R. Jarrow et al., Eds., *Handbooks in OR & MS, Vol. 9*

Chapter 1

Portfolio Theory

G.M. Constantinides

Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637, U.S.A.; and NBER

A.G. Malliaris

School of Business Administration, Loyola University of Chicago, 820 North Michigan Avenue, Chicago, IL 60611, U.S.A.

1. Introduction

Consider a consumer with a given amount of income. Such a consumer typically faces two important economic decisions. First, how to allocate his or her current consumption among goods and services. Second, how to invest among various assets. These two interrelated consumer or household problems are known as the consumption-saving decision and the portfolio selection decision.

Beginning with Adam Smith, economists have systematically studied the first decision. Arguing that a consumer will choose commodities and services that offer the greatest marginal utility relative to price, a theory of value was developed that combines subjective notions from consumer utility with objective notions from the production theory of the firm. By the beginning of the twentieth century, neoclassical economists had developed a static theory of consumer behavior as part of an analysis of market pricing under conditions of perfect competition and certainty.

The asset allocation decision was not adequately addressed by neoclassical economists, probably because they treated savings as the supply of loanable funds in developing a theory of interest rate determination instead of portfolio selection. More importantly, however, these two decisions, although closely interrelated, require substantially different methodologies. The methodology of deterministic calculus is adequate for the decision of maximizing a consumer's utility subject to a budget constraint. Portfolio selection involves making a decision under uncertainty. The probabilistic notions of expected return and risk become very important. Neoclassical economists did not have such a methodology available to them and despite some very early attempts by probabilists, like Bernoulli [1738] to define and measure risk, or Irving Fisher [1906] to describe asset returns in terms of a probability distribution, the twin concepts of expected return and risk had not yet been fully integrated. An early and important attempt to do that was made by

Marschak [1938] who expressed preferences for investment by indifference curves in the mean–variance space.[1]

The methodological breakthrough of treating axiomatically the theory of choice under uncertainty was offered by von Neumann & Morgenstern [1947] and it was only a few years later that Markowitz [1952, 1959] and Tobin [1958], used this theory to formulate and solve the portfolio selection problem.

In this essay we plan to exposit portfolio theory with a special emphasis on its historical evolution and methodological foundations. In Section 2, we describe the early work of Markowitz [1952, 1959] and Tobin [1958] to illustrate the individual contributions of these authors. Following these general remarks about the early beginning of portfolio theory, we define and solve the mean–variance portfolio problem in Section 3 and relate it to its most famous intellectual first fruits, namely the two-fund separation and the capital asset pricing theory of Sharpe [1964] and Lintner [1965] in Sections 4, 5 and 6. In particular, a portion of Section 6 is devoted to the presentation of Roll's [1977] critique of the asset pricing theory's tests and the interplay of analysis and empirical testing. This leads to an analysis of the foundational assumptions of portfolio theory with respect to investor preferences and asset return distributions, both reviewed in Section 7. The contrast of methodologies is illustrated in Sections 8 and 9 where stochastic calculus and stochastic control techniques are used to generalize the consumption-investment problem to an arbitrary number of periods. Market imperfections are addressed in Section 10. The last section identifies several extensions and refers the reader to several articles, some included in this volume. It also contains our summary and conclusions.

2. The early contributions

Markowitz [1952] marks the beginning of modern portfolio theory, where for the first time, the problem of portfolio selection is clearly formulated and solved. Earlier contributions of Keynes [1936], Marschak [1938] and others only tangentially analyze investment decisions. Markowitz's focus is the explanation of the phenomenon of portfolio diversification.

Before Markowitz could propose the "expected returns–variance of returns" rule, he first had to discredit the then widely accepted principle that an investor chooses a portfolio by selecting securities that maximize discounted expected returns.[2] Markowitz points out that if an investor follows this rule, his or her

[1] Marschak [1938, p. 312] recognizes that "the unsatisfactory state of Monetary Theory as compared with General Economics is due to the fact that the principle of determinateness so well established by Walras and Pareto for the world of perishable consumption goods and labor services has never been applied with much consistency to durable goods and, still less, to claims (securities, loans, cash)". In our modern terminology we could replace the names Monetary Theory and General Economics with Financial Economics and Microeconomic Theory, respectively.

[2] Markowitz refers the reader to a standard investments textbook by Williams [1938] that elaborates the notion that portfolio choice is guided by the rule of maximizing the discounted

portfolio will consist of only one stock, namely the one that has the highest discounted expected return which is contrary to the observed phenomenon of diversification. Therefore a rule of investor behavior which does not yield portfolio diversification must be rejected. Furthermore, the rejection of this rule holds no matter how expectations of future returns are formed and how discount rates are selected. Markowitz then proposes the expected mean returns–variance of returns M–V rule. He concludes that the M–V rule not only implies diversification, it actually implies the right kind of diversification for the right reason. In trying to reduce the portfolio variance, it is not enough to just invest in many securities. It is important to diversify across securities with low return covariances. In 1959, Markowitz published a monograph on the same topic. In the last part (consisting of four chapters) and in an appendix, portfolio selection is grounded firmly as rational choice under uncertainty.

In contrast to Markowitz's contributions which may be viewed as microeconomic, Tobin [1958] addresses a standard Keynesian macroeconomic problem, namely liquidity preference. Keynes [1936] used the concept of liquidity preference to describe an inverse relationship between the demand for cash balances and the rate of interest. This aggregative function was postulated by Keynes without a formal derivation. Tobin derives the economy's liquidity preference by developing a theory that explains the behavior of the decision-making units of the economy.[3]

Numerous contributions followed. To mention just a few, Sharpe [1970], Merton [1972], Gonzalez-Gaverra [1973], Fama [1976] and Roll [1977], are important references. Ziemba & Vickson [1975] have collected numerous classic articles on both static and dynamic models of portfolio selection. The recent books by Ingersoll [1987], Huang & Litzenberger [1988], and Jarrow [1988] also contain a useful analysis of the mean–variance portfolio theory. Our exposition relies heavily on Roll [1977].

value of future returns. It is not correct to deduce that earlier economists completely ignored the notion of risk. They simply were unsuccessful in developing a precise microeconomic theory of investor behavior under conditions of risk. The typical way risk was accounted for in Keynes' [1936] marginal efficiency of investment or Hicks' [1939] development of the investment decisions of a firm was by letting expected future returns include an allowance for risk or by adding a risk premium to discount rates.

[3] One may wonder what is the connection between liquidity preference and portfolio theory. You may recall that Keynes identified three motives for holding cash balances: transactions, precautionary and speculative. Furthermore, while the transactions and precautionary motives were determined by income, the amount of cash balances held for speculative purposes was influenced by the rate of interest. Tobin analyzes this speculative motive of investors to offer a theoretically sound foundation of the interest elasticity of the liquidity preference. Because he wishes to explain the demand for cash, he considers an investor whose portfolio selection includes only two assets: cash and consoles. Of course, the yield of cash is zero while the yield of consoles is positive. Tobin posits and solves a two-asset portfolio selection problem using a quadratic expected utility function. He justifies his choice of a quadratic utility function by arguing that the investor considers two parameters in his or her portfolio selection: expected return and risk (measured by the standard deviation of the portfolio return). Finally, having developed his portfolio selection theory, he applies it to show that changes in real interest rates affect inversely the demand for cash, which is what Keynes had conjectured without offering a proof.

3. Mean–variance portfolio selection

In the formulation of the mean–variance portfolio we use the following notation: $\boldsymbol{x}$ is an n-column vector whose components $x_1, \ldots x_n$ denote the weight or proportion of the investor's wealth allocated to the ith asset in the portfolio with $i = 1, 2, \ldots n$. Obviously the sum of weights is equal to 1, i.e. $\sum_1^n x_i = 1$; $\mathbf{1}$ is an n-column vector of ones and superscript T denotes the transpose of a vector or a matrix. $\boldsymbol{R}$ is an n-column vector of mean returns $R_1, \ldots, R_n$ of the n assets, where it is assumed that not all elements of $\boldsymbol{R}$ are equal, and $\mathbf{V}$ is the $n \times n$ covariance matrix with entries σ_{ij}, $i, j = 1, 2, \ldots n$. We assume that $\mathbf{V}$ is nonsingular. This essentially requires that none of the asset returns is perfectly correlated with the return of a portfolio made up of the remaining assets; and that none of the assets or portfolios of the assets is riskless. The case where one of the assets is riskless will be treated separately at a later stage. Observe that $\mathbf{V}$ is symmetric and positive definite being a covariance matrix. We say that an $n \times n$ matrix $\mathbf{V}$ is *positive definite*, if for any nonzero n-vector $\boldsymbol{x}$, it follows that $\boldsymbol{x}^{\mathrm{T}}\mathbf{V}\boldsymbol{x} > 0$. In our case the property of positive definiteness of $\mathbf{V}$ follows from the fact that variances of risky portfolios are strictly positive. The mean returns and covariance matrix of the assets are assumed to be known. We do not specify if n denotes the entire population or just a sample of assets. Finally, for a given portfolio p, its variance, denoted by σ_p^2, is given by $\boldsymbol{x}^{\mathrm{T}}\mathbf{V}\boldsymbol{x}$, while the portfolio mean, denoted by R_p, is given by $R_p = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{R}$.

Much in the spirit of Markowitz's [1952] formulation[4] the portfolio selection problem can be stated as

$$\begin{aligned} \text{minimize} \quad & \sigma_p^2 = \boldsymbol{x}^{\mathrm{T}}\mathbf{V}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{x}^{\mathrm{T}}\mathbf{1} = 1 \\ & \boldsymbol{x}^{\mathrm{T}}\boldsymbol{R} = R_p. \end{aligned} \tag{3.1}$$

In problem (3.1) we minimize the portfolio variance σ_p^2 subject to two constraints: first, the portfolio weights must sum to unity, which means that all the wealth is invested, and second the portfolio must earn an expected rate of return equal to R_p. Technically, we minimize a convex function subject to linear constraints. Observe that $\boldsymbol{x}^{\mathrm{T}}\mathbf{V}\boldsymbol{x}$ is convex because $\mathbf{V}$ is positive definite and also note that the two linear constraints define a convex set. Therefore, the problem has a unique solution and we only need to obtain the first-order conditions.

Two remarks are appropriate. First, the investor's preferences, as represented by a utility function, do not enter explicitly in (3.1). We only assume that a utility function exists which is defined over the mean and variance of the portfolio return and which has the further property of favoring higher mean and smaller variance. Second, unlike Tobin who explicitly considers cash in his portfolio selection

[4] Markowitz [1952] considers only three securities because he solves the same problem as (3.1) using geometric methods. He does not allow short sales in order to simplify the analysis. In (3.1) short sales are permitted, which means that portfolio weights are allowed to be negative.

problem, (3.1) does not include a riskless asset. A riskless asset will be included in Section 5.

Form the Lagrangian function

$$L = \boldsymbol{x}^{\mathrm{T}}\mathbf{V}\boldsymbol{x} - \lambda_1(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{R} - R_p) - \lambda_2(\boldsymbol{x}^{\mathrm{T}}\mathbf{1} - 1). \tag{3.2}$$

The first-order conditions are

$$\frac{\partial L}{\partial \boldsymbol{x}} = 2\mathbf{V}\boldsymbol{x} - \lambda_1\boldsymbol{R} - \lambda_2\mathbf{1} = \mathbf{0}, \tag{3.3}$$

where $\mathbf{0}$ in (3.3) is an n-vector of zeros, and

$$\frac{\partial L}{\partial \lambda_1} = R_p - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{R} = 0, \tag{3.4}$$

$$\frac{\partial L}{\partial \lambda_2} = 1 - \boldsymbol{x}^{\mathrm{T}}\mathbf{1} = 0. \tag{3.5}$$

From equation (3.3) we obtain

$$\boldsymbol{x} = \frac{1}{2}\mathbf{V}^{-1}(\lambda_1\boldsymbol{R} + \lambda_2\mathbf{1}) = \frac{1}{2}\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix}. \tag{3.6}$$

In this last equation the term $\lambda_1\boldsymbol{R}+\lambda_2\mathbf{1}$ is written in a matrix form because we will use (3.4) and (3.5) to solve for $\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix}$. Doing this we write (3.4) and (3.5) as

$$[\boldsymbol{R}\ \ \mathbf{1}]^{\mathrm{T}}\boldsymbol{x} = \begin{bmatrix}R_p\\ 1\end{bmatrix}. \tag{3.7}$$

Premultiply both sides of (3.6) by $[\boldsymbol{R}\ \ \boldsymbol{1}]^{\mathrm{T}}$ and use (3.7) to obtain

$$[\boldsymbol{R}\ \ \mathbf{1}]^{\mathrm{T}}\boldsymbol{x} = \frac{1}{2}[\boldsymbol{R}\ \ \mathbf{1}]^{\mathrm{T}}\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix} = \begin{bmatrix}R_p\\ 1\end{bmatrix}. \tag{3.8}$$

For notational convenience denote by

$$\mathbf{A} \equiv [\boldsymbol{R}\ \ \mathbf{1}]^{\mathrm{T}}\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}] \tag{3.9}$$

the 2×2 symmetric matrix with entries

$$\begin{bmatrix}a & b\\ b & c\end{bmatrix} = \begin{bmatrix}\boldsymbol{R}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{R} & \boldsymbol{R}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{1}\\ \boldsymbol{R}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{1} & \mathbf{1}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{1}\end{bmatrix}. \tag{3.10}$$

We need to establish that $\mathbf{A}$ is positive definite. For any y_1, y_2 such that at least one of the elements y_1, y_2 is nonzero, observe that

$$[\boldsymbol{R}\ \ \mathbf{1}]\begin{bmatrix}y_1\\ y_2\end{bmatrix} = [y_1\boldsymbol{R} + y_2\mathbf{1}]$$

is a nonzero n-vector because, by assumption, the elements of $\boldsymbol{R}$ are not all equal.

Then $\mathbf{A}$ is positive definite because

$$\begin{aligned}[y_1\ \ y_2]\mathbf{A}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} &= [y_1\ \ y_2][\boldsymbol{R}\ \ \mathbf{1}]^{\mathrm{T}}\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \\ &= [y_1\boldsymbol{R} + y_2\mathbf{1}]^{\mathrm{T}}\mathbf{V}^{-1}[y_1\boldsymbol{R} + y_2\mathbf{1}] > 0\end{aligned}$$

by the positive definiteness of $\mathbf{V}^{-1}$.

Substitute the newly defined $\mathbf{A}$ in (3.9) to get

$$\frac{1}{2}\mathbf{A}\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} R_p \\ 1 \end{bmatrix}$$

from which we can immediately solve for the multipliers since $\mathbf{A}$ is nonsingular and its inverse exists. Thus

$$\frac{1}{2}\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \mathbf{A}^{-1}\begin{bmatrix} R_p \\ 1 \end{bmatrix}. \tag{3.11}$$

From these manipulations we obtain the desired result using (3.11) and (3.6). Thus, the n-vector of portfolio weights $\boldsymbol{x}$ that minimizes portfolio variance for a given mean return is

$$\boldsymbol{x} = \frac{1}{2}\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_p \\ 1 \end{bmatrix}. \tag{3.12}$$

The result of this analysis can be stated as:

Theorem 3.1 (Mean–variance portfolio selection). *Let* $\mathbf{V}$ *be the* $n \times n$ *positive definite covariance matrix and* $\boldsymbol{R}$ *be the n-column vector of mean returns of the n assets where it is assumed that not all elements of* $\boldsymbol{R}$ *are equal. Then the minimum variance portfolio with given mean return* R_p *is unique and its weights are given by (3.12).*

Let us compute the variance of any minimum variance portfolio with a given mean R_p. Using the definitions of the variance σ_p^2, matrix $\mathbf{A}$ in (3.9) and the solution of weights in (3.12), calculate

$$\begin{aligned}\sigma_p^2 &= \boldsymbol{x}^{\mathrm{T}}\mathbf{V}\boldsymbol{x} = [R_p\ \ 1]\mathbf{A}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_p \\ 1 \end{bmatrix} \\ &= [R_p\ \ 1]\mathbf{A}^{-1}\begin{bmatrix} R_p \\ 1 \end{bmatrix} \\ &= [R_p\ \ 1]\frac{1}{(ac-b^2)}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}\begin{bmatrix} R_p \\ 1 \end{bmatrix} \\ &= \frac{a - 2bR_p + cR_p^2}{(ac-b^2)}.\end{aligned} \tag{3.13}$$

In (3.13) the relation between the variance of the minimum variance portfolio σ_p^2 for any given mean R_p is expressed as a parabola and is called the *minimum variance portfolio frontier* or *locus*. In mean–standard-deviation space the relation is expressed as a hyperbola.

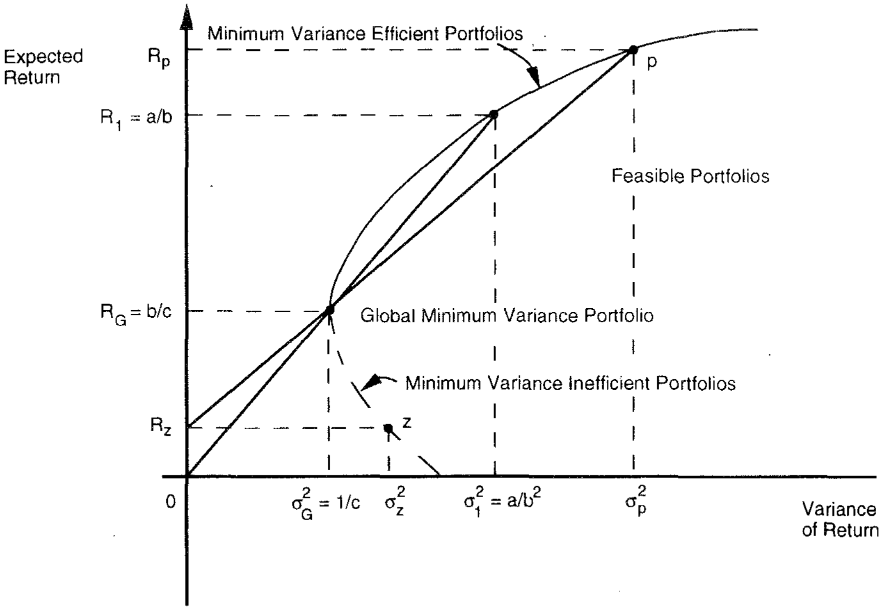

Fig. 1. Portfolios of n risky assets.

Figure 1 graphs equation (3.13) and distinguishes between the upper half (solid curve) and the bottom half (broken curve). The upper half of the minimum variance portfolio frontier identifies the set of portfolios having the highest return for a given variance; these are called mean–variance *efficient portfolios*. The portfolios on the bottom half are called *inefficient portfolios*. The mean–variance efficient portfolios are a subset of the minimum variance portfolios. Portfolios to the right of the parabola are called *feasible*. For a given variance the mean return of a feasible portfolio is less than the mean return of an efficient portfolio and higher than the mean return of an inefficient one, both having the same variance.

Figure 1 also identifies the *global minimum variance portfolio*. This is the portfolio with the smallest possible variance for any mean return. Its mean, denoted by R_{G} is obtained by minimizing (3.13) with respect to R_p, to yield

$$R_{\mathrm{G}} = \frac{b}{c} \tag{3.14}$$

and its variance, denoted by σ_{G}^2, is calculated by inserting (3.14) into the general equation (3.13) to obtain

$$\sigma_{\mathrm{G}}^2 = \frac{a - 2bR_{\mathrm{G}} + cR_{\mathrm{G}}^2}{ac - b^2} = \frac{a - 2b(b/c) + c(b/c)^2}{ac - b^2} = \frac{1}{c}. \tag{3.15}$$

Similarly, by inserting R_G from (3.14) into (3.12) we find the weights of the global minimum variance portfolio, denoted by $\boldsymbol{x}_{\mathrm{G}}$,

$$\boldsymbol{x}_{\mathrm{G}} = \mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_{\mathrm{G}} \\ 1 \end{bmatrix} = \frac{\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}\begin{bmatrix} b/c \\ 1 \end{bmatrix}}{(ac-b^2)} = \frac{\mathbf{V}^{-1}\ \mathbf{1}}{c}. \tag{3.16}$$

An additional notion that will be used later in this section and which is illustrated in Figure 1 also, is the concept of an orthogonal portfolio. We say that two minimum variance portfolios $\boldsymbol{x}_p$ and $\boldsymbol{x}_z$ are *orthogonal* if their covariance is zero, that is,

$$\boldsymbol{x}_z^{\mathrm{T}}\mathbf{V}\boldsymbol{x}_p = 0. \tag{3.17}$$

We want to show that for every minimum variance portfolio, except the global minimum variance portfolio, we can find a unique orthogonal minimum variance portfolio. Furthermore, if the first portfolio has mean R_p, its orthogonal one has mean R_z with

$$R_z = \frac{a - bR_p}{b - cR_p}. \tag{3.18}$$

To establish (3.18), let first p and z be two arbitrary minimum variance portfolios with weights $\boldsymbol{x}_p$ given by (3.12) and $\boldsymbol{x}_z$ given by

$$\boldsymbol{x}_z = \mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_z \\ 1 \end{bmatrix}. \tag{3.19}$$

The covariance between portfolios p and z, being zero implies

$$0 = \boldsymbol{x}_z^{\mathrm{T}}\mathbf{V}\boldsymbol{x}_p = \begin{bmatrix} R_z & 1 \end{bmatrix}\mathbf{A}^{-1}\begin{bmatrix} R_p \\ 1 \end{bmatrix}, \tag{3.20}$$

from which (3.18) follows.

In Figure 1, we also illustrate the geometry of orthogonal portfolios. Given an arbitrary efficient portfolio p on the efficient portfolio frontier, the line passing between p and the global minimum variance portfolio can be shown to intersect the expected return axis at R_z. Once R_z is known, then the orthogonal portfolio z can be uniquely identified on the minimum variance portfolio frontier. Note that if a portfolio p is efficient and therefore lies on the positively sloped segment of the portfolio frontier, as in Figure 1, then its orthogonal portfolio z is inefficient and lies on the negatively sloped segment. In general, orthogonal portfolios lie on opposite-sloped segments of the portfolio frontier.

4. Two-fund separation

We now present the important property of two-fund separation. The mathematics of this property is straightforward; its economic implications however are significant because the following theorem establishes that the minimum variance portfolio frontier can be generated by any two distinct frontier portfolios.

Theorem 4.1 (Two-fund separation). *Let $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ be two minimum variance portfolios with mean returns R_a and R_b respectively, such that $R_a \neq R_b$.*

(a) Then every minimum variance portfolio $\boldsymbol{x}_c$ is a linear combination of $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$.

(b) Conversely, every portfolio which is a linear combination of $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, i.e, $\alpha\boldsymbol{x}_a + (1-\alpha)\boldsymbol{x}_b$, is a minimum variance portfolio.

(c) In particular, if $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ are minimum variance efficient portfolios, then $\alpha\boldsymbol{x}_a + (1-\alpha)\boldsymbol{x}_b$ is a minimum variance efficient portfolio for $0 \leq \alpha \leq 1$.

Proof. (a) Let R_c denote the mean return of the given minimum variance portfolio $\boldsymbol{x}_c$. Choose parameter α such that

$$R_c = \alpha R_a + (1-\alpha)R_b \tag{4.1}$$

that is, choose α given by

$$\alpha = \frac{R_c - R_b}{R_a - R_b}. \tag{4.2}$$

Note that α exists and is unique because by hypothesis $R_a \neq R_b$.

We claim that

$$\boldsymbol{x}_c = \alpha\boldsymbol{x}_a + (1-\alpha)\boldsymbol{x}_b. \tag{4.3}$$

To establish (4.3) use first (3.12) and next (4.1) to write

$$\begin{aligned}
\boldsymbol{x}_c &= \mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_c \\ 1 \end{bmatrix} \\
&= \mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} \alpha R_a + (1-\alpha)R_b \\ \alpha + (1-\alpha) \end{bmatrix} \\
&= \alpha\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_a \\ 1 \end{bmatrix} + (1-\alpha)\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_b \\ 1 \end{bmatrix} \\
&= \alpha\boldsymbol{x}_a + (1-\alpha)\boldsymbol{x}_b.
\end{aligned} \tag{4.4}$$

(b) Consider portfolio $\boldsymbol{x}_c$ which is a linear combination of $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ as in (4.3). Then

$$\begin{aligned}
\boldsymbol{x}_c &= \alpha\boldsymbol{x}_a + (1-\alpha)\boldsymbol{x}_b \\
&= \alpha\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_a \\ 1 \end{bmatrix} + (1-\alpha)\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_b \\ 1 \end{bmatrix} \\
&= \mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} \alpha R_a + (1-\alpha)R_b \\ 1 \end{bmatrix}
\end{aligned}$$

By (3.12) we conclude that $\boldsymbol{x}_c$ is the minimum variance portfolio with expected return $\alpha R_a + (1-\alpha)R_b$.

(c) This is proved as in (b) noting that the restriction $0 \leq \alpha \leq 1$ implies, $R_a \leq \alpha R_a + (1-\alpha)R_b \leq R_b$, if $R_a \leq R_b$.

This completes the proof. $\square$

It is of historical interest that this fact was discovered by Tobin [1958]. Tobin uses only two assets (riskless cash and a risky consol), and demonstrates that nothing essential is changed if there are many risky assets. He argues that the risky assets can be viewed as a single composite asset (mutual fund) and investors find it optimal to combine their cash with a specific portfolio of risky assets. In particular, Theorem 4.1 shows that any mean variance efficient portfolio can be generated by two arbitrary distinct mean–variance efficient portfolios. In other words, if an investor wishes to invest in a mean–variance efficient portfolio with a given expected return and variance, he or she can achieve this goal by investing in an appropriate linear combination of any two mutual funds which are also mean–variance efficient. Practically this means that the n original assets can be purchased by only two mutual funds and investors then can just choose to allocate their wealth, not in the original n assets directly but in these two mutual funds in such a way that the investment results (mean–variance) of the two actions (portfolios) would be identical.

There is, however, an additional implication from part (c) of the two-fund separation theorem. Suppose that utility functions are restricted so that all investors choose to invest in mean–variance efficient portfolios and choose $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ to be the investment proportions of two distinct mean–variance efficient portfolios that generate all the others. In particular $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ can be used to generate the *market portfolio*, that is, the wealth weighted sum of the portfolio holdings of all investors.[5] This implies that the market portfolio is also mean–variance efficient. Black [1972] employs this result in deriving the capital asset pricing model.

Having shown that any two distinct portfolios can generate all other portfolios, it is of practical interest to select two portfolios whose means and variances are easy to compute. One such portfolio is the global minimum variance portfolio with R_G, σ_G^2 and x_G given in the previous section. The other one is identified in Figure 1, with $R_1 = a/b$, $\sigma_1^2 = a/b^2$, and

$$\boldsymbol{x}_1 = \frac{\mathbf{V}^{-1}\boldsymbol{R}}{b}. \tag{4.5}$$

[5] To clarify the concept of market portfolio, it is helpful to proceed inductively. Suppose that investors 1 and 2 have wealth w_1 and w_2 invested in minimum variance efficient portfolios with weights $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Then the sum of their holdings is a portfolio with wealth $w_1 + w_2$ and portfolio weights $\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2$ where $\alpha = w_1/(w_1+w_2)$. Since $0 \leq \alpha \leq 1$, from Theorem 4.1(c), the sum total of their holdings is also an efficient portfolio. Next suppose that the wealth w_n of n investors is invested in an efficient portfolio with weights $\boldsymbol{x}_n$ and investor $n+1$ has wealth w_{n+1} invested in an efficient portfolio with weights $\boldsymbol{x}_{n+1}$. Again from Theorem 4.1(c) the sum total of the holdings of all $n+1$ investors is an efficient portfolio. Proceeding in this manner we conclude that the sum total of all the investors' portfolios is an efficient portfolio. By definition, however, this is the market portfolio. Thus we conclude that the market portfolio is efficient.

Observe from Figure 1 that this second portfolio's orthogonal portfolio has an expected return of zero. Theorem 4.2 below uses these two portfolios $\boldsymbol{x}_{\mathrm{G}}$ and $\boldsymbol{x}_1$.

We state a theorem about the relation of individual asset parameters which will be useful in the analysis of the capital asset pricing model.

Theorem 4.2. *For a given portfolio $\boldsymbol{x}_p$, the covariance vector of individual assets with respect to portfolio p is linear in the vector of mean returns R if and only if p is a minimum variance portfolio.*

Proof. Let $\boldsymbol{x}_p$ be the weights of a minimum variance portfolio which can be written as (3.12). The vector of covariances between individual assets and $\boldsymbol{x}_p$ is given by

$$\mathbf{V}\boldsymbol{x}_p = \mathbf{V}\mathbf{V}^{-1}[\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_p \\ 1 \end{bmatrix} = [\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_p \\ 1 \end{bmatrix} \tag{4.6}$$

which verifies the linearity between the covariance vector and the vector of expected returns, $\boldsymbol{R}$.

Conversely, let the vector of covariances with an arbitrary portfolio $\boldsymbol{x}_p$ be expressed linearly as

$$\mathbf{V}\boldsymbol{x}_p = g\boldsymbol{R} + h\mathbf{1} \tag{4.7}$$

where g and h are arbitrary constants. From (4.7), solving for $\boldsymbol{x}_p$ we get

$$\boldsymbol{x}_p = g\mathbf{V}^{-1}\boldsymbol{R} + \mathbf{V}^{-1}\mathbf{1} = gb\boldsymbol{x}_1 + hc\boldsymbol{x}_{\mathrm{G}}. \tag{4.8}$$

Note that in this last equation $\boldsymbol{x}_p$ is generated by two distinct efficient portfolios $\boldsymbol{x}_1$ and $\boldsymbol{x}_{\mathrm{G}}$. Recall that $\boldsymbol{x}_{\mathrm{G}}$ is the vector of investment proportions of the global minimum variance portfolio and $\boldsymbol{x}_1$ is the vector of investment proportions described in (4.5). Since both $\boldsymbol{x}_{\mathrm{G}}$ and $\boldsymbol{x}_1$ are investment proportions, they satisfy $\boldsymbol{x}_{\mathrm{G}}^{\mathrm{T}}\mathbf{1} = \boldsymbol{x}_1^{\mathrm{T}}\mathbf{1} = 1$ which combined with the property that $\boldsymbol{x}_p^{\mathrm{T}}\mathbf{1} = 1$ allows us to conclude that $gb + hc = 1$. Thus we conclude from Theorem 4.1 that $\boldsymbol{x}_p$ is a minimum variance portfolio. This completes the proof. □

We close this section by expressing (4.6) in a way that will be useful in the discussion of the capital asset pricing model in Section 6. From (4.6) write

$$\begin{aligned} \operatorname{cov}(R_i, R_p) &= [0\ \dots\ 1\ \dots\ 0]\mathbf{V}\boldsymbol{x}_p \\ &= [0\ \dots\ 1\ \dots\ 0][\boldsymbol{R}\ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} R_p \\ 1 \end{bmatrix} \\ &= [R_i\ \ 1]\mathbf{A}^{-1}\begin{bmatrix} R_p \\ 1 \end{bmatrix}, \end{aligned} \tag{4.9}$$

where the 1 in the row vector is placed in the position of the ith asset. Let $\boldsymbol{x}_z$ be orthogonal to $\boldsymbol{x}_p$ and calculate their covariance as in (3.20). Subtract (3.20) from (4.9) to get

$$\operatorname{cov}(R_i, R_p) = [r_i\ \ 0]\mathbf{A}^{-1}\begin{bmatrix} R_p \\ 1 \end{bmatrix} = \gamma r_i \tag{4.10}$$

where the two new variables r_i and γ are defined as

$$r_i = R_i - R_z, \tag{4.11}$$

and

$$\gamma = \frac{cR_p - b}{ac - b^2}. \tag{4.12}$$

Observe that (4.10) holds for each i and must therefore hold for all assets, i.e.

$$\operatorname{cov}(R_p, R_p) = \sigma_p^2 = \gamma r_p, \tag{4.13}$$

where r_p expresses the excess mean return of portfolio p from its orthogonal z. From this last equation obtain $\gamma = \sigma_p^2/r_p$ and substitute in (4.10) to conclude that

$$r_i = \frac{\operatorname{cov}(R_i, R_p)}{\sigma_p^2} r_p = \beta_i r_p \tag{4.14}$$

which expresses the excess mean return of the ith asset as a proportion of its beta, β_i, with respect to portfolio p, where

$$\beta_i = \frac{\operatorname{cov}(R_i, R_p)}{\sigma_p^2}. \tag{4.15}$$

These mathematical manipulations show that (4.14), which has a capital asset pricing appearance, holds true for any minimum variance portfolio, in general, and for any minimum variance efficient portfolio, in particular.

5. Mean–variance portfolio with a riskless asset

The previous two sections presented and solved the portfolio selection problem for n risky assets, and then established the two fund separation theorem. We now return to Tobin's original idea of introducing a riskless asset. The portfolio selection problem with n risky assets and one riskless, i.e. a total of $(n+1)$ assets can easily be formulated and solved. Let there be $n+1$ assets, $i = 0, 1, 2, \ldots, n$, where 0 denotes the riskless asset with return R_0. The vector of expected excess returns has elements defined as $r_i = R_i - R_0$, $i = 1, 2, \ldots, n$, and is denoted by $\boldsymbol{r}$. Wealth is now allocated among $(n+1)$ assets with weights $w_0, w_1, \ldots, w_n$. In the various calculations we denote the vector of weights $w_1, \ldots, w_n$ as $\boldsymbol{w}$ and write $w_0 = 1 - \boldsymbol{w}^{\mathrm{T}}\mathbf{1}$.

For a given portfolio p, the mean excess return is

$$r_p = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{R} + (1 - \boldsymbol{w}^{\mathrm{T}}\mathbf{1})R_0 - R_0 = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{r}. \tag{5.1}$$

The variance of p is

$$\sigma_p^2 = \boldsymbol{w}^{\mathrm{T}}\mathbf{V}\boldsymbol{w}, \tag{5.2}$$

where in (5.1) and (5.2), $\boldsymbol{R}$ and $\mathbf{V}$ are as in Section 3. Note that in (5.2) the riskless asset does not contribute to the variance.

The mean–variance portfolio selection problem with a riskless asset can be stated as

$$\begin{aligned}&\text{minimize} \quad \boldsymbol{w}^{\mathrm{T}}\mathbf{V}\boldsymbol{w}\\ &\text{subject to} \quad \boldsymbol{w}^{\mathrm{T}}\boldsymbol{r} = r_p.\end{aligned} \tag{5.3}$$

In (5.3), the variance of the n-risky assets is minimized subject to a given excess return r_p. Note that $\boldsymbol{w}^{\mathrm{T}}\mathbf{1} = 1$ is not a constraint because the wealth need not all be allocated to the n-risky assets; some may be held in the riskless asset.

Following the method of (3.1) one obtains the solution

$$\boldsymbol{w} = \left(\frac{r_p}{\boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{r}}\right)\mathbf{V}^{-1}\boldsymbol{r} \tag{5.4}$$

which gives the variance of the minimum-variance portfolio with excess mean r_p as

$$\begin{aligned}\sigma_p^2 &= \boldsymbol{w}^{\mathrm{T}}\mathbf{V}\boldsymbol{w}\\ &= \left(\frac{r_p}{\boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{r}}\right)^2 \boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\boldsymbol{r}\\ &= \frac{r_p^2}{\boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{r}}.\end{aligned} \tag{5.5}$$

The Sharpe's measure of portfolio p, defined as the ratio of its excess mean return to the standard deviation of its return, is obtained from (5.5) as

$$\frac{r_p}{\sigma_p} = \left\{\begin{array}{ll} (\boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{r})^{1/2}, & \text{if } r_p \geq 0\\ -(\boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{r})^{1/2}, & \text{if } r_p < 0\end{array}\right\}. \tag{5.6}$$

The tangency portfolio T is the minimum-variance portfolio for which

$$\mathbf{1}^{\mathrm{T}}\boldsymbol{w}_{\mathrm{T}} = 1. \tag{5.7}$$

Combining equations (5.4) and (5.7) we obtain

$$r_{\mathrm{T}} = \frac{\boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{r}}{\mathbf{1}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{r}} \gtrless 0. \tag{5.8}$$

It is economically plausible to assert that the riskless return is lower than the mean return of the global minimum variance portfolio of the risky assets, that is, $R_0 < R_{\mathrm{G}}$. We may then prove that $\mathbf{1}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{r} > 0$. Also $\boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{r} > 0$ by the positive definiteness of the matrix $\mathbf{V}$. It then follows that $r_{\mathrm{T}} > 0$ and the slope of the tangency line in Figure 2 is positive. This positively-sloped line is the capital market line and defines the set of minimum variance efficient portfolios. For an actual calculation of Figure 2, see Ziemba, Parkan & Brooks-Hill [1974].

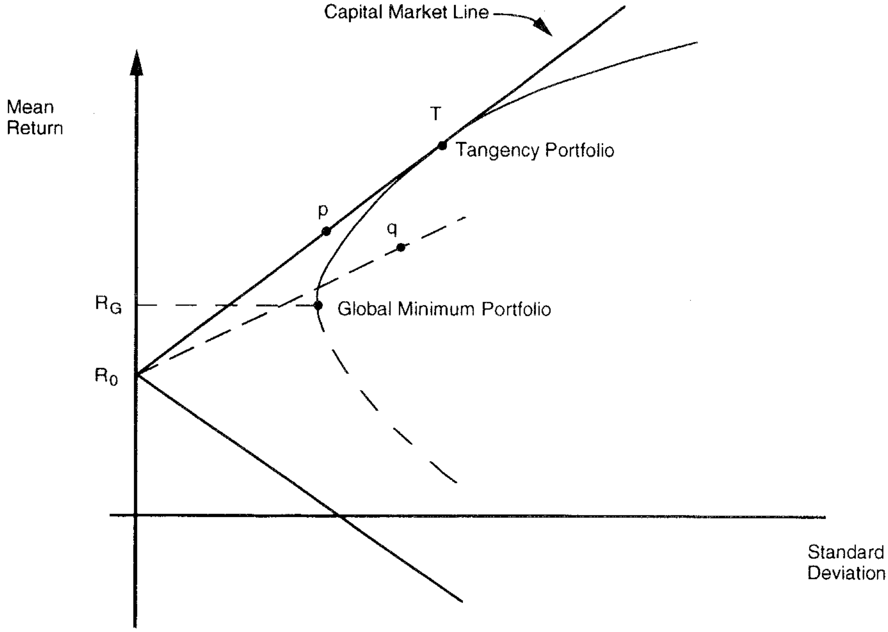

Fig. 2. Portfolios of n-risky assets and a riskless asset.

The correlation coefficient of the return of any portfolio q, with weights w_q, and any portfolio p on the efficient segment of the minimum-variance frontier is

$$\begin{aligned}\rho(p,q) &= \frac{\boldsymbol{w}_q^{\mathrm{T}}\mathbf{V}\boldsymbol{w}_p}{\sigma_q\sigma_p} \\ &= \frac{r_p r_q}{(\boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{r})\sigma_q\sigma_p} \\ &= \frac{r_q/\sigma_q}{r_p/\sigma_p} \\ &= \frac{\text{Sharpe's measure of portfolio } q}{\text{Sharpe's measure of portfolio } p}\end{aligned} \tag{5.9}$$

Referring to Figure 2, the correlation $\rho(p,q)$ is the ratio of the slope of the line from R_0 to q to the slope of the efficient frontier.

6. The capital asset pricing model

Markowitz's approach to portfolio selection may be characterized as normative. The analysis of Sections 3, 4 and 5 concentrates on a typical investor and by making several simplifying assumptions, solves the investor's portfolio selection problem. Recall the assumptions: (i) the investor considers only the first two moments of the probability distribution of returns; (ii) given the mean portfolio return, the investor chooses a portfolio with the lowest variance of returns; and (iii) the investment horizon is one period. There are also a few additional assumptions that are implicit: (i) the investor's individual decisions do not affect market prices; (ii) fractional shares may be purchased (i.e. investments are infinitely divisible); (iii) transaction costs and taxes do not exist, and (iv) investors can sell assets short.

It is historically worth observing that six years had to elapse before the normative results of portfolio selection could be generalized into a positive theory of capital markets. Brennan [1989] claims that "[t]he reason for delay was undoubtedly the boldness of the assumption required for progress, namely that all investors hold the same beliefs about the joint distribution of a security[6]". Indeed, Sharpe [1964] emphasizes that in order to obtain equilibrium conditions in the capital market the *homogeneity of investor expectations*[7] assumption must be made.

Under these assumptions we have demonstrated that all investors hold mean–variance efficient portfolios. With the added homogeneity assumption, Theorem 4.1 shows that a portfolio which consists of two (or more) mean–variance efficient portfolios is mean variance efficient. Therefore the market portfolio is mean variance efficient. Therefore, the mean asset returns are linear in their covariance with the market return as shown in Theorem 4.2. This simple, yet powerful argument due to Black [1972] does not rely on the existence of a riskless asset, unlike the original derivation of the Capital Asset Pricing Model (CAPM) by Sharpe [1964]. From equation (4.14) we may write the CAPM as

$$R_i - R_z = \beta_i (R_M - R_z) \tag{6.1}$$

where R_M is the mean return of the market portfolio, β_i is $\text{cov}(R_i, R_M)/\text{var}(R_M)$ and R_z is the mean return of a minimum variance portfolio which is orthogonal to the market portfolio. In the special case that a riskless asset exists, R_z must equal the riskless rate of return. Ferson [1994] surveys in this volume both the theory and testing of the capital asset pricing model.

Fama [1976] and Roll [1977] pointed out that testing the capital asset pricing model is equivalent to testing the market's mean–variance efficiency. If the only testable hypothesis of the capital asset pricing theory is that the market portfolio is mean–variance efficient, then such testing is infeasible. The infeasibility is due to our ignorance of the exact composition of the true market portfolio. In other words, the capital asset pricing theory is not testable unless all individual assets are included in the market. Using a proxy for the true market portfolio does not solve the problem for two reasons: first, the proxy itself may be mean–variance

[6] See Brennan [1989, p. 93].

[7] Two brief remarks are in order. First, Sharpe attributes the term of homogeneity of investor expectations to one of the referees of his paper. Second, he acknowledges that this assumption is highly restrictive and unrealistic but defends it because of its implication, i.e. attainment of equilibrium. See also Lintner [1965] and Mossin [1966]. Numerous papers have appeared which have relaxed some of the stated assumptions. For example see Levy & Samuelson [1992]

efficient even when the true market portfolio is not; second, the chosen proxy may be inefficient even though the true market portfolio is actually efficient.

We conclude this section by pointing out that the empirical methodologies of testing for the mean–variance efficiency of a given portfolio may be applied in testing a broad class of asset pricing models. Absence of arbitrage among n assets with returns represented by the random variables, $\tilde{R}_i$ $i = 1, \ldots, n$, implies the existence of a strictly positive pricing kernel represented by the random variable $\tilde{m}$ such that

$$E\left[\tilde{m} \quad \tilde{R}_i\right] = 1, \qquad i = 1, \ldots, n. \tag{6.2}$$

For example, in the consumption asset pricing model, $\tilde{m}$ stands for the marginal rate of substitution in consumption between the beginning and end of the period.

Let $\boldsymbol{x}$ denote the weights of a portfolio of n assets which has return maximally correlated with the pricing operator $\tilde{m}$. Then we can write $\tilde{m}$ as

$$\tilde{m} = \alpha \sum_{j=1}^{n} x_j \tilde{R}_j + \tilde{\varepsilon} \tag{6.3}$$

where α is a constant. The property of maximal correlation implies that $\mathrm{cov}(\tilde{\varepsilon}, \tilde{R}_j) = 0$, $j = 1, \ldots, n$. Combining equations (6.2) and (6.3) we obtain

$$1 = E\left[\tilde{m} \quad \tilde{R}_i\right] = E\left[\tilde{m}\right]E\left[\tilde{R}_i\right] + \alpha \,\mathrm{cov}\left(\sum_{j=1}^{n} x_j \tilde{R}_j, \tilde{R}_i\right), \quad i = 1, \ldots, n. \tag{6.4}$$

This implies that the n assets' covariances with the portfolio $\boldsymbol{x}$ are linear in their mean returns. By Theorem 4.2 we conclude that the portfolio $\boldsymbol{x}$ must lie on the minimum-variance frontier of the n assets, a property which can be tested by the methodologies which test for the efficiency of a given portfolio. For further discussion of these issues see the papers of Hansen & Jagannathan [1991] and Ferson [1995].

7. Theoretical justification of mean–variance analysis, mutual fund separation and the CAPM

In this section we first address the following question: what set of assumptions is needed on the investor's utility function or distribution of asset returns so that the investor chooses a mean–variance efficient portfolio?

Tobin [1958] uses a quadratic utility function represented by

$$u(c) = c - B\frac{c^2}{2}, \qquad B > 0 \tag{7.1}$$

and defined only for $c \leq 1/B$, where c denotes consumption. Arrow [1971] has remarked that quadratic utility exhibits increasing absolute risk aversion which implies that risky assets are inferior goods in the context of the portfolio selection problem. It can be easily shown that utility is increasing in the mean and decreasing in the variance, and that moments higher than the variance do not matter. Therefore only mean–variance efficient portfolios will be selected by expected quadratic utility maximizing investors.

Next note that multivariate normality is a special distribution of asset returns for which mean–variance analysis is consistent with expected utility maximization without assuming quadratic utility. To show this recall that the distribution of any portfolio is completely specified by its mean and variance. This follows from the basic property that any linear combination of multivariate normally distributed variables has a distribution in the same family.

Chamberlain [1983a] shows that the most general class of distributions that allow investors to rank portfolios based on the first two generalized moments is the family of *elliptical distributions*. A vector $\boldsymbol{x}$ of n random variables is said to be elliptically distributed if its density function is of the form

$$f(x) = |\Omega|^{-1/2} g[(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \Omega^{-1} (\boldsymbol{x} - \boldsymbol{\mu}); \boldsymbol{x}] \tag{7.2}$$

where Ω is an $n \times n$ positive definite dispersion matrix and $\boldsymbol{\mu}$ is the vector of medians. From (7.1) Ingersoll [1987] obtains as special cases both the multivariate normal and the multivariate Student-t distributions.

Having presented a theoretical justification for mean–variance analysis[8] we can now ask a second and broader question: which is the class of utility functions that imply two-fund separation? Without assuming the existence of a riskless asset, Cass & Stiglitz [1970] prove that a necessary and sufficient condition for two-fund separation is that preferences are either quadratic or of the constant-relative-risk-aversion family, $u(c) = (1 - A)^{-1} c^{1-A}$, $A > 0$, $A \neq 1$ (with $u(c) = \ln c$ corresponding to the case $A = 1$). Actually constant relative risk aversion implies the stronger property of one-fund separation. If a riskless asset is assumed to exist, the necessary and sufficient condition for two-fund separation is either quadratic preferences or HARA preferences defined as $u(c) = (1 - A)^{-1} (c - \hat{c})^{1-A}$, $A > 0$, $A \neq 1$ (with $u(c) = \ln(c - \hat{c})$ corresponding to the case $A = 1$). Their main conclusion is that utility-based conditions under which separation holds are very restrictive. But more to the point, utility-based two-fund separation, with the exception of quadratic utility, does not imply mean–variance choice and does not imply the CAPM.

Ross [1978] establishes the necessary and sufficient conditions on the stochastic structure of asset returns such that two-fund portfolio separation would obtain for any increasing and concave von Neumann–Morgenstern utility function. More specifically, a vector of asset returns $\boldsymbol{R}$ is said to exhibit two-fund separability if

[8] Ingersoll [1975] and Kraus & Litzenberger [1976] address the interesting question of how portfolios are formed when either the utility function or the distribution of returns are not of the type that imply mean–variance analysis. In particular, Kraus & Litzenberger [1976] extend the portfolio selection problem to include the effect of skewness. The rate of return on the investor's portfolio is assumed to be nonsymmetrically distributed and the investor's utility function considers the first three moments of such a distribution. See also Ziemba [1994], Ohlson & Ziemba [1976], and Kallberg & Ziemba [1983].

there are two mutual funds α and β of n assets such that for any portfolio q there exists a portfolio weight λ such that

$$E[u(\lambda R_\alpha + (1-\lambda)R_\beta)] \geq E[u(R_q)] \tag{7.3}$$

for each monotone increasing and concave utility functions $u(\cdot)$. Observe that (7.3) captures analytically the intuitive notion that portfolios generated by the two funds are preferred to arbitrary portfolios. There is an extensive literature that deals with this important issue of comparing portfolios for a class of investor preferences known as *stochastic dominance*. Ingersoll [1987] or Huang & Litzenberger [1988] give a general overview of these ideas and Rothschild & Stiglitz [1970] offer a detailed analysis.

From the above definition, Ross [1978, p. 267] proves that two-fund separability is equivalent to the following conditions: there exist random variables $\tilde{R}$, $\tilde{Y}$ and $\tilde{\varepsilon}$ and weights x_i, x_i^{M} and x_i^z, $i = 1, 2, \ldots, n$, such that

$$\tilde{R}_i = \tilde{R} + b_i\tilde{Y} + \tilde{\varepsilon}_i \qquad \text{for all } i \tag{7.4}$$

$$E[\tilde{\varepsilon}_i \mid \tilde{R} + \xi\tilde{Y}] = 0 \qquad \text{for all } i,\ \xi \tag{7.5}$$

$$\sum_i w_i^{\mathrm{M}} = 1, \qquad \sum_i w_i^z = 1 \tag{7.6}$$

$$\sum_i w_i^{\mathrm{M}}\tilde{\varepsilon}_i = 0, \qquad \sum_i w_i^z\tilde{\varepsilon}_i = 0 \tag{7.7}$$

$$\text{and either } b_i = b \text{ for all } i, \text{ or } \sum_i w_i^{\mathrm{M}} b_i \neq \sum_i w_i^z b_i. \tag{7.8}$$

Observe that conditions (7.4)–(7.8) represent the most general form of distribution of returns which permits two-fund separation. In particular, Ross [1978, p. 273] shows that all multivariate normally distributed random variables satisfy condition (7.7). But, more to the point Ross shows that, if asset returns are drawn from the family of two-fund separating distributions, and if asset variances are finite, then the CAPM holds.

Having reviewed the assumptions needed on asset distributions for mean–variance portfolio theory and two-fund separation to hold, we close with a brief evaluation of these assumptions. Osborne [1959], Mandelbrot [1963], Fama [1965a, b], Boness, Chen & Jatusipitak [1974] and numerous other studies have shown that there are substantial deviations from normality in the distribution of actual stock prices. Although actual returns are not normally distributed and the use of quadratic utility cannot be supported empirically, the mean–variance portfolio theory remains theoretically useful and empirically relevant. Actually, portfolio theory is a prime example of Milton Friedman's assertion that a theory should not be judged by the relevance of its assumptions, but rather, by the realism of its predictions.[9]

[9] Stiglitz [1989] evaluates the various assumptions placed on investor preferences, and Markowitz [1991] in his Nobel Lecture supports the appropriateness of the approximation. See also Levy & Markowitz [1979] and Markowitz [1987].

8. Consumption and portfolio selection in continuous time

Mean–variance portfolio theory addresses the investor's asset selection problem for an investment horizon of one period. Progress in portfolio theory came as financial economists relaxed this restrictive assumption. In so doing, however, they were faced with the twin decisions discussed in the introduction: consumption-saving and portfolio selection. The relaxation of the single-period assumption proceeded along two lines: first, in discrete time multiperiod models by Samuelson [1969], Hakansson [1970], Fama [1970], Rubinstein [1976], Long [1974] and others, and second, in continuous time models by Merton [1969, 1971, 1973], Breeden [1979, 1986], Cox, Ingersoll & Ross [1985a, b], and others. Ingersoll [1987] presents a detailed overview of discrete time models. Here, we follow Merton [1973] to develop and solve a continuous-time intertemporal portfolio selection problem.[10]

Assume that there exist continuously trading markets for all $n+1$ assets and that prices per share $P_i(t)$ are generated by Itô processes, i.e.

$$\frac{\mathrm{d}P_i}{P_i} = \alpha_i(x,t)\,\mathrm{d}t + \sigma_i(x,t)\,\mathrm{d}z_i(t), \qquad i = 1,\ldots,n+1 \tag{8.1}$$

where α_i is the conditional arithmetic expected rate of return and $\sigma_i^2\,\mathrm{d}t$ is the conditional variance of the rate of return of asset i. We either assume zero dividends on the stock or, more plausibly, we assume that the dividends are continuously reinvested in the stock and P_i represents the price of one share plus the value of the reinvested dividends. The random variable $z_i(t)$ is a Wiener process. The variance of the increment of the Wiener process is $\mathrm{d}t$. The processes $z_i(t)$ and $z_j(t)$ have correlated increments and we denote

$$\operatorname{cov}\left[\sigma_i \mathrm{d}z_i(t), \sigma_j \mathrm{d}z_j(t)\right] = \sigma_{ij}\mathrm{d}t.$$

In the particular case (not assumed hereafter) where α_i and σ_i are constants, the price $P_i(t)$ is lognormally distributed.

The conditional mean and variance of the rate of return are functions of the random variable $x(t)$, assumed here to be a scalar solely for expositional ease. The random variable $x(t)$, referred to here as the *state variable*, is an Itô process

$$\mathrm{d}x = m(x,t)\mathrm{d}t + s(x,t)s\,\mathrm{d}\hat{z}(t). \tag{8.2}$$

The covariance $\operatorname{cov}[s\,\mathrm{d}\hat{z}(t), \sigma_i \mathrm{d}z_i(t)]$ is denoted by $\sigma_{ix}\mathrm{d}t$.

[10] The appropriateness of the continuous-time approach to the intertemporal portfolio selection problem in particular, and to problems of financial economics in general, is skillfully evaluated in Merton [1975, 1982]. He argues that the use of stochastic calculus methods in finance allows the financial theorist to obtain important generalizations by making realistic assumptions about trading and the evolution of uncertainty. These methods are briefly exposited in Ingersoll [1987] or more extensively in Malliaris & Brock [1982]. The remainder of this paper assumes some familiarity with these techniques.

An investor has wealth $W(t)$ at time t. The investor consumes $C(t)\mathrm{d}t$ over $[t, t+\mathrm{d}t]$ and invests fraction $w_i(t)$ of the wealth in asset i, $i = 1, \ldots, n, n+1$. The budget constraint, or wealth dynamics, is

$$\mathrm{d}W(t) = \mathrm{d}y(t) - C\mathrm{d}t + \sum_{i=1}^{n+1} w_i \frac{\mathrm{d}P_i}{P_i} W \tag{8.3}$$

where $\mathrm{d}y(t)$ is the labor income, or generally the exogenous endowment income over the infinitesimal interval $[t, t+\mathrm{d}t]$.

For expositional simplicity we assume that the labor income is zero. We also assume that the $(n+1)$st asset is riskless, i.e. $\sigma_{n+1} = 0$ and we denote α_{n+1} by r, the instantaneously riskless rate of interest. Then the wealth dynamics equation simplifies to

$$\begin{aligned} \mathrm{d}W &= -C\mathrm{d}t + rW(1 - \sum_{i=1}^{n} w_i)\mathrm{d}t + \sum_{i=1}^{n} w_i W(\alpha_i \mathrm{d}t + \sigma_i \mathrm{d}z_i) \\ &= -C\mathrm{d}t + rW\mathrm{d}t + \sum_{i=1}^{n} w_i W\left[(\alpha_i - r)\mathrm{d}t + \sigma_i \mathrm{d}z_i\right]. \end{aligned} \tag{8.4}$$

We assume that the investor makes sequential consumption and investment decisions with the objective to maximize the von Neumann–Morgenstern expected utility i.e.

$$\max E_0 \left[\int_0^\infty u(C, x, t)\, \mathrm{d}t \right] \tag{8.5}$$

where u is monotone increasing and concave in the consumption flow C. Note that in the above representation of preferences utility is time-separable but nonstate separable since preferences depend on x. The case of nontime-separable preferences is discussed in Sundaresan [1989], Constantinides [1990], and Detemple & Zapatero [1991].

To derive the optimal consumption and investment policies we define

$$J(W, x, t) = \max_{\{C,w\}} E_t \left[\int_t^\infty u(C, x, \tau)\, \mathrm{d}\tau \right].$$

Assuming sufficient regularity conditions as presented in Fleming & Richel [1975], so that a solution exists, the derived utility of wealth, J, satisfies the equation derived by Merton [1971, 1973]

$$\begin{aligned} 0 = \max_{\{C,w\}} \Bigg[u(C, x, t) &+ \left\{ -C + rW + W \sum_{i=1}^{n} w_i(\alpha_i - r) \right\} J_W + mJ_x + J_t + \\ &+ \frac{1}{2} W^2 J_{WW} \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij} + W J_{Wx} \sum_{i=1}^{n} w_i \sigma_{ix} + \frac{s^2}{2} J_{xx} \Bigg]. \end{aligned} \tag{8.6}$$

The first-order conditions with respect to C and w_i are

$$u_C - J_W = 0 \tag{8.7}$$

and

$$W(\alpha_i - r)J_W + W^2 J_{WW} \sum_{j=1}^{n} w_j \sigma_{ij} + W J_{Wx}\sigma_{ix} = 0, \quad i = 1, \ldots, n. \tag{8.8}$$

The concavity of the utility function implies that J is concave in W; hence the second-order conditions are satisfied.

Under appropriate regularity conditions which are not discussed here a verification theorem can be stated to the effect that the solution of the partial differential equation is unique, and therefore is the solution of the original optimal consumption and investment problem.

Since the topic of this essay is the portfolio problem we focus on the first-order conditions (8.8) implied by optimal investment which we write in matrix notation as

$$(\alpha - r\mathbf{1})J_W + W J_{WW} \mathbf{w}^{\mathrm{T}}\mathbf{V} + J_{Wx}\boldsymbol{\sigma}_x = 0, \tag{8.9}$$

where $\mathbf{V}$ is the $n \times n$ covariance matrix with $i \times j$ element σ_{ij} and $\boldsymbol{\sigma}_x$ is a vector with ith element σ_{ix}. Solving for the optimal portfolio weights we obtain

$$\mathbf{w} = \left(\frac{-J_W}{W J_{WW}}\right)\mathbf{V}^{-1}(\alpha - r\mathbf{1}) - \frac{J_{Wx}}{W J_{WW}}\mathbf{V}^{-1}\boldsymbol{\sigma}_x. \tag{8.10}$$

Before we analyze the optimal portfolio decision in its full generality, consider first the important special case where the term $[J_{Wx}/(W J_{WW})]\mathbf{V}^{-1}\boldsymbol{\sigma}_x$ is a vector of zeros. We will shortly discuss three cases where this occurs. Then we may write equation (8.10) as

$$\mathbf{w} = \left(\frac{-J_W}{W J_{WW}}\right)\left[\mathbf{1}^{\mathrm{T}}\mathbf{V}^{-1}(\alpha - r\mathbf{1})\right]\mathbf{w}_{\mathrm{T}} \tag{8.11}$$

where

$$\mathbf{w}_{\mathrm{T}} = \frac{\mathbf{V}^{-1}(\alpha - r\mathbf{1})}{\mathbf{1}^{\mathrm{T}}\mathbf{V}^{-1}(\alpha - r\mathbf{1})}. \tag{8.12}$$

From our discussion in Section 5, we recognize $\mathbf{w}_{\mathrm{T}}$ as the vector of portfolio weights of the tangency portfolio on the frontier of minimum variance portfolios generated by the n risky assets. We also interpret $(-J_W/W J_{WW})^{-1}$ as the relative risk aversion (RRA) coefficient of the investor. Then equation (8.11) states that the investor invests in just two portfolios, namely the riskless asset and the tangency portfolio. The extent of the investment in the tangency portfolio depends on the investor's RRA coefficient. Thus we have proved that there is two-fund separation with the two funds being the riskless asset and the tangency portfolio. From here it is a small step, outlined in Section 9, to show that the CAPM holds.

We present three sets of conditions each of which implies two-fund separation and the CAPM:

(a) Logarithmic utility. Then we may show that the derived utility $J(W, x)$ is the sum of a function of W and a function of x. Hence the cross-derivative J_{Wx} equals zero and the second term in equation (8.10) becomes a vector of zeros.

(b) All assets' returns are uncorrelated with the change in x, i.e. $\sigma_{ix} = 0$, $i = 1, \ldots, n$.

(c) All assets have distributions of returns which are independent of x, i.e. α_i, σ_i are independent of x for $i = 1, \ldots, n$.

We now return to the general case where none of the assumptions (a)–(c) hold and the term $[J_{Wx}/(WJ_{WW})]\mathbf{V}^{-1}\boldsymbol{\sigma}_x$ is not a vector of zeros. Define by $\boldsymbol{w}_{\mathrm{H}}$ the weights of a portfolio

$$\boldsymbol{w}_{\mathrm{H}} = \frac{\mathbf{V}^{-1}\boldsymbol{\sigma}_x}{\mathbf{1}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{\sigma}_x}.$$

Then we may write equation (8.10) as

$$\boldsymbol{w} = \left(\frac{-J_W}{WJ_{WW}}\right)\left[\mathbf{1}^{\mathrm{T}}V^{-1}(\alpha - r\mathbf{1})\right]\boldsymbol{w}_{\mathrm{T}} + \left(\frac{-J_{Wx}}{WJ_{WW}}\right)\left[\mathbf{1}^{\mathrm{T}}V^{-1}\boldsymbol{\sigma}_x\right]\boldsymbol{w}_{\mathrm{H}}. \quad (8.13)$$

We observe that three-fund portfolio separation obtains: The investor invests in the riskless asset, the tangency portfolio $\boldsymbol{w}_{\mathrm{T}}$ and the hedging portfolio $\boldsymbol{w}_{\mathrm{H}}$. The weights which the investor assigns to each portfolio depend on his/her preferences and are, therefore, investor-specific.

We may further interpret the hedging portfolio by solving the following maximization problem: Choose vector $\boldsymbol{y}$ such that $\mathbf{1}^{\mathrm{T}}\boldsymbol{y} = 1$ (i.e. $\boldsymbol{y}$ is the vector of a portfolio's weights) to maximize the correlation of $\mathrm{d}x$ and $\sum_{i=1}^{n} y_i(\mathrm{d}P_i/P_i)$. The solution to this problem is easily shown to be $\boldsymbol{y} = \boldsymbol{w}_{\mathrm{H}}$. That is, the hedging portfolio is the portfolio of the risky assets with returns maximally correlated with the change in the state variable x.

Note that x enters into the decision problem through α_i and σ_i, that is, it causes changes in the investment opportunity set and through the utility of consumption, $u(C, x, t)$, that is, it causes shifts in tastes. We may interpret the three fund separation result as follows: The investor invests in the riskless asset and in the tangency portfolio, as in the mean–variance case, but modifies his or her portfolio investing in (or selling short) a third portfolio which has returns maximally correlated with changes in the variable x which represents shifts in the investment opportunity set and tastes.

As we stated earlier we have chosen x to be a scalar solely for expositional ease. If instead, $\boldsymbol{x}$ is a vector with m elements we obtain $(m + 2)$-fund separation where the investor invests in the riskless asset, the tangency portfolio and the m hedging portfolios.

In evaluating Merton's [1971, 1973] intertemporal continuous-time portfolio theory at least two important contributions need to be identified: first, its generalization of the static mean–variance theory is achieved by considering both the consumption and portfolio selection over time and by dropping the quadratic utility assumption; and second, its realism and tractability compared to the discrete-time portfolio theories which assume normally distributed asset prices implying a nonzero probability of negative asset prices. By replacing the assumption of normally distributed asset prices with the assumption that prices follow (8.1), the continuous-time portfolio theory becomes more realistic as well as more tractable in view of the extensive mathematical literature on diffusion processes.

Merton's work was extended in several directions. Among them, Breeden [1979] and Cox, Ingersoll & Ross [1985a, b] consider a generalization of the intertemporal continuous-time portfolio theory in a general equilibrium model with production. Another contribution was made by Breeden [1979] who shows that Merton's [1973] multi-beta pricing model can be expressed with a single beta measured with respect to changes in aggregate consumption assuming that consumption preferences are time separable. One interesting result of Breeden's work is that, in an intertemporal economy, the portfolio that has the highest correlation of returns with aggregate real consumption changes is mean–variance efficient.

Several authors have considered equation (8.1) which is the most significant assumption of continuous-time portfolio theory and have asked the question: under what conditions is a price system representable by Itô processes such as (8.1)? Huang [1985a, b] shows that when the information structure is a Brownian filtration then any arbitrage-free price system is an Itô process. The arbitrage-free concept is analyzed in Harrison & Kreps [1979] and Harrison & Pliska [1981] who make a connection to a martingale representation theorem. The role of information is analyzed in Duffie & Huang [1986].

Finally, in contrast to the stochastic dynamic programming approach to the continuous time consumption and portfolio problem, Pliska [1986] and Cox & Huang [1989], among others have used the martingale representation methodology. In the martingale approach, first, the dynamic consumption and portfolio problem is transformed and solved as a static utility maximization problem to find the optimal consumption and, second, the martingale representation theorem is applied to determine the portfolio trading strategy which is consistent with the optimal consumption. It is usually assumed that markets are dynamically complete which allows for the determination of a budget constraint and the solution of the static utility maximization. The case when markets are dynamically incomplete with the dimension of the Brownian motion driving the security prices being greater than the number of risky securities is presented in He & Pearson [1991].

9. The Intertemporal Asset Pricing Model (ICAPM) and the Arbitrage Pricing Theory (APT)

In the last section we solved for the optimal weights of the portfolio of risky assets held by an investor with given preferences. If all consumers in the economy have identical preferences and endowments then the above optimal portfolio may be identified as the market portfolio of risky assets. The condition that consumers have identical preferences and endowments may be relaxed under conditions which imply demand aggregation as in Rubinstein [1974] and Constantinides [1980] or under complete markets as in Constantinides [1982]. Hereafter we assume that either through demand aggregation or through complete markets we can claim that the optimal portfolio in (8.10) is indeed the market portfolio of risky assets. We denote the weights of this portfolio by w^{M} and its return by

$$\frac{\mathrm{d}P_{\mathrm{M}}}{P_{\mathrm{M}}} = \sum_{i=1}^{n} w_i^{\mathrm{M}} \frac{\mathrm{d}P_i}{P_i}$$

We should stress that, in general, the market portfolio does not coincide with the tangency portfolio. In the last section we discussed conditions under which the two portfolios coincide but these conditions will not be imposed here.

To derive the intertemporal capital asset pricing model (ICAPM), we rewrite equation (8.8) as

$$\begin{aligned} \alpha_i - r &= \left(-\frac{W J_{WW}}{J_W}\right) \sum_{j=1}^{n} w_j^{\mathrm{M}} \sigma_{ij} + \left(-\frac{J_{Wx}}{J_W}\right) \sigma_{ix} \\ &= \lambda_{\mathrm{M}} \beta_{iM} + \lambda_x \beta_{ix} \qquad i = 1, \ldots, n. \end{aligned} \tag{9.1}$$

where

$$\begin{aligned} \beta_{iM} &= \frac{\operatorname{cov}(\mathrm{d}P_i/P_i, \mathrm{d}P_{\mathrm{M}}/P_{\mathrm{M}})}{\operatorname{var}(\mathrm{d}P_{\mathrm{M}}/P_{\mathrm{M}})} \\ \lambda_{\mathrm{M}} &= -\frac{W J_{WW}}{J_W} \frac{\operatorname{var}(\mathrm{d}P_{\mathrm{M}}/P_{\mathrm{M}})}{\mathrm{d}t} \\ \beta_{ix} &= \frac{\operatorname{cov}(\mathrm{d}P_i/P_i, \mathrm{d}x)}{\operatorname{var}(\mathrm{d}x)} \end{aligned}$$

and

$$\lambda_x = -\frac{J_{Wx}}{J_W} \frac{\operatorname{var}(\mathrm{d}x)}{\mathrm{d}t}.$$

This result generalizes in a routine fashion to the case where the state variable is a vector.

We conclude this section by discussing the empirically testable implications of the theory, along with the arbitrage pricing theory of Ross [1976a, b]. The common starting point of both the ICAPM and the APT is a linear multivariate regression of the $n \times 1$ vector of asset returns, $\tilde{\boldsymbol{R}}$, on a $k \times 1$ vector of state variables (in the ICAPM) or factors (in the APT), $\tilde{f}$:

$$\tilde{\boldsymbol{R}} = \boldsymbol{R} + B(\tilde{f} - f) + \tilde{\varepsilon} \tag{9.2}$$

where $R \equiv E[\tilde{\boldsymbol{R}}]$, $f \equiv E[\tilde{f}]$ and $E[\tilde{\varepsilon}] = 0$. In both theories the elements of $\tilde{f}$ are assumed to have finite variance. The covariance matrix $\Omega \equiv E[\tilde{\varepsilon}\tilde{\varepsilon}^{\mathrm{T}}]$ is assumed to have finite elements. Furthermore, in the APT the elements of $\tilde{f}$ are assumed to be factors in the sense that the largest eigenvalue of Ω remains bounded as $n \to \infty$ [see Chamberlain, 1983b].

The pricing restriction implied by the ICAPM is that there exist a constant, λ_0, and a $k \times 1$ vector of risk "premia", $\boldsymbol{\lambda}$, such that

$$\boldsymbol{R} = \lambda_0 \mathbf{1} + B\boldsymbol{\lambda} \tag{9.3}$$

where $\mathbf{1}$ is the $n \times 1$ vector of ones as before. The pricing restriction implied by the APT is

$$\lim_{n\to\infty} (\boldsymbol{R} - \lambda_0\mathbf{1} - B\boldsymbol{\lambda})^{\mathrm{T}}(\boldsymbol{R} - \lambda_0\mathbf{1} - B\boldsymbol{\lambda}) = A, \qquad A < \infty \tag{9.4}$$

which, in empirical work (where n is finite), is interpreted to imply (9.3).

If the proxies for state variables in the ICAPM or factors in the APT are portfolios of the n assets, the ICAPM or APT pricing restrictions, (9.3), state that there exists a portfolio of these proxy portfolios which has mean and variance on the mean–variance, minimum-variance frontier. See Jobson & Korkie [1985], Grinblatt & Titman [1987] and Huberman, Kandel & Stambaugh [1987]. Therefore the econometric methods for testing that a given portfolio lies on the minimum-variance frontier may be extended to test the ICAPM and the APT. See Kandel & Stambaugh [1989] and the Connor & Korajczyk [1995] essay in this volume.

10. Market imperfections

Market imperfections were suppressed in our earlier discussion by implicitly assuming that (i) transaction costs are zero, (ii) the capital gains tax is zero (or, capital gains and losses are realized and taxed in every period), and (iii) the assets may be sold short with full use of the proceeds which, in the case of a riskless asset, implies that the borrowing rate equals the lending rate. How sensitive are our conclusions on portfolio selection and equilibrium asset pricing to the presence of these imperfections? Whereas a comprehensive discussion of these issues is beyond the scope of this essay, we discuss briefly one instance of market imperfections.

Consider first the discrete-time intertemporal investment and consumption problem with proportional transaction costs. The agent maximizes the expectation of a time-separable utility function where the period utility is of the convenient power form. The agent consumes in every period and invests the remaining wealth in only two assets. The agent enters period t with x_t units of account of the first asset and y_t units of account of the second asset. If the agent buys (or, sells) v_t units of account of the second asset, the holding of the first asset becomes $x_t - v_t - \max[k_1 v_t, -k_2 v_t]$, net of transaction costs where the constants k_1, k_2 satisfy $0 \le k_1 \le 1$ and $0 \le k_2 \le 1$. The optimal investment policy, described in terms of two parameters $\underline{\alpha}_t$ and $\overline{\alpha}_t$, $\underline{\alpha}_t \le \overline{\alpha}_t$, is to refrain from transacting as long as the portfolio proportions, x_t/y_t, lie within the interval $[\underline{\alpha}_t, \overline{\alpha}_t]$; and transact to the closer boundary, $\underline{\alpha}_t$ or $\overline{\alpha}_t$, of the region of no transactions whenever the portfolio proportions lie outside this interval (provided, of course, that this is feasible). The parameters $(\underline{\alpha}_t, \overline{\alpha}_t)$ are functions of time and of the state variables which define the conditional distribution of the assets' return. This general form of the optimal portfolio policy also holds in a model with continuous trading under additional assumptions on the distribution of asset returns. See Kamin [1975], Constantinides [1979], Taksar, Klass & Assaf [1988] and Davis & Norman [1990].

In numerical solutions of the portfolio problem with even small proportional transaction costs one finds that the region of no transactions is wide. We conclude from these examples and extrapolate in more general cases with transaction costs that even small transaction costs distort significantly the optimal portfolio policy which is optimal in the absence of transaction costs. See Constantinides [1986], Dumas & Luciano [1991], Fleming, Grossman, Vila & Zariphopoulou [1990] and Gennotte & Jung [1991]. An encouraging finding, however, is that transaction costs have only a second-order effect on equilibrium asset returns: investors accommodate large transaction costs by drastically reducing the frequency and volume of trade. It turns out that the agent's utility is insensitive to deviations of the asset proportions from those proportions which are optimal in the absence of transaction costs. Therefore, a small liquidity premium is sufficient to compensate an agent for deviating significantly from the target portfolio proportions. These results need to be qualified as they apply to the case where the only motive for trade is portfolio rebalancing. Transaction costs may have a first-order effect on equilibrium asset returns in cases where the investors receive exogenous income or trade on the basis of inside information.

11. Concluding remarks

Portfolio theory is the analysis of the real world phenomenon of diversification. This paper has exposited this theory in its historical evolution, from the early work on static mean–variance mathematics to its generalization of dynamic consumption and portfolio rules. In its intellectual development portfolio theory has benefitted from empirical work which came from capital asset pricing tests and from statistical investigations of the distributions of asset prices. Furthermore, as more powerful techniques were developed, such as stochastic calculus, portfolio theory became dynamic and many results were generalized.

Because the topic of our paper is theoretical, we have not mentioned any issues related to real world portfolio management. Interested readers can find such topics in standard graduate textbooks such as Lee, Finnerty & Wort [1990] or papers in this volume on performance evaluation by Grinblatt & Titman [1995], on market microstructure by Easley & O'Hara [1995], and on world wide security market regularities by Hawawini & Keim [1995], among others. Although our topic was on portfolio theory, numerous important theoretical developments are not mentioned. Fortunately again, some are treated in this volume such as futures and options markets by Carr & Jarrow [1995], market volatility by LeRoy

& Steigerwald [1995], and the extension of portfolio theory from national to international markets by Stulz [1995]. A useful companion survey is presented in Constantinides [1989], where theoretical issues of financial valuation are presented in a unified way.

Acknowledgements

We thank Wayne Ferson, Brian Kennedy and Bill Ziemba for several useful comments.

References

Arrow, K.J. (1971). *Essays on the Theory of Risk-Bearing*, Markham, Chicago.

Bernoulli, D. (1738). Exposition of a new theory of the measurement of risk, translated by L. Sommer and published in 1954 in *Econometrica* 22, 23–36.

Black, F. (1972). Capital market equilibrium with restricted borrowing. *J. Bus.* 45, 444–455.

Boness, A.J., A. Chen and S. Jatusipitak (1974). Investigations on nonstationarity in prices. *J. Bus.* 47, 518–537.

Breeden, D.T. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities. *J. Financ. Econ.* 7, 265–296.

Breeden, D.T. (1986). Consumption, production, inflation, and interest rates: A synthesis. *J. Financ. Econ.* 16, 3–39.

Brennan, M.J. (1989). Capital asset pricing model, in: J. Eatwell, M. Milgate and P. Newman (eds.), *The New Palgrave Dictionary of Economics*, Stockton Press, New York, NY.

Carr, P., and R.A. Jarrow (1995). A discrete time synthesis of derivative security valuation using a term structure of futures prices, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Elsevier, Amsterdam, pp. (this volume).

Cass, D., and J. Stiglitz (1970). The structure of investor preferences and asset returns, and separability in portfolio selection: A contribution to the pure theory of mutual funds. *J. Econ. Theory* 2, 122–160.

Chamberlain, G. (1983a). A characterization of the distributions that imply mean–variance utility functions. *J. Econ. Theory* 29, 185–201.

Chamberlain, G. (1983b). Funds, factors, and diversification in arbitrage pricing models. *Econometrica* 51, 1305–1323.

Connor, G, and R. Korajczyk (1995). The arbitrage pricing theory, and multifactor models of asset returns, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Elsevier, Amsterdam, pp. 65–122 (this volume).

Constantinides, G.M. (1979). Multiperiod consumption and investment behavior with convex transaction costs. *Manage. Sci.* 25, 1127–1137.

Constantinides, G.M. (1980). Admissible uncertainty in the intertemporal asset pricing model. *J. Financ. Econ.* 8, 71–86.

Constantinides, G.M. (1982). Intertemporal asset pricing with heterogeneous consumers and without demand aggregation. *J. Bus.* 55, 253–267.

Constantinides, G.M. (1986). Capital market equilibrium with transaction costs. *J. Polit. Econ.* 94, 842–862.

Constantinides, G.M. (1989). Theory of valuation: Overview and recent developments, in: S. Bhattacharya and G.M. Constantinides (eds.), *Theory of Valuation*, Rowman & Littlefield, Totowa, NJ.

Constantinides, G.M. (1990). Habit formation: A resolution of the equity premium puzzle. *J. Polit. Econ.* 98, 519–543.

Cox, J.C, and C.F. Huang (1989). Optimal consumption and portfolio policies when asset prices follow a diffusion process. *J. Econ. Theory* 49, 33–83.

Cox, J.C., J.E. Ingersoll, and S.A. Ross (1985a). A theory of the term structure of interest rates. *Econometrica* 53, 385–407.

Cox, J.C., J.E. Ingersoll, and S.A. Ross (1985b). An intertemporal general equilibrium model of asset prices. *Econometrica* 53, 363–384.

Davis, M.H.A., and A.R. Norman (1990). Portfolio selection with transactions costs. *Math. Oper. Res.* 15, 676–713.

Detemple, J.B., and F. Zapatero (1991). Asset prices in an exchange economy with habit formation. *Econometrica* 59, 1633–1657.

Dumas, B., and E. Luciano (1991). An exact solution to a dynamic portfolio choice problem under transactions costs. *J. Finance* 46, 577–595.

Duffie, D, and C. Huang (1986). Multiperiod securities markets with differential information: Martingales and resolution times. *J. Math. Econ.* 15, 283–303.

Easley, M, and M. O'Hara (1995). Market microstructure, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Elsevier, Amsterdam, pp. 357–384 (this volume).

Fama, E.F. (1965a). Portfolio analysis in a stable Paretian market. *Manage. Sci.* 11, 409–419.

Fama, E.F. (1965b). The behavior of stock market prices. *J. Bus.* 38, 34–105.

Fama, E.F. (1970). Multiperiod consumption-investment decisions. *Am. Econ. Rev.* 60, 163–174.

Fama, E.F. (1976). *Foundations of Finance*, Basic Books, New York, NY.

Ferson, W. (1995). Theory and empirical testing of asset pricing models, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Elsevier, Amsterdam, pp. 145–200 (this volume).

Fisher, I. (1906). *The Nature of Capital and Income*, Macmillan, New York, NY.

Fleming, W, and R. Richel (1975). *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, NY.

Fleming, W.H.S., S.G. Grossman, J.L. Vila, and T. Zariphopoulou (1990). Optimal portfolio rebalancing with transactions costs. Working paper, Brown University.

Gennotte, G, and A. Jung (1991). Investment strategies under transaction costs: The finite horizon case. Working paper, University of California at Berkeley.

Gonzalez-Gaverra, N. (1973). Inflation and capital asset market prices: Theory and tests. Unpublished PhD Dissertation, Stanford University.

Grinblatt, M, and S. Titman (1987). The relation between mean–variance efficiency and arbitrage pricing. *J. Bus.* 60, 97–112.

Grinblatt, M, and S. Titman (1995). Performance evaluation, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Elsevier, Amsterdam, pp. 581–610 (this volume).

Hakansson, N. (1970). Optimal investment and consumption strategies under risk for a class of utility functions. *Econometrica* 38, 587–607.

Hakansson, N. (1974). Convergence in multiperiod portfolio choice. *J. Financ. Econ.* 1, 201–224.

Hansen, L.P, and R. Jagannathan (1991). Implications of security market data for models of dynamic economies. *J. Polit. Econ.* 99, 225–262.

Harrison, M, and D. Kreps (1979). Martingales and arbitrage in multiperiod securities markets. *J. Econ. Theory* 20, 381–408.

Harrison, M, and S. Pliska (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stoch. Process Appl.* 11, 215–260.

Hawawini, G, and D. Keim (1995). On the predictability of common stock returns: world-wide evidence, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Elsevier, Amsterdam, pp. (this volume).

He, H, and N.D. Pearson (1991). Consumption and portfolio policies with incomplete markets and short-sale constraints: The infinite dimensional case. *J. Econ. Theory* 54, 259–304.

Hicks, J.R. (1939). *Value and Capital*, Oxford University Press, New York, NY.

Huang, C. (1985a). Information structure and equilibrium asset prices. *J. Econ. Theory* 35, 33–71.

Huang, C. (1985b). Information structure and viable price systems. *J. Math. Econ.* 14, 215–240.

Huang, C.F, and R.H. Litzenberger (1988). *Foundations for Financial Economics*, North Holland, Amsterdam.

Huberman G., S. Kandel, and R.F. Stambaugh (1987). Mimicking portfolios and exact arbitrage pricing. *J. Finance* 42, 1–9.

Ingersoll, Jr., J.E. (1975). Multidimensional security pricing. *J. Financ. Quant. Anal.* 10, 785–798.

Ingersoll, Jr., J.E. (1987). *Theory of Financial Decision Making*, Rowman and Littlefield, Totowa, NJ.

Jarrow, R.A. (1988). *Finance Theory*, Prentice Hall, Englewood Cliffs, NJ.

Jobson, J.D., and B. Korkie (1985). Some tests of linear asset pricing with multivariate normality. *Can. J. Adm. Sci.* 2, 114–138.

Kallberg, J.G, and W.T. Ziemba (1983). Comparison of alternative utility functions in portfolio selection problems. *Manage. Sci.* 29, 1257–1276.

Kamin, J. (1975). Optimal portfolio revision with a proportional transactions cost. *Manage. Sci.* 21, 1263–1271.

Kandel, S., and R.F. Stambaugh (1989). A mean-variance framework for tests of asset pricing models. *Rev. Financ. Studies* 2, 125–156.

Keynes, J.L. (1936). *The General Theory of Employment, Interest and Money*, Harcourt Brace and Company, New York, NY.

Kraus, A, and R. Litzenberger (1976). Skewness preference and the valuation of risk assets. *J. Finance* 31, 1085–1100.

Lee, C.F., J.E. Finnerty and D.H. Wort (1990). *Security Analysis and Portfolio Management*, Scott, Foresman and Company, Glenview, IL.

LeRoy, S, and D. Steigerwald (1995). Volatility, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Elsevier, Amsterdam, pp. 411–434 (this volume).

Levy, H, and H.M. Markowitz (1979). Approximating expected utility by a function of mean and variance. *Am. Econ. Rev.* 69, 308–317.

Levy, H, and P.A. Samuelson (1992) The capital asset pricing model with diverse holding periods. *Manage. Sci.* 38, 1529–1542.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital assets. *Rev. Econ. Stat.* 47, 13–37.

Long, J.B. (1974). Stock prices, inflation, and the term structure of interest rates. *J. Financ. Econ.* 2, 131–170.

Malliaris, A.G, and W.A. Brock (1982). *Stochastic Methods in Economics and Finance*, North Holland, Amsterdam.

Mandelbrot, B. (1963). The variation of certain speculative prices. *J. Bus.* 36, 394–419.

Markowitz, H. (1952). Portfolio selection. *J. Finance* 7, 77–91.

Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investments*, Wiley, New York, NY.

Markowitz, H.M. (1987). *Mean–Variance Analysis in Portfolio Choice and Capital Markets*, Basil Blackwell, Oxford.

Markowitz, H.M. (1991). Foundations of portfolio theory. *J. Finance* 46, 469–477.

Marschak, J. (1938). Money and the theory of assets. *Econometrica* 6, 311–325.

Merton, R.C. (1969). Lifetime portfolio selection under uncertainty: The continuous time case. *Rev. Econ. Stat.* 51, 247–257.

Merton, R.C. (1971). Optimum consumption and portfolio rules in a continuous time model. *J. Econ. Theory* 3, 373–413.

Merton, R.C. (1972). An analytical derivation of the efficient portfolio frontier. *J. Financ. Quant. Anal.* 7, 1851–1872.

Merton, R.C. (1973). An intertemporal capital asset pricing model. *Econometrica* 41, 867–887.

Merton, R.C. (1975). Theory of finance from the perspective of continuous time. *J. Financ. Quant. Anal.* 10, 659–674.

Merton, R.C. (1980). On estimating the expected return on the market: An explanatory investigation. *J. Financ. Econ.* 8, 323–361.

Merton, R.C. (1982). On the mathematics and economics assumptions of continuous-time models, in: W.F. Sharpe and C.M. Cootner (eds.), *Financial Economics: Essays in Honor of Paul Cootner*, Prentice Hall, Englewood Cliffs, NJ.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica* 35, 768–783.

Mossin, J. (1969). Optimal multiperiod portfolio policies. *J. Bus.* 41, 215–229.

Ohlson, J.A, and W.T. Ziemba (1976). Portfolio selection in a lognormal market when the investor has a power utility function. *J. Financ. Quant. Anal.* 11, 393–401.

Osborne, M.F.M. (1959). Brownian motion in the stock market. *Oper. Res.* 7, 145–173.

Pliska, S. (1986). A stochastic calculus model of continuous trading: Optimal portfolios. *Math. Oper. Res.* 11, 371–382.

Roll, R. (1977). A critique of the asset pricing theory's tests: Part I. *J. Financ. Econ.* 4, 129–176.

Rothschild, M, and J. Stiglitz (1970). Increasing risk. I. A definition. *J. Econ. Theory* 2, 225–243.

Ross, S.A. (1976a). Return, risk and arbitrage, in: I. Friend and J. Bicksler (eds.), *Risk and Return in Finance*, Ballinger, Cambridge, MA.

Ross, S.A. (1976b). The arbitrage theory of capital asset pricing. *J. Econ. Theory* 13, 341–360.

Ross, S.A. (1978). Mutual fund separation in financial theory — The separating distributions. *J. Econ. Theory* 17, 254–286.

Rubinstein, M. (1974). An aggregation theorem for security markets. *J. Financ. Econ.* 1, 225–244.

Rubinstein, M. (1976). The valuation of uncertain income streams and the pricing of options. *Bell J. Econ. Manage. Sci.* 7, 407–425.

Samuelson, P. (1969). Lifetime portfolio selection by dynamic stochastic programming. *Rev. Econ. Stat.* 57, 239–246.

Sharpe, W.F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *J. Finance* 19, 425–442.

Sharpe, W.F. (1970). *Portfolio Theory and Capital Markets*, McGraw-Hill, New York, NY.

Stiglitz, J.E. (1989). Discussion: Mutual funds, capital structure, and economic efficiency, in: S. Bhattacharya and G.M. Constantinides (eds.), *Theory of Valuation*, Rowman & Littlefield Publishers, Totowa, NJ.

Stulz, R. (1995). International portfolio choice and asset pricing: an integrative survey, in: R. Jarrow, V. Maksimovic and W.T. Ziemba (eds.), *Finance*, Elsevier, Amsterdam, pp. 201–224 (this volume).

Sundaresan, S.M. (1989). Intertemporal dependent preferences and the volatility of consumption and wealth. *Rev. Financ. Studies* 2, 73–90.

Taksar, M., M. Klass, and D. Assaf (1988). A diffusion model for optimal portfolio selection in the presence of brokerage fees. *Math. Oper. Res.* 13, 277–294.

Tobin, J. (1958). Liquidity preference as behavior toward risk. *Rev. Econ. Studies* 25, 65–86.

Von Neumann, J, and O. Morgenstern (1947). *Theory of Games and Economic Behavior*, 2nd edition, Princeton University Press, Princeton.

Williams J.B. (1938). *The Theory of Investment Value*, Harvard University Press, Cambridge, MA.

Ziemba W.T., C. Parkan and R. Brooks-Hill (1974). Calculation of investment portfolios with risk free borrowing and lending. *Manage. Sci.* 21, 209–222.

Ziemba, W.T. (1974). Choosing investment portfolios when returns have stable distributions, in: P.L. Hammer and G. Zoutendijk (eds.), *Mathematical Programming in Theory and Practice*, North-Holland, Amsterdam, pp. 443–482.

Ziemba, W.T, and R.G. Vickson, editors (1975). *Stochastic Optimization Models in Finance*, Academic Press, New York, NY.