NOTE: Cowles Commission Discussion Papers are preliminary materials circulated privately to stimulate private discussion and are not ready for critical comment or appraisal in publications. References in publications to Discussion Papers (other than mere acknowledgment by a writer that he has had access to such unpublished material) should be cleared with the author to protect the tentative character of these papers.

Contributions to the Econometrics of Financial Behavior

Harry Markowitz

July 12, 1951
Introduction

This is the first part of a two part final report on the writer's "Financial Behavior" project. The second part will be presented after necessary computations, now in progress, have been completed.

In this work formal, mathematical methods are applied to problems relating to portfolio management and investment behavior. Two main types of problems are considered. One type takes the viewpoint of the portfolio manager and asks what are efficient allocations of funds. The other takes the viewpoint of the investigator of economic phenomena, and asks what objective relationships are to be observed in the behavior of "the investor."

The subject matter of our empirical study is the "regulated diversified open end investment company." Our theoretical results apply to various "investing institutions" of which the open end investment company is an example. These institutions represent an important part of the capital market.

This paper presents various theoretical results and consists of Chapters II to VII of the completed report.
Table of Contents

1. The Investment Company and Similar Institutions
2. Implications of Diversification
3. Assumptions
4. Efficient Surfaces
5. Behavior Equations
6. Segregate and Aggregate Relations
7. Time Preference
8. An Experiment: Hypotheses
9. Design of Experiment
10. Results and Evaluation
11. Belief Formation Functions
12. On the Application of Our Results

* To be presented subsequently.
II. Implications of Diversification

It is frequently stated by writers on financial matters that the value of a security is its discounted future returns (see for example Graham and Dodd [5], Williams [15], and Dowling [3]). Since the future is uncertain it must be estimated or "anticipated" returns that we discount. Some authorities require us to subtract an allowance for "risk" from each expected return. Some authorities require the rates at which particular securities are discounted to vary with "risk." Some authorities tell us that the investor does (or should) maximize the discounted value of his portfolio. Others simply speak of the (intrinsic) value of the security and leave us to presume that the investor does (or should) maximize the (intrinsic) value of his portfolio.

The hypothesis (or maxim) that the investor does (or should) maximize discounted return must be rejected. If we ignore market imperfections, the above hypothesis never implies that there is a diversified portfolio which is preferable to all non-diversified portfolios. Diversification is both observed and sensible; a rule of behavior which does not imply the superiority of diversification must be rejected both as a hypothesis and as a maxim.

The above hypothesis fails to imply diversification no matter how the anticipated returns are formed; whether the same or different discount rates are used for different securities; no matter how these discount rates are decided upon or how they vary over time.\(^1\) The hypothesis implies that the investor places all his funds in the security with the

\(^{1}\) The results depend on the assumption that the anticipated returns and discount rates are independent of the particular investor's portfolio.
greatest discounted value. If two or more securities have the same value then any of these or any combination of these is as good as any other.

We can see this analytically: suppose there are \( N \) securities; let \( r_{it} \) be the anticipated return (however decided upon) at time \( t \) per dollar invested in security \( i \); let \( d_{it} \) be the rate at which the return on the \( i^{th} \) security at time \( t \) is discounted back to the present; let \( x_i \) be the relative amount invested in security \( i \). \( \sum x_i = 1 \). We exclude short sales thus relative return (per dollar invested) \( x_i \geq 0 \) for all \( i \). Then the discounted anticipated/ of the portfolio is

\[
(1) \quad R = \sum_{t=1}^{N} \sum_{i=1}^{N} d_{it} r_{it} x_i
\]

\[
= \sum_{i=1}^{N} x_i (\sum_{t=1}^{N} d_{it} r_{it})
\]

\[
(2) \quad r_{i} = \sum_{t=1}^{N} d_{it} r_{it} \text{ is the discounted return of the } i^{th} \text{ security},
\]

therefore

\[
(3) \quad R = \sum x_i r_i \quad \text{where } r_i \text{ is independent of } x_i. \quad \text{Since } x_i \geq 0
\]

for all \( i \) and \( \sum x_i = 1 \), \( R \) is a weighted average of \( r_i \) with the \( x_i \) as non-negative weights. To maximize \( R \), we let \( x_i = 1 \) for \( i \) with maximum \( r_i \). If several \( r_i \), \( a = 1, \ldots, K \) are maximum then any allocation with

\[
K \sum_{a=1}^{K} x_{i a} = 1 \text{ maximizes } R. \text{ In no case is a diversified portfolio preferred to all non-diversified portfolios.}^2
\]

---

2. A more formal proof seems unnecessary here. One can be easily built almost identical to that of footnote , p. .

3. If short sales were allowed an infinite amount of money would be placed in the security with highest \( r_i \).
The discounted anticipated returns hypothesis is a heritage of classical economic theory. The classical model assumes

(a) a perfect capital market, including no costs of transactions and the opportunity to lend or borrow at the same market rate of interest; it also assumes

(b) certainty with respect to all relevant variables. Given these conditions the value of an object is its discounted anticipated returns (where the discount rate is the market rate of interest, and the anticipated returns are the returns foreseen with certainty) (Fisher [4]). In this situation theory and common sense tell us that diversification is unnecessary.

In applying the classical model to the real world, "returns" became "anticipated returns" and the discount rate was required, by some, to vary with risk. But these adaptations are clearly inadequate to deal with risk, as is evidenced by their failure to imply diversification.

Let us consider hypotheses which imply diversification. It will be convenient, for the moment, to consider a static model. Instead of speaking of the time series of returns from the $i^{th}$ security, $(r_{it}, \ldots, r_{it})$, we will speak of "the flow of returns" $(r_{i})$ from the $i^{th}$ security. The flow of returns from the portfolio as a whole is $R = \sum X_{i} r_{i}$. As in the dynamic case, if the $r_{i}$ were known with certainty the investor would place all his funds in that security with maximum $r_{i}$. Similarly, if he wished to maximize "anticipated" return from the portfolio, he would place all his funds in that security with greatest anticipated return (whatever be the way these figures were derived).

We will argue in the next chapter that a reasonable working hypothesis is that the investor does (or should) maximize a utility function depending
on expected return \((E)\) and variance of return \((V)\): \(U = U (E, V)\), where \(E\) is desirable \(\left( \frac{\partial U}{\partial E} > 0 \right)\) and \(V\) undesirable \(\left( \frac{\partial U}{\partial V} < 0 \right)\). We do not claim that no better hypothesis or maxim will ever be found; we only claim that there is good reason to start with this as a working hypothesis and working maxim.

One of the reasons that recommend this hypothesis is that it implies diversification; it implies the right kind of diversification and it implies it for the right reasons. The adequacy of diversification is not thought by investors to depend solely on the number of different securities held. A portfolio with 60 different railway securities, for example, would not be as well diversified as the same size portfolio with some railroad, some public utility, mining, various sort of manufacturing, etc. An inspection of investment company portfolios and conversations with their management, reveal that they deliberately diversify across a wide variety of industries. The reason is that it is generally more likely for firms within the same industry to do poorly at the same time, than for firms in dissimilar industries.

Similarly in trying to make variance small it is not enough to invest in many securities. It is necessary to avoid investing in securities with high covariances among themselves. We should diversify across industries because firms in different industries, especially different kinds of industries, have lower covariances than firms within an industry.

There is a hypothesis which implies both that the investor should diversify and that he should maximize expected return. This hypothesis has been used as a defense of the maxim that the investor should maximize
expected return (Williams [15]). The hypothesis (or maxim) states the investor does (or should) diversify his funds among all those securities which give maximum expected return. The law of large numbers will insure that the actual yield of the portfolio will be almost the same as its expected yield.

This hypothesis can be considered as a special case of the hypothesis that \( J = U (E, V) \). It assumes that there is a portfolio which gives both maximum mean and minimum variance and it commends this portfolio to the investor.

This hypothesis cannot be accepted. First, the law of large numbers does not apply to a portfolio of securities, no matter how large. The returns from securities are too intercorrelated. Second, the portfolio with maximum mean is not the same as that with minimum variance. The investor can choose among portfolios which have various expected returns and subject the investor to varying degrees of risk (variance).

We saw that the classical assumptions, including a perfect capital market and certainty, failed to imply the superiority of diversification. Uncertainty is generally recognized as the main reason for diversification. We cannot say, however, that the existence of diversification proves the existence of uncertainty. By suitably changing the assumption concerning a perfect capital market we can obtain a model which both assumes certainty and implies the superiority of "diversification" (i.e., the superiority of a portfolio with more than two securities). Two examples in the footnote show that the superiority of diversification may be implied if we have (1) imperfect competition or, alternatively, (2) costs of transactions.4

4. Example (1), imperfect competition: Assume as a convenient example that
But, in fact, costs of transactions are small; market imperfections are usually negligible. The problems of "risk" and "risk bearing" play an important part in financial considerations. Diversification is generally put forth as a method of reducing "risk." As a point of logic we must admit that the existence of diversification does not, in itself, prove the existence of uncertainty. Nevertheless, we know from other sources that the diversification of portfolios is principally due to uncertainty rather than market imperfections.

In summary, diversification is an observed and sensible phenomena. The discounted anticipated returns hypothesis seems untenable both as a hypothesis

\[ R = X_1 R_1 (X_1) + X_2 R_2 (X_2) \]

then the Lagrangian equations (footnote ) are

\[ R_1 (X_1) - R_2 (X_2) \quad \text{or} \quad R_1 (X_1) - R_2 (X_2) = 0 \]

A sufficient condition for the superiority of diversification is that

\[ R_1 (X_1) - R_2 (X_2) = 0 \]

\[ X_1 + X_2 = 1 \]

be uniquely satisfied by a solution with \( X_1 > 0, X_2 > 0 \).

Example (2), perfect competition but costs of transactions: Suppose for example, that the investor foresees the following history for two securities:

<table>
<thead>
<tr>
<th>Security</th>
<th>t=0</th>
<th>t=1</th>
<th>t=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price per share</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>cash dividend</td>
<td></td>
<td>( d_1 )</td>
<td>0</td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price per share</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>cash dividend</td>
<td></td>
<td>0</td>
<td>( d_2 )</td>
</tr>
</tbody>
</table>

Suppose the investor seeks to maximize the value of assets at time 2. Suppose the cost of buying or selling a share of stock is \( c \). If \( d_2 < c < d \), the optimal strategy of the investor is to buy only security one at time 0, use the dividend on security 1 to buy security 2 at time 1. At time one he has a "diversified portfolio," i.e., he holds both securities one and two.
and a maxim; since (ignoring market imperfections) it never implies the
superiority of a diversified portfolio. The hypothesis that utility depends
on expected return and variance of return gains in plausibility because
it implies the right kind of diversification for the right reasons. The
existence of diversification does not prove the existence of uncertainty;
but other information makes it plausible that diversification is due
principally to uncertainty rather than market imperfection.
III. Assumptions

In this Chapter we will introduce and evaluate hypotheses to be used subsequently.

Speculation vs. Investment

Financial literature (see Graham & Dodd [5], and Williams [15]) distinguishes two types of security buyers: the investor and the speculator. We are told that the investor is interested in income and long run appreciation; the speculator, in short run capital gains. The investor looks to the soundness of the firm which stands behind a security; the speculator tries to foresee the short run changes in market price. The investor buys to hold as "an investment"; the speculator buys to sell again when the market moves. The investor may also sell when the price of a security rises; not because he had sought a short run gain, but because the security, attractive at the lower price, is no longer so at the higher price.

The open end investment company, the trust fund, the endowed institution (charitable, religious, educational) are thought of as investors, as distinguished from speculators. We will make this distinction. Specifically we will assume that the investment company acts, in purchasing securities, as if it were going to hold these for their investment return (dividends, interest). 1/ 2/ Three reasons (besides its occurrence in the literature) lead us to make the assumption: (1) Open end investment company portfolios

1. Long run capital gains (growth) is conceived as a future increase in dividends.

2. Investment, in the sense described here, is a policy rather than a goal. It would be more satisfactory if we could deduce this policy from the goals of the investor, rather than make it an assumption of the analysis.
are generally managed in a way which typifies the investor as distinguished from the speculator; (2) there is good reason for them being managed in this manner; and (3) by ignoring speculative considerations we attain a tremendous simplification of our model.

(1) The research done by the portfolio supervisor concerns the strong and weak points of firms, rather than attempts to predict short run market movements. Wide diversification and only moderate turnover of securities is suggestive of the investor as opposed to the speculator. Open end investment companies state their goals in terms of yield, growth, and stability rather than short run capital gains.

(2) There is good reason for persons in "trust" positions to abstain from playing the turns of the market. The market is capricious; attempts to systematically predict its course have generally proved unsuccessful (Cowles [1], Keynes [9]).

(3) By ignoring "speculative" considerations we attain a tremendous simplification of our models. The problem of the rational investor is an allocation problem. He must allocate funds among various securities on the basis of his beliefs about their future returns. The problem of the rational speculator is a problem of strategy. If he were to "rationally" attempt to gain certain goals concerning his venture as a whole he would have to plan out a strategy giving his actions over time for every possible contingency/ Optimal strategies are far more complicated than optimal allocations.

In the following chapters we deal with problems of allocation. An example of an optimal strategy can be found in Markowitz [11].


Probability Beliefs

We will assume that the investor acts as if he had subjective probability beliefs about future variables (Savage [13]).

If we asked an investment manager what probability he attached to an event (say, default by railroad A in 1952) he would probably be unable to answer. We might however, expect him to express an opinion if we asked him which he personally considered more likely, that U.S. Steel will pass a dividend or that Coca Cola will. In general we would expect him to be able to tell us whether he personally considered event A more likely than B, B more likely than A or both equally likely. In particular, he could tell us which, if either, he considered more likely, that (say) U.S. bonds will default or that a poker player would draw a royal flush. The latter event has an objective probability. If the investment manager were consistent in his answers those answers would reveal a system of probability beliefs.

We can not expect the investor to be consistent in every detail. We can, however, expect his probability beliefs to be roughly consistent on important matters that have been carefully considered. We should also expect that he will base his actions upon these probability beliefs—even in part though they be subjective.

In the next three chapters we will assume that the investor has static beliefs; i.e., we will assume the investor has beliefs as to "the" probability distribution of "the flow" of returns from the various securities. In Chapter 7 we will let this probability distribution be a function of time.

There are several reasons for giving first and careful consideration to the static case. We often think and speak in terms of the static case. We discuss "the" return from and "the" variability of various securities.
For many purposes this static viewpoint is sufficient. For other purposes—e.g., in the evaluation of "growth" vs. present "yield"—some dynamic elements must be considered. The static case is considered first because of its simplicity, as well as its importance.

Theoretical Results

We will seek theoretical results of two sorts: "behavior equations" and "efficient surfaces."

Behavior equations are equations which relate human actions to observable variables. They are not expected to hold exactly, but only "on the average." Behavior equations typically—though not necessarily—have unknown parameters which can be estimated with the aid of past observations.

We will derive systems of behavior equations by combining assumptions as to behavior under uncertainty with descriptions of "the investor." Our equations thus will be consequences of economic hypotheses and could be used to test these hypotheses.

Our equations will be based on two sorts of assumptions: (a) assumptions as to the behavior of the investor; and (b) assumptions as to the knowledge of the investigator.

(a) Concerning the investor we will assume (i) that the investor acts as if he maximizes utility \( U \) which depends on a finite number of moments \( (M_1, \ldots, M_N) \) of the probability distribution of returns from the portfolio as a whole: \( U = U (M_1, \ldots, M_N) \); (ii) if a portfolio \( P_o \) is chosen by the investor (i.e., if \( P_o \) maximizes \( U \) subject to the constraint that \( P \) be in an "attainable set" \( S \)) then \( P_o \) either minimizes or maximizes the value of some moment \( M_j \) for all \( P \) in \( S \) with the same \( M_j \) \( j \neq i \) as \( P_o \).
(b) One system of behavior equations can be derived if the investigator can specify a set of moments which include the set of moments that enter the utility function. If the investigator leaves out a relevant moment, he will obtain false equations. If he includes an irrelevant moment he will not thereby get false equations, but will incur other penalties.

Another system of equations requires both (1) knowledge of (or beliefs about) the moments which enter the utility function, and (2) some knowledge (or beliefs) as to how the investor's subjective probability beliefs are formed. The derived equations could be used to test the investor's "knowledge" and beliefs.

The assumption that action depends (at least to a good approximation) on a finite number of moments, is not un plausible. The moments are descriptive statistics. Particular moments and combinations of them are universally used to describe the average, the variability, the skewness, etc., of a distribution which interest the investor. For a large class of distributions, the distribution is determined by its moments (infinite in number). Under certain conditions a probability distribution can be approximated to any degree of accuracy if a large enough number of its moments are known (see Gramé [2]).

Assumption (a) is hardly subject to doubt. For example, suppose the investor had the choice of two probability distributions \( P_1(x) \) or \( P_2(x) = P_1(x - \lambda) \) where \( \lambda \) is a positive constant. (In other words \( P_2 \) is \( P_1 \) shifted to the right: \[ P_1(x) \quad P_2(x) \])

Certainly \( P_2 \) is preferred to \( P_1 \). This implies that expected return is one of the moments which enters the utility function and that it is maximized for given values of the other moments. This implies that assumption (a)
is satisfied.

Perhaps the most popular alternative hypothesis is that the investor attaches a utility \( U(x) \) to each possible outcome \( x \), and acts so as to maximize expected utility (von Neumann [14], Marschak [12]). If the investor did maximize expected utility, the function \( U(x) \) could be approximated, over the relevant range, by a polynomial: \( U^* = \alpha_0 + \alpha_1 x + \cdots + \alpha_n x^n \). Then we approximately maximize \( E U(x) \) if we maximize \( U^* = \sum_{i=0}^{N} \alpha_i E x^i \).

\( E x^i \) is a non-central moment; the first \( n \) non-central moments are uniquely related to the first \( n \) central moments \( M_1, \ldots, M_n \). To maximize \( U^* \) we must maximize \( \alpha_n m_n \) given \( m_i \), \( i < n \) and consequently must extremize \( M_n \) given \( M_i \), \( i < n \). Thus even if the expected utility hypothesis is true our hypothesis is approximately correct.

But our hypothesis is more than the expected utility hypothesis. The expected utility hypothesis says that (approximately) \( U = U(M_1, \ldots, M_n) = u(m_1, \ldots, m_n) = \sum_{i=0}^{m} \alpha_i E x^i \). Our hypothesis is that \( U = U(M_1, \ldots, M_n) \) where the only restriction is (a1)\(^{3/2} \).

\[ 3. \] The expected utility hypothesis applies to various aspects of behavior under uncertainty—insurance, lotteries, commercial risks—as well as financial behavior. One phenomena which apparently contradicts this hypothesis are lotteries with many prizes of different size (e.g., one large grand prize plus various smaller prizes). If this forces us to admit that the lottery purchaser would prefer, say, some fair multiple prized lottery to any fair single prize lottery, then we have contradicted the expected utility hypothesis. For, as we shall show shortly the expected utility hypothesis never implies the superiority of a multiple prize lottery, just as the expected income hypothesis never implied the superiority of a diversified portfolio. This does not prove that the expected utility hypothesis will lead to contradictions in the field of financial behavior. It does serve to make desirable the added generality of the hypothesis we use in deriving behavior equations.

Suppose there are a finite number of possible prizes \(-K, -K+1, \ldots, -1 0, \ldots, 1, \ldots, L \) where (for generality) the unit may be very small (say, one
penny) and \( K \) and \( L \) are very large. Suppose each outcome \( X \) has a utility \( U_X \). Let \( p_X \) be the probability of obtaining prize \( X \). For a fair lottery

\[
\sum_{X=1}^{L} p_X X = 0.
\]

Expected utility is

\[
E = \sum_{X=1}^{L} U_X p_X.
\]

By the nature of probability, \( \sum_{X=1}^{L} p_X = 1 \), \( p_X \geq 0 \). Suppose the lottery buyer could construct any fair lottery as wished. We must find \( (p_1, \ldots, p_0, \ldots, p_L) \) so as to maximize \( E \) subject to the constraints. If a multiple prize lottery, with at least three \( p_a \) not zero were at least as good as any other lottery then \( E = \sum_{i=1}^{\alpha} p_{a_i} U_{a_i} \) would have a relative maximum subject to

\[
\sum_{i=1}^{\alpha} p_{a_i} = 1
\]

\[
\sum_{i=1}^{\alpha} p_{a_i} X_{a_i} = 0
\]

where \( a_1, \ldots, a_\alpha \) are just those indices such that \( p_{a_i} \neq 0 \). For \( E \) to be a constrained relative maximum we must have

\[
\frac{\partial}{\partial p_{a_j}} \left( \sum_{i=1}^{\alpha} p_{a_i} U_{a_i} + \lambda_1 \left( \sum_{i=1}^{\alpha} p_{a_i} - 1 \right) + \lambda_2 \left( \sum_{i=1}^{\alpha} p_{a_i} X_{a_i} \right) \right) = 0 \quad \text{for } j = 1, \ldots, \alpha.
\]

Specifically,

\[
U_{a_j} + \lambda_2 X_{a_j} + \lambda_1 = 0, \quad j = 1, \ldots, \alpha.
\]

or

\[
U_{a_j} - U_{a_1} = -\lambda_2 (X_{a_j} - X_{a_1}), \quad j = 2, \ldots, \alpha.
\]

Since this gives us \( \alpha - 2 \) equations in no unknowns it is either satisfied by no set of \( p_{a_1} \)'s or by every set of \( p_{a_1} \)'s. In either case no multiple prized lottery is superior to all single prize lotteries. (In a single prize lottery, two \( p_{a_1} \)'s (say \( p_{a_1} \) and \( p_{a_2} \)) are not zero. \( a_1 \) is the loss; \( p_{a_1} \), its probability; \( a_2 \) is the win, \( p_{a_2} \), its probability.)

It may be objected that our argument does not take into account the other prospects of the lottery ticket buyer. He may already expect \( X_1 \) with probability \( p_1 \). If the outcome of the lottery is independent \( 1 \) of his other prospects, the probability of ending with \( X_1 \), taking into account his initial prospects \( p_1^* \) and the lottery prospect \( p_i \), is \( p_1^* = \sum_{a} p_a p_{i-a} \). His expected
Efficient Sets

Suppose we knew that the investor's actions were based on expected return and variance; that he liked expected return and disliked variance of return. Suppose we also knew his probability beliefs about each security in which he could invest. What portfolios give minimum variance for given expected return and maximum expected return for given variance? What combinations of expected return (E) and variance of return (V) are associated with these portfolios? We will call these portfolios and these pairs (E, V) the E-V efficient sets in the portfolio space and in the (E, V) space respectively.

We could similarly define efficient sets for other combinations of moments.

We will characterize and find ways of computing the efficient sets when the investor's action depends on E and V. Many of our results will apply when action depends on E and any measure of risk satisfying certain properties. We have a limited knowledge about efficient surfaces when action depends on combinations of moments other than E and V.

The calculation of efficient surfaces might possibly be of practical use. Perhaps there are ways, by combining statistical techniques and the judgment of experts, to form reasonable probability beliefs (e.g., means, variances, covariances) concerning available securities. We could use these

\[
\text{utility is } U = \sum_{i} p_i^0 U_i \\
= \sum_{i} \sum_{a} p_{i-a}^* U_i
\]

We can see that E is linear in the p_i; therefore the argument used previously still holds. A multiple prized lottery is never superior to all single prize lotteries.
to compute the attainable combinations of \((E, V)\). The investor, being then informed of what \((E, V)\) combinations were attainable could state which he desired. We could then find the portfolio which gave this desired combination. The question of practical use will be discussed further in Chapter 12.

Here we will point out reasons which recommend the use of expected return and variance (besides the transient fact that the \(E, V\) efficient surfaces are the only ones we have worked out in detail). The reasons apply both to the use of \(E, V\) efficiency as a working maxim and a working hypothesis. We do not claim that no better hypothesis or maxim will ever be found; but that here is a sensible place to start.

We have already noted that the minimizing of variance implies diversification; implies the right kind of diversification and implies it for the right reasons. The observed policy of diversifying across a wide range of industries may be thought of as an attempt to reduce variance by seeking securities with low covariances.

The concepts "yield" and "risk" appear frequently in financial writings. As a rule if the term "yield" were replaced by "expected yield" or "expected return," and "risk" by "variance of return" little change of apparent meaning would occur.

The third moment \((M_3)\) may be associated with a "propensity to gamble." For example, if \(U = M_1 + \alpha M_2, \quad \alpha < 0\), then the individual would never take a fair bet. If \(U = M_1 + \alpha M_2 + \beta M_3, \quad \alpha < 0, \beta > 0\), then the individual would take some fair bets.

Perhaps—for institutions in trust positions, where yield is considered to be a good thing, risk, is a bad thing, gambling, to be avoided—\(E, V\) efficiency is reasonable as a working hypothesis and a working maxim.
IV. Efficient Sets

In this chapter when we refer to efficiency we mean $E$, $V$ efficiency, unless we specify otherwise. We will first present the efficient sets geometrically in the three and four security cases. Then we will derive efficient sets in the $N$-security case. We will see that the main characteristics illustrated by the geometrical examples remain true more generally.

Let the expected returns per dollar invested in securities $1, \ldots, N$ be $\mu_1, \ldots, \mu_N$; let variances and covariances be $\sigma_{11}, \ldots, \sigma_{ij}, \ldots, \sigma_{NN}$; the amount allocated to security $i$ is $x_i$; since we exclude short sales $x_i$ is non-negative; let $A$ be the total amount of assets (given exogenously).

In the three security cases we have

(1) $E = \frac{3}{2} \sum_{i=1}^{3} x_i \mu_i$

(2) $V = \frac{3}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} x_i x_j \sigma_{ij}$

(3) $\sum_{i=1}^{3} x_i = A$

(4) $x_i \geq 0$ for $i = 1, \ldots, 3$.

From (3) we know (3') $x_3 = 1 - x_1 - x_2$. If we substitute (3') in equations (1) and (2) we get $E$ and $V$ as functions of $x_1$ and $x_2$. The exact formulae are

1. $E = \mu_3 + x_1 (\mu_2 - \mu_3) + x_2 (\mu_2 - \mu_3)$

   $V = x_1^2 (\sigma_{11} - 2x_1 \sigma_{13} + \sigma_{33}) + x_2^2 (\sigma_{22} - 2 \sigma_{23} + \sigma_{33}) + 2x_1 x_2 (\sigma_{12} + \sigma_{13} - \sigma_{23} + \sigma_{33}) + 2x_2 (\sigma_{23} - \sigma_{33}) + \sigma_{33}$
not important here; we can simply write

(a) \( E = E(x_1, x_2) \)

(b) \( V = V(x_1, x_2) \)

(c) \( x_1 \geq 0, x_2 \geq 0, A = x_1 - x_2 \geq 0 \)

The **attainable set** consists of all portfolios satisfying constraints (3) and (4) (or alternatively (3) and (c)). An **iso-mean line** consists of all points (portfolios) with the same expected return. An **iso-variance line** consists of all portfolios with the same variance. The attainable set, iso-mean lines and iso-variance lines can be represented geometrically in the \( x_1, x_2 \) space. See figure 1. The attainable set is the triangle \( a b c \).

![Figure 1](image-url)
Typically\(^2\), the iso-mean lines are parallel straight lines; the iso-variance lines are a system of concentric ellipses. The center of the system is the point (degenerate ellipse) at which \(V(x_1, x_2)\) is a minimum. The further an iso-variance line is from this center, the greater the variance it represents. We will label the center of the iso-variance system, \(\bar{x}\); its mean and variance, \(\bar{E}\) and \(\bar{V}\) respectively. \(\bar{x}\) may fall either inside or outside the attainable set. Figure 2 illustrates a case wherein \(\bar{x}\) falls within a b c. In this case:

\(\bar{x}\) is efficient. For no other portfolio has a \(V\) as low as \(\bar{V}\); therefore no portfolio can have either smaller \(V\) (with the same or greater \(E\)) or greater \(E\) with the same or smaller \(V\).

No point (portfolio) with expected return \(E\) less than \(\bar{E}\) is efficient. For we have \(\bar{E} > E\) and \(\bar{V} < V\).

Consider all points with a given expected return \(E\); i.e., all points on the iso-mean line associated with \(E\). The point of the iso-mean line at which \(V\) takes on its least value is the point at which the iso-mean line is tangent to an iso-variance curve. We call this point \(\hat{x}(E)\). If we let \(E\) vary \(\hat{x}(E)\) traces out a curve. Algebraic considerations, presented later, show us that this curve is a straight line. We will call it the critical line \(\ell\). The critical line passes through \(\bar{x}\); for this point minimizes \(V\) for all points with \(E(x_1, x_2) = E\). As we go along \(\ell\) in either direction from \(\bar{x}\), \(V\) increases. The segment of the critical line from \(\bar{x}\) to the point where the critical line crosses the a b boundary line is part of the efficient set. The rest of the efficient set is the segment of the a b line from d to b. b is the point of maximum attainable \(E\).

2. We will discuss a-typical cases later.
Figures 3 to 5 illustrate various other cases. In Figure 3 \( X \) lies outside the admissible area but the critical line cuts the admissible area.
The efficient line begins at the point of (constrained) minimum variance (in this case on the \( a \ b \) line). It moves towards \( b \) until it intersects the critical line, moves along the critical line until it intersects a boundary and finally moves along the boundary to \( b \). In Figure 4, \( X \) is exterior to \( a \ b \ c \) and does not penetrate \( a \ b \ c \). In this case the efficient set is the same as it would be if the admissible set had been the line \( a \ b \); i.e., if only securities 1 and 2 had been available.

Illustrated in Figure 5, is the case in which two securities have the same \( \mu_1 \).

Just as we used the equation \( \sum_{i=1}^{N} x_i = 1 \) to reduce the dimensionality in the three security case, we can use it to represent the four security case in 3-dimensional space. Eliminating \( x_4 \) we get \( E = E(x_1, x_2, x_3), V = V(x_1, x_2, x_3) \).

The attainable set is represented, in three-space, by the tetrahedron with vertices \((0,0,0), (0,0,A), (0,A,0), (A,0,0)\), representing portfolios with, respectively, \( x_4 = A, x_1 = A, x_2 = A, x_3 = A \).

Let \( s_{1,2,3} \) be the subspace consisting of all points with \( x_4 = 0 \).

Similarly we can define \( s_{a_1,\ldots,a_\alpha} \) to be the subspace consisting of all points with \( x_4 = 0 \) if \( a_1,\ldots,a_\alpha \). For each subspace \( s_{a_1,\ldots,a_\alpha} \) we can define a critical line \( \ell_{a_1,\ldots,a_\alpha} \). This line is the locus of points \( P \), where \( P \) minimizes \( V \) for all points in \( s_{a_1,\ldots,a_\alpha} \) with the same \( E \) as \( P \).

If a point is in \( s_{a_1,\ldots,a_\alpha} \) and is efficient it must be on \( \ell_{a_1,\ldots,a_\alpha} \). The efficient set may be traced out by stating at the point of minimum available variance, moving continuously along various \( \ell_{a_1,\ldots,a_\alpha} \) according to definite rules, ending in a point which gives maximum \( E \). As in the two dimensional
The point with minimum available variance may be in the interior of the available set or on one of its boundaries. Typically we proceed along a given critical line until either this line intersects one of a larger subspace or meets a boundary (and simultaneously the critical line of a lower dimensional subspace). In either of these cases the efficient line turns and continues along the new line. The efficient line terminates when a point with maximum $E$ is reached. Examples are given in Figures 6 and 7.
Now that we have seen the nature of the set of efficient portfolios, it is not difficult to see the nature of the set of efficient \((E, V)\) combinations. In the three security case

\[
E = a_1 x_1 + a_2 x_2 \quad \text{is a plane}
\]

\[
V = a_{11} x_1^2 + a_{12} x_1 x_2 + a_{22} x_2^2 \quad \text{is a paraboloid. As shown in Figure 8, the section of the E-plane over the efficient portfolio set, is a series of connected line segments.}
\]

The section of the \(V\)-paraboloid over the efficient portfolio set is a series of connected parabola segments. If we plotted \(V\) against \(E\) for efficient portfolios we would again get a series of connected parabola segments. This result obtains for any number of securities.

![Figure 8.](image)

![Figure 9.](image)
The N Security Case

We will see that in the N security case, as in the simpler cases treated above, the efficient surface is a series of connected line segments. We will give the formulae for these segments; we will show how the efficient set can be traced out, and how the set of efficient \((E,V)\) combinations can be computed. Our discussion will require more mathematical background of the reader than previously.

Let \(\Sigma\) be a covariance matrix; \(\mu\), a vector expected returns; \(\sigma=(1,\ldots,1)\); \(X=(x_1,\ldots,x_N)\), a portfolio; \(R_N\) euclidean N-space.

\[
\begin{align*}
(1) & \quad V = XX' = V(X) \\
(2) & \quad E = \mu X' = E(X)
\end{align*}
\]

**Definition 1.** A portfolio is **attainable** if it satisfies

\[
(3) \quad \sigma X' = \Lambda
\]

\[
(4) \quad X \geq 0
\]

**Definition 2.** An attainable portfolio \(X_0\) is **efficient** — for given \(\Sigma\) and \(\mu\) — if there does not exist \(X_1\) such that either

\[
\begin{align*}
V(X_1) & \leq V(X_0) \\
E(X_1) & > E(X_0)
\end{align*}
\]

or

\[
\begin{align*}
V(X_1) & < V(X_0) \\
E(X_1) & \geq E(X_0)
\end{align*}
\]

**Definition 3.** A couple \((E,V)\) is efficient if \(E=E(X_0), V=V(X_0)\) where \(X_0\) is efficient.
Definition 4. The subspace $a_1, \ldots, a_\alpha$ is the set of all points such that $x_j = 0$, for all $j \neq a_1, \ldots, a_\alpha$.

Definitions 5, 6, 7. A set $S$ is convex if when points $X_1$ and $X_2$ are in $S$, the point $\alpha X_1 + (1-\alpha) X_2$ is also in $S$ where $0 \leq \alpha \leq 1$. A function $f(X)$ is convex over a (convex) set $S$ if for any $\alpha$, $0 \leq \alpha \leq 1$ and any two points $X_1, X_2$ in $S$, $f(\alpha X_1 + (1-\alpha) X_2) \leq \alpha f(X_1) + (1-\alpha) f(X_2)$. A function is strictly convex over the set $S$ if for any $\alpha$, $0 < \alpha < 1$ and $X_1, X_2$ in $S$, $f(\alpha X_1 + (1-\alpha) X_2) < \alpha f(X_1) + (1-\alpha) f(X_2)$.

Lemma 1. If a function is convex on a set $S$, it is also convex on any convex subset of $S$. Similarly for strict convexity.

Lemma 2. The admissible set is convex.

Proof if $X_1 = (x_{11}, x_{21}, \ldots, x_{N1})$

and $X_2 = (x_{12}, x_{22}, \ldots, x_{N2})$

satisfy $\sum_{i=1}^{N} x_{ij} = 1$

and $x_{ij} \geq 0$ for all $ij$

then we also have

$\sum \alpha x_{i1} + (1-\alpha) x_{i2} = \alpha \sum x_{i1} + (1-\alpha) \sum x_{i2} = 1$

and $x_{i1} + x_{i2} \geq 0$

Theorem 1. Variance is always convex over $R_N$; it is strictly convex over $R_N$ if and only if the covariance matrix $\Sigma$ is non-singular.
Proof

\[ f(\alpha x_1 + (1-\alpha) x_2) = \]

\[ (\alpha x_1 + (1-\alpha) x_2) \Sigma (\alpha x_1 + (1-\alpha) x_2)' = \]

\[ \alpha^2 x_1 \Sigma x_1' + 2 \alpha^2 (1-\alpha) x_1 \Sigma x_2' + (1-\alpha)^2 x_2 \Sigma x_2' \]

\[ \alpha f(x_1) + (1-\alpha) f(x_2) = \]

\[ \alpha x_1 \Sigma x_1' + (1-\alpha) x_2 \Sigma x_2' \]

(multiply by \((\alpha + (1-\alpha))\))

\[ \alpha^2 x_1 \Sigma x_1' + (1-\alpha)^2 x_2 \Sigma x_2' + \]

\[ \alpha(1-\alpha) \left\{ x_1 \Sigma x_1' + x_2 \Sigma x_2' \right\} \]

Thus \( f(\alpha x_1 + (1-\alpha) x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2) \)

when and only when

\[ x_1 \Sigma x_1' + x_2 \Sigma x_2' \geq 2 x_1 \Sigma x_2' \]

\[ (x_1 \Sigma x_1' - x_1 \Sigma x_2') + (x_2 \Sigma x_2' - x_1 \Sigma x_2') \geq 0 \]

\[ (x_1 - x_2) \Sigma (x_1 - x_2)' \leq 0 \]

Since \( \Sigma \) is positive definite if and only if it is non-singular, and positive semi-definite otherwise, our theorem follows.

We will not derive efficient sets for all \( \Sigma \), but will restrict our attention to the case in which \( V \) is a strictly convex function over the set \( \{ x \mid e x' = 1 \}^3 \).

---

3. To be read: "the set of all \( x \)'s such that \( e x' = 1 \)."
By theorem 1 and lemma 1 we know that this is so if $V$ is non-singular.
It is not necessary, however, for $V$ to be strictly convex over $R^n$ to
be strictly convex over $\left\{ x \mid e^x = A \right\}$. If we substitute $x = \sum_{i=1}^{N-1} x_i$ in
$V(x_1, \ldots, x_N)$ we get $V^*(x_1, \ldots, x_{N-1})$. It is necessary and sufficient for
$V^*(x_1, \ldots, x_{N-1})$ to be strictly convex over $R_{N-1}$. Specifically let
\[
\begin{pmatrix}
x_1 \\
\vdots \\
x_N
\end{pmatrix} =
\begin{pmatrix}
0 \\
\vdots \\
0
\end{pmatrix} +
\begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
-1 & \cdots & 0 & 1
\end{pmatrix}
\begin{pmatrix}
x_1 \\
\vdots \\
x_{N-1}
\end{pmatrix}
\]
which we will write more briefly
\[
x^{(N)} = e_N + D x^{(N-1)},
\]
Then
\[
V = x^{(N)} \Sigma x^{(N)} = \left\{ e_N + x^{(N-1)}_D \right\} \Sigma \left\{ e_N + D x^{(N-1)} \right\} 
\]
\[
= e_N \Sigma e_N + 2 e_N \Sigma D x^{(N-1)} + x_N \Sigma D x^{(N-1)}
\]
The first term of this sum is a constant, the second is a linear function
of $x_1, \ldots, x_{N-1}$, the third term is a quadratic in $x_1, \ldots, x_{N-1}$, where $D$ is
either positive definite or indefinite.
$V = K + L(x) + Q(x)$. In such situations we have

**Lemma 3.** If $g(x) = K + L(x) + f(x)$ where $K$ is a constant and $L$ is linear,
g(x) is strictly convex if and only if $f(x)$ is strictly convex. For
\[
g(\alpha x + (1-\alpha)y) - \alpha g(x) - (1-\alpha)g(y) > 0
\]
if and only if
\[
K + L(\alpha x + (1-\alpha)y) + f(\alpha x + (1-\alpha)y) - K + L(x) + (1-\alpha)L(y)
+ \alpha f(x) + (1-\alpha)f(y) > 0
\]
which is equivalent to

\[ f(\alpha X + (1-\alpha)Y) - \alpha f(X) - (1-\alpha)f(Y) > 0 \]

From lemma 3 and the preceding discussion we can easily see

**Theorem 2.** \( V \) is convex over \( \{X|Xe=1\} \) if and only if \( \Sigma D \Sigma' \) is non-singular.

\( \Sigma \) is an \( N \times N \) matrix. \( D \) is \( N \times 1 \times N \). \( \Sigma D \Sigma' \) may be non-singular, even though \( \Sigma \) is singular. The geometry of the situation is best seen in the three security space. When the rank of \( \Sigma \), written \( \rho(\Sigma) \), is 3, the iso-variance surfaces in \( R_3 \) are ellipsoids. When intersected with the plane \( eX' = 1 \) and projected on the \( (x_1, x_2) \) plane, the iso-variance regions are ellipses. When \( \rho(\Sigma) = 2 \) the iso-variance surfaces in \( R_3 \) are elliptical cylinders with common axes. If the axis of this system of cylinders is not parallel to the plane \( eX' = A \), then, the iso-variance regions, intersected with \( eX' = A \) and projected on the \( x_1, x_2 \) space, are still ellipses. If the axis of the system is parallel to the plane, then the projected intersections are no longer ellipses but pairs of parallel lines. If \( \rho(\Sigma) = 1 \), the iso-variance surfaces in \( R_3 \) are planes, the projected intersections are lines.

The assumption that \( |\Sigma| \neq 0 \) is often made and is usually considered quite general. The added generality of \( |D\Sigma D'| \neq 0 \) is valuable, for frequently in our analysis there is a security with no variance. In this case, if

\[ \rho(\Sigma) = 1, |D\Sigma D'| \neq 0; \text{ for we can write } D\Sigma D' = \begin{pmatrix} \Sigma & 0 \\ \vdots & \ddots \\ 0 & 0 \\ 0 & \ddots & 0 \end{pmatrix} \text{ where } \Sigma, \]

and therefore \( D\Sigma D' \), is non-singular. It would be desirable, of course, to have an analysis for all \( \Sigma \).
Lemma 4. If \(|D^2D| \neq 0\), there exists one and only one point \(\bar{X}\) at which \(V\) assumes its minimum on the set \(\{x \mid x \geq 0, e^x = \epsilon\}\).

Existence follows from the continuity of \(V\) and the compactness of the attainable set. Uniqueness follows from the strict convexity of \(V\) over \(\{x \mid e^{x'} = \bar{A}\}\) and the convexity of the attainable set.

We will let \(\bar{E} = E(x)\), \(\bar{V} = V(x)\).

Lemma 5. If \(|D^2D| \neq 0\), there exists a set of points which maximize \(E\) on the attainable set. There is one and only one point \(\bar{x}\) which minimizes \(V\) for all points which maximize \(E\).

We will let \(\bar{E} = E(\bar{x})\), \(\bar{V} = V(\bar{x})\).

Compactness and continuity assures us that \(\max E\) is attained. The set of attainable points which maximize \(E\) is \(\{x \mid x \geq 0, e^{x'} = \bar{A}, x' \geq 0\}\) and is convex. This, and the strict convexity of \(V\) imply the uniqueness of \(\bar{x}\).

Lemma 6. \(x\) and \(\bar{x}\) are efficient.

Lemma 7. No \(x\) with \(E(x) < \overline{E}\) is efficient.

Lemma 8. If \(x_0\) minimizes \(V\) for all attainable \(x\) with given \(E\) or more, i.e., for all \(x\) in \(\{x \mid x \geq 0, e^{x'} = \bar{A}, x' \geq \bar{E}\}\) where \(E \geq \bar{E}\), then \(\bar{X} = x_0\) (i.e., then \(x_0\) has that \(E\)).

Suppose otherwise. Then there exists \(x_1 < x_2 > E\), which minimizes \(V\) for all \(x\) such that \(\bar{X} \geq E\). The plane \(\bar{X} = E\) lies between \(x_1\) and \(x_2\). A straight line drawn between \(x_1\) and \(x_2\) cuts this plane (say at \(x\)). By lemma 4 \(V(x_1) > V(x)\); by strict convexity \(V(x_1) > V(x) > V(\bar{x})\).

Theorem 3. Let \(F\) be the efficient portfolio set. \(F\) is the set of all attainable \(x\)'s which minimize \(V\) for given \(E\), where \(\bar{E} \leq E \leq \overline{E}\).

Let \(\bar{x}(E) \leq E \leq \overline{E}\) be the attainable \(x\) which minimizes \(V\). Our theorem
states that
\[ \mathcal{F} = \left\{ X \mid \exists E \in \mathcal{E}, \exists x^* \in \mathcal{X}(E), E \not\subseteq E \right\} \]

Proof:

Suppose \( x_0 \) is efficient; then, among other things, it minimizes \( V \) for
\[ \left\{ X \mid x \in E - x_0, E \not\subseteq E \right\} \] by lemmas 4 and 6. Suppose \( x_0 \) minimizes \( V \) for given \( E_0, E \not\subseteq E_0 \), By lemma 7 no \( x \) with higher \( E \) gives as low a \( V \). Also since \( x_0 \) minimizes \( V \) for \( \{ X \mid \text{attainable}, E \not\subseteq E_0 \} \) any \( x \) with lower \( V \) must give lower \( E \).

Lemma 9. \( \mathcal{X}(E) \) is single valued. It seems self-evident that \( \mathcal{X}(E) \) is continuous. The writer has been unable to find a simple proof of this. Continuity of \( \mathcal{X}(E) \) follows as a special case of:

Theorem 5. Let \( S \) be a compact, convex subset of \( \mathbb{R}^n \). Let \( f(y) \) be continuous and strictly convex on \( S \). Let \( y(b) \) maximize \( f \) for all \( y \in S \cap \{ y | ay = b \} \) where \( a \) is a fixed vector. Then \( y(b) \) is single valued and continuous.

Proof: That \( y(b) \) is single valued follows immediately from the convexity of \( S \cap \{ y | ay = b \} \) and the strict convexity of \( b \).

Continuity is not so simple. We will show that given any \( b_0 \) and \( \epsilon \),
\[ \exists \delta > 0 \quad b_0 - \delta < b < b + \delta \implies |f(y) - f(y_0)| < \epsilon \] where \( f(y_1, y_2) \) is the distance between \( y_1 \) and \( y_2 \). Let \( C(\alpha) \) be the closed sphere with center \( w \) and radius \( \alpha \); i.e., all points \( y \in C(\alpha) \). We must show that
\[ \exists \delta > 0 \quad |b - b_0| < \delta \implies y(b) \in C(\epsilon) \]

\( B = C(y_0) \cap \{ y | ay = b \} \) is a compact subset of \( \{ y | ay = b \} \). \( f(y) > f(y_0) \) for all \( y \in B \), since \( y_0 \) minimizes \( f \) for \( \{ y | ay = b \} \). \( \min f(y) = f(y_0) \) for \( y \in B \) (by compactness of \( B \) as a subset of \( \{ y | ay = b \} \)). Since \( S \) is compact \( f \) is uniformly continuous. \( \exists \delta > 0 \quad |f(y_1, y_2)| < \delta \implies |f(y_1) - f(y_2)| \leq \frac{\epsilon}{3} \). We can choose \( \delta \) so that any point which is on the sphere \( C(\epsilon) \) and is between
the planes $ay' = E_0 + \delta$ and $ay' = E_0 - \delta$ is within $w$ of some point on $B$. $w > \delta$; both $ay' = E_0 + \delta$ and $ay' = E_0 - \delta$ intersect the sphere $C_E(y_0)$. Now we must show that $|E - E_0|<\delta$ if $f(y(E), E_0) \leq t$. If this were not so then $f(E)$ would not be in $C_E(y_0)$. A line between $y_0$ and $f(E)$ would cut $C_E(y_0)$, say at $y_0$.

$$f(y_0) \geq f(y_0) + 1/3 \gamma = f(y_0) + 2/3 \gamma.$$ 

$$f(f(E)) > f(y_0) \geq f(y_0) + 2/3 \gamma > f(y_0).$$

There is at least one point on $ay + E$ and in $C_E(y_0)$; call this point $y_w$.

$$f(y_w) \leq f(y_0) + 1/3 \gamma \leq f(y_0) + 2/3 \gamma \leq f(f(E)).$$ Therefore $f(E)$ could not have minimized $f$ for all $y$ in $\{ y | ay = E \}$.

**Corollary.** $f(E)$ is continuous.

In the theorems and lemmas from lemma 1 onward the only information used concerning $V$ is that it is continuous and strictly convex. These results apply whenever the investors decisions depend on $E(X)$ and $r(X)$ where $r$ is a strictly convex "risk function." There is a close connection between diversification to eliminate risk and the convexity of the risk function. For example, if we have two portfolios $X_1, X_2$ with equal risk $r(X_1) = r(X_2)$ then strict convexity implies that a diversification among these portfolios gives a portfolio of less risk; i.e., $r(\alpha X_1 + (1-\alpha) X_2) \leq \alpha r(X_1) + (1-\alpha) r(X_2)$

$$= r(X_1) = r(X_2)$$

If the risk function were only convex then combining two portfolios with the same risk at least could not increase risk. While our theory so far applies to any strictly convex risk function this does not imply that any such function will give "the right kind of diversification for the right reasons." This much, at least, can be said for $V(X)$.  

---

4. The only information used concerning the available set is that it is convex and compact.
We define the critical set \( l_{a_1, \ldots, a_\alpha} \) to be the set of points \( X_0 \) such that there exists an \( E_0 \) such that \( V(X_0) \leq V(X) \) for all points in \( s_{a_1, \ldots, a_\alpha} \) with \( E(X) = E_0 \).

**Theorem 6.** If \( |DZD| \neq 0 \), \( l_{a_1, \ldots, a_\alpha} \) is i) a point \( x_{a_1, \ldots, a_\alpha} \) when \( a_1 = \ldots = a_\alpha \); and is ii) a line when the \( a_i \) are not all equal. In the second case the critical line is the set of all points which satisfy the following equations for some \( E \):

\[
\begin{pmatrix}
\sigma_{a_1a_1} & \ldots & \sigma_{a_1a_\alpha} & \lambda_{a_1} & 1 \\
\vdots & & \vdots & \vdots & \vdots \\
\sigma_{a_\alpha a_1} & \ldots & \sigma_{a_\alpha a_\alpha} & \lambda_{a_\alpha} & 1 \\
1 & \ldots & 1 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
x_{a_1} \\
\vdots \\
x_{a_\alpha} \\
\lambda_1 \\
\lambda_2
\end{pmatrix} =
\begin{pmatrix}
0 \\
\vdots \\
0 \\
E \\
A
\end{pmatrix}
\]

or briefly \( x_{a_1, \ldots, a_\alpha} (x_{a_1, \ldots, a_\alpha}, \lambda)^T = (0, E, A)^T \), where

\( l_{a_1, \ldots, a_\alpha} \) is non-singular.

**Proof.**

i) When all \( a_i \) are the same only one \( E \) is attainable; the point which minimizes \( V \) over \( s_{a_1, \ldots, a_\alpha} \) is \( x_{a_1, \ldots, a_\alpha} \).

ii) When \( \exists i, j \) such that \( a_i \neq a_j \), then \( \{X \mid \exists X' = a_i X = E, X \in s_{a_1, \ldots, a_\alpha}\} \) defines a family of parallel hyperplanes with parameter \( E \). On each hyperplane one and only one point minimizes \( V \) (by strict convexity). A necessary condition for \( V \) to be minimum is found by the method of Lagrangian multipliers. This
yields the equations of the theorem. Since we know that one and only one point minimizes \( V \) for given \( E \), we also know that \( \Sigma_{a_1, \ldots, a_\alpha} \) is nonsingular.

Let us note the relations between the equations describing the various critical lines and the case with which these equations may be derived. The equations of the critical line of \( R_N = a_1, \ldots, N \) is

\[
\begin{pmatrix}
\sigma_{11} & \cdots & \sigma_{1N} & \mu_1 & 1 \\
\vdots & \ddots & \vdots & \vdots & \vdots \\
\sigma_{N1} & \cdots & \sigma_{NN} & \mu_N & 1 \\
\mu_1 & \cdots & \mu_N & 0 & 0 \\
1 & \cdots & 1 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_N \\
\lambda_1 \\
\lambda_2
\end{pmatrix}
= 
\begin{pmatrix}
0 \\
0 \\
0 \\
E \\
A
\end{pmatrix}
\]

where \( E \) is a parameter. To obtain \( a_1, \ldots, a_\alpha \), we delete, in (1), all equations and columns of the matrix except \( a_1, \ldots, a_\alpha, N+1, N+2 \).

**Lemma 10.** All efficient points are on critical lines. The efficient surface may be traced out by starting at \( \bar{X} \) and moving continuously along efficient surfaces until \( \bar{X} \) is reached. As we trace out the efficient set \( \hat{X}(E) \) for \( E \leq E \leq E \), \( \hat{X} \) can not "jump" from one efficient set to another, but can turn at intersections. We must look more closely at these intersections and turnings.

It can be shown that the critical line of one subspace \( a_1, \ldots, a_\alpha \) intersects the critical lines of every subspace which contains \( a_1, \ldots, a_\alpha \) and is one dimension larger than it. Similarly the critical line of \( a_1, \ldots, a_\alpha \) intersects the critical lines of all subspaces which are contained by \( a_1, \ldots, a_\alpha \) and are one dimension smaller than it. In particular when the critical line crosses a boundary of the attainable set (i.e., when some \( x_{a_1} \) becomes zero) it
intersects another critical line. \( \hat{X}(E) \) then turns and proceeds along the new critical line. Intersections other than those described above are accidental, but must be provided for.

The fact that a critical line meets another critical line whenever it meets a boundary, is important. But we need not prove it here since we have already established continuity.

If \( \hat{X}(E) \), traveling along \( \mathcal{E}_1, \ldots, a \), as \( E \) increases, meets another critical line, we distinguish four cases: i) at the intersection \( E=\bar{E} \) and the efficient surface ends; ii) as \( E \) increases only one line remains in the region; then \( \hat{X}(E) \) moves along this line. iii) More than one line remains attainable; but one line is of a subspace which contains the subspaces of the other lines. In this case \( \hat{X}(E) \) moves along the line of the largest subspace.

For minimum \( V \) for given \( E \) is at least as small if we minimize it over the larger space rather than the smaller space. iv) If none of the above hold we must use the crude rule: we compute the \( (E,V) \) combination associated with each line and choose the one which gives smallest \( V \) for given \( E \). (Strict convexity assures that one will give smallest \( V \) for all points except the point of intersection.)

**Theorem 7.** On \( \mathcal{E}_1, \ldots, a \), the \( (E,V) \) combinations are

\[
V=(0, \ldots, 0, E, 1)
\[
\begin{pmatrix}
\sigma_{a_1 a_1} & \sigma_{a_1 a_\mu} / a_1 & \mu_{a_1} / a_1 \\
\sigma_{a_2 a_2} & \sigma_{a_2 a_\mu} / a_2 & \mu_{a_2} / a_2 \\
\vdots & \vdots & \vdots \\
\sigma_{a_\mu a_\mu} & \sigma_{a_\mu a_\mu} / a_\mu & \mu_{a_\mu} / a_\mu \\
\mu_{a_1} & \mu_{a_2} & 0 & 0 \\
1 & \ldots & 1 & 0 & 0
\end{pmatrix}
\]

(equation continued)
or \( V = (00, \ldots, 0, E, 1) \mu^{-1}_{a_1, \ldots, a} z^{00} \mu^{-1}_{a_1, \ldots, a} \)
Our theorems show us the nature of the efficient set F. Methods exist for finding \( \mathbb{S} \). Given \( \mathbb{S} \), our theorems show how to trace out F. Given F theorem 7 enables us to get the set of efficient couples (\( E, V \)).

**Generalization**

When the investors' decisions depend on two moments \( M_1, M_j \) we can define critical sets for each subspace by the Lagrange equations. All efficient points lie on these curves. When \( M_1 = M_1 = E \) and when \( M_j \) is strictly convex, our previous discussion applies with little modification. Even when \( M_1, M_j \) are not thus restricted, it seems plausible that under general conditions the efficient set is continuous, and therefore can be traced out continuously along critical curves. When action is based on three or more moments, the critical regions satisfying the Lagrangian equations are surfaces of 2 or more dimensions. This raises a number of unsolved problems.
V. Behavior Equations

Computations based on statistical data have had, broadly, two uses: description and inference. In its purest form the use of statistical computations for purposes of inference is based on the explicit use of probability models. No sharp line, however, can be drawn between the fitting of curves and other relationships for descriptive purposes or for inferential purposes. Often relations which are found in data without the aid of formal inference, are extrapolated and used as a basis of prediction. Such extrapolations seem implicitly to assume an underlying probability model.

Statistical inference, in its explicit form, assumes the existence of an underlying probability model which generates the observations in question. Typically, something but not everything, is assumed to be known about this model. These a-priori assumptions are combined with observation to estimate unknown properties of the model.

A type of model frequently found in econometric work is the system of (one or more) equations. These equations express hypothetical, physical or human relations. They are only expected to hold "on the average"; observations are assumed to be a combination of these equations and random elements. The investigator makes assumptions as to the variables which enter each equation, the forms of the equations, and properties of the random elements. The values of parameters are left to be inferred from data. For certain sets of such assumptions, mathematical statisticians have derived "good" or "best" ways of using observations to make inferences and predictions.

Two uses of systems of equations are policy recommendations and prediction. If the investigator's assumptions were correct and the mathematical statisticians'
prescriptions followed, then good estimates of unknown parameters could be obtained. Questions as to the effects of policy and the course of future events could be answered with accuracy as great as the data permitted. If the investigator's assumptions were correct, such equations might be extremely useful.

But if wrong assumptions were made and acted upon the results might well be unreliable or punicious. Serious errors may result because the equations include the wrong variables, are of the wrong form, or because wrong assumptions are made as to the random elements. The effect of even small errors in equations is illustrated by a comparison of the Keynesian and Classical economic models (see Hicks [7]). These models involve apparently small differences in premises but large differences in important conclusions. The effects of errors in assumptions about random elements is illustrated by the results obtained when least squares methods are applied to systems of simultaneous equations (Haavelmo [6]).

The variables that we would most like to predict or influence—such as national income and employment—are generated by a complex web of relations involving many types of persons and institutions. We cannot expect—it seems to me—to build an adequate model for predicting or controlling such variables until we have formally "charted out" these relations and have formal analyses of the various types of persons and institutions.

Another use of stochastic equations is to test hypotheses. If verifiable equations can be derived from plausible or traditional economic hypotheses, then these equations can be fitted and tested as a test of the underlying hypotheses. If the equations are tested and rejected, an interesting economic

1. If not enough is known about the basic model, it may be impossible to make inferences from data concerning some of the unknown aspects of the model. This is the identification problem; see Koopmans [10].
hypothesis must be rejected or modified. If after rigid tests the equations are not rejected, the underlying hypotheses become, in some sense, more plausible. In either case the equations have served one purpose and may serve more.

When our short run goal is to test hypotheses we can seek out situations which yield the simplest equations for given economic hypotheses. From the information got from such situations we perhaps can sort successful from unsuccessful hypotheses, the former to be used in building the grander models relevant to policy.

When we test stochastic equation systems, we test the hypothesis as a whole including assumptions concerning the unexplained residual. It is only the assumptions concerning the unexplained residuals which make the equations something other than tautologies.²

Investment company equations have two uses: First, the investment company and institutions with similar investment problems have an important place in the web of relations which generate important economic variables. Here is one set of institutions whose opportunities and behavior must be formalized before we can hope to build our grander models.

Second, the investment company allows us to test, in a simple context, hypotheses concerning behavior under uncertainty. In particular it serves to test hypotheses as to motivation and belief formation.

The behavior equations we derive for the investment company are not of the simple form generally found in econometric models. The usual equations are

² For example if we asserted that \( y - 5x + \epsilon \) where \( y \) and \( x \) are observable and no assumptions are made about \( \epsilon \), then the equation is only a (tautological) definition of \( \epsilon \). If we make assumptions about \( \epsilon \) our equation becomes a verifiable hypothesis. Similarly if we assert that there exists an \( \alpha \) such that \( y - \alpha x = \epsilon \).
linear with additive random variables. Our simplest equations are bilinear; others are far more complex. Our equations are derived from quite general assumptions. They are derived for a simple institution. It is hard to believe that the non-linearities of our model will disappear as we add problems of liquidity, solvency, a more complicated production function and divergence of interest between owners and management.

We also consider hypotheses concerning randomness in human behavior, and from these derive assumptions concerning the random variables which enter our equations. None of our hypotheses lead to the usual additive random variables.

**Derivation of Equations**

At first we will ignore random elements. We will state our assumptions and derive our equations as if they held exactly, without unexplained residuals. We apply the same behavior equation to equations all of whose variables (except the random variables) are observable. Behavior equations will be distinguished from subsidiary equations which contain unobservable variables such as subjective probability beliefs.

As was discussed in Chapter 3, in deriving investment company behavior equations we make assumptions as to a) the behavior of the investor, and b) the knowledge of the investigator. a) We assume that i) the investor maximizes 

\[ U = U(W_1, \ldots, W_n) \]

where \( W_i \) is the \( i \)th moment of the probability distribution of the returns from the portfolio as a whole; and ii) at every point chosen by

---

3. I.e., it is assumed that

\[
\sum_{j=1}^{n} \alpha_{ij} y_i + \sum_{j=1}^{n} \beta_{ij} z_i = u_j \quad i=1, \ldots, n
\]

where \( \alpha_{ij}, \beta_{ij} \) are parameters

\( y_i \) indogenous variables (with values determined by the system)

\( z_i \) exogenous variables (with values given from outside the system).

\( u_i \) random variables (which, among other things, are independent of the \( z \)'s).
the investor some moment is either maximized or minimized for given values of the other moments and subject to the constraints defining the attainable set. b) In one case we assume that the investigator can specify a finite set of moments which include $M_{a_1}, \ldots, M_{a_k}$. In another case we also assume that the investigator can specify the forms of "belief formation functions" (b.f.f.'s) which relate relevant moments (concerning returns on individual securities) to observable variables.

Let us first derive behavior equations when action depends on expected return and variance of return. We will use the same notation as in Chapter 4.

Let $(x^0_1, \ldots, x^0_N)$ be a point chosen by the investor. Arrange the subscripts so that $x^*_1 > 0, \ldots, x^*_K > 0, x^*_K, \ldots, x^*_N = 0$. Suppose, for the moment, that at this point $V$ is minimized for the given value of $E$. For this to be so we must not be able to reduce $V$ by small changes in $x^*_i$ $i \leq K$, such that

$$\frac{K}{i=1} x^*_i = A, \quad \frac{K}{i=1} \frac{x^*_i}{\mu^*_i} = E(= \Sigma x^0_i / \mu^*_i).$$

A necessary condition for this to be so is given by the Lagrangian equations

1) $$\begin{cases} \frac{\partial V + 2 \lambda_1 (\Sigma \mu^*_1 x^*_1 - E) + 2 \lambda_2 (\Sigma x^*_1 - 1)}{\partial x^*_j} = 0 & j = 1, \ldots, K \\ \frac{K}{i=1} x^*_i = A \\ \frac{K}{i=1} \frac{x^*_i}{\mu^*_i} = A \end{cases}$$

2) $$\begin{cases} \frac{K}{i=1} x^*_i \sigma^*_i + \lambda_1 x^*_i + \lambda_2 = 0 & j = 1, \ldots, K \\ \frac{K}{i=1} x^*_i = A \\ \frac{K}{i=1} \frac{x^*_i}{\mu^*_i} = A \end{cases}$$

If we assume that, at the point $(x^0_1, \ldots, x^0_N)$, $V$ is maximized for the given value of $E$, then $V$ must have a relative maximim in the subspace $x^*_1, \ldots, x^*_K$, subject to the constraints $\frac{K}{i=1} x^*_i / \mu^*_i = E(= \Sigma x^0_i / \mu^*_i)$ and $\frac{K}{i=1} x^*_i = A$. Once again we get the Lagrangian equations above.
If we assume that E is minimized or maximized for given V we get the Lagrangian equations

$$\frac{\partial E + \frac{1}{2} \delta_1 (\Sigma \sigma_{ij} x_i x_j V) + \delta_2 (\Sigma x_i - 1)}{\partial x_j} = 0 \quad j=1, \ldots, n$$

or

$$\lambda_j + \delta_1 \frac{\sigma_{ij}}{\delta_1} x_i + \delta_2 = 0.$$ 

If \( \delta_1 \neq 0 \) (i.e., if the marginal utility of variance is not zero) we can divide by \( \delta_1 \) and let \( \lambda_1 = \frac{1}{\delta_1} \lambda_2 = \frac{\delta_2}{\delta_1} \) then

$$\sum_{i=1}^k \sigma_{ij} x_i + \lambda_1 \mu_j + \lambda_2 = 0 \quad \Sigma x_i = A$$

Thus: if action depends on E and V; if, for any chosen portfolio, one of these is either maximized or minimized for given values of the other; and if \( \delta_1 \neq 0 \) then equations 2 hold.

This can be stated more symmetrically without assuming \( \delta \neq 0 \): If action depends on E and V, and if at any chosen point one of these is extremized for a given value of the other then there exists a vector \( (\gamma_2, \gamma_1, \gamma_0) \) such that

3) \( \gamma_2 \frac{\partial V}{\partial x_j} + \gamma_1 \frac{\partial E}{\partial x_j} + \gamma_0 = 0 \quad j=1, \ldots, K \)

or

4) \( \gamma_2 \frac{\sigma_{ij}}{\sigma_{ii}} x_i + \gamma_1 \mu_j + \gamma_0 = 0 \quad j=1, \ldots, K \)

\( \Sigma X_j = A \).

Assumption a.ii) assures us that at least one \( \gamma \) is not zero.

To get behavior equations we eliminate the \( \gamma \)'s from 3). To eliminate \( \gamma_0 \) from the \( j \)th equation subtract the first equation from the \( j \)th and get

$$\gamma_2 \frac{1}{\sigma_{ii}} (\sigma_{ij} - \sigma_{il}) x_i + \gamma_1 (\mu_j - \mu_l) = 0 \quad j=2, \ldots, K$$

4. In the E, V case it is plausible to assume that \( \delta_1 \neq 0 \) (see Chapter 2). In the general case the symmetric statement is needed.
To eliminate \( \gamma_1 \) subtract \( (\mu_j - \mu_1) \) times equation \( 2 \) from \( (\mu_2 - \mu_1) \) times equation \( 1 \). This gives

\[
\gamma_2 \left\{ (\mu_2 - \mu_1)(\sigma_{1j} - \sigma_{11}) - (\mu_j - \mu_1)(\sigma_{12} - \sigma_{11}) \right\} x_1 = 0
\]

If \( \gamma_2 \neq 0 \), then divide by \( \gamma_2 \) and get

\[
\gamma \left\{ (\mu_2 - \mu_1)(\sigma_{1j} - \sigma_{11}) - (\mu_j - \mu_1)(\sigma_{12} - \sigma_{11}) \right\} x_1 = 0 \quad j = 3, \ldots, n
\]

If we had eliminated any \( \gamma_i \) last we would have got equations 5) except with \( \gamma_i \) instead of \( \gamma_2 \). Since at least one \( \gamma_i \) is not zero, equations 6) can always be obtained; i.e., equations 6) are true for any observed portfolio.

Suppose the investigator knew the forms of the investor's "belief formation functions" (b.f.f.'s):

\[
\mu_1 = \mu_1(z_1, \ldots, z_S; \alpha_{11}, \ldots, \alpha_{1L_1})
\]

\[
\sigma_{ij} = \sigma_{ij}(z_1, \ldots, z_S; \beta_{11}, \ldots, \beta_{ijL_j})
\]

where \( z_1, \ldots, z_S \) are observable variables and where the \( \alpha_k \)'s and \( \beta_k \)'s are unknown parameters. We can substitute 7) in 6) to get behavior equations containing observable variables \( X_1, z_S \) and unknown parameters \( \alpha_1, \beta_{ij} \).

In particular suppose that the b.f.f.'s are linear:

\[
\mu_1 = \sum_{s=1}^{S} \alpha_s z_s
\]

\[
\sigma_{ij} = \sum_{s=1}^{S} \beta_{ij} z_s
\]

\( z_s \), itself, may be a product, ratio, logarithm or other combination containing no unknown parameters.

Substituting 7a) in 6 we get equations which are bilinear in \( X \) and \( z \), with coefficients \( \gamma_{ia,jkb} = \alpha_{i,a} \beta_{j,k,b} \).
If the $z$'s vary through time while the $\alpha$'s and $\beta$'s remain constant, the $y$'s can typically be estimated. As the $z$'s vary the set of non-zero $X$'s change. If an $X_j \neq 1, 2$ becomes zero, the equations $3, \ldots, j-1, j+1, \ldots, K$ remain valid. If an $X_j \neq 1, 2$ becomes zero, all equations become invalid.

Thus, in fitting our equations over a period it is expedient to let securities $1, \ldots, K$ include all securities held at any time during the period and choose as 1 and 2 securities which were held throughout the period.

Even if we know the $y$'s we can not necessarily infer the values of the $\alpha$'s and $\beta$'s. Specifically if we let $\alpha_{ia} = \lambda \alpha_{i1}$ and let $\beta_{jkb} = \frac{1}{\lambda} \beta_{jkb}$ then $\gamma_{ia, jkb} = \gamma_{ia, jkb}$. If we know a priori the value of one $\alpha$ or one $\beta$ then, it can be shown by construction, that the values of the rest can be inferred.

The behavior equations derived above require us to know the investor's b.f.f.'s. They are to be fit to the portfolio of a single investment company over a period of time. If we assume that differences in portfolios, held by different investors at a given time, are due to differences in preference rather than differences in beliefs, then behavior equations can be found which apply to a cross-section of portfolios, as of a particular time.

If the various investors have the same probability beliefs then equations 6) hold for each. Equations 6), as they stand are unidentifiable; but from these we can get equations of the following form:

3) $X_j + \alpha_j X_2 = 0 \quad j=3, \ldots, n$

$(X_1 - A - \sum_{j=2}^{n} X_j)$
Given a sufficient number of portfolios containing the same number of securities, equations 8) could be fit.

The General Case

When action depends on moments $M_{d_1}, \ldots, M_{d_\ell}$, and when assumption a.ii) is satisfied, there exists a vector $y_1, \ldots, y_\ell$ such that

$$\sum_{i=1}^{\ell} y_i \frac{\partial M_{d_i}}{\partial x_j} = 0$$

where at least one $y_1$ is not zero. $\frac{\partial M_{d_i}}{\partial x_j}$ is a homogeneous polynomial of degree $a_i-1$. We can eliminate the $y_1$ and substitute b.f.f.'s to get behavior equations which are homogeneous polynomials in the $X$'s of degree $\sum_{i=1}^{\ell} (a_i-1)$.

We have $K-\ell$ equations in $K$ unknowns. This reflects the fact that we have an $\ell$-dimensional efficient surface. A moment $\sum_{i=1}^{\ell} \left( \prod_{k=1}^{K} (x_{ki} - \bar{x}_{ki})^e_{i} \right)$ will be said to be of degree $\sum e_i$. To infer the values of the parameters of the b.f.f.'s from those of the behavior equations we must know the values of $\ell-1$ moments, including at most one moment of each degree. If the investigator does not know the forms of b.f.f.'s, but believes that, as of a given time, differences in portfolios are due to differences in preferences rather than differences in beliefs, cross section equations can be found.

All operations used in deriving our equations are valid even if some $y_1 = 0$. Thus if the investigator specifies some irrelevant variables the equations do not become wrong. There are other penalties however. The equations increase in degree; the efficient surfaces increase in dimensionality, rendering our equations less specific; there is a greater lack of identifiability of the parameters of the b.f.f.'s.
Random variables

When we use data to estimate the parameters of behavior equations, the procedures for obtaining "good" estimates (and therefore the estimates obtained) depend on the assumptions we make concerning the random variables.

When we test a set of equations, we test the hypothesis as a whole including assumptions as to random variables. Just as it was desirable to derive our equations from economic hypotheses, it is also desirable to derive our assumptions concerning random variables from economic hypotheses.

We will derive assumptions concerning random variables for our time series equations when the b.f.f.'s are linear. The same or similar hypothesis could be used to add random variables to our cross section equations or to equations obtained when the b.f.f.'s are not linear.

In general when the b.f.f.'s are linear we have equations of the form

\[ \sum_{j,k,\ldots,m} \xi_{j,k,\ldots,m} x_j x_k \cdots x_m w_{i,j,\ldots,m} = 0 \]

\[ \sum x_j = A \]

where the \( w \)'s are combinations of the variables which enter the b.f.f.'s; the \( \xi \)'s are combinations of the parameters which enter the b.f.f.'s; where certain \( \xi \)'s are known a priori to be zero; and where \( A \) is predetermined (in the sense of Koopmans [10]). When \( u = u(E, V) \) we have equations

\[ \sum_{j=1}^{n} x_j \left\{ \left[ \mu_j(z) - \mu_1(z) \right] [\sigma_{j1}(z) - \sigma_{11}(z)] \right\} = 0 \quad 0=1,\ldots,n \]

\[ \sum x_j = A \]

1. One hypothesis, as to the source of randomness, is that equations 9) characterize a rational or optimal portfolio \( x^0 = (x^0_1, \ldots, x^0_N) \), but because
the human being is an inaccurate or lazy mechanism, the actual portfolio 
\( X = (X_1, \ldots, X_N) \) deviates from the optimal by some random term \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_N) \). 
Thus in the general case \( \sum \beta \ v \ x_j \ v \ \sum x_i = 0, \ \sum x_i = A \)
\( X = X^0 + \varepsilon \ \ \ \sum X = A \) where \( X \) and \( z \) are observable. It may be reasonable to assume that the deviations from optimality are independent of the variables that determine which portfolios are optimal: \( \varepsilon z \ v \ = 0 \). If we believe that the investor is right on the average, we have \( \varepsilon z \ = 0 \).

2. We may hypothesize that the probability beliefs are random variables. In particular:

A. We may believe that the "weights" attached to the \( z_i \) vary from time to time and case to case. This variation, we may hypothesize, can be accounted for by assuming that the parameters of the \( b, f, e, s \) are random variables (presumably independent of the \( z_i \)s). Thus when \( u = u(E, V) \)

\[ \delta = \alpha \beta = (\alpha^0 + \varepsilon) (\beta^0 + \gamma) = \alpha^0 \beta^0 + \beta^0 \gamma + \alpha^0 \varepsilon + \gamma \delta^0 + \epsilon \]
where \( \alpha^0, \beta^0, \delta^0 \) are constants and \( \varepsilon, \gamma, \delta, \epsilon \) are random variables.

B. We may believe that the fluctuations in subjective probability beliefs are due to variations in "\( z_i \)s" which are not taken into account in our analysis. Thus the "truth" may be, for example,

\[ \mu_1 = \alpha_0 + \alpha_1 z_1 + \ldots + \alpha_S z_S + \alpha_{S+1} z_{S+1} + \ldots + \alpha_t z_t \]
while our analysis takes into account only the first \( z_i \)s and ascribes variation in the rest to a random variable \( \gamma \)

\[ \mu_1 = \alpha_0 + \alpha_1 z_1 + \ldots + \alpha_S z_S + \gamma \]

Our equations in the \( E, V \) case would then be
\[ \sum_{j=1}^{k} L_j \left\{ \left[ \sigma_j(z) - \eta_j(z) - \gamma_j \right] - \left[ \sigma_{j11} - \sigma_{j1} - \gamma_{j1} \right] \right\} = 0 \quad i=3, \ldots, n \]

There are, of course, other hypotheses as to sources of randomness.

We might hypothesize, for example, that our observations of, or knowledge about, some z's differ from that of the investor. Or we may find reasons for doubting the independence of the z's and the random variables.

We see that plausible hypotheses as to the sources of randomness in behavior lead to assumptions concerning random variables different from those customarily used when fitting curves. The implications of these assumptions have not been well explored by statisticians.\(^5\)

---

\(^5\) On non-additive random variables see Hurwicz [8].
VI. Aggregation and Segregation

Aggregate relations concern the allocation of funds between groups of securities, no predictions or recommendations being made for the allocation of funds within these groups. Segregate equations concern the allocation of funds within a particular group of securities, no predictions or recommendations being made for the allocation of funds between this group and the rest of the portfolio or within the rest of the portfolio.

Various amounts of knowledge concerning the parts of the portfolio not directly under consideration, can be postulated when seeking aggregate and segregate relations. For example we might seek efficient allocations of funds among different industries, the proportions devoted to firms within each industry being given (and known). In deriving behavior equations for the purpose of prediction, we can not assume that these within industry allocations are given a priori.

In the last chapter we saw that behavior equations could be used to test economic hypotheses. These equations were not simple; if applied to a portfolio of 60, 80, 100 or more securities our model would be unmanageable, at least with conventional computing machines. If we could find segregate or aggregate relations, the fitting of behavior equations to test hypotheses might become practicable.

In Chapter IV we showed how efficient surfaces could be computed. Such computations would be unmanageable, at least with conventional computing machines, if we applied our analysis to all securities in the portfolio, or all those given consideration by the investment management. For example, if our analysis included 100 securities, the covariance matrix would include 5,050 distinct entries. The search for aggregate and segregate efficient

1. Excluding repetitions since $\sigma_{ij} = \sigma_{ji}$
surfaces is one way, though not necessarily the only way, to meet this
problem of complexity.

When the relative allocation of funds within industries is given,
we can find aggregate relations which treat each industry as if it were a
single firm.

Thus let \( w_{ik} \) be the relative amount allocated to the \( k \)th security
in the \( i \)th industry (\( \sum_{k=1}^{K} w_{ik} = 1 \)); let \( X_i \) be the amount devoted to the \( i \)th
industry (\( \sum X_i = A \), where \( A \) is total assets) \( X_i w_{ik} \) is the absolute amount
allocated to the \( k \)th security in the \( i \)th industry. Let \( \mu_{ik} \) be the expected
return per dollar of \( X_i w_{ik} \); \( \sigma_{ik,jl} \) is the covariance of return between the
\( k \)th security in the \( i \)th industry and the \( l \)th security in the \( j \)th industry.

\[
E = \sum_{i=1}^{m} \frac{k_i}{\sum_{k=1}^{K} k_i} \mu_{ik} w_{ik} X_i
\]

\[
E = \sum_{i=1}^{m} \left( \frac{k_i}{\sum_{k=1}^{K} k_i} \mu_{ik} \right) X_i
\]

if we let

\[
\mu_{i} = \sum_{k=1}^{K} k_i w_{ik}
\]

then

\[
E = \sum_{i=1}^{m} \mu_{i} X_i
\]

\[
V = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{K} \sum_{l=1}^{L} k_i k_j X_i X_j \left( \sum_{k=1}^{K} \sum_{l=1}^{L} w_{ik} w_{jl} \sigma_{ik,lj} \right)
\]

If we let

\[
\sigma_{ij} = \sum_{k=1}^{K} \sum_{l=1}^{L} w_{ik} w_{jl} \sigma_{ik,lj}
\]

then

\[
V = \sum_{i=1}^{m} \sum_{j=1}^{m} \sigma_{ij} X_i X_j.
\]
Thus when the relative amounts \( w_{ik} \) allocated among securities within an industry are known the amounts \( x_i \) allocated between industries can be treated as if they were allocations in particular securities, where \( \mu_i \) and \( \sigma_{ij} \) are suitably defined weighted averages.

If the \( w_{ik} \) were not known we could seek equations that were independent of the \( w_{ik} \). \( \mu_i \) is independent of the \( w_{ik} \) if and only if \( \mu_i = \mu_i \). \( \sigma_{ij} \) is independent of the \( w_{ik} \) if and only if \( \sigma_{ik} = \sigma_{ik} \). for all \( a, c \) in one group and \( b, d \) in another, not necessarily different, group of securities. This implies, in particular, that the correlation \( \rho_{ij} = 1 \). These conditions are too stringent to be of practical use.

When \( U = U(E, V) \) an assumption from which segregate relations can be derived is the following: Let securities \( 1, \ldots, k \) be those of, say, one industry; let securities \( k+1, \ldots, N \) be those outside this industry. Let \( \rho_{ij} \) be the correlation between the return on the \( i \)th and \( j \)th securities. Our assumption is that

\[
\rho_{ik} = \rho_{jk} \quad \text{for } i, j \leq k \\
\quad \text{and } k > k
\]

In words, if \( i \) and \( j \) are securities within the given industry, and \( k \) is a security outside the industry, then the correlation between \( i \) and \( k \) is the same as that between \( j \) and \( k \).

This assumption does not imply that \( \rho_{ij} = 1 \). We will shortly present a model which may give some insight into this assumption and perhaps make it somewhat plausible. For the moment let us assume it and see what segregate relations follow from it.
We saw in Chapter V that behavior equations could be derived from the Lagrangian equations

1) \[ \sum_{j=1}^{n} x_j \sigma_{ij} + \lambda_1 \mu_1 + \lambda_2 = 0, \quad i=1, \ldots, n \]

(We assume for convenience that the first \( n \) securities are observed, and that the first \( K \) of these are from a particular "industry." ) From (1), and our assumption we get

2) \[ \sum_{j=1}^{K} x_j \sigma_{ij} + \sigma_1 \left( \sum_{j=K+1}^{n} \sigma_{ij} \right) + \lambda_1 \mu_1 + \lambda_2 = 0, \quad i=1, \ldots, n \]

Since \( P_{ik} = P_{jk} \quad i, j \leq K < k \)

we can write

3) \[ \sum_{j=1}^{K} x_j \sigma_{ij} + \sigma_1 \left( \sum_{j=K+1}^{n} \sigma_{ij} \right) + \lambda_1 \mu_1 + \lambda_2 = 0, \quad i=1, \ldots, K \]

If we eliminate \( \lambda_1, \lambda_2 \) and \( \sigma_1 \) we get \( K-3 \) equations in \( X_1 / \mu_1 \) and \( \sigma_{ij} \) with all \( i, j \leq K \). That is we get equations which only involve the allocations among and probability beliefs about securities within the given industry. These can be used as the basis of segregate behavior equations.

Segregate efficient surfaces can be derived with the aid of the following relations:

\[ E = \sum_{i=1}^{K} \frac{X_i}{\mu_1} + \sum_{i=K+1}^{N} \frac{X_i}{\mu_1} \]

\[ = \sum_{i=1}^{K} \frac{X_i}{\mu_1} + \sigma_1 \]

\[ V = \sum_{i=1}^{K} \sum_{j=1}^{K} \sigma_{ij} X_i X_j + 2 \sum_{i=1}^{K} \sum_{j=K+1}^{N} \sigma_{ij} X_i X_j \]

\[ + \sum_{i=K+1}^{N} \sum_{j=K+1}^{N} \sigma_{ij} X_i X_j \]

Since \( P_{ik} = P_{jk} \)

\[ \sum_{i=1}^{K} \sum_{j=K+1}^{N} \sigma_{ij} X_i X_j = \sum_{i=1}^{K} \sum_{j=1}^{N} \sigma_{ij} \left( \sum_{j=K+1}^{N} P_{ij} \sigma_j X_j \right) \]
\[ V = \sum_{i=1}^{K} \sum_{j=1}^{K} \sigma_{ij} X_i X_j + 2 \sum_{i=1}^{K} X_i \sigma_i (C_2) + C_3 \]

The constants \( C_1 \) and \( C_3 \) are not needed to derive the segregate efficient sets. \( C_2 \) can be computed when \( \mu_i, \sigma_{ij} (i=K+1, \ldots, N; j=1, \ldots, N) \) are known.

If it is expensive to obtain the \( \mu_i \) and \( \sigma_{ij} \), \( C_2 \) may be estimated from a sample of them (see Chapter 9).

Let \( R_i \) be the return on the \( i \)th security. We present a model showing how the \( R_i \) might be generated in such a way that \( R_{ik} = R_{jk} \quad i,j \leq K < k \).

In this model we ascribe deviations of \( R_i \) from its expected value \( \mu_i \) to two sources: an industrywise source and a personal source. The industrywise source of fluctuation includes such things as changes in product and factor prices which affect all firms in the industry. The personal source includes such things as good will or skill of management which affect the position of individual firms. We assume that the return on a security of a firm outside the industry is independent of the personal component of the return of a firm within the industry (although it may be correlated with the industrywise component, and the personal components of firms within the industry may be correlated).

\[ R_i = \mu_i + P_i + \delta_i, \quad i=1, \ldots, K \]

\( \delta_i \) a constant

\( P_i, \delta_i \) random variables (the effects, respectively, of the personal and industrywide sources of fluctuation).

\[ R_i = \mu_i + D_i, \quad i=K+1, \ldots, N \]

(it is not necessary to break down the deviations \( D_i \) for firms outside the industry).
\[ E_{P_1} = E_C = E_{D_1} = 0 \]

\[ E_{P_{i \neq j}} = 0, \quad i \leq K < j \]

\[ \frac{E_{(P_1 + \delta_1 C)^2}}{E_{(\delta_1 C)^2}} = \frac{E_{(P_j + \delta_j C)^2}}{E_{(\delta_j C)^2}}, \quad \text{for } i, j \leq K \]

From this it follows that (for i, j \leq K < k)

\[ P_{ik} = \frac{E_{(P_1 + \delta_1 C)^2} D_k}{\sqrt{E_{(P_1 + \delta_1 C)^2}} \sqrt{E_{D_k}^2}} \]

\[ = \frac{E_{\delta_1 C} D_k}{\sqrt{E_{(P_1 + \delta_1 C)^2}} \sqrt{E_{D_k}^2}} \]

\[ = \frac{E_{\delta_j C} D_k}{\sqrt{E_{(P_j + \delta_j C)^2}} \sqrt{E_{D_k}^2}} \]

\[ = P_{jk} \]
VII. Time Preference

In Chapters IV to VI we have assumed that the investor has "static" beliefs. For some purposes static models are convenient and revealing. We frequently speak in terms of static models, but the world is not static: some firms, for example, must look forward to increasing expected return; others to decreasing expected return. These differences in outlook for various firms are recognized by the investment manager. He must choose between portfolios which give more or less preference to the future. If we hope to advise the investor this must be taken into account in our efficient surfaces. Similarly, it must be taken into account in deriving behavior equations.

Suppose the investor's utility depended on the expected return $E_t$, variance of return $C_{tt}$ at time $t = 1, \ldots, T$, and covariance of return $C_{st}$ between time $s$ and $t$, $s, t = 1, \ldots, T$. $U = U (E_1, \ldots, E_T, C_{11}, C_{12}, \ldots, C_{tt})$. Necessary conditions for $U$ to have a maximum at a point $(x_1^0, \ldots, x_K^0, 0, \ldots, 0)$ $x_k^0 > 0$, $k \leq K$, subject to the constant that $\Sigma x = A$, is that

$$
(1) \quad \sum_{t=1}^{T} \frac{\partial U}{\partial E_t} \frac{\partial E_t}{\partial x_i} + \sum_{s=1}^{T} \sum_{t=1}^{T} \frac{\partial U}{\partial C_{st}} \frac{\partial C_{st}}{\partial x_i} + \lambda = 0 \quad i = 1, \ldots, K
$$

$$
(2) \quad \sum_{t=1}^{K} \frac{\partial U}{\partial E_t} \frac{\partial E_t}{\partial x_i} + \sum_{s=1}^{T} \sum_{t=1}^{T} \frac{\partial U}{\partial C_{st}} \frac{\partial C_{st}}{\partial x_i} + \lambda \leq 0 \quad i = K+1, \ldots, N
$$

(1) and (2) can be rewritten as follows:

---

1. For mathematical convenience we deal with only a finite number ($T$) of time periods. For generality we can think of this number as being very large, say $10^9$ years.
\[
\begin{aligned}
&\left\{ \frac{\partial E_t}{\partial x_1} + \sum_{t=2}^{T} \frac{\partial U}{\partial E_t} \frac{\partial E_t}{\partial x_1} \right\} \\
&+ \left\{ \frac{\partial U}{\partial c_{1,1}} \frac{\partial c_{1,1}}{\partial x_1} + \sum_{(s,t) \neq (1,1)} \frac{\partial U}{\partial c_{st}} \frac{\partial c_{st}}{\partial x_1} \right\}
\end{aligned}
\]

\[+ \lambda = 0 \quad \text{for } i = 1, \ldots, K\]
\[\leq 0 \quad \text{for } i = K + 1, \ldots, N.\]

In the language of the economist, \(\frac{\partial U}{\partial E_t} / \frac{\partial U}{\partial x_1}\) is the marginal rate of substitution between expected return in the \(t^{th}\) and first periods; \(\frac{\partial U}{\partial c_{st}} / \frac{\partial U}{\partial c_{1,1}}\) is the marginal rate of substitution between covariance \(c_{st}\) and variance in the first period. \(\frac{\partial U}{\partial c_{1,1}} / \frac{\partial U}{\partial x_1}\) is the marginal rate of substitution between variance in the first period and expected return in the first period.

Compare (3) with the results we get if we assume that the investor acts on the basis of discounted mean and discounted variance. Let \(\delta_t\) be the rate at which \(E_t\) is converted into \(E_1\); \(\delta_{st}\), the rate at which \(c_{st}\) is converted into \(c_{1,1}\).

\[
E = \sum_{t=1}^{T} \delta_t E_t = E_1 + \sum_{t=2}^{T} \delta_t E_t
\]
\[
v = \sum_{s,t} \delta_{st} c_{st} = c_{1,1} + \sum_{(s,t) \neq (1,1)} \delta_{st} c_{st}
\]

\[U = U(E, v)\]
To maximize $U$ we must have:

$$
\left\{ \frac{\partial E_i}{\partial x_1} + \sum_{t=2}^{T} \delta_t \frac{\partial E_t}{\partial x_1} \right\} + \frac{\partial U}{\partial \xi_{1,1}} \frac{\partial \xi_{1,1}}{\partial E_1} \left\{ \frac{c_{1,1}}{x_1} \sum_{(s,t) \neq (1,1)}^N \delta_t \frac{c_{st}}{x_1} \right\} 
$$

$$
+ \lambda = 0 \quad \text{for } i = 1, \ldots, K
$$

$$
\leq 0 \quad \text{for } i = K + 1, \ldots, N
$$

A portfolio which maximizes of the form $U(E_1, \ldots, E_T, c_{1,1}, \ldots, c_{st})$ also maximizes a utility function of the form $U(E, V)$ when $E$ is discounted expected returns and where $V$ is discounted covariance of return.

We may define the discounted expected return on the $i^{th}$ security as

$$
\mu_i = \sum_{t=1}^\infty \mu_{it} \delta_t
$$

The discounted return from the portfolio as a whole is

$$
E = \sum_{i=1}^N \sum_{t=1}^\infty \mu_{it} \delta_t x_i
$$

$$
= \sum_{i=1}^N x_i \sum_{t=1}^\infty \mu_{it} \delta_t
$$

Therefore

$$
E = \sum_{i=1}^N x_i \mu_i
$$

The discounted covariance of return of between the $i^{th}$ and $j^{th}$ security may be defined as $\sigma_{ij} = \sum_{s,t}^\infty \delta_{st} \sigma_{ijst}$. The discounted variance of return from the portfolio as a whole is

$$
V = \sum_{s,t}^\infty \sum_{i,j}^\infty \delta_{st} \sigma_{ijst} x_i x_j
$$

$$
= \sum_{i,j}^\infty x_i x_j (\sum_{s,t}^\infty \delta_{st} \sigma_{ijst})$$
Therefore

\[ V = \sum_{i} \sum_{j} X_i X_j \sigma_{ij} \]

Thus the discounted mean and discounted variance of the portfolio as a whole are the same function of the portfolio and the discounted probability beliefs as the mean and variance were of the portfolio and probability beliefs in the static case.

Thus \( \mathbf{x} \) is efficient if and only if, for some \( \delta_{ts}, s, t = 1, \ldots, T \), it gives minimum \( V \) for given \( E \) or more and maximum \( E \) for given \( V \) or less, where

\[ E = \sum_{i=1}^{N} \mu_i X_i \]

\[ V = \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{ij} X_i X_j \]

\[ X_i \geq 0 \]

\[ \sum X_i = 1 \]

and where \( \mu_i \) and \( \sigma_{ij} \) are the discounted mean and covariance of individual securities. For given \( \delta_{ts}, \delta_{st} \), the methods of Chapter IV can be applied.

The idea of discounting is imbedded in financial tradition and may be used to advantage. We might inquire of the investor at what rates he wished to discount expected returns and covariances of return. Given these discount rates we derive the efficient (discounted) \( E, V \) surface. When the investor chose the particular efficient portfolio he desired, we could plot out for him the time pattern of \( E_t \) and \( V_t \) (and perhaps \( C_{st} \)). If, given the rates at which substitution could take place, he wished to alter this time pattern the analysis could be done with revised discount rates. Each portfolio
obtained in this way would be an efficient portfolio. A short number of trials might bring us near the efficient portfolio which best suits the investor's taste.

What pattern of discount rates is it plausible to start with? Suppose we were interested in the mean and variance of discounted returns (as distinguished from the discounted mean and variance of return). Suppose further that we discounted (continuously) by a single interest rate \( r \).

Let \( \gamma_t \) be the return in the \( t^{th} \) period; then the discounted return and its mean and variance would be:

\[
D = \sum_{t=1}^{\infty} \frac{\gamma_t}{e^{rt}}
\]

\[
\mathbb{E}(D) = \sum_{t=1}^{\infty} \frac{\mathbb{E}\gamma_t}{e^{rt}}
\]

\[
\mathbb{V}(D) = \sum_{s=1}^{\infty} \sum_{t=1}^{\infty} \frac{\text{Cov}(\gamma_s, \gamma_t)}{e^{r(s+t)}}
\]

In this case \( \delta_t = \left(\frac{1}{e^r}\right)^t \) and \( \delta_{st} = \left(\frac{1}{e^r}\right)^{s+t} \). It seems reasonable to begin with a set of \( \delta_t \) and \( \delta_{st} \) of this structure.

Above we show how the methods of Chapter IV can be applied to the multitemporal case when we introduced discount rates \( \delta_t \) and \( \delta_{st} \). Similarly, if in deriving behavior equations our b.f.f.'s refer to discounted probability beliefs, the methods of Chapter V can be applied.

In Chapter II we rejected the hypothesis of discounted expected returns because of its absurd implications. These implications followed because \( U = \sum \delta_t E_t \) is a linear function of the \( X \)'s. This is not so when we discount mean and variance and/or higher moments.
REFERENCES


