An Inequality for Detecting Financial Fraud, Derived from the Markowitz Optimal Portfolio Theory

Gregory V. Bard[1]

[1]*Dept. of Math., Stat., and Comp. Sci., Jarvis Hall Science Wing,*
The University of Wisconsin—Stout, Menomonie, Wisconsin, 54751, USA

bardg@uwstout.edu

Abstract. The Markowitz Optimal Portfolio Theory, published in 1952, is well-known, and was often taught because it blends Lagrange Multipliers, matrices, statistics, and mathematical finance. However, the theory faded from prominence in American investing, as Business departments at US universities shifted from techniques based on mathematics, finance, and statistics, to focus instead on leadership, public speaking, interpersonal skills, advertising, etc. . .

The author proposes a new application of Markowitz's Theory: the detection of a fairly broad category of financial fraud (called "Ponzi schemes" in American newspapers) by looking at a particular inequality derived from the Markowitz Optimal Portfolio Theory, relating volatility and expected rate of return. For example, one recent Ponzi scheme was that of Bernard Madoff, uncovered in December 2008, which comprised fraud totaling 64,800,000,000 US dollars [23].

The objective is to compare investments with the "efficient frontier" as predicted by Markowitz's theory. Violations of the inequality should be impossible in theory; therefore, in practice, violations might indicate fraud.

Introduction

In 1952, Harry Markowitz published a paper [18] that established a model, wherein the selection and weighting of a portfolio of stocks is represented as an optimization problem solved with Lagrange[1] Multipliers. That paper was the birth of an entire field, often called "Markowitz Portfolio Theory" or "Optimal Portfolio Theory" [18] [19]. Of course, later research has in many ways superseded the original work, as one would expected in most branches of the mathematical sciences. Nonetheless, the original theory is very respected and widely[2] taught. In the following work, the author proposes an entirely different use of Markowitz Portfolio Theory—the detection of fraud.

A central idea in Markowitz Portfolio Theory is that the volatility and risk of a financial instrument are well modeled by measuring the variance of that financial instrument (or its square root—the standard deviation). In addition to considering the variances and expected values of stocks, the theory also considers the covariance of each stock with each other stock. In fact, the covariance matrix, computed from historical data (e.g. [16]), will play a central role.

In Markowitz's first model [18] [19], one seeks either to find a portfolio of stocks that maximizes the expected rate of return for a fixed variance (i.e. risk tolerance), or to find a portfolio that minimizes the variance (i.e. the risk) for a desired expected rate of return. Instead, this second notion—the objective of minimizing the variance subject to a desired expected rate of return—will be turned around to construct an inverse problem. This inverse problem creates a binding inequality (see Theorem 2, below), connecting the expected rate of return and the variance. It should be impossible, in the strongest sense of the word impossible, to violate this inequality. When violations are detected in practice, they should be investigated for possible fraud.

By fraud, let us consider mutual funds and hedge funds, such as the Ponzi[3] scheme of Bernard Madoff [7] [11]

[1]A reader who desires to review the theory of Lagrange Multipliers is encouraged to consult [12], a short paper which presents multiple equivalent versions of the technique, and points out many false statements that appear inside American textbooks about "Business Calculus," as well as providing interesting geometric and intuitive justifications for the technique and why it works.

[2]Readers new to the topic might want to consult the survey articles [20] and [25], as well as the historical retrospective [21].

[3]While it is unimportant to this paper, for a discussion of what makes a fraudulent fund a Ponzi scheme, including a history from Charles Ponzi himself to Bernard Madoff, the reader can see [2].

[27], that claim to have high rates of return with low levels of risk. In fact, Bernard Madoff's fund exhibited a relatively high rate of return (12%–18%) and extremely low variance [4] [17], and is a spectacular violation of the inequality. Using much simpler techniques, this incompatibility was detected by mathematical financier Harry Markopolos, who in turn alerted the United States Security and Exchange Commission (SEC) on several occasions and also alerted the Attorney General of New York State once [8] [9] [13] [17]. Here is a quotation from Harry Markopolos's book about his attempts to expose Madoff.

> At the bottom of the page, a chart of Madoff's return stream *rose steadily at a 45-degree angle*, which simply doesn't exist in finance. Within five minutes I told Frank, "There's no way this is real. This is bogus." As I continued examining the numbers, the problems with them began popping out as clearly as a red wagon in a field of snow. There was a stunning lack of financial sophistication. Anyone who *understood the math of the market would have seen these problems immediately*. A few minutes later I laid the papers down on my desk. "This is a fraud, Frank," I told him.
>
> —Harry Markopolos [17]

Another quote from a TIME magazine interview [4] explicitly highlights the relationship between expected rate of return, variance, and covariance, three hallmarks of the Markowitz Portfolio Theory.

> Markopolos said Madoff was earning 82% of the S&P 500's return with less than 22% of the risk, but his returns only had a 6% correlation when Markopolos expected "something like a 50%" correlation. "If your returns are coming from the S&P 100 stock index, you better at least resemble that stock's performance," he said [4].

In the case of Bernard Madoff, the money was not invested at all. Some suspected Madoff of fraud but many others simply imagined him to be a financial genius. We know from several interviews of Harry Markopolos [6] [13] [26], and his book [17], that he used much simpler techniques than what is proposed here. For example, Markopolos wrote a memo the SEC in 2005 that contained 29 different "red flags" or statistical impossibilities that suggested Madoff's fraud [4] [8]. In contrast, the power of a mathematical theorem is that it removes the role of opinion, and provides a clearly and precisely testable criterion.

Had this inequality (Theorem 2, below) been known to fraud investigators at the time, Bernard Madoff's fraud would have been detected years, if not decades, earlier—saving the US Economy considerable damage (64,800,000,000 US dollars [23]) and the catastrophic loss of the life-savings of his many famous investors. Of course, Madoff's scheme is only one example. In the year after Madoff was exposed, the SEC identified three other, totally unrelated, Ponzi schemes [5]. Since the resources of the SEC (and the equivalent body in other nations) are finite, it makes sense to have some instrument which identifies a few mutual funds worthy of investigation for fraud.

Notation

Consider a market with n available stocks. For $i = 1, 2, 3, \ldots, n$, let us define the following variables:

- The vectors $\vec{0}$ and $\vec{1}$ are vectors comprised entirely of zeros, and of ones, respectively.
- Let S_i represent the rate of return of Stock i, considered as a random variable.
- Let $\mu_i = E(S_i)$ represent the expected value of S_i.
- For $i \neq j$ let $M_{ij} = \text{Cov}(S_i, S_j)$, the covariance of S_i and S_j.
- Let $M_{ii} = \text{Var}(S_i) = \text{Cov}(S_i, S_i)$, the variance of S_i, because it is well-known that the covariance of a variable with itself equals its variance. Thus, the standard deviation of S_i, a common measure of its volatility, is $\sqrt{M_{ii}}$, because the standard deviation is the square root of the variance.
- The matrix M is a covariance matrix. Like any covariance matrix, it is symmetric. We will frequently make use of the fact that $M_{ij} = M_{ji}$ during the derivation without commenting on it.
- The column vector $\vec{\mu} = \langle \mu_1, \mu_2, \ldots, \mu_n \rangle$, is simply the vector of the expected rates of returns.
- Any particular portfolio can be defined as a collection of weights $\vec{w} = \langle w_1, w_2, \ldots, w_n \rangle$. Specifically, w_i is the amount invested in security i divided by the total value of the portfolio.
- Because of this definition of the w_is, we know that $\sum_{i=1}^{i=n} w_i = 1$.
- The symbol Ω will be used to mark the end of a proof.

There is a special note to be made here. The inexperienced reader might imagine that $w_i \geq 0$ for all $i \in \{1, 2, \ldots, n\}$. However, this is not true, because of the phenomena of "short selling," which is equivalent to owning a negative number of shares. There is also a complication in that fractional shares of stock are not usually possible—one must buy an integer number of shares. We will deal with these matters in the section "Reflections on the Realism of the Model" below.

About the Appendices: The author has written some appendices to this paper, which will explain some of the concepts to mathematicians who might not be as familiar as others with this specific branch of mathematical finance. For example, readers unfamiliar with the concept of a short sale can find an explanation in Appendix 3: "What is a Negative w_i? What is a "Short Position?" The appendices are given in the online version of this paper, on the author's webpage[4] but are too large to be included in the published paper.

On the Probability Distribution of the Rates of Return

As you can see, there is no inclusion of the skewness, kurtosis, or higher-order moments in this model. That means, in some sense, that the model assumes the rate of return is normally distributed (distributed according to the Gaussian Distribution), with mean μ_i and standard deviation $\sqrt{M_{ii}}$. More precisely, the rates of return following the Gaussian Distribution is *sufficient* for this theory to be effective, but it is not necessary.

For example, a very interesting possibility is to model the logarithms of the rates of return, instead of the actual rates of return. This would imply that the rates of return are log-normally distributed (sometimes called the Gibrat distribution). In such a case, the rates of return are modeled by e^X, where X is a random variable from the Gaussian distribution. We will do this in Appendix 5: "Modeling with the Logarithms of the Rates of Return." Amazingly, all the formulas are left unchanged, except that M is replaced by a different matrix. (Actually, the idea to consider the logarithms of the rates of return, rather than the rates of return themselves, was already proposed by Markowitz in 1959 [19] and he proposed other possible changes as well, such as replacing the variance with the semi-variance.) A discussion on the merits and demerits of taking the logarithms of rates of return can be found in [1], based on data from Norway, Germany, Japan, and the USA.

A great deal of literature has been published relating to underlying assumptions of the Markowitz model, but most especially about the utility functions and about the probability distributions of the rates of return—so much, that it would be infeasible to cite it all in this paper. Just as an example, there have been attempts to include skewness [15]. A recent paper by Markowitz himself surveys the literature dealing with the challenges to these underlying assumptions, with a focus on how well the model works in practice on actual historical data [22].

Helpful Lemmas

Lemma 1 *The expected value of a portfolio (μ_p) with weights given by $\vec{w}$ can be computed with*

$$\mu_p = \vec{w} \circ \vec{\mu}$$

signifying the ordinary dot-product of vectors.

Proof: Using the fact that the expected value of a sum is the sum of the expected values, and that $E(rV) = rE(V)$ for any constant real number r and any random variable V, we can simply calculate the following

$$\mu_{\text{p}} = \text{E}\left(\sum_{i=1}^{i=n} w_i K_i\right) = \sum_{i=1}^{i=n} E(w_i K_i) = \sum_{i=1}^{i=n} w_i E(K_i) = \sum_{i=1}^{i=n} w_i \mu_i = \vec{w} \circ \vec{\mu}$$

Ω

Lemma 2 *The variance of the a portfolio (σ_p^2) with weights given by $\vec{w}$ can be computed with*

$$\sigma_p^2 = \vec{w}^T M \vec{w}$$

where ^{T} is the transpose.

[4] `http://www.gregorybard.com/` under "Publications."

Proof: First, we use the property that the variance of a random variable is equal to the covariance of that random variable with itself.

$$\sigma_{\text{p}}^2 = \text{Var}\left(\sum_{i=1}^{i=n} w_i S_i\right) = \text{Cov}\left(\sum_{i=1}^{i=n} w_i S_i, \sum_{j=1}^{j=n} w_j S_j\right)$$

While this looks like a mess, we must use the property called "the covariance of sums" or "the linearity of covariance" in various textbooks, to transform this.

$$Cov\left(\sum_{i=1}^{i=n} w_i S_i, \sum_{j=1}^{j=n} w_j S_j\right) = \sum_{i=1}^{i=n}\sum_{j=1}^{j=n} w_i w_j Cov(S_i, S_j) = \sum_{i=1}^{i=n}\sum_{j=1}^{j=n} w_i w_j M_{ij}$$

Last but not least, one can algebraically verify that

$$\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} w_i w_j M_{ij} = \vec{w}^T M \vec{w}$$

Ω

A Technical Note on the Invertibility of M

The derivation in the following section will require the use of the inverse of M. Therefore, one must attempt to argue that the inverse of M actually exists. Moreover, even if M is not singular, it might be the case that M is nearly-singular, in the sense of having a large "condition number." Such ill-conditioned matrices have inverses that are difficult to compute numerically. This topic has been researched thoroughly by Pappas, Kiriakopoulos, and Kaimakamis in [24]. For example, when analyzing the trading of 43 stocks traded on the Athens and Frankfurt stock exchanges, they have a case where $\det M = 1.9272 \times 10^{-50}$. Since $\det M \neq 0$, strictly speaking, the inverse does exist. Yet, the condition number in this case is 126.65, which means that accurate and fast computation of M^{-1} will be somewhat challenging. There is another example, involving the trading of seven currencies,[5] where $\det M = 1.76 \times 10^{-77}$ and the condition number is 5.9323×10^{16}.

In [24], those authors entirely circumvent these numerical challenges by substituting the Moore-Penrose Pseudo-Inverse of M in place of M^{-1}. Since the present paper is a theoretical derivation, we will maintain the notation with M^{-1} representing the inverse of M. However, anyone who wishes to implement this in practice would be advised to use the Moore-Penrose Pseudo-Inverse for M^{-1}, instead. When the Moore-Penrose Pseudo-Inverse is computed via the singular-value decomposition, the rounding error is considerably reduced compared to Gaussian Elimination [24].

It should be noted that the Moore-Penrose Pseudo-Inverse of M exists for all M, even when M is singular. Therefore, the derivation that follows will not separately consider those cases where M is singular.

The Traditional Derivation of the Markowitz-Optimal Portfolio

Theorem 1 *For a given expected rate of return μ_D, the minimum variance portfolio to achieve this expected rate of return is given by*

$$\vec{w} = \mu_D\left(\frac{k_2}{2(k_1^2 - k_2k_3)}M^{-1}\vec{\mu} + \frac{-k_1}{2(k_1^2 - k_2k_3)}M^{-1}\vec{1}\right) + \left(\frac{-k_1}{2(k_1^2 - k_2k_3)}M^{-1}\vec{\mu} + \frac{k_3}{2(k_1^2 - k_2k_3)}M^{-1}\vec{1}\right)$$

where

$$k_1 = -\frac{1}{2}\left(\vec{1} \circ M^{-1}\vec{\mu}\right) \qquad k_2 = -\frac{1}{2}\left(\vec{1} \circ M^{-1}\vec{1}\right) \qquad k_3 = -\frac{1}{2}\left(\vec{\mu} \circ M^{-1}\vec{\mu}\right)$$

and where M and μ are defined as in the "Notation" section of this paper, provided that $k_1^2 - k_2k_3 \neq 0$. (Note, if M is singular, then replace M^{-1} with the Moore-Penrose Pseudo-Inverse.)

[5] In all fairness, one should mention that the trading of currencies is somewhat different than the trading of stocks. If one believes that the currency market is efficient in the sense of the "efficient market hypothesis" then it should be impossible to expect a profit by trading currencies, as a positive expected value would represent arbitrage.

A proof of this theorem can be found in many financial mathematics textbooks, including [3, Ch 5]. However, to ensure that this paper is self-contained, a full proof, including all possible details, is given in Appendix 1: "A Full Derivation of the Classical Markowitz Optimal Portfolio Theorem." Every algebraic step is carefully included. Also, Appendix 2: "Viewing the Optimal Portfolio Theory as an Algorithm" contains a specification of this method as an algorithm, suitable for computer science students.

As you can see, the weights of the optimal portfolio are an affine function of μ_D, which is amazing, given how complicated the derivation was. Since we know $\vec{w}$ explicitly now, we know the optimal portfolio structure for any μ_D, where optimal means "lowest variance."

It is only natural to consider when $k_1^2 - k_2k_3 = 0$. This will occur if $\vec{1}\vec{\mu}^T - \vec{\mu}\vec{1}^T = 0$ (i.e. the outer product), which implies that all entries of $\vec{\mu}$ are equal. In other words, that all stocks in the data set have equal expected rates of return. Clearly, that would never actually happen.

Further Insight and the Derivation of the Inequality

Lemma 3 *For any μ_D, the variance of the minimum variance portfolio, using the notation of Theorem 1, will obey the following equation.*

$$\sigma_P^2 = \mu_D^2 \frac{k_2}{2(k_1^2 - k_2k_3)} + \mu_D \frac{-k_1}{k_1^2 - k_2k_3} + \frac{k_3}{2(k_1^2 - k_2k_3)}$$

Proof: Multiplying Equation 6 by M gives us our starting point.

$$\begin{aligned}
M\vec{w} &= M\left(\mu_D \frac{k_2}{2(k_1^2 - k_2k_3)} M^{-1}\vec{\mu} + \mu_D \frac{-k_1}{2(k_1^2 - k_2k_3)} M^{-1}\vec{1} + \frac{-k_1}{2(k_1^2 - k_2k_3)} M^{-1}\vec{\mu} + \frac{k_3}{2(k_1^2 - k_2k_3)} M^{-1}\vec{1}\right) \\
&= \mu_D \frac{k_2}{2(k_1^2 - k_2k_3)} MM^{-1}\vec{\mu} + \mu_D \frac{-k_1}{2(k_1^2 - k_2k_3)} MM^{-1}\vec{1} + \frac{-k_1}{2(k_1^2 - k_2k_3)} MM^{-1}\vec{\mu} + \frac{k_3}{2(k_1^2 - k_2k_3)} MM^{-1}\vec{1} \\
&= \mu_D \frac{k_2}{2(k_1^2 - k_2k_3)} \vec{\mu} + \mu_D \frac{-k_1}{2(k_1^2 - k_2k_3)} \vec{1} + \frac{-k_1}{2(k_1^2 - k_2k_3)} \vec{\mu} + \frac{k_3}{2(k_1^2 - k_2k_3)} \vec{1} \\
&= \left(\mu_D \frac{k_2}{2(k_1^2 - k_2k_3)} + \frac{-k_1}{2(k_1^2 - k_2k_3)}\right)\vec{\mu} + \left(\mu_D \frac{-k_1}{2(k_1^2 - k_2k_3)} + \frac{k_3}{2(k_1^2 - k_2k_3)}\right)\vec{1} \\
\vec{w} \circ M\vec{w} &= \vec{w} \circ \left[\left(\mu_D \frac{k_2}{2(k_1^2 - k_2k_3)} + \frac{-k_1}{2(k_1^2 - k_2k_3)}\right)\vec{\mu} + \left(\mu_D \frac{-k_1}{2(k_1^2 - k_2k_3)} + \frac{k_3}{2(k_1^2 - k_2k_3)}\right)\vec{1}\right] \\
&= \left(\mu_D \frac{k_2}{2(k_1^2 - k_2k_3)} + \frac{-k_1}{2(k_1^2 - k_2k_3)}\right)\vec{w} \circ \vec{\mu} + \left(\mu_D \frac{-k_1}{2(k_1^2 - k_2k_3)} + \frac{k_3}{2(k_1^2 - k_2k_3)}\right)\vec{w} \circ \vec{1} \\
&= \left(\mu_D \frac{k_2}{2(k_1^2 - k_2k_3)} + \frac{-k_1}{2(k_1^2 - k_2k_3)}\right)(\mu_D) + \left(\mu_D \frac{-k_1}{2(k_1^2 - k_2k_3)} + \frac{k_3}{2(k_1^2 - k_2k_3)}\right)(1) \\
&= \mu_D^2 \frac{k_2}{2(k_1^2 - k_2k_3)} + \mu_D \frac{-k_1}{2(k_1^2 - k_2k_3)} + \mu_D \frac{-k_1}{2(k_1^2 - k_2k_3)} + \frac{k_3}{2(k_1^2 - k_2k_3)} \\
&= \mu_D^2 \frac{k_2}{2(k_1^2 - k_2k_3)} + \mu_D \frac{-k_1}{k_1^2 - k_2k_3} + \frac{k_3}{2(k_1^2 - k_2k_3)}
\end{aligned}$$

Since $\vec{w} \circ M\vec{w} = \vec{w}^T M\vec{w}$ is the variance of the portfolio defined by $\vec{w}$ (see Lemma 2), then we have computed the optimal portfolio's variance. Ω

Theorem 2 *For any portfolio the expected rate of return μ_P and the standard deviation σ_P will obey the following inequality (using the notation of Theorem 1)*

$$\sigma_P \geq \sqrt{\mu_P^2 \left(\frac{k_2}{2(k_1^2 - k_2k_3)}\right) + \mu_P \left(\frac{-k_1}{k_1^2 - k_2k_3}\right) + \frac{k_3}{2(k_1^2 - k_2k_3)}}$$

Proof: The standard deviation of the optimal portofolio

$$\sigma_{\text{p}} = \sqrt{\mu_{\text{p}}^2 \left(\frac{k_2}{2(k_1^2 - k_2 k_3)} \right) + \mu_{\text{p}} \left(\frac{-k_1}{k_1^2 - k_2 k_3} \right) + \frac{k_3}{2(k_1^2 - k_2 k_3)}}$$

was proven to be the minimum possible, in Lemma 3.

Of course, whatever the standard deviation of any particular portfolio, it cannot be better than (less than) the optimal. Therefore, it is greater than or equal to the optimal. Ω

One can represent Theorem 2 graphically. Classically, the μ_{p} is on the y-axis and the σ_{p} is on the x-axis. Any particular portfolio can be thought of as a point inside the coordinate plane. The inequality then becomes $x \geq \sqrt{\#y^2 + \#y + \#}$ where the # symbol represents the coefficients found in Theorem 2. Squaring this equation gives a hyperbola in the coordinate plane. This hyperbola is called "the efficient frontier" or the "Markowitz Bullet" in textbooks (e.g. [3, Ch 5]). In Theorem 2, we have proven that any such feasible portfolio will be a point on, or inside, the Markowitz Bullet.

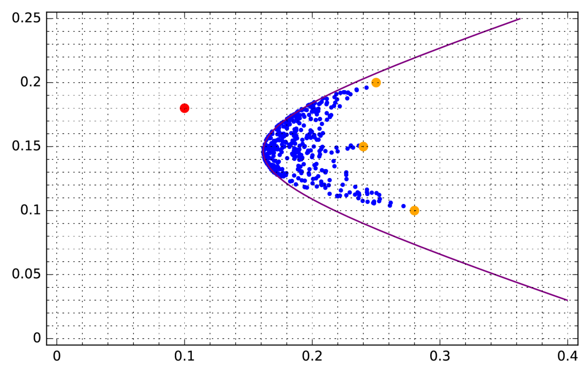

FIGURE 1. A Sample Plot of 500 Random Portfolios, and the Markowitz Bullet, Generated from Three Hypothetical Stocks. (The blue dots are the random portfolios; the purple curve is the Markowitz Bullet; the orange large dots, inside the curve, are the three original stocks; the large red dot, outside the curve, is one case of Madoff's reported "performance.")

An example is plotted in Figure 1, from hypothetical data. While actual market data would have been preferable, such data sources are not free, and the importation of the data into computer algebra systems is a highly non-trivial task. This figure represents a continuation of Example 5.10 from [3], a case three stocks, so that M is a 3×3 matrix. The values are realistic, and represent a good year for the stock market. The Markowitz Bullet (the purple curve) was computed in Sage

$$x = \sqrt{9.85134y^2 - 2.88543y + 0.237392} \qquad \text{implying} \qquad \sigma_p \geq \sqrt{9.85134\mu_p^2 - 2.88543\mu_p + 0.237392}$$

and matches the textbook example. To this were added 500 random portfolios, each without short sales ($w_i \geq 0$), and a red dot reporting the performance of Madoff's fund in a good year: $\mu = 0.18$ and $\sigma = 0.1$. As the reader can see, the red dot (Madoff) is no where near the Markowitz Bullet.

Reflections on the Realism of the Model

The Markowitz Optimal Portfolio Model, like any model in the mathematical sciences, makes some assumptions which must be compared to reality. In order to show the impact of these flaws in the model, it is necessary to make an observation that might be obvious to some readers but not others.

Suppose that in two optimization problems, Problem A and Problem B, one is trying to minimize the same objective function. Further suppose that the feasible set of Problem A, denoted $\mathcal{A}$, and the feasible set of Problem B, denoted $\mathcal{B}$, are both subsets of the same universal set $\mathcal{U}$. It is clear that if $\mathcal{A} \subset \mathcal{B} \subset \mathcal{U}$, then the minimum feasible value of the objective function in Problem A is greater than or equal to the minimum feasible value of the objective function in Problem B.

In plainer words, if you have solved a optimization problem, minimizing some function on a set $\mathcal{B}$, and then remove points from the set $\mathcal{B}$ to create a new set $\mathcal{A}$, then the new feasible minimum value of the objective function might remain the same, or it might increase, but it definitely will not decrease.

Integer Shares of Stock In many situations, one can buy only an integer number shares. This does not mean that the w_is are integers. If a fund has F dollars to invest, and Stock j costs c dollars per share, then w_j must be an integer multiple of c/F. If we restrict ourselves to integer shares of stock, then we are (drastically) shrinking the feasible set of the original optimization problem. Therefore the minimum variance achievable in practice will either be equal to, or greater than, the minimum variance proven during Lemma 3.

Management Dislikes of Particular Companies It may come to pass that one of the fund's managers might have a strong dislike for a particular industry, or a particular company. For example, some funds refuse to invest in companies that engage in environmental abuses, because they are afraid that some customers would withdraw their investments entirely if the abuses are made known in the popular press. In any case, ruling a company out is clearly an example of making a subset of the available choices, by forcing some particular w_i to be zero. Since this shrinks the feasible set of the original optimization problem, the minimum variance achievable in practice will either be equal to, or greater than, the minimum variance proven during Lemma 3.

Bounded Short Positions Negative w_is represent "short positions." These can be thought of as borrowing someone else's shares, or as buying a negative number of shares (see Appendix 3: "What is a Negative w_i? What is a "Short Position?"). The sum of the absolute values of those w_is which happen to be negative should be limited by some fixed constant. Call this sum s. (For example, some investment bank might impose $s \leq 1/3$ as a fixed constraint.) Many investors might require $s = 0$. For sure, s cannot be unbounded in a realistic setting. Portfolios with s less than or equal to this constant are feasible, but portfolios with s greater than this constant must be ruled infeasible. This shrinks the feasible set of the original optimization problem. As before, the minimum variance achievable in practice will either be equal to, or greater than, the minimum variance proven during Lemma 3.

Observe that in all of these cases, we conclude that the minimum variance achievable in practice will either be equal to, or greater than, the minimum variance proven during Lemma 3. Since we are using that minimum variance as our lower bound, the inequality at the heart of this paper (given in Theorem 2), remains absolutely true. None of these cases should cause the inequality to be violated under any circumstances.

Another set of flaws in the model involve lowering μ_p, the yield of the fund. The Markowitz model cannot take into account the following expenses: brokerage and transaction fees, wages of the fund management, taxes of all kinds, interest/fees on short positions. These are not a problem, because they cost the fund money. This actually makes the inequality stronger. Thinking graphically, because these expenses lower the expected rate of return of the portfolio, they move the portfolio downward on the graphs that comprise Figure 1. If a fund violates the inequality from Theorem 2 after these expenses, then that means it was even more in violation (that is to say, violating by a larger quantity) before such expenses. Moreover, it is not possible for a legitimate fund, inside the Markowitz Bullet before these adjustments, to end up outside of it after adjustment.

Next, the sum of the $\vec{w}$s should not be 1 in reality. Some money must be left over as a "cash position," somewhat due to the requirement of only owning an integer number of shares, and somewhat due to the need to pay the four types of expenses listed above. Since this money will earn extremely meager returns, it lowers the overall yield of the fund, μ_p. Therefore, it makes the inequality stronger, as in the above paragraph.

It is always possible that a hedge fund might invest in a bizarre asset, like a diamond mine or Hollywood movies. The covariances of those assets would not be available to regulators. However, the vast majority of mutual funds restrict themselves to some category of assets, and often a set of stocks or bonds. For example, Bernard Madoff's fund restricted itself to the stocks of the S&P100, a subset of the S&P500, and options on those stocks [17] [27].

Last but not least, there are the non-financial rewards, in some cases, of funds that own shares in a company. These can including voting rights or even control of some company, even possibly acquiring and closing a competitor. For example, a fund that owns large parts of a traditional publishing house, that purchases a small but successful new online publishing company, only to close it down and prevent competition with the traditional publishers. This author knows of no way to model such rewards mathematically, but they are rare events.

Acknowledgements

First, this paper is especially indebted to Prof. Angel Marchev, Jr., of the Department of Management, at the University of National and World Economy, in Sofia, Bulgaria. His criticism of the model during the "42nd International Conference on the Applications of Mathematics in Engineering and Economics" was extremely welcome and constructive. The paper is much improved as a result, particularly the section "Reflections on the Realism of the Model." Second, another attendee of that conference suggested examining logarithmic prices, but regrettably, the author did not catch the commentator's name. Third, the author would like to thank the organizers of that conference, for both allowing this unusual paper to be presented and for modifying the schedule of the day to allow more conference attendees to be present during the report. Fourth, the author would like to thank Prof. Dan Kalman of American University, in Washington, DC, USA, for writing the paper [12]. Fifth, Prof. Shapoor Vali of Fordham University, New York City, USA, has given much encouragement to the author regarding the study of the overlaps between mathematics, economics and finance. Yet above all others, Prof. Harry Markowitz, of the Rady School of Management at the University of California at San Diego, USA, is to be thanked for inventing this model, and his continued work on all topics related to mathematical models of risk.

Finally, the author must thank Harry Markopolos for his bravery, persistence, and unrelenting pursuit of justice.

REFERENCES

[1] K. Aas, *To log or not to log: the Distribution of Asset Returns*, Tech. Rep. SAMBA/03/04, Norwegian Computing Center, September 2004.

[2] A. Altman, *A Brief History Of Ponzi Schemes*, TIME Magazine, (2008).

[3] M. Capinski and T. Zastawniank, *Mathematics for Finance: An Introduction to Financial Engineering*, SUMS: Springer Undergraduate Mathematics Series, Springer, 2003.

[4] R. Chew, *A Madoff Whistle-Blower Tells His Story*, TIME Magazine, (2009).

[5] ——, *Beyond Madoff, Ponzi Schemes Proliferate*, TIME Magazine, (2009).

[6] DealBOOK-Staff, *Madoff Whistleblower Assails S.E.C. for Ignoring Him*, The New York Times, (2009).

[7] A. Efrati, T. Lauricella, and D. Searcey, *Top Broker Accused of Fraud: Madoff, Money Manager for the Wealthy, Said to Have Run '$50 Billion Ponzi Scheme'*, The Wall Street Journal, (2008).

[8] J. Fox, *Harry Markopolos really did have the goods on Bernie Madoff*, TIME Magazine, (2008).

[9] D. Henriques, *At Madoff Hearing, Lawmakers Lay Into S.E.C.*, The New York Times, (2009).

[10] M. R. Hestenes, *Optimization Theory: the Finite Dimensional Case*, Wiley, 1975.

[11] E. Honan and D. Wilchins, *Former NASDAQ chair arrested over alleged 33 billion [pound] fraud*, International Herald Tribune, (2008).

[12] D. Kalman, *Leveling with Lagrange: An Alternate View of Constrained Optimization*, Mathematics Magazine, 82 (2009), pp. 186–196.

[13] R. Kerber, *The Whistleblower: Dogged pursuer of Madoff wary of fame*, The Boston Globe, (2009).

[14] S. Kotz, N. Balakrishnan, and N. L. Johnson, *Continuous Multivariate Distributions, Vol. 1: Models and Applications*, Wiley Interscience, second ed., 2000.

[15] T.-Y. Lai, *Portfolio selection with skewness: A multiple-objective approach*, Review of Quantitative Finance and Accounting, 1 (1991), pp. 293–305.

[16] O. Ledoit and M. Wolf, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection*, Journal of empirical finance, 10 (2003), pp. 603–621.

[17] H. Markopolos, *No One Would Listen: A True Financial Thriller*, Wiley, 2011.

[18] H. Markowitz, *Portfolio Selection*, The Journal of Finance, 7 (1952), pp. 77–91.

[19] ——, *Portfolio Selection: Efficient Diversification of Investment*, no. 16 in Cowles Foundation Monograph, John Wiley and Sons, 1959.

[20] ——, *Foundations of Portfolio Theory*, The Journal of Finance, 46 (1991), pp. 469–477.

[21] ——, *The Early History of Portfolio Theory: 1600–1960*, Financial Analysts Journal, 55 (1999), pp. 5–16.

[22] ——, *The "Great Confusion" Concerning MPT [Markowitz Portfolio Theory]"*, Aestimatio, the IEB International Journal of Finance, (2012), pp. 8–27.

[23] G. McCool and M. Graybow, *Madoff pleads guilty, is jailed for $ 65 billion fraud*, Reuters, (2009).

[24] D. Pappas, K. Kiriakopoulos, and G. Kaimakamis, *Optimal portfolio selection with singular covariance matrix*, International Mathematical Forum, 5 (2010), pp. 2305–2318.

[25] M. Rubinstein, *Markowitz's "Portfolio Selection:" A Fifty-Year Retrospective*, The Journal of Finance, 57 (2002), pp. 1041–1045.

[26] D. Solomon, *Math is Hard: Questions for Harry Markopolos*, The New York Times Magazine, (2010).

[27] The Office of the US Attorney of the Southern District of New York, *Bernard l. madoff charged in 11-count criminal information.* Federal Bureau of Investigation Press Release, March 2009.

1: A Full Derivation of the Classical Markowitz Optimal Portfolio Theorem

Here is a full proof, showing all possible details, of Theorem 1.

Proof: First, we recall the general technique of Lagrange Multipliers, in the case of two constraints. In the case when f, g_1 and g_2 are suitably[6] well-behaved, the minimum value of the function

$$f(x_1, x_2, \ldots, x_n) \text{ subject to } g_1(x_1, x_2, \ldots, x_n) = 0 \text{ as well as } g_2(x_1, x_2, \ldots, x_n) = 0$$

will be a point in the set of points that makes the gradient of the following function equal to $\vec{0}$.

$$L(x_1, x_2, \ldots, x_n, \lambda_1, \lambda_2) = f(x_1, x_2, \ldots, x_n) + \lambda_1 g_1(x_1, x_2, \ldots, x_n) + \lambda_2 g_2(x_1, x_2, \ldots, x_n)$$

where the new variables λ_1 and λ_2 are the Lagrange Multipliers.

Second, suppose that an investor desires an expected rate of return of μ_D but wishes to achieve that return with the minimum possible risk. Mathematically, suppose that we wish to minimize the variance of the portfolio, subject to the constraint that $\mu_{\text{p}} = \mu_D$ as well as the trivial constraint that $\sum_{i=1}^{i=n} w_i = 1$.

We can pose the investor's desires in terms of f, g_1, and g_2, as follows.

- $f(w_1, w_2, \ldots, w_n) = \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} w_i w_j M_{ij} = \sigma_{\text{p}}^2$, to be minimized.
- $g_1(w_1, w_2, \ldots, w_n) = -\mu_D + \sum_{i=1}^{i=n} w_i \mu_i$, thus forcing $\mu_{\text{p}} = \mu_D$.
- $g_2(w_1, w_2, \ldots, w_n) = -1 + \sum_{i=1}^{i=n} w_i$, thus forcing $\sum_{i=1}^{i=n} w_i = 1$.
- As the reader can see, these three functions are polynomials, and are therefore have continuous second derivatives everywhere. In fact, their second derivatives are constants.

This gives us the following L function:

$$L(w_1, w_2, \ldots, w_n, \lambda_1, \lambda_2) = \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} w_i w_j M_{ij} + \lambda_1 \left(-\mu_D + \sum_{i=1}^{i=n} w_i \mu_i \right) + \lambda_2 \left(-1 + \sum_{i=1}^{i=n} w_i \right)$$

whose gradient must be zero for the optimal portfolio.

Of course, the gradient being the zero vector requires each partial derivative to be zero. In order to compute the partial derivative of L with respect to some w called w_k, we should first consider that double summation. Any terms where both $i \neq k$ and $j \neq k$ are going to become zero. Clearly, we are concerned with three groups of terms: those entries where $i = k$ but $j \neq k$; where $i \neq k$ but $j = k$; and where $i = k$ and $j = k$ simultaneously. Thus we obtain:

$$\frac{\partial}{\partial w_k} L = \frac{\partial}{\partial w_k} \left(\sum_{j=1, j\neq k}^{j=n} w_k w_j M_{kj} + \sum_{i=1, i\neq k}^{i=n} w_i w_k M_{ik} + w_k^2 M_{kk} \right) + \lambda_1 (0 + \mu_k) + \lambda_2 (0 + 1)$$

which simplifies to

$$\frac{\partial}{\partial w_k} L = \left(\sum_{j=1, j\neq k}^{j=n} w_j M_{kj} + \sum_{i=1, i\neq k}^{i=n} w_i M_{ik} + 2 w_k M_{kk} \right) + \lambda_1 \mu_k + \lambda_2$$

Splitting $2w_k M_{kk} = w_k M_{kk} + w_k M_{kk}$ allows us to put one of each into those first two sums, allowing us to remove the restriction that $j \neq k$ on the first sum and $i \neq k$ on the second sum.

$$\frac{\partial}{\partial w_k} L = \left(\sum_{j=1}^{j=n} w_j M_{kj} + \sum_{i=1}^{i=n} w_i M_{ik} \right) + \lambda_1 \mu_k + \lambda_2 \tag{1}$$

However,

$$\sum_{j=1}^{j=n} w_j M_{kj} = \sum_{j=1}^{j=n} M_{kj} w_j = entry_k(M\vec{w}) \text{ while } \sum_{i=1}^{i=n} w_i M_{ik} = \sum_{i=1}^{i=n} M_{ki}^T w_i = entry_k(M^T \vec{w}) = entry_k(M\vec{w})$$

[6] In this case, it is sufficient that each of f, g_1 and g_2 have continuous second derivatives everywhere [10, pp. 255–261].

Now we have

$$\frac{\partial}{\partial w_k} L = 2entry_k(M\vec{w}) + \lambda_1 \mu_k + \lambda_2$$

For the gradient of L to be zero, it is necessary that the partial derivative of L with respect to w_k must equal zero for all $k \in \{1, 2, \dots, n\}$, yielding

$$\begin{array}{rclcl} 0 & = & \frac{\partial}{\partial w_1} L & = & 2entry_1(M\vec{w}) + \lambda_1 \mu_1 + \lambda_2 \\ 0 & = & \frac{\partial}{\partial w_2} L & = & 2entry_2(M\vec{w}) + \lambda_1 \mu_2 + \lambda_2 \\ \vdots & = & \vdots & = & \vdots \\ 0 & = & \frac{\partial}{\partial w_n} L & = & 2entry_n(M\vec{w}) + \lambda_1 \mu_n + \lambda_2 \end{array}$$

Combining these, we get

$$\vec{0} = 2(M\vec{w}) + \lambda_1 \vec{\mu} + \lambda_2 \vec{1}$$

which means that

$$M\vec{w} = -\frac{\lambda_1}{2}\vec{\mu} - \frac{\lambda_2}{2}\vec{1} \tag{2}$$

Now, assuming M is invertible (see "A Technical Note on the Invertibility of M", earlier) then

$$\vec{w} = M^{-1}\left(-\frac{\lambda_1}{2}\vec{\mu} - \frac{\lambda_2}{2}\vec{1}\right) = -\frac{\lambda_1}{2}M^{-1}\vec{\mu} - \frac{\lambda_2}{2}M^{-1}\vec{1} \tag{3}$$

At this moment, we do not yet know λ_1 or λ_2 but we have not used the partial derivatives $\partial L/\partial \lambda_1$ or $\partial L/\partial \lambda_2$. We can compute the latter,

$$\frac{\partial}{\partial \lambda_2} L = \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} 0 + \lambda_1 0 + 1\left(-1 + \sum_{i=1}^{i=n} w_i\right) = -1 + \sum_{i=1}^{i=n} w_i = -1 + \vec{w} \circ \vec{1}$$

and setting it to zero, we just recover one of our constraints, namely that the sum of the weights must equal one.

Plugging in Equation 3 as the definition of $\vec{w}$ we obtain

$$0 = -1 + \vec{1} \circ \vec{w} = -1 + \vec{1} \circ \left(-\frac{\lambda_1}{2}M^{-1}\vec{\mu} - \frac{\lambda_2}{2}M^{-1}\vec{1}\right) = -1 - \vec{1} \circ \left(\frac{\lambda_1}{2}M^{-1}\vec{\mu}\right) - \vec{1} \circ \left(\frac{\lambda_2}{2}M^{-1}\vec{1}\right)$$

which can be rearranged into

$$1 = -\frac{\lambda_1}{2}\left(\vec{1} \circ M^{-1}\vec{\mu}\right) - \frac{\lambda_2}{2}\left(\vec{1} \circ M^{-1}\vec{1}\right) \tag{4}$$

We can compute the remaining partial derivative,

$$\frac{\partial}{\partial \lambda_1} L = \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} 0 + 1\left(-\mu_D + \sum_{i=1}^{i=n} w_i \mu_i\right) + \lambda_2 0 = -\mu_D + \sum_{i=1}^{i=n} w_i \mu_i = -\mu_D + \vec{w} \circ \vec{\mu}$$

and setting it also to zero, we now recover the other of our constraints, namely $\mu_D = \vec{w} \circ \vec{\mu}$.

As before, plugging in Equation 3 as the definition of $\vec{w}$ we obtain

$$\mu_D = \vec{\mu} \circ \left(-\frac{\lambda_1}{2}M^{-1}\vec{\mu} - \frac{\lambda_2}{2}M^{-1}\vec{1}\right) = \vec{\mu} \circ \left(-\frac{\lambda_1}{2}M^{-1}\vec{\mu}\right) - \vec{\mu} \circ \frac{\lambda_2}{2}\left(M^{-1}\vec{1}\right)$$

which can be rearranged into

$$\mu_D = -\frac{\lambda_1}{2}\left(\vec{\mu} \circ M^{-1}\vec{\mu}\right) - \frac{\lambda_2}{2}\left(\vec{\mu} \circ M^{-1}\vec{1}\right) \tag{5}$$

Now Equation 4 and Equation 5 can be regarded as two linear equations in two unknowns

$$\begin{array}{rcl} 1 & = & \left(-\frac{1}{2}\left(\vec{1} \circ M^{-1}\vec{\mu}\right)\right)\lambda_1 + \left(-\frac{1}{2}\left(\vec{1} \circ M^{-1}\vec{1}\right)\right)\lambda_2 \\ \mu_D & = & \left(-\frac{1}{2}\left(\vec{\mu} \circ M^{-1}\vec{\mu}\right)\right)\lambda_1 + \left(-\frac{1}{2}\left(\vec{\mu} \circ M^{-1}\vec{1}\right)\right)\lambda_2 \end{array}$$

and it would be useful to recall Cramer's Rule:

$$\begin{aligned} a &= bx + cy \\ d &= ex + fy \end{aligned}$$

has, as its solutions

$$x = \frac{\det\begin{bmatrix} a & c \\ d & f \end{bmatrix}}{\det\begin{bmatrix} b & c \\ e & f \end{bmatrix}} = \frac{af-cd}{bf-ce} \qquad y = \frac{\det\begin{bmatrix} b & a \\ e & d \end{bmatrix}}{\det\begin{bmatrix} b & c \\ e & f \end{bmatrix}} = \frac{bd-ae}{bf-ce}$$

provided that the determinant in the denominator is not zero.

Yet, in our case

$$\begin{aligned} a &= 1 & b &= -\tfrac{1}{2}\left(\vec{1} \circ M^{-1}\vec{\mu}\right) \\ c &= -\tfrac{1}{2}\left(\vec{1} \circ M^{-1}\vec{1}\right) & d &= \mu_D \\ e &= -\tfrac{1}{2}\left(\vec{\mu} \circ M^{-1}\vec{\mu}\right) & f &= -\tfrac{1}{2}\left(\vec{\mu} \circ M^{-1}\vec{1}\right) \end{aligned}$$

Oddly enough, $b = f$, because

$$\begin{aligned} b &= -\frac{1}{2}\left(\vec{1} \circ M^{-1}\vec{\mu}\right) = -\frac{1}{2}\left(M^{-1}\vec{\mu} \circ \vec{1}\right) = -\frac{1}{2}\left(\left(M^{-1}\vec{\mu}\right)^T \vec{1}\right) = -\frac{1}{2}\left(\left(\vec{\mu}^T M^{-T}\right)\vec{1}\right) \\ &= -\frac{1}{2}\left(\vec{\mu}^T\left(M^{-T}\vec{1}\right)\right) = -\frac{1}{2}\left(\vec{\mu}^T\left(M^{-1}\vec{1}\right)\right) = -\frac{1}{2}\left(\vec{\mu} \circ \left(M^{-1}\vec{1}\right)\right) = f \end{aligned}$$

To keep the formulas human readable, we will define

$$k_1 = b = f = -\frac{1}{2}\left(\vec{1} \circ M^{-1}\vec{\mu}\right) \qquad k_2 = c = -\frac{1}{2}\left(\vec{1} \circ M^{-1}\vec{1}\right) \qquad k_3 = e = -\frac{1}{2}\left(\vec{\mu} \circ M^{-1}\vec{\mu}\right)$$

and then plug these into Cramer's Rule to obtain the values of the λs. Note that $c = k_2$ is just the sum of the entries of M^{-1}, multiplied by $-1/2$.

$$\begin{aligned} \lambda_1 &= x = \frac{af-cd}{bf-ce} = \frac{(1)k_1 - k_2\mu_D}{k_1^2 - k_2k_3} = \mu_D\left(\frac{-k_2}{k_1^2 - k_2k_3}\right) + \frac{k_1}{k_1^2 - k_2k_3} = \mu_D\left(\frac{-k_2}{k_1^2 - k_2k_3}\right) + \frac{k_1}{k_1^2 - k_2k_3} \\ \lambda_2 &= y = \frac{bd-ae}{bf-ce} = \frac{k_1\mu_D - (1)k_3}{k_1^2 - k_2k_3} = \mu_D\left(\frac{k_1}{k_1^2 - k_2k_3}\right) + \frac{-k_3}{k_1^2 - k_2k_3} = \mu_D\left(\frac{k_1}{k_1^2 - k_2k_3}\right) + \frac{-k_3}{k_1^2 - k_2k_3} \end{aligned}$$

Last but not least, we can plug the λs back into Equation 3 to obtain

$$\vec{w} = -\frac{\lambda_1}{2}M^{-1}\vec{\mu} - \frac{\lambda_2}{2}M^{-1}\vec{1} = \left(\mu_D\frac{k_2}{2(k_1^2 - k_2k_3)} + \frac{-k_1}{2(k_1^2 - k_2k_3)}\right)M^{-1}\vec{\mu} + \left(\mu_D\frac{-k_1}{2(k_1^2 - k_2k_3)} + \frac{k_3}{2(k_1^2 - k_2k_3)}\right)M^{-1}\vec{1}$$

but the relationship with μ_D can be made more explicit by rearranging the terms thusly:

$$\vec{w} = \mu_D\frac{k_2}{2(k_1^2 - k_2k_3)}M^{-1}\vec{\mu} + \mu_D\frac{-k_1}{2(k_1^2 - k_2k_3)}M^{-1}\vec{1} + \frac{-k_1}{2(k_1^2 - k_2k_3)}M^{-1}\vec{\mu} + \frac{k_3}{2(k_1^2 - k_2k_3)}M^{-1}\vec{1} \tag{6}$$

and finally, factoring out the μ_D gives us the statement of the theorem. Ω

2: Viewing the Optimal Portfolio Theory as an Algorithm

Some students, particularly those from computer science, might benefit from seeing the optimal portfolio theory presented as pseudocode. We can define the Markowitz Optimal Portfolio as an algorithm as follows. To simplify notation, we define $M^{-1}\vec{1} = \vec{v_1}$ and $M^{-1}\vec{\mu} = \vec{v_2}$, since those appeared so often in the proof of Theorem 1.

Inputs: The matrix M, the vector $\vec{\mu}$, and some $\vec{\mu_D}$, as defined in the section "Notation."

1. Compute M^{-1}. If M is singular, then compute the Moore-Penrose Pseudo-Inverse instead.
2. Compute $M^{-1}\vec{1} = \vec{v_1}$.
3. Compute $M^{-1}\vec{\mu} = \vec{v_2}$.
4. Compute $k_1 = -\frac{1}{2}\left(\vec{1} \circ \vec{v_2}\right)$.
5. Compute $k_2 = -\frac{1}{2}\left(\vec{1} \circ \vec{v_1}\right)$.
6. Compute $k_3 = -\frac{1}{2}\left(\vec{\mu} \circ \vec{v_2}\right)$.
7. If $k_1^2 - k_2 k_3 = 0$ then abort.
8. Compute

$$\vec{w} = \mu_D\left(\frac{k_2}{2(k_1^2 - k_2k_3)}\vec{v_2} + \frac{-k_1}{2(k_1^2 - k_2k_3)}\vec{v_1}\right) + \left(\frac{-k_1}{2(k_1^2 - k_2k_3)}\vec{v_2} + \frac{k_3}{2(k_1^2 - k_2k_3)}\vec{v_1}\right)$$

Output: The vector $\vec{w}$, which contains the weights for constructing the minimum variance portfolio that achieves expected return $\vec{\mu_D}$.

3: What is a Negative w_i? What is a "Short Position?"

Suppose the stock of an imaginary company, ABC industries, is currently selling for 50 Euros per share. In return for 200 Euros today, Investor Alpha promises Investor Beta that he will deliver 4 shares of ABC industries, on a date exactly one year from the present. Consider three scenarios, for what the shares of ABC have done, during that year.

- The price rises to 60 Euros per share. On the delivery date, Investor Alpha buys the shares at a cost of 240 Euros, and delivers them to Investor Beta. Investor Alpha has a net profit of $200 - 240 = -40$ Euros, before transaction costs.
- The price remains steady at 50 Euros per share. On the delivery date, Investor Alpha buys the shares at a cost of 200 Euros, and delivers them to Investor Beta. Investor Alpha has a net profit of $200 - 200 = 0$ Euros, before transaction costs.
- The price falls to 40 Euros per share. On the delivery date, Investor Alpha buys the shares at a cost of 160 Euros, and delivers them to Investor Beta. Investor Alpha has a net profit of $200 - 160 = 40$ Euros, before transaction costs.

The financial community describes this sequence of events as "Investor Alpha buys -4 shares of ABC stock" for two reasons. First, the initial "cost" to Investor Alpha is $(-4)(50) = -200$ Euros, in the sense that Investor Alpha receives 200 Euros at the start of the transaction. Second, if the price changes from 50 to $50 + \Delta$, the profit of Investor Alpha is given by

$$(4)(50) - (4)(50 + \Delta) = 200 - 200 - 4\Delta = -4\Delta$$

in Euros, before transaction costs. As the reader can see, this really is mathematically equivalent to Investor Alpha buying -4 shares of ABC stock.

In reality, there is (usually) no fixed delivery date; that was added to this explanation to ease understanding. Often the way this is carried out in practice is that Investor Alpha might borrow shares of ABC stock from some Investor Gamma, who would charge a fee. Alternatively, Investor Alpha's bank might extend him a credit line, perhaps equal to about 1/4th of his portfolio, or 1/3rd, and the total value of the short positions must be less than this amount. This is so that in the event that all of Investor Alpha's short positions turn out badly, the bank can simply confiscate part of Alpha's portfolio, buy shares of ABC stock, and make sure that Investor Beta gets his shares.

4: Reasons to Believe M is Always Invertible in Practice

An explanatory note is needed here. Some users of the Markowitz model will make an $(n+1)$th asset, called "cash reserves," as a risk-free asset. Other authors place the risk-free asset outside the set of n assets modeled by M and μ. If this $(n+1)$th asset is included, it might yield a fixed rate of return. In that case, its variance is zero, and its covariance with all other stocks is zero. This means that there is an all-zero row and an all-zero column in M and consequently M will be singular. The majority of authors (including this author) therefore put only risk-bearing assets in M and μ, and model the risk-free asset separately, viewing the entire fund as a linear combination of the risk-bearing assets and the risk-free asset.

Let us the follow the convention of grouping the n risk-bearing assets together, and excluding any risk-free asset. The following sequence of lemmas proves that, in this model, if M were to be singular, then something economically absurd would happen. More precisely, one of the three scenarios would occur, each of which is very unlikely.

1. There is a collection of stocks that results in a portfolio with zero variance. That means the rate of return is entirely constant.
2. One can take a feasible portfolio $\vec{w}$, and make a new portfolio $\vec{w}'$, whose variance is unchanged, but whose expected rate of return can be as high as desired (e.g. 800% or 900% per year would not be a problem).
3. There is a stock that is actually a linear combination of other stocks. This would be analogous to a perfectly efficient mutual fund with no inefficiency due to fees, transaction costs, or the salaries of the fund employees, and no time lag.

While these three situations are possibly valid mathematically, they are economically absurd. In practice, they should never occur. Therefore, in practice, M is in all likelihood going to be invertible, provided that the n assets which M describes are all risk-bearing and not risk-free.

Lemma 4 *Suppose a vector $\vec{n}$ exists in the nullspace of M, with $\sum_{i=1}^{i=n} n_i \neq 0$. Then there exists a set of weights that creates a portfolio whose rate of return is absolutely constant.*

Proof: Define $k = \sum_{i=1}^{i=n} n_i$, and note $k \neq 0$ by assumption. Then let $\vec{w} = \frac{1}{k}\vec{n}$. Of course, now $\sum_{i=1}^{i=n} w_i = 1$, and thus $\vec{w}$ can be thought of as the weights of some portfolio. Moreover,

$$\sigma_{\text{p}}^2 = \vec{w}^T M \vec{w} = \vec{w}^T M \left(\frac{1}{k}\vec{n}\right) = \frac{1}{k}\vec{w}^T (M\vec{n}) = \frac{1}{k}\vec{w}^T\vec{0} = 0$$

and therefore this portfolio has zero variance. That means its rate of return is entirely constant. Ω

Lemma 5 *Suppose a vector $\vec{n}$ exists in the nullspace of M, with $\sum_{i=1}^{i=n} n_i = 0$, but $\vec{n} \neq 0$ and $\vec{n} \circ \vec{\mu} \neq 0$. Then for any feasible portfolio $\vec{w}$, there exists a another feasible portfolio $\vec{w}' = \vec{w} + \kappa\vec{n}$ which achieves any desired expected rate of return, no matter how high, with the same variance as $\vec{w}$. (The value κ is easily computed from the desired expected rate of return.)*

Proof: Let $\vec{w}$ be some portfolio, and thus $\sum_{i=1}^{i=n} w_i = 1$.

Let us construct a new portfolio, where the weights are $\vec{w}' = \vec{w} + \kappa\vec{n}$ for some κ. First, let us compute the variance of this new portfolio.

$$\begin{aligned}
\sigma_{\text{p}}^2 &= \vec{w}'^T M \vec{w}' = (\vec{w} + \kappa\vec{n})^T M (\vec{w} + \kappa\vec{n})^T = (\vec{w} + \kappa\vec{n})^T M\vec{w} + (\vec{w} + \kappa\vec{n})^T M\kappa\vec{n} \\
&= (\vec{w} + \kappa\vec{n})^T M\vec{w} + (\vec{w} + \kappa\vec{n})^T \kappa(M\vec{n}) = (\vec{w} + \kappa\vec{n})^T M\vec{w} + (\vec{w} + \kappa\vec{n})^T \kappa\vec{0} = (\vec{w} + \kappa\vec{n})^T M\vec{w} + \vec{0} = \vec{w}^T M\vec{w} + \kappa\vec{n}^T M\vec{w} \\
&= \vec{w}^T M\vec{w} + \kappa\vec{n}^T (M^T)\vec{w} = \vec{w}^T M\vec{w} + \kappa(M\vec{n})^T\vec{w} = \vec{w}^T M\vec{w} + \kappa(\vec{0})^T\vec{w} = \vec{w}^T M\vec{w} + \vec{0} = \vec{w}^T M\vec{w}
\end{aligned}$$

as you can see, the variance of the new portfolio is exactly the same as the variance of the old portfolio. Now let's calculate the expected rate of return.

$$\vec{w}' \circ \vec{\mu} = (\vec{w} + \kappa\vec{n}) \circ \vec{\mu} = \vec{w} \circ \vec{\mu} + \kappa(\vec{n} \circ \vec{\mu})$$

If $\vec{n} \circ \vec{\mu}$ is positive, then we can make the expected rate of return as high as we want, by choosing a very large κ. Likewise, if $\vec{n} \circ \vec{\mu}$ is negative, then we can still make the expected rate of return as high as we want, by choosing a very negative κ. Recall, we assumed $\vec{n} \circ \vec{\mu} \neq 0$. So we can, for sure, make the expected rate of return arbitrarily high. Of course, the idea of a portfolio whose expected rate of return can be made arbitrarily high is economically absurd, even if it is mathematically valid. Ω

Lemma 6 *Suppose a vector $\vec{n}$ exists in the nullspace of M, with $\sum_{i=1}^{i=n} n_i = 0$ and $\vec{n} \circ \vec{\mu} = 0$, but $\vec{n} \neq 0$. Then one asset can be written as a linear combination of the $n-1$ other assets.*

Proof: While $\vec{n}$ does not represent a portfolio, because $\sum_{i=1}^{i=n} n_i = 0 \neq 1$, we can still calculate the variance of investing according to the vector $\vec{n}$. This represents purchasing n_i shares in Stock i, for example.

$$\sigma_{\mathrm{p}}^2 = \vec{n}^T M \vec{n} = \vec{n}^T \vec{0} = 0$$

The variance is zero. Next, the expected rate of return is $\vec{n} \circ \vec{\mu} = 0$, by assumption. However, since the variance is zero, that means that the rate of return is constantly precisely zero, and not merely a random value that has zero as its expected value. Recall that S_i is the random variable representing the rate of return for Stock i. We now have

$$n_1 S_1 + n_2 S_2 + n_3 S_3 + \cdots + n_n S_n = 0$$

Because $\vec{n} \neq \vec{0}$, there is some entry of $\vec{n}$ which is non-zero. Call that entry n_a. We can write

$$S_a = -\sum_{i=1, i\neq a}^{i=n} \frac{n_i}{n_a} S_i \tag{7}$$

and this means that Stock a is really a linear combination of the other $n-1$ stocks. One could think of it as a mutual fund, composed of a portfolio made up of those $n-1$ stocks, with weights computable from (but not identical to) the coefficients of Equation 7. However, it would have to be an idealized mutual fund, with no inefficiency due to fees, transaction costs, or the salaries of the fund employees, and no time lag. This is economically absurd. Ω

Some further notes on the S_a at the end of the last proof are worthwhile. Even if such an S_a could exist, one can prove that any objectives that can be accomplished using it can be accomplished without it. Regardless of the objective, suppose that a feasible portfolio with weights $\vec{w}$ achieve that objective. Replacing each share of S_a with shares of the other $n-1$ stocks, according to Equation 7 would produce a new portfolio, with identical expected rate of return and identical variance, but with zero shares of Stock a. Pragmatically, this means that one can delete Stock a from the problem, and start over from the very beginning, without any loss of optimality.

Theorem 3 *If M is the covariance matrix of the expected rates of return of risk-bearing assets, and yet singular, then one of the three economically absurd cases (listed above) will occur.*

Proof: If M is singular, then there is some vector $\vec{n}$ in the nullspace of M with $\vec{n} \neq 0$. If $\sum_{i=1}^{i=n} n_i \neq 0$, then Lemma 4 provides the absurd situation. If $\sum_{i=1}^{i=n} n_i = 0$ and $\vec{n} \circ \vec{\mu} \neq 0$, then Lemma 5 provides the absurd situation. Finally, if both $\sum_{i=1}^{i=n} n_i = 0$ and $\vec{n} \circ \vec{\mu} = 0$, then Lemma 6 provides the absurd situation. Ω

For this reason, the author believes that M would always be non-singular in practice, with stocks. Currency trading or options trading might be more complicated.

5: Modeling with the Logarithms of the Rates of Return

Recall that S_i is the rate of return of Stock i, taken as a random variable. Let's assume that the S_is are log-normally distributed, meaning that $S_i = e^{X_i}$ where X_i is Gaussian (normally distributed). From historical data, we can compute covariances, $M'_{ij} = \mathrm{Cov}(X_i, X_j)$. We will retain the definition that $\mu_i = E[S_i]$, the expected value of S_i, not of X_i.

Lemma 7 *If $S_i = e^{X_i}$ and $S_j = e^{X_j}$ are log-normally distributed random variables, meaning that X_i and X_j are Gaussian, then*

$$Cov(S_i, S_j) = E[S_i]E[S_j]\left(\exp(Cov(X_i, X_j) - 1\right)$$

where $E[S_i]$ is the expected value of S_i and $E[S_j]$ is the expected value of S_j.

Proof: The expected value of a log-normal random variable $S_i = e^{X_i}$ is given by

$$E[S_i] = \exp\left(\zeta_i + \frac{1}{2}\sigma_i^2\right)$$

where ζ_i is the expected value of X_i, and σ_i is the standard deviation of X_i. The covariance of two log-normal random variables $S_i = e^{X_i}$ and $S_j = e^{X_j}$ with correlation ρ_{ij}, as taken from [14, Ch. 5, Eq 44.35], is the first equation below.

$$\begin{aligned}
\text{Cov}(S_i, S_j) &= \left(\exp\left(\rho_{ij}\sigma_i\sigma_j\right) - 1\right)\exp\left(\zeta_i + \zeta_j + \frac{1}{2}(\sigma_i^2 + \sigma_j^2)\right) \\
\text{Cov}(S_i, S_j) &= \left(\exp\left(\text{Cov}(X_i, X_j)\right) - 1\right)\exp\left(\zeta_i + \frac{1}{2}\sigma_i^2 + \zeta_j + \frac{1}{2}\sigma_j^2)\right) \\
\text{Cov}(S_i, S_j) &= \left(\exp\left(\text{Cov}(X_i, X_j)\right) - 1\right)\exp\left(\zeta_i + \frac{1}{2}\sigma_i^2\right)\exp\left(\zeta_j + \frac{1}{2}\sigma_j^2)\right) \\
\text{Cov}(S_i, S_j) &= \left(\exp\left(\text{Cov}(X_i, X_j)\right) - 1\right)E[S_i]E[S_j]
\end{aligned}$$

Ω

Next, we should use this formula for the covariance to find the variance of the entire portfolio. The first part of Lemma 2 is unchanged, because it uses no facts about the Gaussian distribution. We can reuse

$$\sigma_{\text{p}}^2 = \text{Var}\left(\sum_{i=1}^{i=n} w_i S_i\right) = \text{Cov}\left(\sum_{i=1}^{i=n} w_i S_i, \sum_{j=1}^{j=n} w_j S_j\right) = \sum_{i=1}^{i=n}\sum_{j=1}^{j=n} w_i w_j \text{Cov}(S_i, S_j)$$

but now things will turn temporarily a bit ugly. Substituting Lemma 7, we have

$$\sigma_{\text{p}}^2 = \sum_{i=1}^{i=n}\sum_{j=1}^{j=n} w_i w_j \text{Cov}(S_i, S_j) = \sum_{i=1}^{i=n}\sum_{j=1}^{j=n} w_i w_j E[S_i]E[S_j]\left(\exp\left(\text{Cov}(X_i, X_j)\right) - 1\right) = \sum_{i=1}^{i=n}\sum_{j=1}^{j=n} w_i w_j \mu_i \mu_j \left(\exp(M'_{ij}) - 1\right)$$

which looks a bit intimidating. Let's define a matrix Ψ, given by

$$\Psi_{ij} = \mu_i \mu_j \left(\exp(M'_{ij}) - 1\right)$$

and substitute this to obtain

$$\sigma_{\text{p}}^2 = \sum_{i=1}^{i=n}\sum_{j=1}^{j=n} w_i w_j \Psi_{ij} = \vec{w}^T \Psi \vec{w}$$

Amazingly, the formula for the variance is unchanged, except we replace M with Ψ, but no other changes.

Since this author's proof of Theorem 1 (as given in Appendix 1: "A Full Derivation of the Classical Markowitz Optimal Portfolio Theorem") was computed from three functions, f, g_1, and g_2, then we should consider how each one changes.

- As we noted, $f(\vec{w}) = \vec{w}^T M \vec{w}$ becomes $f(\vec{w}) = \vec{w}^T \Psi \vec{w}$.
- Even better, $g_1(\vec{w})$ is totally unchanged, because we chose to keep $\mu_i = E[S_i]$ and not change it to X_i.
- Of course, $g_2(\vec{w})$ is totally unchanged also, because we merely require the w_is to sum to 1.

In short, the only change to *all of the theorems in this paper* would be that M is replaced by Ψ.