefficient, τ: the higher τ, the higher the discount. We shall elaborate on this risk aversion coefficient below.

These conclusions are the utility counterparts to those of the CAPM in the previous section. According to the CAPM, the risk premium relates to the project exposure to market movements (see eq. (1.29)). In the model of this section, the risk premium arises to compensate a risk averse investor for the risk they would undertake in this specific project (see eq. (1.32)). Granted, in equilibrium, market movements may well be related to the investors' attitude vis-à-vis risk. However, the link to risk aversion is direct in this model. Section 1.3.5 clarifies how asset prices link to risk aversion in an equilibrium model—one in which agents choose amongst several alternatives, as in the CAPM.

Apart from exceptions (e.g., the previous Gaussian case), no closed-form solutions are available for V. We can, however, approximate Λ when the risks regarding the fair lottery are "small"; that is, when the variance of ϵ, σ_ϵ^2, say, is small. Appendix 1.C provides the following approximation to Λ in this small risks case:

$$\Lambda \approx \hat{\Lambda} \equiv -\frac{1}{2}\frac{u''(w)}{u'(w)}\sigma_\epsilon^2. \tag{1.33}$$

For example, if the representative investor has exponential utility, then $\hat{\Lambda} = \frac{1}{2}\tau\sigma_\epsilon^2$, which is the exact risk premium in the previous Gaussian case (see eq. (1.32)). Appendix 1.C. provides one additional example of evaluation models, in which no closed-form solutions are available for V, and discusses the quality of the approximations to V based on eq. (1.33).

1.3.4.3 With utility functions II: Certainty equivalents

An alternative albeit closely related notion of risk premium relies on the following reasoning. Assume that an individual *is entering* into a gamble—that is, they have no choice but to experience a random perturbation to their wealth, leaving them with $w + \tilde{x}$ dollars.[8]

8. As an example, think of these wealth fluctuations as those arising during the business cycles.

To illustrate, suppose that the individual expects a loss from this lottery, $E(\tilde{x}) < 0$. We ask, how much of their wealth would they be willing to give up if they were hypothetically asked to *avoid* the gamble?

If the individual were risk neutral, they would be glad to give up the expected loss. But if they are risk averse, they would be willing to give up more than the expected loss, say $-E(\tilde{x}) + \Pi$, for some positive Π interpreted as a risk premium. In other words, the individual would be glad to accept a *sure* reduction of their wealth to $w + E(\tilde{x}) - \Pi$, in exchange for avoiding to gamble,

$$u(w + E(\tilde{x}) - \Pi) = E(u(w + E(\tilde{x}) + \epsilon)) \leq u(w + E(\tilde{x})), \tag{1.34}$$

where the inequality is Jensen's inequality and shows that Π is positive. We refer to $\text{CE}(w, \tilde{x}) \equiv E(\tilde{x}) - \Pi$ as the *certainty equivalent* for $\tilde{x}$ for an initial wealth equal to w. Naturally, the sign of $E(\tilde{x})$ can be anything.

The risk premium Π has an "insurance flavor." It relates to an economic premium that the individual is willing to pay to avoid a loss rather than to an economic incentive to undertake a risky project. However, it closely relates to the risk premium Λ defined in eq. (1.31). Consider, for example, the Gaussian project case in the previous subsection: if $\tilde{C}$ is Gaussian and the utility is exponential with risk aversion τ, the risk premium in this section is $\Pi = \frac{1}{2}\tau\sigma_\epsilon^2$, that is, the same as Λ in the previous section (see eq. (1.32)). Further, and just as with the approximation in eq. (1.33), the risk premium Π relates to the curvature of the utility function u, in that, in the small risks case, it can be approximated as follows (see appendix 1.C):

$$\Pi \approx \hat{\Pi} \equiv -\frac{1}{2}\frac{u''(w + E(\tilde{x}))}{u'(w + E(\tilde{x}))}\sigma_\epsilon^2. \tag{1.35}$$

Note that $\hat{\Pi}$ and $\hat{\Lambda}$ are the same as soon as the initial lottery is fair, $E(\tilde{x}) = 0$ (see eq. (1.33)). We now discuss alternative measures of risk premiums.

1.3.4.4 Constant and relative risk aversion

The notions of risk premiums investigated so far relate to *dollar* incentives to undertake (or avoid) risks. The expressions in eqs. (1.33) and (1.35) reveal that they are proportional to these risks (i.e., σ_ϵ^2) and that they increase with the "normalized curvature" of the utility function, $A(\cdot) \equiv -u''(\cdot)/u'(\cdot)$, a function we refer to as *absolute risk aversion*. The numerator of this function naturally measures how far the investor is from being risk neutral. The denominator of A makes this notion invariant to any positive and affine transformation of the original utility function: if two risk-averse investors have utilities u_1 and u_2, they have the same absolute risk aversion if and only if $u_1 = \varphi(u_2)$ and φ is an affine function. The exponential utility function hypothesized in the previous section is one for which absolute risk aversion is constant and equal to τ; it is known as constant absolute risk aversion (CARA) utility for this reason.

These notions can be extended to cover *relative* incentives to undertake (or avoid) risks. Consider, for example, the insurance risk premium of the previous section—one, however, regarding a multiplicative risk, that is, a lottery with an outcome equal to $(w + E(\tilde{x}))(1 + \tilde{g})$ for some zero-mean random variable $\tilde{g}$, not $w + E(\tilde{x}) + \epsilon$ (an additive risk). The question we ask now is, what *proportion* of their wealth would the investor be

willing to give up if hypothetically asked to avoid this lottery? The indifference condition in (1.34) is now replaced with

$$u\left(w+E(\tilde{x})\left(1-\Pi_r\right)\right)=E\left(u\left(\left(w+E(\tilde{x})\right)\left(1+\tilde{g}\right)\right)\right)\leq u\left(w+E(\tilde{x})\right), \tag{1.36}$$

where Π_r is a constant, relative (or percentage) risk premium. The previous inequality reveals that Π_r is positive. In appendix 1.C, we show that this relative risk premium can approximated as follows:

$$\Pi_r \approx \hat{\Pi}_r \equiv -\frac{1}{2}\frac{u''\left(w+E(\tilde{x})\right)\left(w+E(\tilde{x})\right)}{u'(w+E(\tilde{x}))}\sigma_g^2, \tag{1.37}$$

where σ_g^2 denotes the variance of $\tilde{g}$.

We refer the function $R_r(x)\equiv -u''(x)\,x/u'(x)$ to as *relative risk aversion*. Consider the utility function $u(x)=\frac{x^{1-\gamma}}{1-\gamma}$ for some constant $\gamma\neq 1$, or $u(x)=\ln x$ when $\gamma=1$. In this case, relative risk aversion is constant and equal to γ; it is known as constant relative risk aversion (CRRA) utility for this reason. Naturally, a CARA investor has increasing relative risk aversion. The implication of this property is that an investor keeps the same amount of dollars invested in risky assets as their wealth increases. However, the relative proportion of risky assets in their portfolio decreases with their wealth, as explained next. These notions of risk aversion are known since at least Pratt (1964) and Arrow (1965).

1.3.5 Back to CAPM: Equilibrium with Expected Utility

We explained that the investors' risk aversion determines the proportion of the market portfolio to hold and achieve any desired point of the risk-return trade-off. We formalize these explanations. We assume that our investors are risk averse: they prefer higher expected returns than lower, but for any expected portfolio return, they like portfolios that provide less volatility. We assume that investors are all the same and that the representative investor chooses portfolio holdings that maximize their utility derived from the expected returns and variance, namely,

$$\pi^*=\arg\max_{\pi} U\left(E\left[w'(\pi)\right],var\left[w'(\pi)\right]\right), \tag{1.P3}$$

where $U(\cdot,\cdot)$ is a concave utility function satisfying $U_1(\cdot,\cdot)>0$ and $U_2(\cdot,\cdot)<0$, subscripts denote partial derivatives, and the expressions for the portfolio expected return and variance are given in eq. (1.3).

Rearranging the first-order conditions to problem [1.P3] leads to the following portfolio choice:

$$\pi^*=\tau_*^{-1}\Sigma^{-1}\left(\mu-\mathbf{1}_m r\right),\quad \tau_*\equiv\frac{-2U_2^*}{U_1^*}, \tag{1.38}$$

where $U_j^*\equiv U_j\left(E\left[w'(\pi^*)\right],var\left[w'(\pi^*)\right]\right)$ and subscripts denote partial derivatives. That is, the investor's choice is a proportion of the market portfolio (see eq. (1.18)). Moreover, the more risk averse the investor is (i.e., the higher τ_*), the lower the proportion they desire to hold of the market portfolio.

The investor wishes to lend or to borrow according to whether $w-\mathbf{1}_m^\top\pi^*$ is positive or negative. If the riskless security is in zero net supply, then, in equilibrium,

$w - \mathbf{1}_m^\top \pi^* = 0$—that is, the representative investor holds exactly the market portfolio, which is the value-weighted portfolio, $\pi_i^* = \bar{\theta}_i S_i$, where $\bar{\theta}_i$ is the number of the ith outstanding securities. Note, indeed, that if the riskless security is in zero net supply,

$$w = \mathbf{1}_m^\top \pi^* = \tau_*^{-1} \mathbf{1}_m^\top \Sigma^{-1} (\mu - \mathbf{1}_m r) = \tau_*^{-1} (D - Cr) . \tag{1.39}$$

Replacing the previous relation into eq. (1.38) gives exactly the market portfolio, $\frac{\pi^*}{w} = \frac{\pi_M}{w}$, where π_M is as in eq. (1.18). Note that eq. (1.39) places a restriction on the safe interest rate, a restriction which we shall return below. Further, chapter 4 provides many additional instances of restrictions on the interest rate that link to investors' preferences (see, e.g., eq. (4.98)).

1.3.5.1 Equilibrium expected returns

What are the equilibrium expected returns? By eq. (1.38), we have that

$$\mu - \mathbf{1}_m r = \tau_* w \cdot \Sigma \frac{\pi^*}{w} = \tau_* w \cdot cov(\tilde{\mu}, \tilde{\mu}_M), \tag{1.40}$$

where $\tilde{\mu}_M = \tilde{\mu}^\top \frac{\pi^*}{w}$ denotes the market return, and the second equality follows by results in appendix 1.B (see eq. (1.A12)). Moreover, note that, in equilibrium,

$$\mu_M = \frac{\pi^{*\top}}{w} \mu = r + \tau_* w \cdot \frac{\pi^{*\top}}{w} \Sigma \frac{\pi^*}{w} = r + \tau_* w \cdot v_M^2, \tag{1.41}$$

where the second equality follows by replacing μ from eq. (1.40). Eqs. (1.40) and (1.41) do obviously lead to the CAPM in eq. (1.21),

$$\mu - \mathbf{1}_m r = \frac{cov(\tilde{\mu}, \tilde{\mu}_M)}{v_M^2} (\mu_M - r) ,$$

but they now also lead to a determination of the overall system of the expected excess returns: by eqs. (1.40) and (1.41), the expected excess returns on all assets (including the market portfolio) link to the representative agent's attitudes vis-à-vis risk.

1.3.5.2 A parametric example: Exponential utility

A simple example helps to illustrate the previous conclusions. Suppose the investor maximizes their expected utility of wealth, taken to be negative exponential, $u(w) = -e^{-\tau w}$, where τ is the absolute risk aversion. Assume that the asset returns are normally distributed, such that

$$\pi^* = \arg\max_\pi E\left[u\left(w'(\pi)\right)\right] = \arg\max_\pi \left(E\left[w'(\pi)\right] - \frac{\tau}{2} var\left[w'(\pi)\right]\right) . \tag{1.42}$$

The solution to this problem is the same as that in eqs. (1.38), where τ^* collapses to the constant τ.[9] In equilibrium, the safe interest rate satisfies eq. (1.39), such that now the investor's equilibrium asset holdings coincide with the market portfolio.

9. Note that the investor desires to hold relative proportions of the risky assets (π^*/w) that are decreasing in w. This property follows by the assumption that this investor has constant absolute risk aversion.

We may elaborate on these results with a further simplification of this example, one with two risky assets that have uncorrelated payoffs. Assume also that the asset return volatilities are entirely driven by the asset payoffs, in that $\sigma_i^2 S_i^2 = \sigma_{x_i}^2$, where $\sigma_{x_i}^2$ is the variance of the ith asset payoff. It is easy to see that each asset demand is $\pi_i^* = (\tau\sigma_i^2)^{-1}(\mu_i - r) = \bar{\theta}_i S_i$, where the second equality is the equilibrium condition and, by the definition of μ_i, the equilibrium price is then

$$S_i = S_0\left(E(\tilde{x}_i) - \tau\bar{\theta}_i\sigma_{x_i}^2\right), \tag{1.43}$$

where we have set $x_0 \equiv 1$, such that $S_0 \equiv (1+r)^{-1}$. We assume that these prices are strictly positive.

To determine the equilibrium riskless rate r (or, equivalently, S_0), we use the zero net supply condition for the safe asset, that is, we aggregate the value of the asset holdings and set the result equal to the initial wealth, $\bar{\theta}_1 S_1 + \bar{\theta}_2 S_2 = w$, leaving us with

$$S_0 = \frac{w}{\bar{\theta}_1 E(\tilde{x}_1) + \bar{\theta}_2 E(\tilde{x}_2) - \tau(\bar{\theta}_1^2\sigma_{x_1}^2 + \bar{\theta}_2^2\sigma_{x_2}^2)}. \tag{1.44}$$

The single risky asset case is equally instructive. Assume that the second asset is not available for trading. In this case, the wealth of the representative investor equals $w = \bar{\theta}_1 S_1$, where S_1 is as in eq. (1.43) and the price of the safe asset is given by eq. (1.44), with $\bar{\theta}_2 \equiv 0$. That is, the equilibrium riskless rate simply forces the value of the risky asset holdings to equal initial wealth, $S_0 : \bar{\theta}_1 S_1 = w$.

1.3.5.3 Mean-variance utility

Next, suppose that, given a coefficient of variance aversion equal to τ, the agent maximizes a mean-variance criterion applying to portfolio *returns*, not to dollar positions, as in the case of the exponential utility in the previous section (see eq. (1.42)). That is, the investor chooses the following portfolio:

$$\pi_{\text{mv}} = \arg\max_{\pi}\left(E\left[w'(\pi)/w\right] - \frac{\tau}{2}var\left[w'(\pi)/w\right]\right) = \frac{w}{\tau}\Sigma^{-1}(\mu - \mathbf{1}_m r), \tag{1.45}$$

where the last equality follows by a simple calculation. This solution differs from that in the previous section because the relative *proportions* of wealth invested in risky assets are now independent of w. Of course, exponential utility and mean-variance criterions would lead to the same outcome if exponential utility was referenced to portfolio returns, that is, if $u(w') = -e^{-\tau w'/w}$, in which case,

$$\pi^* = \arg\max_{\pi}\left(E\left[u\left(w'(\pi)/w\right)\right]\right) = \pi_{\text{mv}}, \tag{1.46}$$

where π_{mv} is as in eq. (1.45).

Note that, in these cases, and in terms of the notation in (1.38), $\tau_* = \frac{\tau}{w}$, such that, by eq. (1.39), the equilibrium interest rate satisfies

$$r : \tau = D - Cr. \tag{1.47}$$

Section 1.3.7 relies on eq. (1.47) while discussing some foundational issues regarding risk parity portfolio allocations in equilibrium.

1.3.5.4 *Tobin's reinterpretation of Keynesian money demand*

One of the earliest applications of the ideas of portfolio selection occurred in monetary economics. In 1958, Tobin provided a new framework to think about money demand, aimed to provide microeconomic foundations to some traits of the monetary theory of Keynes (1936). In this novel framework, money demand arises as the agent invests their savings while optimizing over their risk-returns trade-offs based on their risk aversion. Tobin's work would actually inspire the very same Sharpe's ideas, leading to the CAPM. Let us review the main points underlying Tobin's contribution.

Tobin explains that taken at face value, Keynes's explanations of money demand would imply that agents end up making dichotomic choices: they would hold either money or bonds. That is, at the individual level, each agent holds money or bonds based on their own expectations about future interest rate levels. Yet at the aggregate level, money demand does inversely relate to nominal interest rates while being flat in correspondence of small rate values—a liquidity trap. Appendix 1.D provides a parametric example that clarifies the details of how these mechanisms operate.

Tobin formulates a theory of money demand in which agents do not make previous dichotomic choices. Consider the following specification of the examples in section 1.2.4. We interpret the safe asset as money, which is therefore such that its return and volatility are $\mu_1 \equiv 0$ and $\sigma_1 \equiv 0$; instead, bonds are risky, in that they provide a superior return but at the cost of some volatility. Therefore, the expected return and volatility of a portfolio comprising money and bonds are $\mu_p = \mu_2 \pi_2$ and $\nu_p = \pi_2 \sigma_2$, with straightforward notation, and initial wealth normalized to one.

An hypothetical representative agent optimizes a risk-return trade-off identical to that underlying eq. (1.42). Thus, their money demand is

$$1 - \pi_2 = 1 - \frac{\mu_2}{\tau \sigma_2^2}$$

and obviously decreases with the bond expected return. This simple model leads to some predictions regarding the effects of a decreased interest rate volatility. Suppose, for example, that a central bank has the power to lower both interest rates *and* interest rate volatility in such a way to keep the ratio μ_2 / σ_2 unchanged. Money demand decreases as a result. More generally, money demand is decreasing with interest rate volatility, σ_2. According to this very simple model, low interest rates may be targeted by merely lowering interest rate volatility.

1.3.6 The Black-Litterman Model

The mean-variance approach to asset allocation obviously requires knowledge of the assets expected returns. However, it has always been clear that the predictions of the mean-variance model were very sensitive to the assumptions underlying the assets expected returns. The main issue is that expected returns are very difficult to estimate. Moreover, even a small estimation error has the potential to lead to sizeable uncertainty around the accuracy of the model portfolio recommendations. This section provides a succinct description of a simple Bayesian approach to addressing these problems, originally developed by Black and Litterman (1991). Its key element is to acknowledge that the assets' expected returns, μ, are *unknown* and need to be estimated while having regard not only to data but also to the subjective views that investors might have about them. Fundamentally, this

approach takes parameter uncertainty into account by allowing the model user to formulate their own *views*. We present succinct derivations of this approach. Meucci (2005) provides a comprehensive account of portfolio allocation in the presence of statistical risk.

1.3.6.1 Views

How do we estimate the assets' expected returns? A basic answer is to ask investors' opinions. Consider the following example. While the vector of expected returns, μ, is unknown, the model user may have views about some of its constituents. For example,

$$E(P\mu) = \mathcal{V}, \quad P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \quad \mathcal{V} = \begin{bmatrix} 6\% \\ 1\% \end{bmatrix},$$

where the expectation is taken because μ is not known anymore and P is a matrix (also known as a pick matrix), which enables one to incorporate their own views across the whole spectrum of the expected returns. In this example with six assets, the first asset is expected to return 6% (an absolute view), and the fifth asset is expected to outperform the sixth by 1% (a relative view). We now discuss how the investor should update the estimation of the expected returns based on their views.

1.3.6.2 Updating the estimates of the expected returns

We assume that the assets expected returns satisfy

$$\begin{cases} \mu &= \bar{\mu} + \epsilon_\mu, \quad \epsilon_\mu \sim N(0, C), \quad C = c\Sigma \\ P\mu &= \mathcal{V} + \epsilon_v \quad \epsilon_v \sim N(0, \Omega), \end{cases} \tag{1.48}$$

where the second equation describes the investors' views; Σ is the usual variance-covariance matrix of the asset returns; $\bar{\mu}$, P, C, and Ω are vector or matrices of constants; and finally, c is a constant that measures the degree of intolerance to the investors' own prespecified belief, $\bar{\mu}$: the higher c, the higher the intolerance. We assume that Σ is known by the investors. A possible parametrization of Ω is $\Omega = P\Sigma P^\intercal/\psi$, where the constant ψ is a measure of the investors' confidence in their views.

We wish to determine the conditional distribution of μ given $\mathcal{V}$, say $\mu|\mathcal{V}$, that is, how the estimation of the expected returns changes with the views. This distribution can be obtained by an application of Bayesian methods that are reviewed in more advanced parts of the book (see, e.g., appendix 10.A in chapter 10). Appendix 1.E contains details of the proof that

$$\mu|\mathcal{V} \sim N(\mu_{bl}, \Sigma_*), \tag{1.49}$$

where

$$\mu_{bl} = \Sigma_* \left((c\Sigma)^{-1} \bar{\mu} + P^\intercal \Omega^{-1} \mathcal{V} \right), \quad \Sigma_* = c \left(\Sigma^{-1} + cP^\intercal \Omega^{-1} P \right)^{-1}. \tag{1.50}$$

A possible choice for the prior $\bar{\mu}$ is that implied by the mean-variance portfolio, obtained by inverting (1.45) for the expected excess returns,

$$\bar{\mu} = \mathbf{1}_m r + \tau \Sigma \frac{\pi_{\text{mv}}}{w}. \tag{1.51}$$

That is, we take the CAPM as the prior and update the assets expected returns by incorporating the views. Note that the asset returns, $\tilde{R} \sim N\left(\mu_{bl}, \Sigma_{bl}\right)$, where $\Sigma_{bl} = \Sigma + \Sigma_*$. A mean-variance portfolio decision consistent with the investors' views is

$$\begin{aligned}\frac{\pi_{bl}}{w} &\equiv \frac{1}{\tau}\Sigma_{bl}^{-1}\left(\mu_{bl} - \mathbf{1}_m r\right) \\ &= \Sigma_{bl}^{-1} c^{-1} \Sigma_* \frac{\pi_{\text{mv}}}{w} + \frac{1}{\tau}\Sigma_{bl}^{-1}\left(\left(c^{-1}\Sigma_*\Sigma^{-1} - \mathbf{I}_m\right)\mathbf{1}_m r + \Sigma_* P^\intercal \Omega^{-1} \mathcal{V}\right),\end{aligned} \tag{1.52}$$

where $\mathbf{I}_m$ denotes the m-dimensional identity matrix, the second line is obtained by replacing the expression for the CAPM prior $\bar{\mu}$ in (1.51) into (1.50) and, finally, by rearranging terms. Eq. (1.52) shows that the Black-Litterman portfolio with CAPM prior expected returns is a combination of the mean-variance portfolio (which obtains as a limit for $c \to 0$) and a term that tracks the investors' views.

1.3.7 Knightian Uncertainty and Global Minimum Variance Portfolio

This section describes a model by which investors react to parameter uncertainty by making conservative decisions. Compared with the Black-Litterman analysis, investors acknowledge that they might never learn about the true distribution affecting asset returns and are averse to this lack of knowledge. This situation is known as one of "Knightian uncertainty." Chapter 9 analyzes models with Knightian uncertainty, in which agents are unable to assess the probability distribution of asset returns and fundamentals (see section 9.6). In some cases, they take robust decisions; that is, decisions based on worst-case scenarios. In terms of the problems of this chapter, an investor lacks knowledge regarding the assets' expected returns but takes decisions *as if* a malevolent Nature had chosen the worst outcome for them. For example, the worst outcome could be the lowest expected return when the investor goes long and the highest expected return when they go short. Section 9.6 explains that these problems were introduced in finance since at least Dow and Werlang (1992) and provides a list of references and surveys, including the seminal contribution of Gilboa and Schmeidler (1989) on "maxmin" preferences (i.e., optimizing behavior under worst-case scenarios). This section is a slight variation on Garlappi, Uppal, and Wang (2007). It shows that, in equilibrium, investors tend to tilt their portfolios toward the global minimum variance portfolio defined in section 1.2.5, thereby providing some economic foundations to risk parity.

Consider the following very simple case first. An investor (with initial wealth $w = 1$) would like to build up a portfolio according to a mean-variance criterion, but they are concerned about parameter estimation error. Specifically, suppose the investor knows the asset returns' volatilities, and given a sample size T, they have obtained estimates of the asset expected returns μ_i, say $\hat{\mu}_i$, $i, \cdots, m$. They then solve the following program

$$\max_{\pi} \min_{\mu \in M} \left(\pi^\top (\mu - \mathbf{1}_m r) + R - \frac{\tau}{2}\pi^\top \Sigma \pi\right), \tag{1.P4}$$

where, for some prespecified values ϵ_i, $M = M_1 \times \cdots \times M_m$ and $M_i = \left\{\mu_i : \left(\mu_i - \hat{\mu}_i\right)^2 \leq \epsilon_i \sigma_i^2 / T\right\}$. In words, the investor selects their portfolio weights while assigning the

worst-case outcome to each μ_i within some confidence band

$$\mu_i \in \left(\hat{\mu}_i - \tfrac{\sigma_i}{\sqrt{T}}\sqrt{\epsilon_i}, \hat{\mu}_i + \tfrac{\sigma_i}{\sqrt{T}}\sqrt{\epsilon_i}\right).$$

We shall soon return to explain the economic significance of this band while dealing with a similar decision problem (see (1.53) below). For now, note that the program [1.P4] can equivalently be cast as

$$\max_{\pi}\left(\pi^{\top}\left(\hat{\mu}_* - \mathbf{1}_m r\right) + R - \frac{\tau}{2}\pi^{\top}\Sigma\pi\right),$$

where each element of $\hat{\mu}_*$ is $\hat{\mu}_{*,i} = \hat{\mu}_i - \text{sign}\,(\pi_i)\,\tfrac{\sigma_i}{\sqrt{T}}\sqrt{\epsilon_i}$. That is, the investor acts so as to tilt the expected return toward the less favorable case (low when they buy, and high when they sell).

Next, consider a slightly different problem, where the set M in [1.P4] is, for a given constant η,

$$M = \left\{\mu : \left(\hat{\mu} - \mu\right)^{\top}\Sigma^{-1}\left(\hat{\mu} - \mu\right) \leq \eta\right\}, \tag{1.53}$$

a joint restriction on the admissible values for the unknown expected returns. The constant η may be interpreted in *cognitive terms*, in that the investor has information on the boundaries of their knowledge on the expected returns, μ, which is summarized by the set M. Another interpretation is that η is the agent's *aversion to parameter uncertainty*: the higher η, the more acute the worst-case scenarios they would implement while building up their portfolio choice.[10] Chapter 9 contains additional discussions on these topics (see section 9.6).

We assume that $\eta < \text{Sh}_*^2 \equiv \left(\hat{\mu} - \mathbf{1}_m r\right)^{\top}\Sigma^{-1}\left(\hat{\mu} - \mathbf{1}_m r\right)$. In appendix 1.E, we show that the solution to this problem is

$$\hat{\pi} = \arg\max_{\pi}\left(\pi^{\top}\left(\hat{\mu} - \mathbf{1}_m r\right) + R - \frac{\tau}{2}\pi^{\top}\Sigma\pi - \sqrt{\eta\pi^{\top}\Sigma\pi}\right) = \frac{1}{\tau}\left(1 - \Omega\left(\eta\right)\right)\Sigma^{-1}\left(\hat{\mu} - \mathbf{1}_m r\right), \tag{1.54}$$

where $\Omega\left(\eta\right) \equiv \sqrt{\frac{\eta}{\text{Sh}_*^2}}$. We determine the equilibrium interest rate. Note that, by the expression in (1.54) and the definitions of C and D in section 1.2.3, the equilibrium condition, $1 = w = \mathbf{1}_m^{\top}\hat{\pi}$, leads to

$$r = \hat{r}_{\text{mv}} - \frac{\tau}{C}\frac{\Omega\left(\eta\right)}{1 - \Omega\left(\eta\right)}, \quad \hat{r}_{\text{mv}} \equiv \frac{\hat{D} - \tau}{C}, \tag{1.55}$$

where $\hat{D}$ is defined as D but with $\hat{\mu}$ replacing μ.

Based on this expression of the equilibrium interest rate, the solution (1.54) may be expressed as a weighted average of the mean-variance portfolio in (1.46) and the global

10. Assume, for example, that returns are normally distributed. Then $\frac{T(T-m)}{m(T-1)}\left(\hat{\mu} - \mu\right)^{\top}\Sigma^{-1}\left(\hat{\mu} - \mu\right) \sim \chi^2\left(m\right)$, i.e., a chi-square variate with m degrees of freedom. In this case, the constraint on the expected returns may be $\Pr(\left(\hat{\mu} - \mu\right)^{\top}\Sigma^{-1}\left(\hat{\mu} - \mu\right) \leq \eta) = 1 - p$, where $\eta = \epsilon\frac{m(T-1)}{T(T-m)}$, and ϵ is chosen so as to ensure a given probability $1 - p$ that the quadratic form lies within some chosen quintile.

minimum variance portfolio in (1.13),

$$\hat{\pi} = (1 - \Omega(\eta))\,\hat{\pi}_{\text{mv}} + \Omega(\eta)\,\pi_{\text{gmv}}, \quad \hat{\pi}_{\text{mv}} = \frac{1}{\tau}\Sigma^{-1}\left(\hat{\mu} - \mathbf{1}_m \hat{r}_{\text{mv}}\right). \tag{1.56}$$

Note that the mean-variance portfolio, $\hat{\pi}_{\text{mv}}$, is obtained while using estimated (not the true, unknown) expected returns.

Eq. (1.55) says that the equilibrium interest rate is equal to the interest rate in a mean-variance economy without parameter uncertainty, $\hat{r}_{\text{mv}}$ (i.e., r in (1.47) but with $\hat{D}$ replacing D), minus a wedge arising due to parameter uncertainty. In other words, agents react to parameter uncertainty by increasing their demand for safe assets, thereby decreasing the equilibrium interest rate. Thus, the model may be interpreted as one predicting flight-to-quality effects: the demand for safe assets increases with an hypothetical increase in η (interpreted as, say, an increase in uncertainty). Moreover, eq. (1.56) clearly shows that an increase in η results in a higher proportion of holdings in the global minimum variance portfolio: as uncertainty grows, agents tilt their portfolios toward asset compositions that are independent of the assets' expected returns; in the extreme case in which assets were uncorrelated, agents would tilt their portfolios toward risk parity after an increase in η.

1.4 The Arbitrage Pricing Theory

1.4.1 Exact APT

1.4.1.1 No-arb restrictions on expected returns

Suppose that asset returns are generated by the following *linear factor model*,

$$\underset{m\times 1}{\tilde{\mu}} = \underset{m\times 1}{\mu} + \underset{m\times k}{B} \cdot \underset{k\times 1}{f} \equiv \mu + cov(\tilde{\mu}, f)[var(f)]^{-1} \cdot f, \tag{1.57}$$

where μ and B are a vector and a matrix of constants and f is a k-dimensional vector of factors with zero mean, which are supposed to affect the asset returns, with $k \leq m$.

Let us normalize $[var(f)]^{-1} = I_{k\times k}$, so that $B = cov(\tilde{\mu}, f)$. With this normalization, we have

$$\tilde{\mu} = \mu + \begin{bmatrix} cov(\tilde{\mu}_1, f) \\ \vdots \\ cov(\tilde{\mu}_m, f) \end{bmatrix} \cdot f = \mu + \begin{bmatrix} \sum_{j=1}^{k} cov(\tilde{\mu}_1, f_j) f_j \\ \vdots \\ \sum_{j=1}^{k} cov(\tilde{\mu}_m, f_j) f_j \end{bmatrix}. \tag{1.58}$$

Next, consider a portfolio of m risky assets and a riskless asset. The wealth generated by this portfolio is given by eq. (1.2), and in this model is

$$w' = \pi^{\top}(\mu - \mathbf{1}_m r) + Rw + \pi^{\top} B f. \tag{1.59}$$

An arbitrage opportunity arises, in this context, if there exists some portfolio π such that the wealth generated by this portfolio, w' in eq. (1.59), is certain and different from the safe gross interest rate R, that is, if $\exists \pi : \pi^{\top} B = 0$ and $\pi^{\top}(\mu - \mathbf{1}_m r) \neq 0$. Mathematically, this is ruled out whenever $\exists \lambda \in \mathbb{R}^k : \mu = \mathbf{1}_m r + B\lambda$. Substituting this relation into eq. (1.57) leaves

$$\tilde{\mu} = \mathbf{1}_m r + B\lambda + Bf = \mathbf{1}_m r + cov(\tilde{\mu}, f)\lambda + cov(\tilde{\mu}, f) f.$$

Taking expectations, we have for each asset return

$$\mu_i = r + (B\lambda)_i = r + \sum_{j=1}^{k} \underbrace{cov(\tilde{\mu}_i, f_j)}_{\equiv \beta_{i,j}} \lambda_j, \qquad i = 1, \cdots, m. \tag{1.60}$$

Let f_j be the jth component of the vector of risks in f. The economic interpretation of eq. (1.60) is that, to be induced to invest in a risky market, we need a risk premium that compensates us beyond the risk-free rate. This risk premium is $\sum_{j=1}^{k} \beta_{i,j}\lambda_j$. For each asset i, it is a linear combination of the betas $\beta_{i,j}$, which are the exposures of the asset return i to the sources of risks f_j times the unit risk premiums λ_j relating to each f_j. So naturally, the unit risk premiums are common to all asset returns, although of course the return exposures, $\beta_{i,j}$, can vary across i.

1.4.1.2 Project evaluation

Project evaluation under the exact APT is obtained under a straightforward generalization of eq. (1.28). For any project with random cash flow equal to $\tilde{C}_i$, we have that its random return is $\tilde{\mu}_i = \frac{\tilde{C}_i}{V_i} - 1$, such that the value, V_i, is given by

$$V_i = \frac{E(\tilde{C}_i)}{1 + \mu_i}, \tag{1.61}$$

where μ_i is as in eq. (1.60). In section 1.4.2, we shall explain how to represent the value of any project in terms of an alternative probability.

1.4.1.3 APT and CAPM

The APT collapses to the CAPM when the market portfolio is the only factor affecting asset returns. This property is easily seen. First, normalize the market portfolio return such that its variance equals one, consistent with eq. (1.60). Accordingly, let $\tilde{r}_M$ be the normalized market return, defined as $\tilde{r}_M \equiv v_M^{-1}\tilde{\mu}_M$, so that $var(\tilde{r}_M) = 1$. We have

$$\tilde{\mu}_i = \mu + \beta_i \tilde{r}_M, \qquad i = 1, \cdots, m,$$

where $\beta_i = cov(\tilde{\mu}_i, \tilde{r}_M) = v_M^{-1} cov(\tilde{\mu}_i, \tilde{\mu}_M)$. Then

$$\mu_i = r + \beta_i \lambda, \qquad i = 1, \cdots, m. \tag{1.62}$$

In particular, $\beta_M = cov(\tilde{\mu}_M, \tilde{r}_M) = v_M^{-1} var(\tilde{\mu}_M) = v_M$, and so, by eq. (1.62),

$$\lambda = \frac{\mu_M - r}{v_M}, \tag{1.63}$$

which is known as the *Sharpe ratio* on the market portfolio or the market price of risk.

By replacing $\beta_i = v_M^{-1} cov(\tilde{\mu}_i, \tilde{\mu}_M)$ and the expression for λ above into eq. (1.62), we obtain

$$\mu_i = r + \frac{cov(\tilde{\mu}_i, \tilde{\mu}_M)}{v_M^2} (\mu_M - r), \qquad i = 1, \cdots, m.$$

This is simply the SML in eq. (1.21).

1.4.2 Risk Neutral Tilts or the Fundamental Theorem of Asset Pricing

Only if the market were "risk neutral," would eq. (1.60) predict that the expected return on any risky asset collapse to the safe interest rate, $\mu_i = r$. But, is there a way to construct a risk neutral market—that is, one with a zero risk premium—from the original market where the risk premium $\sum_{j=1}^{k} \beta_{i,j}\lambda_j$ is different from zero? The answer is in the affirmative. It is a quite fundamental theme in financial economics. It links to the celebrated *fundamental theorem of asset pricing*, sometimes referred to as the "FTAP."

1.4.2.1 Probability twists

Let us develop intuition on such a beautiful result by elaborating a simple example. Assume that each of the risks f_j is standard normal, $P\left(df_j\right) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}f_j^2)df_j$, and next, that we tilt their densities by a factor equal to $\zeta\left(f_j\right) = \exp(-\frac{1}{2}\lambda_j^2 - \lambda_j f_j)$, where λ_j is the unit risk premium, as defined in section 1.4.1. This tilt defines a new probability under which each factor f_j is distributed. Let us determine this probability by tilting P through ζ,

$$Q(df_j) = \zeta(f_j)P(df_j) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}f_j^2 - \frac{1}{2}\lambda_j^2 - \lambda_j f_j\right) df_j = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\tilde{f}_j^2\right) d\tilde{f}_j, \tag{1.64}$$

where

$$\tilde{f}_j = f_j + \lambda_j. \tag{1.65}$$

Note that the new density, Q, is still that of standard normal variate. Yet under Q, it is $\tilde{f}_j$ that has zero expectation, not f_j. In other words, we have that under Q, (i) $\tilde{f}_j$ is standard normal and (ii) f_j is normal with unit variance but with an expectation equal to $-\lambda_j$. That is, assuming that $\lambda_j > 0$, f_j has a lower expectation under Q than under P—under P, this expectation is zero, and under Q, it is $-\lambda_j$.

We label the new probability Q risk neutral probability, for the following reasons. Consider eqs. (1.58) and (1.60), which say that under P,

$$\tilde{\mu}_i = r + \sum_{j=1}^{k} \beta_{i,j}\lambda_j + \sum_{j=1}^{k} \beta_{i,j} f_j \,. \tag{1.66}$$

We also know that under the new probability Q, each factor f_j is normally distributed with mean $-\lambda_j$. That is, replacing eq. (1.65) into eq. (1.66), we have that under Q,

$$\tilde{\mu}_i = r + \sum_{j=1}^{k} \beta_{i,j} \tilde{f}_j \,.$$

To summarize, the return on each assets $\tilde{\mu}_j$ has the following distributions under P and under Q:

$$\tilde{\mu}_j \sim N_P\left(r + \textstyle\sum_{j=1}^{k} \beta_{i,j}\lambda_j, \sigma_j^2\right), \quad \tilde{\mu}_j \sim N_Q\left(r, \sigma_j^2\right), \tag{1.67}$$

where $\sigma_j^2 = \Sigma_{jj}$, $\Sigma \equiv BB^\top$, as in eq. (1.3), and N_P and N_Q denote the Normal densities under P and under Q. That is, under Q, the expected return on each asset equals r, whence the risk neutral probability label. In later chapters, we shall use the celebrated Girsanov's

theorem to elaborate on these topics and label the tilt ζ in eq. (1.64) as the Radon-Nikodym derivative of P against Q (see chapter 4, section 4.3.3).

1.4.2.2 Evaluation of derivatives

Why do we need to complicate everything with the previous probability changes? In fact, the entire building block underlying asset evaluation relies on similar risk neutral tilts. Consider the following example of derivative evaluation. We wish to price a quadratic derivative, that is, one that pays off the square of the cash flow promised by the first asset, $\tilde{C}_1^2$. It is challenging to evaluate this derivative through "APT software" in this Gaussian market because $\tilde{C}_1^2$ is obviously not normally distributed, which complicates the reasoning underlying its exposure to the factors f_j. In fact, in this Gaussian market, we cannot restrict the behavior of the expected return on this derivative without assuming something more. Let us explain.

Suppose we want to construct a portfolio of the existing assets to replicate the payoff of the quadratic derivative for each possible value this derivative could take. Can we do this? The answer is in the negative. We cannot use a finite number of assets to span an asset payoff, which could take a continuum of values, such as $\tilde{C}_1^2$. We say markets are *incomplete* in this context. In chapter 2, we shall see that the price of such and related derivatives can be found in a preference-free format as soon as the number of assets is at least as large as the number of states—markets are, then, *complete*. In this case, a portfolio that replicates the derivative's payoff can be found, and its value is the same as the derivative's, for two assets are worth the same whenever they promise the same payoff.

Chapter 2 explains that, in a world with complete markets, the price of the existing traded assets can be inverted for the shadow value of some elementary assets—those that pay off one unit of numéraire in a given state of the world and zero otherwise. The price of these elementary securities can then be used to price any derivative, which is redundant indeed. Chapter 4 explains how these results can be generalized to markets with a continuum of states, as soon as we assume that there exist a number of sufficiently diverse elementary securities, which guarantee a payoff could be delivered for each state of nature. To illustrate through the example of this section, consider the following elementary security, which promises the following payoff

$$\Pi(s) = \begin{cases} 1, & \text{if } \tilde{C}_1 \in (s, s+ds) \\ 0, & \text{otherwise} \end{cases}, \tag{1.68}$$

and let $\phi(s)ds$ be its current price. We shall refer these securities as Arrow-Debreu securities, for reasons explained in chapter 2.

We could utilize all of these Arrow-Debreu securities, that is, for all $s \in \mathbb{R}$, and replicate any generic function of the state, $\psi(s)$, including our original payoff, $\psi(s) = s^2$, $s \in \mathbb{R}$. Indeed, note that by purchasing $\psi(s)$ units of the security that pays off $\Pi(s)$ in eq. (1.68), we pay $\psi(s)\,\phi(s)\,ds$ today and are guaranteed to receive $1 \times \psi(s)$ tomorrow in state $\tilde{C}_1 \in (s, s+ds)$ and zero otherwise. Therefore, by purchasing all the securities that span $\mathbb{R}$, in proportion of $\psi(\cdot)$, we shall receive, for sure, $\psi(\tilde{C}_1)$ for any possible value of $\tilde{C}_1$ tomorrow and today pay

$$C_\psi \equiv \int_{-\infty}^{\infty} \psi(s)\,\phi(s)\,ds. \tag{1.69}$$

We call P_ψ such a portfolio. We claim that the value of the derivative, say V_ψ, is just $\mathcal{C}_\psi$. For suppose not, and assume, for instance, that $V_\psi > \mathcal{C}_\psi$. Then we could short sell the derivative for V_ψ, invest $\mathcal{C}_\psi$ into the portfolio P_ψ, and retain an arbitrage profit equal to $V_\psi - \mathcal{C}_\psi$. It is an arbitrage profit because the portfolio P_ψ delivers the exact payoff we need to honor the short sale of the derivative.

The crucial point is to determine the value of ϕ. We claim that

$$\phi(s) = \frac{1}{1+r} q\left(s; V_1(1+r), \bar{\sigma}^2\right) \tag{1.70}$$

for some $\bar{\sigma}^2$, where $q\left(s; m, \bar{\sigma}^2\right)$ denotes the density of a normal distribution with mean and variance m and $\bar{\sigma}^2$. Indeed, let us take expectations of $\tilde{C}_1 = V_1(1+\tilde{\mu}_1)$ under Q, such that, by eq. (1.67),

$$V_1 = \frac{E_Q(\tilde{C}_1)}{1+r}. \tag{1.71}$$

On the other hand, let us apply eq. (1.69) to determine the value of the derivative that pays off, $\psi(\tilde{C}_1) = \tilde{C}_1$, which is

$$V_1 = \int_{-\infty}^{\infty} s\phi(s)\, ds. \tag{1.72}$$

Comparing eq. (1.71) and eq. (1.72) yields eq. (1.70).[11]

We are now ready to evaluate the quadratic derivative, by relying on eq. (1.69), and the expression of ϕ in eq. (1.70). We have for $\psi(s) = s^2$, that the value of the derivative, say $V_{\mathcal{C}}$, can be expressed as a risk neutral expectation:

$$V_{\mathcal{C}} = \int_{-\infty}^{\infty} s^2 \phi(s)\, ds = \frac{E_Q(\tilde{C}_1^2)}{1+r}. \tag{1.73}$$

The expression in eq. (1.73) tells us that all we have to do is to recast the initial evaluation problem in terms of this new risk neutral setup. To fix ideas, suppose that the unit prices of risk λ_j are all positive. In this setup, the discounting factor is the safe interest rate, r. To compensate for such generous discounting, the expectation in the numerator of eq. (1.73) is lower than that under P because the distribution of $\tilde{C}_1$ is more skewed to the left under Q than under P, due to the factors driving the value of $\tilde{C}_1 = V_1(1+\tilde{\mu}_1)$ being skewed to take more pessimistic values under Q, on average, by a factor equal to $-\lambda_j$.

Let us proceed with the determination of $V_{\mathcal{C}}$, by using eq. (1.73). We have that $E_Q(\tilde{C}_1^2) = V_1^2 E_Q((1+\tilde{\mu}_1)^2)$, where $1+\tilde{\mu}_1 \sim N_Q\left(1+r, \sigma_1^2\right)$ such that, by a direct calculation,

$$V_{\mathcal{C}} = V_1^2\left(\frac{\sigma_1^2}{1+r} + 1 + r\right),$$

where V_1 is the value of the first asset, determined as usual through eq. (1.61). Note how simple this formula is. It links the value of the derivative to the square of the value of the

11. We can check that eq. (1.70) is consistent with the pricing of a pure discount bond. Such a bond has a payoff equal to $\psi(s) = 1$ for all s such that by eq. (1.69), $\frac{1}{1+r} = \int_{-\infty}^{\infty} \phi(s)\, ds$, which it does, by eq. (1.70).

underlying risk, V_1^2, and the discounted value of σ_1^2, reflecting that, after all, a quadratic derivative is about a play in volatility.

1.4.3 Uncertainty and Asset Evaluation

How asset prices relate to the volatility of fundamentals? It is a very old issue in financial economics.[12] Consider a two-period market for a cash ψ to be paid in the second period where we assume that $\psi = \psi(\tilde{C})$ for some random variable $\tilde{C}$. Accordingly, and generalizing eq. (1.73), $V_{\mathcal{C}} \equiv (1+r)^{-1} E_Q[\psi(\tilde{C})]$ is the current value of a derivative that pays $\psi(\tilde{C})$.

We ask: How does the value $V_{\mathcal{C}}$ change when uncertainty regarding $\tilde{C}$ increases? For example, how does $V_{\mathcal{C}}$ change after a change in the variance of $\tilde{C}$? It is a complex question, and the answers to it depend on the assumptions we make about the uncertainty changes surrounding $\tilde{C}$. In this chapter, we begin with the simplest situation, one in which a change in the *uncertainty* of $\tilde{C}$ does not entail a change in the *expected value* of $\tilde{C}$. This case is actually very relevant. Suppose that that $\tilde{C}$ is the cash flow of a traded security with a given value V, such that its current value is, consistently with results in the previous sections, $E_Q(\tilde{C}) = V(1+r)$, which is clearly independent of the uncertainty surrounding $\tilde{C}$ once we take V to be given, just as we are doing.

Situations in which changes in the volatility of a random variable do not affect its mean are known as *mean-preserving spreads*. Appendix 1.C contains a succinct summary of these situations and develops results that predict that $V_{\mathcal{C}}$ does indeed increase after a mean-preserving spread in $\tilde{C}$, provided ψ is a convex function of $\tilde{C}$. Intuitively, a convex function exaggerates favorable realizations of $\tilde{C}$ and dampens the poor ones. The theory of mean-preserving spreads was introduced by Rothschild and Stiglitz (1970, 1971) and will be used at different junctures of this book.

1.4.4 The APT with Idiosyncratic Risk and a Large Number of Assets

Can idiosyncratic risk be eliminated? This section reviews conclusions on this question contained in the seminal work of Ross (1976) and its initial refinements made by Connor (1984) and Huberman (1983), among others. Consider eq. (1.22). Intuitively, we may form portfolios with a large number of assets so as to make idiosyncratic risk negligible by the law of large numbers. But would the APT relation in eq. (1.60) be still valid? The answer is in the affirmative, although it deserves some qualifications.

Consider the APT equation (1.57), and add a vector of idiosyncratic returns, ε, which are independent of f and have mean zero and variance σ_ε^2:

$$\tilde{\mu} = \mu + B \cdot f + \varepsilon.$$

We wish to show that, in the absence of some appropriate notion of arbitrage (to be defined below), it must be that the number of assets such that eq. (1.60) does *not* hold, $N(m)$ say, is bounded as m gets large, that is,

$$\left|\mu_i - \left((B\lambda)_i + r\right)\right| > 0, \quad i = 1, \cdots, N(m), \tag{1.74}$$

12. See, e.g., the classical work of Malkiel (1979), Pindyck (1984), Poterba and Summers (1985), Abel (1988), Barsky (1989), among others.

where

$$\lim_{m\to\infty} N(m) < \infty. \tag{1.75}$$

In other words, we wish to show that in a large market, eq. (1.60) does indeed hold for most of the assets, an approach close to that in Huang and Litzenberger (1988, 106–108).

By the same arguments leading to eq. (1.1), the wealth generated by a portfolio of the assets satisfying (1.74), $w'_{N(m)}$ say, is

$$w'_{N(m)} = \pi^{\top}_{N(m)} \left(\mu_{N(m)} - \mathbf{1}_{N(m)} r\right) + R w_{N(m)} + \pi^{\top}_{N(m)} \left(B_{N(m)} f + \varepsilon_{N(m)}\right),$$

where μ_N, B_N and ε_N are (i) the vector of the expected returns, (ii) the factor exposures matrix, and (iii) the vector of idiosyncratic return components affecting these assets and, finally, where π_N and w_N are the portfolio and the initial wealth invested in these assets.

In this context, we may define an arbitrage as the portfolio $\pi_{N(m)}$ that in the limit, as the number of the existing assets m gets large, is riskless and yet delivers an expected return strictly larger than the safe interest rate, namely,

$$\lim_{m\to\infty} \frac{E[w'_{N(m)}]}{w_{N(m)}} > R, \quad \text{and} \quad \lim_{m\to\infty} var[w'_{N(m)}] \to 0. \tag{1.76}$$

We want to show that this situation does not arise under the condition in (1.75), thereby establishing that the linear APT relation in eq. (1.60) is valid for most of the assets in a large market.

So suppose the linear relation, $\mu_N - \mathbf{1}_N r = B_N \lambda$, doesn't hold. Then there exists a portfolio $\underline{\pi}$ such that

$$\underline{\pi}^{\top} B_N = 0 \quad \text{and} \quad \underline{\pi}^{\top} (\mu_N - \mathbf{1}_N r) \neq 0. \tag{1.77}$$

Consider the portfolio

$$\hat{\pi}_N = \frac{1}{N} \cdot sign\left(\underline{\pi}^{\top} (\mu_N - \mathbf{1}_N r)\right) \cdot \underline{\pi},$$

where $\underline{\pi}$ is as in (1.77). With this portfolio we clearly have $E[w'_N] = \hat{\pi}^{\top}_N (\mu_N - \mathbf{1}_N r) + R w_N > R w_N$, for each N and even for N large. That is, $\lim_{m\to\infty} E[w'_{N(m)}]/w_{N(m)} > R$, which is the first condition in (1.76). Regarding the second condition in (1.76), we have

$$var[w'_N] = \hat{\pi}^{\top}_N \left(B_N B^{\top}_N + \sigma^2_{\varepsilon} I_{N\times N}\right) \hat{\pi}_N = \sigma^2_{\varepsilon} \hat{\pi}^{\top}_N \hat{\pi}_N,$$

where the second equality follows by the first relation in (1.77). Clearly, $\lim_{m\to\infty} var[w'_{N(m)}] \to 0$ as $N(m) \to \infty$. Hence, in the absence of arbitrage, the condition in (1.75) must hold.

1.5 Empirical Evidence

This section contains a very succinct account regarding the empirical evidence on the CAPM, its factor extensions, and some of the market practice arising from it.

1.5.1 Fama-MacBeth Two-Step Regressions

How do we estimate eq. (1.22)? Consider a slightly more general version of eq. (1.22), one in which the riskless interest rate is time-varying:

$$\tilde{\mu}_{i,t} - r_t = \beta_i(\tilde{\mu}_{M,t} - r_t) + \varepsilon_{i,t}, \quad i = 1, \cdots, m,$$

where $\varepsilon_{i,t}$ denotes time-series residuals. Fama and MacBeth (1973) consider an estimation procedure based on two steps. In a first step, one obtains estimates of the market exposures for all stocks, $\hat{\beta}_i$ say, relying on monthly returns (or returns on other frequencies); we may approximate the market portfolio with some broad stock market index.[13] In the second step, we runs cross-sectional regressions, one for each month,

$$\tilde{\mu}_{i,t} - r_t = \alpha_t + \lambda_t \hat{\beta}_i + \eta_{i,t}, \quad t = 1, \cdots, T,$$

where T is the sample size and $\eta_{i,t}$ denotes "cross-sectional residuals." The time-series of cross-sectional estimates of the intercept α_t and the price of risk λ_t ($\hat{\alpha}_t$ and $\hat{\lambda}_t$, say) are then used to make statistical inference. For example, time-series averages and standard errors of $\hat{\alpha}_t$ and $\hat{\lambda}_t$ may lead to point estimates and standard errors for α (the unconditional expected return unexplained by the model) and λ (the price of risk predicted by the CAPM; see eqs. (1.62)–(1.63)). If the CAPM holds, estimates of α should not be significantly different from zero.

1.5.2 Macroeconomic Forces

Chen, Roll, and Ross (1986) use the Fama-MacBeth two-step estimation procedure and estimate a multifactor APT model, one built up along the lines of section 1.4. They identify macroeconomic forces driving asset returns with innovations in variables such as the term spread, expected and unexpected inflation, industrial production growth, or the corporate spread. They find that these sources of variation in the cross-section of asset returns are significantly priced.

The empirical literature subsequent to this work relies on testing procedures leading to mixed results. For example, Shanken and Weinstein (2006) find that industrial production growth seems to survive to their own testing methodology, although their evidence is less convincing regarding the previous factors. Part II of the book deals with related asset pricing puzzles and, more generally, with how financial markets relate to macroeconomic developments. The main objectives of part II are to strive against data mining (the process of searching for variables that explain asset returns) and, alternatively, to put economic reasoning at the core of the process that leads to suggestions for empirical relations. Section 1.5.5 provides further details on how data mining has been addressed in the empirical literature.

1.5.3 The Fama and French Model

Consider the SML introduced in section 1.3 (see eq. (1.21)). The CAPM predicts that each asset has an average excess return lying precisely on the SML. Assets delivering average

13. In tests of the CAPM, one uses proxies of the market portfolio, such as, say, the S&P 500. However, the market portfolio is unobservable. Roll (1977) points out that, as a result, the CAPM is inherently untestable as any test of the CAPM is a joint test of the model itself and of the closeness of the proxy to the market portfolio.

excess returns significantly away from the SML would point to evidence that this single factor version of the APT does not work; points A, B, C, and D in figure 1.4 illustrate this situation. Consider, for example, the asset corresponding to point A. A regression of the excess return of this asset onto the excess return on the market would produce a positive intercept, some $\alpha > 0$, such that its average excess return would equal $\alpha + \beta_i (\mu_M - r)$, thereby invalidating the SML in eq. (1.21).

Fama and French (1992, 1993) provide evidence that the standard CAPM is invalidated by the presence of two additional factors, size and value. Further, there is strong evidence of at least a third factor against the standard CAPM, momentum. These factors are discussed next.

(1) *Size* (Banz, 1981): Average returns for "small firms" are too high given their beta. Firm size is measured in terms of market capitalization: stock price times outstanding shares.

(2) *Value* (Stattman, 1980; Rosenberg, Reid, and Lanstein, 1985): Average returns on stocks with high book-to-market (BTM) ratios or "value stocks" are too high given their beta. In general, average returns on value stocks are higher than those on stocks with low BTM ratios or "growth stocks." The points A, B and C in figure 1.4 may identify examples of value stocks, with the average excess returns corresponding to point A relating to the firm with the highest BTM ratios; point D may be an example of a growth stock.[14]

(3) *Momentum* (Jegadeesh and Titman, 1993): Stocks with the highest returns in the previous 12 months tend to outperform in the next future.

What is the economic intuition for expecting these factors to explain the cross-section of asset returns? Fama and French (1992, 1993) provide empirical evidence that returns on both size and value are proxies for common risk factors. They argue that small and, especially, value stocks seem to display persistently low earnings. Size and value can then be proxies for profitability and sources of priced risk. However, there is still considerable debate regarding the economic interpretation of these factors. The interpretation of momentum as a factor is also very subtle. A natural explanation relies on the sensitivity of asset prices to news. For example, Pedersen (2015, 139) reviews the simple idea that prices underreact. If, for example, an asset price underreacts to good news, its subsequent increases lead to momentum. Chapter 10 provides additional explanations for underreaction based on "limits of arbitrage." Further, Chapter 8 reviews models by which volatility and expected returns actually increase on the upside (see section 8.6.3): when growth is very strong, asset price volatility and then premiums increase after good news, producing momentum; that is, returns are expected to increase precisely in good times.

The standard CAPM does not have the power to explain the cross-section of asset returns that are sorted by size, BTM, or momentum. Assets sorted in this way command

14. This example is highly hypothetical: it is well known that stocks with higher BTM ratios have higher betas (see Fama and French, 1993).

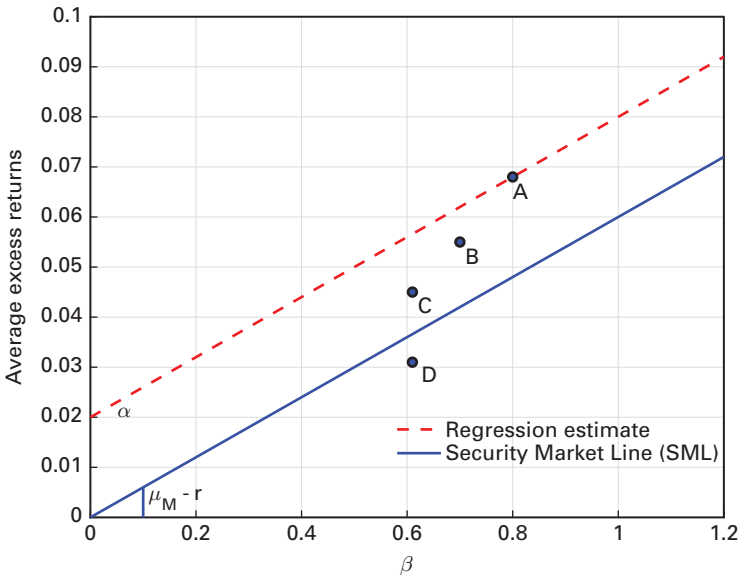

Figure 1.4
The solid line depicts the SML based on the assumption that the excess return on the market $\mu_M - r = 6\%$. The four points, *A* through *D*, identify hypothetical pairs of β estimates and average excess returns for four assets. The α on asset *A* is obtained while identifying the intercept of the line that passes through point *A* and is parallel to the SML.

a size premium, a value premium, and a momentum premium. Let us explain these facts relying on the original examples developed in the Fama and French (1993) paper. The original focus of this paper was the time series estimate of exposures to market, size and value. The returns to be explained relate to portfolios created as follows. First, portfolios are formed while sorting assets in a given universe through size and BTM. In each year, stocks are allocated into 5 size quintiles and 5 BTM quintiles, leading to a total of 25 portfolios obtained as intersections of the 5 size and the 5 BTM groups. Associated with each of these 25 portfolios are value-weighted monthly returns. The puzzle then, at least from the standard CAPM perspective, is that the CAPM cannot explain the expected returns on these 25 portfolios.

Fama and French (1993) show that the returns on these portfolios can be very much better understood by means of a multifactor model, in which both "size premiums" and "value premiums" are taken as "factor premiums." Fama and French consider three factors: (i) the *market premium*, defined as the monthly excess return on the market; (ii) the *small minus big (SMB) premium*, defined as the monthly difference between returns on assets with small and big size; (iii) the *high minus low (HML) premium*, defined as the monthly difference between returns on assets with high and low BTM ratios.

Precisely, stocks ranked by size are split into two groups (S: small; B: big) according to whether their market capitalization is below or above the median, and stocks ranked by BTM are broken into three groups based on the breakpoints for the bottom 30%

(L: low), middle 40% (M: medium) and top 30% (H: high) of the BTM ratios distribution. The result is a 2×3 matrix.

	Book-to-Market		
Size	L	M	H
S	(S,L)	(S,M)	(S,H)
B	(B,L)	(B,M)	(B,H)

The SMB premium is obtained as the difference between the average returns on small stocks minus that on big stocks; the HML premium is the difference between the average return on stocks with high BTM minus that on low; with obvious notation,

$$\text{SMB} = \frac{1}{3}\left[(\text{S,L}) + (\text{S,M}) + (\text{S,H})\right] - \frac{1}{3}\left[(\text{B,L}) + (\text{B,M}) + (\text{B,H})\right]$$
$$\text{HML} = \frac{1}{2}\left[(\text{S,H}) + (\text{B,H})\right] - \frac{1}{2}\left[(\text{S,L}) + (\text{B,L})\right].$$

That is, the two portfolios, SMB and HML, are *factor mimicking portfolios* and are recalculated at a regular frequency. Moreover, these portfolios are *dynamic*: some assets may exit from these portfolios and be replaced by new ones, whenever a criterion determines that they do not fall under the appropriate category at the date portfolios are recalibrated. The exposures to these portfolios are estimated while regressing each excess return on all factors:

$$\mu_i - r = \alpha_i + \beta_i\left(\mu_{Mt} - r\right) + \beta_{i,smb} SMB_t + \beta_{i,shml} HML_t + \varepsilon_{it}.$$

The model can be used to determine the expected return on each asset based on the predictive part of the previous regression. Carhart (1997) extends this model to one with a fourth factor: momentum. Returns on a momentum portfolio are the difference between monthly returns on assets with recent better performance minus returns on assets with recent worst performance. After the publication of Fama and MacBeth (1973), the literature has produced an enormous amount of work pointing to evidence of many additional factors, such as those identified in the Fama and French pieces; Bali, Engle, and Murray (2016) provide a review of this literature. Before we discuss some of this evidence in section 1.5.5, we provide a succinct review on how the "discovery" of factors has shaped investment practice.

1.5.4 "Smart Beta," or Factor Investing

The merit of the CAPM lies in its simple prediction that the risk of an asset and thus its premium relates to the asset returns' exposure to market fluctuations: its beta. But the empirical evidence reviewed in the previous section suggests the presence of additional factors, which the cross-section of expected returns may be exposed to. Thus, in the world of factor models, assets do not really have a meaning on their own in that their returns merely capture the factors they are exposed to. For example, in his extensive review of these topics, Ang (2014, 194) explains that "factors are to assets what nutrients are to food." In this world, investing is thus a means to capture variation in some of these factors and the premiums arising therefrom.

Market practice that emphasizes these ideas is known as *factor investing* or *smart beta*. This phenomenon is so important that smart beta is now part of "commoditization" efforts, notably occurring through smart beta exchange traded funds (ETFs).[15] These ETFs track smart beta indices—that is, indices that reflect the performance of an investment strategy that seeks exposure to some factors that are alternative to standard market capitalization indices. To simplify, while standard beta is exposure to market, smart beta is exposure to "something else." Which factors—or nutrients—and then premiums would any investor like to be exposed to? There is no theory or guidance to the variety of available products. The only assumptions underlying these products are that asset returns incorporate premiums related to their exposure to factors and then that there are dedicated portfolios of them that may isolate specific premiums. Some of these factors may be *macroeconomic factors*, such as economic growth, inflation, or volatility of financial markets; these factors are not traded and are difficult to mimic, with the exception of volatility in some cases as explained in chapter 11. But factors may be simply *investment styles*, such as value or momentum (see the previous section), or additional factors such as quality (embedded into stocks with stable earnings and balance sheets) or low volatility stocks (consistent with the risk parity examples of section 1.2.5). These factors are often dynamic: the investment strategy underlying them involves buying and selling over time, just as in the examples of the previous section.

1.5.5 "Lucky Factors"

How many factors have been reported in the literature, which are potentially able to explain the cross-section of expected returns? Harvey, Liu, and Zhu (2016) point to the existence of more than 300 such factors. This fact naturally raises a "data mining" concern: by force of trying with potentially new explanations of asset prices, we might end up finding some that work. How sure are we that the successful factors only worked by pure chance? This problem can be regarded as one of multiple testing: when we want to examine how plausible a new factor is, we are actually testing several null hypotheses (that the first factor is not significant, that the second is not significant, etc.). In this context, multiple testing is related to data mining simply because the significance of a new factor must be assessed while many variables (factors) are being tested on the same dataset.

Further, we may suspect that, in the underground, researchers have been trying with thousands of additional potential explanations and that none of them have seemed to work. Financial economists typically publish papers with positive results. Moreover, in the failed attempts, it is likely that some true explanations of asset markets were disregarded only because they were not successful given the particular samples on which they were tested. Intuitively, then, the higher the number of "hidden trials," the more likely it is that

15. An ETF is a portfolio of securities that may be traded on an exchange just like a common stock. Unlike a closed-end fund (that issues a fixed number of nonredeemable shares), which only trades once a day, an ETF trades all the time. While closed-end funds may trade at a discount or a premium, arbitrage forces tend to make an ETF more closely reflect its Net Asset Value. ETFs differ from open-end funds (which issue an open-ended number of redeemable shares) for a variety of reasons, such as the possibility to be traded on margin or to be sold short.

the discovered factors come out from pure data mining. This problem is one of sample selection bias of the type addressed by Heckman (1979).

More formally, let H_0 denote the null hypothesis that a given factor is not significant. According to standard terminology (see, e.g., Gouriéroux and Monfort, 2008), we may incur into two types of errors while in the process of statistical testing:

- *Type-I error*: rejecting H_0 when H_0 is true, leading to false positives outcomes;
- *Type-II error*: failing to reject H_0 when H_0 is false, leading to false negatives outcomes.

Associated with each of these errors is a probability of occurrence. Let $t_T(x)$ denote a statistic—that is, a function—of the sample of T observations, $(x_t)_{t=1}^T$. For example, $t_T(x)$ may be the t-statistic for the significance of a factor in a cross-sectional regression. The probabilities of making a Type-I and a Type-II error are, resp.,

$$\alpha_T(\theta) \equiv \Pr\left(t_T(x) \in \mathcal{C}_1 \,|\, H_0\right), \quad \beta_T(\theta) \equiv \Pr\left(t_T(x) \in \bar{\mathcal{C}}_1 \,\middle|\, H_1\right),$$

where $\mathcal{C}_1$ denotes the critical region (and $\bar{\mathcal{C}}_1$, its complement), that is, the set of values such that H_0 is rejected if the statistic takes values belonging to it. The hypothesis H_1 is an alternative to H_0. The two probabilities are a function of the parameter θ underlying the data generating process. The probability α_T is known as the "size" of a test and $1-\alpha_T$ is its "significance." Instead, $1-\beta_T$ is the "power" of the test.

The critical region is chosen so as to represent the values of the statistics that we consider significantly different from those we would have observed under the null, H_0. Naturally, the higher $\alpha_T(\theta)$, the lower $\beta_T(\theta)$. How do we proceed in practice? In science, it seems reasonable to dislike Type-I errors more than Type-II. Typically, the null is the old theory, which we will retain until strong evidence mounts such that we may not maintain it anymore—"strong" meaning that the significance should be quite high. In a way, when abandoning a theory is costly, it might be better to guess that we are making errors without modifying our systems rather than that we are making errors while modifying it. This reasoning seems to be consistent with the Popperian account of science proceeding through falsifications of old hypotheses (Popper, 1959). In our context, Type-I errors lead to false discoveries, and Type-II errors lead to miss out true factors. Because we cannot simultaneously minimize the probabilities of both errors, one compromising approach is to fix a significance level, $1-\alpha_T$, and devise a testing procedure that maximizes power (the probability that true factors are not rejected) for the given significance level. However, note that power depends on the unknown parameters governing the alternative hypothesis.

To summarize, Harvey, Liu, and Zhu (2016) provide a thorough discussion of the statistical literature and the ensuing variety of tests that seem to work as appropriate benchmark for our testing concerns. Their conclusions are quite stringent: a t-statistic of at least three is required as evidence of a significant factor; many of the published successful factors are actually likely not true factors.

Appendix 1.A Portfolio Choice

We derive eq. (1.9), that is, the optimal portfolio allocation in the case without the safe asset. We solve two programs: (i) the primal program [1.P2] in the main text (maximize portfolio expected return for a given variance) and (ii) a dual program, to be introduced below, by which we minimize the portfolio return variance for a given expected return.

These derivations are in appendixes 1.A.1 and 1.A.2. Appendixes 1.A.3 and 1.A.4 provide derivations of selected results given in the main text.

1.A.1 The Primal Program

Given eq. (1.8), the Lagrangian function associated to [1.P2] is

$$\mathcal{L} = \pi^\top \mu + w - \nu_1(\pi^\top \Sigma \pi - w^2 v_p^2) - \nu_2(\pi^\top \mathbf{1}_m - w),$$

where ν_1 and ν_2 are two Lagrange multipliers. The first-order conditions are

$$\hat{\pi} = \frac{1}{2\nu_1}\Sigma^{-1}(\mu - \nu_2 \mathbf{1}_m), \quad \hat{\pi}^\top \Sigma \hat{\pi} = w^2 v_p^2, \quad \hat{\pi}^\top \mathbf{1}_m = w. \tag{1.A1}$$

Using the first and the third conditions leaves

$$w = \mathbf{1}_m^\top \hat{\pi} = \frac{1}{2\nu_1}(\underbrace{\mathbf{1}_m^\top \Sigma^{-1}\mu}_{\equiv D} - \nu_2 \underbrace{\mathbf{1}_m^\top \Sigma^{-1}\mathbf{1}_m}_{\equiv C}) \equiv \frac{1}{2\nu_1}(D - \nu_2 C).$$

We solve for ν_2, obtaining

$$\nu_2 = \frac{D - 2w\nu_1}{C}.$$

By replacing the solution for ν_2 into the first condition in (1.A1), we get

$$\hat{\pi} = \frac{w}{C}\Sigma^{-1}\mathbf{1}_m + \frac{1}{2\nu_1}\Sigma^{-1}\left(\mu - \frac{D}{C}\mathbf{1}_m\right). \tag{1.A2}$$

Next, we derive the value of the program [1.P2]:

$$E\left[w'(\hat{\pi})\right] - w = \hat{\pi}^\top \mu = \frac{w}{C}D + \frac{1}{2\nu_1}\left(A - \frac{D^2}{C}\right), \quad A \equiv \mu^\top \Sigma^{-1}\mu. \tag{1.A3}$$

It is easy to show that

$$\begin{aligned} var\left[w'(\hat{\pi})\right] &= w^2 v_p^2 \\ &= \hat{\pi}^\top \Sigma \hat{\pi} \\ &= \left[\frac{w}{C}\mathbf{1}_m^\top \Sigma^{-1} + \frac{1}{2\nu_1}\left(\mu^\top - \frac{D}{C}\mathbf{1}_m^\top\right)\Sigma^{-1}\right]\left[\frac{w}{C}\mathbf{1}_m + \frac{1}{2\nu_1}\left(\mu - \frac{D}{C}\mathbf{1}_m\right)\right] \\ &= \frac{w^2}{C} + \left(\frac{1}{2\nu_1}\right)^2\left(A - \frac{D^2}{C}\right). \end{aligned} \tag{1.A4}$$

Let us gather eqs. (1.A3) and (1.A4):

$$\begin{cases} \mu_p(v_p) \equiv \dfrac{E\left[w'(\hat{\pi})\right] - w}{w} = \dfrac{D}{C} + \dfrac{1}{2\nu_1 w}\left(A - \dfrac{D^2}{C}\right) \\ v_p^2 = \dfrac{1}{C} + \left(\dfrac{1}{2\nu_1 w}\right)^2\left(A - \dfrac{D^2}{C}\right) \end{cases} \tag{1.A5}$$

where we have emphasized the dependence of μ_p on v_p, which arises through the Lagrange multiplier v_1 as formally seen below (see eq. (1.A7)).

The first equation in (1.A5) can be solved for v_1 as follows:

$$\frac{1}{2v_1 w} = \left(AC - D^2\right)^{-1} \left(C\mu_p(v_p) - D\right). \tag{1.A6}$$

We use eq. (1.A6) and express $\hat{\pi}$ in eq. (1.A2) in terms of the portfolio expected return, $\mu_p(v_p)$. We have

$$\frac{\hat{\pi}}{w} = \frac{\Sigma^{-1}\mathbf{1}_m}{C} + \left(AC - D^2\right)^{-1} \left(C\mu_p(v_p) - D\right) \Sigma^{-1} \left(\mu - \frac{D}{C}\mathbf{1}_m\right).$$

By rearranging terms in the previous equation, we obtain eq. (1.9) in the main text.

Finally, we substitute eq. (1.A6) into the second equation of (1.A5), obtaining

$$v_p^2 = \frac{1}{C}\left[1 + \left(AC - D^2\right)^{-1} \left(C\mu_p(v_p) - D\right)^2\right],$$

which is eq. (1.10) in the main text. Note also that the second condition in (1.A5) reveals that

$$\left(\frac{1}{2v_1 w}\right)^2 = \frac{Cv_p^2 - 1}{AC - D^2}. \tag{1.A7}$$

Given that $AC - D^2 > 0$, the previous equation confirms the properties of the *global minimum variance portfolio* stated in the main text.

1.A.2 The Dual Program

Next, consider the following program:

$$\hat{\pi} = \arg\min_{\pi \in \mathbb{R}^m} var\left[\frac{w'(\pi)}{w}\right] \quad \text{s.t. } E\left[w'(\pi)\right] = E_p \text{ and } w = \pi^\top \mathbf{1}_m \qquad \text{[1.AP2-D]}$$

for some constant E_p. The first-order conditions are

$$\frac{\hat{\pi}}{w} = \frac{\eta_1 w}{2}\Sigma^{-1}\mu + \frac{\eta_2 w}{2}\Sigma^{-1}\mathbf{1}_m, \quad \hat{\pi}^\top \mu = E_p - w, \quad w = \hat{\pi}^\top \mathbf{1}_m, \tag{1.A8}$$

where η_1 and η_2 are two Lagrange multipliers.

Combining the first and second conditions in (1.A8) leaves

$$E_p - w = \hat{\pi}^\top \mu = w^2 \left(\frac{\eta_1}{2}A + \frac{\eta_2}{2}D\right), \tag{1.A9}$$

and combining the first and third conditions in (1.A8) produces

$$w = \hat{\pi}^\top \mathbf{1}_m = w^2 \left(\frac{\eta_1}{2}D + \frac{\eta_2}{2}C\right). \tag{1.A10}$$

Next, let $\mu_p \equiv \frac{E_p - w}{w}$. By eqs. (1.A9) and (1.A10), the solutions for η_1 and η_2 are

$$\frac{\eta_1 w}{2} = \frac{C\mu_p - D}{AC - D^2}, \qquad \frac{\eta_2 w}{2} = \frac{A - D\mu_p}{AC - D^2}.$$

Therefore, the solution for the portfolio in eq. (1.A8) is

$$\frac{\hat{\pi}}{w} = \frac{C\mu_p - D}{AC - D^2}\Sigma^{-1}\mu + \frac{A - D\mu_p}{AC - D^2}\Sigma^{-1}\mathbf{1}_m.$$

Finally, the value of the program is

$$\begin{aligned} var\left[\frac{w'(\hat{\pi})}{w}\right] &= \frac{\hat{\pi}^\top}{w}\Sigma\frac{\hat{\pi}}{w} = \frac{\hat{\pi}^\top}{w}\left(\frac{C\mu_p - D}{AC - D^2}\mu + \frac{A - D\mu_p}{AC - D^2}\mathbf{1}_m\right) \\ &= \frac{C\mu_p^2 - 2D\mu_p + A}{AC - D^2}, \end{aligned}$$

where we have used the fact that (i) $\frac{\hat{\pi}^\top}{w}\mu = \mu_p$ and (ii) $\frac{\hat{\pi}^\top}{w}\mathbf{1}_m = 1$. Rearranging terms leaves eq. (1.10) in the main text.

To check that the second-order conditions apply, consider the bordered Hessian

$$H = \begin{bmatrix} \frac{\partial^2\mathcal{L}}{\partial\pi\partial\pi} & \frac{\partial^2\mathcal{L}}{\partial\pi\partial\eta_1} & \frac{\partial^2\mathcal{L}}{\partial\pi\partial\eta_2} \\ \frac{\partial^2\mathcal{L}}{\partial\eta_1\partial\pi} & \frac{\partial^2\mathcal{L}}{\partial\eta_1\partial\eta_1} & \frac{\partial^2\mathcal{L}}{\partial\eta_1\partial\eta_2} \\ \frac{\partial^2\mathcal{L}}{\partial\eta_2\partial\pi} & \frac{\partial^2\mathcal{L}}{\partial\eta_2\partial\eta_1} & \frac{\partial^2\mathcal{L}}{\partial\eta_2\partial\eta_2} \end{bmatrix} = \begin{bmatrix} -2\eta_1\Sigma & -2\Sigma\pi & -\mathbf{1}_m \\ -2\pi\Sigma & 0 & 0 \\ \mathbf{1}_m^\top & 0 & 0 \end{bmatrix}.$$

It is negative (semi-) definite whenever the leading principal minors (formed through the last k columns and corresponding rows, for $k = 4, \cdots, m+2$) have determinants with signs that alternate, with the first one (formed with the last four rows and corresponding columns) having the sign of $(-1)^2 = +1$. This is possible whenever $\eta_1 > 0$, which is true by eq. (1.A6), whenever $AC > D^2$.

1.A.3 Efficient Portfolios Generate Efficient Portfolios

Let π_1 and π_2 be two portfolios on the efficient portfolio frontier. Then, $\pi_i = \ell_i\pi_d + (1 - \ell_i)\,\pi_{\text{gmv}}$ for some ℓ_i, and $i = 1, 2$. Solving for π_d and π_{gmv} leaves

$$\pi_d = \frac{1 - \ell_1}{\ell_2 - \ell_1}\pi_2 - \frac{1 - \ell_2}{\ell_2 - \ell_1}\pi_1, \quad \pi_{\text{gmv}} = \frac{\ell_2}{\ell_2 - \ell_1}\pi_1 - \frac{\ell_1}{\ell_2 - \ell_1}\pi_2.$$

Replacing these expressions for π_d and π_{gmv} into (1.12) leaves, for any given weight ℓ,

$$\hat{\pi} = \ell\pi_d + (1 - \ell)\,\pi_{\text{gmv}} = \omega(\ell_1, \ell_2)\,\pi_1 + (1 - \omega(\ell_1, \ell_2))\,\pi_2, \quad \omega(\ell_1, \ell_2) \equiv \frac{\ell_2 - \ell}{\ell_2 - \ell_1}.$$

That is, any arbitrary efficient portfolio can be generated by any other two arbitrary efficient portfolios, $\ell \neq \ell_1 \neq \ell_2$.

1.A.4 Covariance of Global Minimum Variance and Efficient Portfolio Returns

Let $\tilde{\mu}^p = \frac{\pi_p^\top}{w}\tilde{\mu}$ and $\tilde{\mu}^g = \frac{\pi_{\text{gmv}}^\top}{w}\tilde{\mu}$ be the returns that can be obtained by any portfolio π_p and π_{gmv}. The covariance of the global minimum variance portfolio with any other portfolio is

$$cov\left(\frac{\pi_p^\top}{w}\tilde{\mu}, \frac{\pi_{\text{gmv}}^\top}{w}\tilde{\mu}\right) = \frac{\pi_p^\top}{w} cov(\tilde{\mu}\tilde{\mu}^\top)\frac{\pi_{\text{gmv}}}{w} = \frac{\pi_p^\top}{w}\frac{\mathbf{1}_m}{C} = \frac{1}{C}.$$

Appendix 1.B The Market Portfolio and the Security Market Line

1.B.1 The Tangent Portfolio Is the Market Portfolio

Let us define the market capitalization for any asset i as the value of all the assets i that are outstanding in the market, namely,

$$\text{Cap}_i \equiv \bar{\theta}_i S_i, \quad i = 1, \cdots, m,$$

where $\bar{\theta}_i$ is the number of assets i outstanding in the market. The market capitalization of all the assets is simply

$$\text{Cap}_M \equiv \sum_{i=1}^{m} \text{Cap}_i.$$

The market portfolio, then, is the portfolio with relative weights given by

$$\bar{\pi}_{M,i} \equiv \frac{\text{Cap}_i}{\text{Cap}_M}, \quad i = 1, \cdots, m.$$

Next, suppose there are N investors and that each investor j has wealth w_j, which they invest in two funds: a safe asset and the tangent portfolio. Let w_j^f be the wealth investor j invests in the safe asset and $w_j - w_j^f$ the remaining wealth the investor invests in the tangent portfolio. The tangent portfolio is defined as $\bar{\pi}_T \equiv \frac{\pi_T}{w_j}$ for some π_T solution to [1.P2] and is obviously independent of w_j (see eq. (1.18) in the main text). The equilibrium in the stock market requires that

$$\text{Cap}_M \cdot \bar{\pi}_M = \sum_{j=1}^{N} \left(w_j - w_j^f\right)\bar{\pi}_T = \sum_{j=1}^{N} w_j \cdot \bar{\pi}_T = \text{Cap}_M \cdot \bar{\pi}_T,$$

where the second equality follows because the safe asset is in zero net supply and, hence, $\sum_{j=1}^{N} w_j^f = 0$ and the third equality holds because all the wealth in the economy is invested in stocks in equilibrium.

1.B.2 Tangency

We check that the CML and the efficient portfolio frontier have the same slope in correspondence of the market portfolio. Let us impose the following tangency condition of the CML to the efficient portfolio frontier in figure 1.2, AMC, at the point M:

$$\sqrt{\text{Sh}} = \frac{AC - D^2}{C\mu_M - D} v_M. \tag{1.A11}$$

The left-hand side (LHS) of this equation is the slope of the CML, obtained through eq. (1.6). The right-hand side (RHS) is the slope of the efficient portfolio frontier, obtained by differentiating $\mu_p(v)$ in the expression for the portfolio frontier in eq. (1.11) and setting $v = v_M$ in

$$\frac{d\mu_p(v)}{dv} = \sqrt{\left(Cv^2 - 1\right)^{-1}\left(AC - D^2\right)}v = \frac{AC - D^2}{C\mu_p(v) - D}v,$$

and where the second equality follows, again, by eq. (1.11). By eqs. (1.A11) and (1.17), we need to show that

$$\frac{C\mu_M - D}{AC - D^2} = \frac{1}{D - Cr}.$$

By plugging $\mu_M = r + \sqrt{\text{Sh}} \cdot v_M$ into the previous equality and rearranging terms,

$$v_M = \frac{\sqrt{\text{Sh}}}{D - Cr},$$

where we have made use of the equality $\text{Sh} = A - 2Dr + Cr^2$, obtained by elaborating on the definition of the Sharpe market performance Sh given in eq. (1.4). This is indeed the variance of the market portfolio given in eq. (1.17).

1.B.3 Alternative Derivation of the SML

The vector of covariances of the m asset returns with the market portfolio are

$$cov(\tilde{\mu}, \tilde{\mu}_M) \equiv cov(\tilde{\mu}, \tfrac{\pi_M^\top}{w}\tilde{\mu}) = cov(\tilde{\mu}\tilde{\mu}^\top)\tfrac{\pi_M}{w} = \Sigma\frac{\pi_M}{w} = \frac{1}{D - Cr}(\mu - \mathbf{1}_m r), \tag{1.A12}$$

where we have used the expression for the market portfolio given in eq. (1.18). Next, premultiply the previous equation by $\frac{\pi_M^\top}{w}$ in eq. (1.16), and obtain

$$v_M^2 = \frac{\pi_M^\top}{w}\Sigma\frac{\pi_M}{w} = \underbrace{\frac{(\mu - \mathbf{1}_m r)^\top \Sigma^{-1}}{\sqrt{\text{Sh}}}v_M}_{=\frac{\pi_M^\top}{w}} \cdot \underbrace{\frac{1}{D - Cr}(\mu - \mathbf{1}_m r)}_{=\Sigma\frac{\pi_M}{w}} \tag{1.A13}$$

or $v_M = \frac{\sqrt{\text{Sh}}}{D - Cr}$, which confirms eq. (1.17).

Let us rewrite eq. (1.A12) component by component. That is, for $i = 1, \cdots, m$,

$$\sigma_{iM} \equiv cov(\tilde{\mu}_i, \tilde{\mu}_M) = \frac{1}{D - Cr}(\mu_i - r) = \frac{v_M}{\sqrt{\text{Sh}}}(\mu_i - r) = \frac{v_M^2}{\mu_M - r}(\mu_i - r),$$

where the last two equalities follow by eq. (1.A13) and by the CML relation $\sqrt{\text{Sh}} = \frac{\mu_M - r}{v_M}$. By rearranging terms, we obtain eq. (1.21).

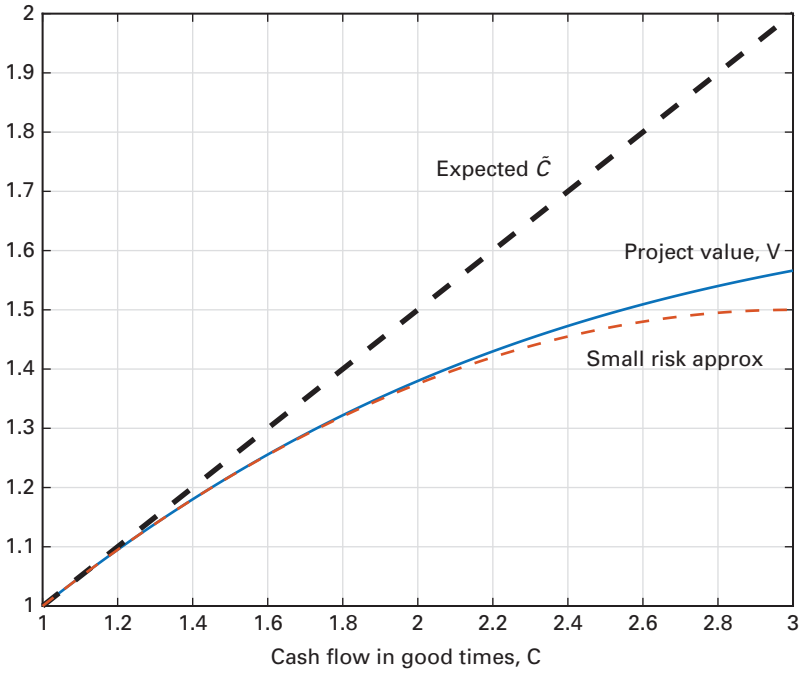

Figure 1.A1
The solid line depicts the value of the project, V, as a function of C, the project cash flow in good times. The thick dashed line depicts the expected value of the project, $E(\tilde{C}) = \frac{1}{2}(1 + C)$, and the thin dashed line is the "small risk approximation" to V in this example, obtained as $\hat{V} \equiv E(\tilde{C}) - \hat{\Lambda}$, where $\hat{\Lambda} \equiv \frac{1}{8}(C-1)^2$. The true risk premium is the difference $\Lambda = E(\tilde{C}) - V$.

Appendix 1.C Risk and Risk Aversion

1.C.1 Modeling and Approximating Risk Premiums

Project value in a representative investor model. We determine the value of the project discussed in section 1.3.4 in a simple model with a representative investor. This value is defined as the solution V to eq. (1.30), and we assume that the representative investor has initial wealth $w = 1$, exponential utility $u(w) = -e^{-\tau w}$ (with τ set equal to 1), and that the random cash flow $\tilde{C}$ is either 1 or $C > 1$ with equal probability. The value C is interpreted as the cash flow realization in "good times." In this example, a closed-form solution is not available for the value of V that satisfies eq. (1.30).

Figure 1.A1 depicts the numerical solution for V as a function of C. Note that due to risk aversion, V is lower than the expected cash flow, $E(\tilde{C}) = \frac{1}{2}(1 + C)$. An analytical approximation to the value of V based on eq. (1.33) is $\hat{V} \equiv E(\tilde{C}) - \hat{\Lambda}$, where the approximated premium, $\hat{\Lambda}$, is defined as

$$\Lambda \approx \hat{\Lambda} \equiv \frac{1}{2}\sigma_\epsilon^2 = \frac{1}{8}(C-1)^2$$

and the second equality follows by a simple calculation. Figure 1.A1 shows that the accuracy of the approximation, $\hat{V}$, deteriorates as the cash flow risk increases, that is,

as the deviation of the cash flow in the good state of nature increases from that in the bad (i.e., one). We now turn to provide the rationale underlying the approximation in eq. (1.33).

The approximation in eq. (1.33). Consider a second-order Taylor's approximation of the RHS of the equality in eq. (1.31) around $\epsilon = \Lambda = 0$,

$$E\left(u\left(w+\Lambda+\epsilon\right)\right) \approx u\left(w\right) + u'\left(w\right)\Lambda + \frac{1}{2}u''\left(w\right)\sigma_{\epsilon}^{2},$$

where σ_{ϵ}^{2} denotes the variance of ϵ and where we have disregarded second-order terms in Λ. Using this approximation in eq. (1.31) leaves eq. (1.33).

The approximation in eq. (1.35). The rationale behind this approximation is similar to that underlying eq. (1.33). By a second-order Taylor's approximation of the RHS of the equality in eq. (1.34) around $\epsilon = 0$,

$$E\left(u\left(w+E(\tilde{x})+\epsilon\right)\right) \approx u\left(w+E(\tilde{x})\right) + \frac{1}{2}u''\left(w+E(\tilde{x})\right)\sigma_{\epsilon}^{2},$$

and by a first-order approximation of the LHS,

$$u\left(w+E(\tilde{x})-\Pi\right) \approx u\left(w+E(\tilde{x})\right) - u'\left(w+E(\tilde{x})\right)\Pi.$$

Using these approximations in eq. (1.34) produces eq. (1.35).

The approximation in eq. (1.37). Approximate the RHS of the equality in eq. (1.36) with a second-order Taylor's expansion around $\tilde{g} = 0$,

$$E\left(u\left(\left(w+E(\tilde{x})\right)\left(1+\tilde{g}\right)\right)\right) \approx u\left(w+E(\tilde{x})\right) + \frac{1}{2}u''\left(w+E(\tilde{x})\right)\left(w+E(\tilde{x})\right)^{2}\sigma_{g}^{2},$$

and by a first-order approximation of the LHS,

$$u\left(w+E(\tilde{x})\left(1-\Pi_{r}\right)\right) \approx u\left(w+E(\tilde{x})\right) - u'\left(w+E(\tilde{x})\right)\left(w+E(\tilde{x})\right)\Pi_{r}.$$

Using these approximations in eq. (1.36) leaves eq. (1.37).

Certainty equivalents: examples. Consider a risk averse individual who is about to enter a gamble by which they will receive the outcome of rolling a dice. We assume that they have an initial wealth $w = 1$ and utility function $u\left(w\right) = \ln w$. The expect payoff from the gamble is $E\left(\tilde{x}\right) = \frac{1}{6}\sum_{j=1}^{6} j = 3.5$, and the variance is $\sigma_{\epsilon}^{2} = 2.9167$. The certainty equivalent is defined as

$$\mathrm{CE}\left(w,\tilde{x}\right) : u(w+\mathrm{CE}\left(w,\tilde{x}\right)) = E\left(u\left(w+\tilde{x}\right)\right) = \frac{1}{6}\sum_{j=1}^{6}\ln\left(1+j\right) = 1.4209,$$

where the last equality holds due to the assumption that $w = 1$. That is, $\mathrm{CE}\left(1,\tilde{x}\right) = 3.1408$. Accordingly, the (insurance) risk premium is $\Pi = E\left(\tilde{x}\right) - \mathrm{CE}\left(1,\tilde{x}\right) = 0.3592$, whereas its approximation in eq. (1.35) is

$$\hat{\Pi} = -\frac{1}{2}\frac{u''\left(w+E(\tilde{x})\right)}{u'\left(w+E(\tilde{x})\right)}\sigma_{\epsilon}^{2} = \frac{1}{2}\frac{1}{1+3.5}2.9167 = 0.3241.$$

Alternatively, assume the individual has exponential utility $u(w) = -e^{-\tau w}$, with $\tau = 1$, and initial wealth still equal to one. The certainty equivalent is now solution to

$$\mathrm{CE}(w, \tilde{x}) : u(w + \mathrm{CE}(w, \tilde{x})) = E(u(w + \tilde{x})) = \frac{1}{6} \sum_{j=1}^{6} (-e^{-(1+j)}) = -3.5594 \times 10^{-2}.$$

That is, $\mathrm{CE}(1, \tilde{x}) = 2.3356$, such that the risk premium is now equal to $\Pi = E(\tilde{x}) - \mathrm{CE}(1, \tilde{x}) = 3.5 - 2.3356 = 1.1644$, whereas its approximation through eq. (1.35) is $\hat{\Pi} = 1.4584$.

1.C.2 Stochastic Dominance and Mean-Preserving Spreads

Variance is the notion of risk underlying the classical CAPM and other models in this chapter. But there are situations studied in this book where the choices of an expected utility maximizer are well understood while relying on a generalized notion of risk due to Rothschild and Stiglitz (1970, 1971). Moreover, this notion of risk can be utilized to characterize quite simply the relation between asset prices and the volatility of fundamentals described in section 1.4.3.

We begin with the following definition.

Definition 1.A1 (First-order stochastic dominance). *$\tilde{x}_2$ dominates $\tilde{x}_1$ if, for any utility function u increasing and concave, we have that $E[u(\tilde{x}_2)] \geq E[u(\tilde{x}_1)]$.*

We have:

Theorem 1.A1. *The following statements are equivalent: (a) $\tilde{x}_2$ dominates $\tilde{x}_1$, or $E[u(\tilde{x}_2)] \geq E[u(\tilde{x}_1)]$ for every increasing function u; (b) for each $x > 0$, $F_2(x)$ is more likely than $F_1(x)$ to pay more than x, that is, $F_1(x) \geq F_2(x)$, where $F_i(x)$ denotes the cumulative distribution of x_i.*

Proof. We prove this result in the case the support is compact, say $[a, b]$. First, we show that $(b) \Rightarrow (a)$. By integrating by parts,

$$E[u(x)] = \int_a^b u(x) dF(x) = u(b) - \int_a^b u'(x) F(x) dx,$$

where we have used the fact that $F(a) = 0$ and $F(b) = 1$. Therefore,

$$E[u(\tilde{x}_2)] - E[u(\tilde{x}_1)] = \int_a^b u'(x) [F_1(x) - F_2(x)] dx.$$

Next, we show that $(a) \Rightarrow (b)$. Indeed (a) implies that $\int_a^b u(x) (dF_2(x) - dF_1(x)) \geq 0$ for $u(x) = \mathbb{I}_{x \geq y}$ and $y \in (a, b)$, or $0 \leq \int_a^b \mathbb{I}_{x \geq y} (dF_2(x) - dF_1(x)) = \int_y^b (dF_2(x) - dF_1(x)) = F_1(y) - F_2(y)$. □

There is an alternative characterization of first-order stochastic dominance. Suppose there exists a strictly positive random variable η by which $\tilde{x}_2$ exceeds $\tilde{x}_1$, namely,

$$\eta > 0 : \tilde{x}_2 = \tilde{x}_1 + \eta. \tag{1.A14}$$

Then, we have that $\forall t \in [a,b]$, $F_1(t) \equiv \Pr(\tilde{x}_1 \leq t) = \Pr(\tilde{x}_2 \leq t+\eta) \geq \Pr(\tilde{x}_2 \leq t) \equiv F_2(t)$. That is, $\tilde{x}_2$ dominates $\tilde{x}_1$ if it can be expressed as in eq. (1.A14). It is quite an intuitive property, but at the same time, it does not insulate the pure component of risk. Instead, we would like to perform the thought experiment to ask every expected utility maximizer to choose between two distributions with the same mean. Consider the following definition of second-order stochastic dominance.

Definition 1.A2 (Second-order stochastic dominance). *Let $\tilde{x}_1$ and $\tilde{x}_2$ be two random variables with the same expectation. We say that $\tilde{x}_1$ is more risky than $\tilde{x}_2$ if, for every concave function u, we have that $E[u(\tilde{x}_1)] \leq E[u(\tilde{x}_2)]$.*

Next, consider a case in which the distribution of one variable $\tilde{x}_1$ is obtained from that of another variable $\tilde{x}_2$, as follows: we take weights from the middle part of the density and move them toward the tails by making sure that the new density has the same mean as the initial. This is equivalent to requiring that condition (b) in the following theorem holds true. We have:

Theorem 1.A2. *The following statements are equivalent: (a) $\tilde{x}_1$ is more risky than $\tilde{x}_2$; (b) $\tilde{x}_1$ has more weight in the tails than $\tilde{x}_2$, i.e. $\forall t$, $\int_{-\infty}^{t} [F_1(x) - F_2(x)]\,dx \geq 0$.*

Proof. As for $(a) \Rightarrow (b)$, consider the function $-b_y(x) = -\max\{y-x, 0\}$. It is increasing and concave and, hence, a candidate utility function. Therefore, it satisfies

$$\int_a^b (-b_y(x))\left[dF_1(x) - dF_2(x)\right] \leq 0.$$

That is, using the definition of b_y,

$$\begin{aligned} 0 &\leq \int_a^y (y-x)\left[dF_1(x) - dF_2(x)\right] \\ &= y\left[F_1(y) - F_2(y)\right] - \int_a^y x\left[dF_1(x) - dF_2(x)\right] \\ &= \int_a^y \left[F_1(x) - F_2(x)\right] dx, \end{aligned}$$

where the last equality follows by an integration by parts. Next we prove that $(b) \Rightarrow (a)$. We have

$$\begin{aligned} E[u(\tilde{x}_1)] - E[u(\tilde{x}_2)] &= \int_a^b u(x)\left[f_1(x) - f_2(x)\right] dx \\ &= u(x)\left[F_1(x) - F_2(x)\right]\Big|_a^b - \int_a^b u'(x)\left[F_1(x) - F_2(x)\right] dx \\ &= -\int_a^b u'(x)\left[F_1(x) - F_2(x)\right] dx \\ &= -\left[u'(x)\left[\bar{F}_1(x) - \bar{F}_2(x)\right]\Big|_a^b - \int_a^b u''(x)\left[\bar{F}_1(x) - \bar{F}_2(x)\right] dx\right] \end{aligned}$$

$$= \int_a^b u''(x)\left[\bar{F}_1(x) - \bar{F}_2(x)\right]dx - u'(b)\left[\bar{F}_1(b) - \bar{F}_2(b)\right]$$
$$= \int_a^b u''(x)\left[\bar{F}_1(x) - \bar{F}_2(x)\right]dx,$$

where $\bar{F}_i(x) = \int_a^x F_i(u)du$. The last equality follows because, by integrating by parts,

$$-\int_a^b \left[F_1(x) - F_2(x)\right]dx = \int_a^b x\left[dF_1(x) - dF_2(x)\right] = 0,$$

where the last equality follows by the assumption that F_1 and F_2 have the same mean. Now, by $u'' < 0$ and $\bar{F}_1(x) > \bar{F}_2(x)$, the previous relation implies that $E[u(\tilde{x}_1)] < E[u(\tilde{x}_2)]$, that is, $\tilde{x}_1$ is more risky than $\tilde{x}_2$. □

Finally, we explain a link between the previous notions of risk and that of mean-preserving spreads utilized in section 1.4.4 of this chapter.

Mean-preserving spreads. We can now consider random variables that add up risk without affecting the mean: suppose that there exists a random variable $\epsilon : \tilde{x}_1$ has the same distribution as $\tilde{x}_2 + \epsilon$ and $E(\epsilon | \tilde{x}_2 = x_2) = 0$. We can think of an experiment in which after receiving a random payoff $\tilde{x}_2$, another random payoff could be offered to us, which has conditional expectation zero, thereby adding randomness without boosting the overall expected return. That is, $\tilde{x}_1$ is a *mean-preserving spread* of $\tilde{x}_2$. A variable that is a mean-preserving spread of another displays the properties stated in theorem 1.A2. Indeed,

$$\begin{aligned} E[u(\tilde{x}_1)] &= E[u(\tilde{x}_2 + \epsilon)] \\ &= E[E(u(\tilde{x}_2 + \epsilon) | \tilde{x}_2 = x_2)] \\ &\leq E[u(E(\tilde{x}_2 + \epsilon | \tilde{x}_2 = x_2))] \\ &= E[u(E(\tilde{x}_2 | \tilde{x}_2 = x_2))] \\ &= E[u(\tilde{x}_2)], \end{aligned}$$

where the inequality follows by concavity of u and by Jensen's inequality.

Appendix 1.D Money Demand and Liquidity Traps

We develop an example mentioned in the main text (see section 1.3.5.4) by which agents make dichotomic choices while either holding money or bonds based on their expectations on future interest rates. Assume the nominal yield curve is flat at i_0, and consider a perpetual bond with coupons equal to one, the price of which is $b_0 = \sum_{h=1}^{\infty} 1 \cdot (1 + i_0)^{-h} = i_0^{-1}$. There is a continuum of agents on $[0, 1]$, and each agent $j \in [0, 1]$ holds the belief that within a given period, i_0 will converge to a "normal" rate, say $i_e(j) \equiv 1/b_e(j)$. Therefore,

the agent believes that the return on this bond over this period is

$$R_j(i_0) \equiv \frac{b_e(j) - b_0 + 1}{b_0} = \frac{i_0}{i_e(j)} - 1 + i_0,$$

such that there exists a value of i_0 for each j, such that $R_j(i_0) = 0$ and given by

$$\hat{i}_0(j) \equiv \frac{i_e(j)}{1 + i_e(j)}.$$

This rate is monotonically increasing in the normal rate for agent j and is critical in that agent j holds money if $i_0 < \hat{i}_0(j)$ and invests in bonds if $i_0 > \hat{i}_0(j)$. In other words, an investor invests in bonds when they believe that interest rates will fall by an amount sufficient to generate positive profits. Without loss of generality, assume that $\hat{i}_0(j)$ is monotonically decreasing.

How do open market operations affect interest rates in this example? For a given i_0, set $\bar{j} : \hat{i}_0(\bar{j}) = i_0$, such that investors $j < \bar{j}$ (resp. $j > \bar{j}$) only hold money (resp. bonds). Next, consider a central bank bond purchase operation. Given the current level of i_0, investors who are currently holding bonds are obviously *not* willing to tender any of their holdings: bondholders would be incentivized to sell at a price higher than b_0. Consider, for example, the price $b_0' \equiv (i_0 - \epsilon)^{-1}$, for some $\epsilon > 0$; then, the investors' break-even condition becomes $R_j(i_0 - \epsilon) = 0$. Now, would-be bondholders are those j such that $j > \bar{j}' > \bar{j}$, that is, a mass of investors equal to $\bar{j}' - \bar{j}$ are willing to sell at b_0': money demand decreases with the nominal interest rates.

Note that if the current interest rate i_0 was so low that $i_0 = \hat{i}_0(1) \equiv \xi$, no agent would currently hold any bonds: everyone would now hold the belief that that interest rates may only rise. This situation is one of a *liquidity trap*: when $i_0 = \xi$, an increase in money supply would not affect interest rates; there are no investors in the market to purchase bonds from, except the marginal investor $j = 1$. So if the central bank increases money supply, the marginal investor is willing to accept this new money and tender the bonds as they are obviously indifferent between investing in bonds or hoarding money. Actually, if the central bank decreases money supply, the marginal investor would also be willing to buy these bonds.

This description illustrates the very well-known Keynes's (1936) point that money demand should be negatively sloped at an aggregate level. Consider the following simple example, in which $i_e(j)$ is uniformly distributed, that is,

$$i_e(j) = \bar{\xi} - (\bar{\xi} - \underline{\xi})j, \quad j \in [0, 1]$$

for some two constants $\underline{\xi}$ and $\bar{\xi} > \underline{\xi}$. Then,

$$\hat{i}_0(j) = \frac{\bar{\xi} - (\bar{\xi} - \underline{\xi})j}{1 + \bar{\xi} - (\bar{\xi} - \underline{\xi})j}.$$

The jth agent money demand, $m^d(j)$ say, is dichotomic, in that

$$m^d(j) = \begin{cases} 1, & \text{if } \hat{i}_0(j) > i_0 \\ 0, & \text{otherwise} \end{cases}.$$

Yet aggregate money demand is

$$M^d \equiv \int_0^1 m^d(j)\,dj = \int_{j:\hat{i}_0(j)>i_0} dj = \frac{\bar{\xi} - (1+\bar{\xi})\,i_0}{(\bar{\xi} - \underline{\xi})(1-i_0)},$$

Provided $i_0 < \frac{\bar{\xi}}{1+\bar{\xi}}$, aggregate money demand is positive. Moreover, it is always decreasing in i_0.

Appendix 1.E Parameter Uncertainty

Proofs of eq. (1.49). By the projection theorem (see appendix 10.A in chapter 10),

$$E(\mu|\mathcal{V}) = E(\mu) + cov(\mu,\mathcal{V})\,var(\mathcal{V})^{-1}(\mathcal{V} - E(\mathcal{V})),$$

and

$$var(\mu|\mathcal{V}) = var(\mu) - cov(\mu,\mathcal{V})\,var(\mathcal{V})^{-1}\,cov(\mu,\mathcal{V})^{\intercal},$$

where, by (1.48), $cov(\mu,\mathcal{V}) = cov(\mu, P\mu - \epsilon_v) = CP^{\intercal}$ and $var(\mathcal{V}) = PCP^{\intercal} + \Omega$. Therefore,

$$E(\mu|\mathcal{V}) = \bar{\mu} + CP^{\intercal}\left(PCP^{\intercal} + \Omega\right)^{-1}(\mathcal{V} - P\bar{\mu}), \quad var(\mu|\mathcal{V}) = C - CP^{\intercal}\left(PCP^{\intercal} + \Omega\right)^{-1}PC^{\intercal}.$$

The distribution in (1.49) follows by rearranging terms and using the definition $C = c\Sigma$. □

Proofs of eq. (1.54). Consider the Lagrangian function for the program [1.P4], where M is as in (1.53):

$$\mathcal{L} = \pi^{\top}(\mu - \mathbf{1}_m r) + R - \frac{\tau}{2}\pi^{\top}\Sigma\pi - \lambda\left(\eta - \left(\hat{\mu} - \mu\right)^{\top}\Sigma^{-1}\left(\hat{\mu} - \mu\right)\right), \tag{1.A15}$$

where λ is the Lagrange multiplier. The first-order conditions for μ leads to $\hat{\mu} - \mu_* = \frac{1}{2\lambda}\Sigma\pi$, where μ_* is the optimal value for μ; replacing this expression and the first-order condition for λ into (1.A15) leaves

$$\mathcal{L} = \pi^{\top}\left(\hat{\mu} - \mathbf{1}_m r\right) + R - \frac{\tau}{2}\pi^{\top}\Sigma\pi - \frac{1}{2\lambda}\pi^{\top}\Sigma\pi. \tag{1.A16}$$

The expression for the Lagrange multiplier, λ_* say, is obtained while using the constraint

$$\eta = \left(\hat{\mu} - \mu_*\right)^{\top}\Sigma^{-1}\left(\hat{\mu} - \mu_*\right) = \left(\frac{1}{2\lambda}\right)^2 \pi^{\top}\Sigma\pi,$$

leaving $2\lambda^* = \sqrt{\frac{\pi^{\top}\Sigma\pi}{\eta}}$. The objective function in eq. (1.54) in the main text follows by replacing λ_* into (1.A16). The second equality follows because the first-order conditions for this problem lead to

$$\hat{\pi} = \frac{\sigma_P}{\tau\sigma_P + \sqrt{\eta}}\Sigma^{-1}\left(\hat{\mu} - \mathbf{1}_m r\right), \quad \sigma_P \equiv \sqrt{\pi^{\top}\Sigma\pi}.$$

Replacing the expression for $\hat{\pi}$ into that for σ_P leads to the following equation satisfied by σ_P^2,

$$\sigma_P^2 = \left(\frac{\sigma_P}{\tau\sigma_P + \sqrt{\eta}}\right)^2 \mathrm{Sh}_*,$$

where Sh_* is defined in the main text. That is, $\tau\sigma_P + \sqrt{\eta} = \sqrt{\mathrm{Sh}_*}$. □

References

Abel, A. B. (1988). "Stock Prices under Time-Varying Dividend Risk: An Exact Solution in an Infinite-Horizon General Equilibrium Model." *Journal of Monetary Economics* 22, 375–393.

Ang, A. (2014). *Asset Management: A Systematic Approach to Factor Investing*. Oxford: Oxford University Press.

Arrow, K. J. (1965). "The Theory of Risk Aversion." In *Aspects of the Theory of Risk Bearing*. Helsinki: Yrjö Jahnssonin Säätiö.

Bali, T. G., R. F. Engle, and S. Murray (2016). *Empirical Asset Pricing: The Cross Section of Stock Returns*. Hoboken, NJ: John Wiley.

Banz, R. W. (1981). "The Relationship Between Return and Market Value of Common Stocks." *Journal of Financial Economics* 9, 3–18.

Barsky, R. B. (1989). "Why Don't the Prices of Stocks and Bonds Move Together?" *American Economic Review* 79, 1132–1145.

Black, F. (1972). "Capital Market Equilibrium with Restricted Borrowing." *Journal of Business* 45, 444–454.

Black, F., and R. Litterman (1991). "Asset Allocation: Combining Investor Views with Market Equilibrium." *Journal of Fixed Income* 1, no. 2, 7–18.

Black, F., M. C. Jensen, and M. Scholes (1972). "The Capital Asset Pricing Model: Some Empirical Tests." In *Studies in the Theory of Capital Markets*, edited by M. C. Jensen, 79–121. New York: Praeger.

Campbell, J. Y., and L. M. Viceira (2002). *Strategic Asset Allocation*. Oxford: Oxford University Press.

Carhart, M. (1997). "On Persistence of Mutual Fund Performance." *Journal of Finance* 52, 57–82.

Chen, N.-F., R. Roll, and S. A. Ross (1986). "Economic Forces and the Stock Market." *Journal of Business* 59, 383–403.

Connor, G. (1984). "A Unified Beta Pricing Theory." *Journal of Economic Theory* 34, 13–31.

Dow, J., and S. Werlang (1992). "Uncertainty Aversion, Risk Aversion, and the Optimal Choice of Portfolio." *Econometrica* 60, 197–204.

Fama, E. F., and K. R. French (1992). "The Cross-Section of Expected Stock Returns." *Journal of Finance* 47, 427–465.

Fama, E. F., and K. R. French (1993). "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33, 3–56.

Fama, E. F., and J. D. MacBeth (1973). "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 38, 607–636.

Frazzini, A., and L. H. Pedersen (2014). "Betting Against Beta." *Journal of Financial Economics* 111, 1–25.

Garlappi, L., R. Uppal, and T. Wang (2007). "Portfolio Selection with Parameter and Model Uncertainty: A Multi-Prior Approach." *Review of Financial Studies* 20, 41–81.

Gilboa, I., and D. Schmeidler (1989). "Maxmin Expected Utility with a Non-Unique Prior." *Journal of Mathematical Economics* 18, 141–153.

Gouriéroux, C., and A. Monfort (2008). *Statistics and Econometric Models*. Vol. 2, *Themes in Modern Econometrics*. Cambridge, UK: Cambridge University Press.

Grossman, S. J., and J. E. Stiglitz (1980). "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70, 393–408.

Harvey, C. R., Y. Liu, and H. Zhu (2016). "... and the Cross-Section of Expected Returns." *Review of Financial Studies* 29, 5–68.

Heckman, J. (1979). "Sample Selection Bias as a Specification Error." *Econometrica* 47, 153–161.

Huang, C.-F., and R. H. Litzenberger (1988). *Foundations for Financial Economics*. New York: North-Holland.

Huberman, G. (1983). "A Simplified Approach to Arbitrage Pricing Theory." *Journal of Economic Theory* 28, 1983–1991.

Jegadeesh, N., and S. Titman (1993). "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *Journal of Finance* 48, 65–91.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. London: Palgrave Macmillan.

Kyle, A. S. (1985). "Continuous Auctions and Insider Trading." *Econometrica* 53, 1335–1335.

Lintner, J. (1965). "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *Review of Economics and Statistics* 47, 13–37.

Malkiel, B. (1979). "The Capital Formation Problem in the United States." *Journal of Finance* 34, 291–306.

Markovitz, H. (1952). "Portfolio Selection." *Journal of Finance* 7, 77–91.

Meucci, A. (2005). *Risk and Asset Allocation*. New York: Springer Verlag.

Mossin, J. (1966). "Equilibrium in a Capital Asset Market." *Econometrica* 34, 768–783.

Pedersen, L. H. (2015). *Efficiently Inefficient*. Princeton, NJ: Princeton University Press.

Pindyck, R. (1984). "Risk, Inflation and the Stock Market." *American Economic Review* 74, 335–351.

Popper, K. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

Poterba, J., and L. Summers (1985). "The Persistence of Volatility and Stock Market Fluctuations." *American Economic Review* 75, 1142–1151.

Pratt, J. W. (1964). "Risk Aversion in the Small and in the Large." *Econometrica* 32, 122–136.

Roll, R. (1977). "A Critique of the Asset Pricing Theory's Tests Part I: On Past and Potential Testability of the Theory." *Journal of Financial Economics* 4, 129–176.

Roncalli, T. (2014). *Introduction to Risk Parity and Budgeting*. London: Chapman & Hall.

Rosenberg, B., K. Reid, and R. Lanstein (1985). "Persuasive Evidence of Market Inefficiency." *Journal of Portfolio Management* 11, 9–17.

Ross, S. (1976). "Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory* 13, 341–360.

Rothschild, M., and J. Stiglitz (1970). "Increasing Risk: I. A Definition." *Journal of Economic Theory* 2, 225–243.

Rothschild, M., and J. Stiglitz (1971). "Increasing Risk: II. Its Economic Consequences." *Journal of Economic Theory* 5, 66–84.

Shanken, J., and M. I. Weinstein (2006). "Economic Forces and the Stock Market Revisited." *Journal of Empirical Finance* 13, 129–144.

Sharpe, W. F. (1964). "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance* 19, 425–442.

Stattman, D. (1980). "Book Values and Stock Returns." *The Chicago MBA: A Journal of Selected Papers* 4, 25–45.

Tobin, J. (1958). "Liquidity Preference as Behavior Towards Risk." *Review of Economic Studies* 25, 65–86.