OPERATIONS RESEARCH
Vol. 53, No. 4, July–August 2005, pp. 586–599
ISSN 0030-364X | EISSN 1526-5463 | 05 | 5304 | 0586

DOI 10.1287/opre.1050.0212

Portfolio Optimization with Factors, Scenarios, and Realistic Short Positions[1]

Bruce I. Jacobs, Kenneth N. Levy
Jacobs Levy Equity Management, 100 Campus Drive, P.O. Box 650, Florham Park, New Jersey 07932-0650
{bruce.jacobs@jlem.com, ken.levy@jlem.com}

Harry M. Markowitz
Harry Markowitz Company, 1010 Turquoise Street, Suite 245, San Diego, California 92109, harryhmm@aol.com

This paper presents fast algorithms for calculating mean-variance efficient frontiers when the investor can sell securities short as well as buy long, and when a factor and/or scenario model of covariance is assumed. Currently, fast algorithms for factor, scenario, or mixed (factor and scenario) models exist, but (except for a special case of the results reported here) apply only to portfolios of long positions. Factor and scenario models are used widely in applied portfolio analysis, and short sales have been used increasingly as part of large institutional portfolios. Generally, the critical line algorithm (CLA) traces out mean-variance efficient sets when the investor's choice is subject to any system of linear equality or inequality constraints. Versions of CLA that take advantage of factor and/or scenario models of covariance gain speed by greatly simplifying the equations for segments of the efficient set. These same algorithms can be used, unchanged, for the long-short portfolio selection problem provided a certain condition on the constraint set holds. This condition usually holds in practice.

Subject classifications: finance, portfolio: optimization with short sales.
Area of review: Financial Engineering.
History: Received April 2002; revisions received June 2003, May 2004; accepted July 2004.

1. Introduction

This paper presents fast methods for computing the set of "mean-variance efficient" portfolios for an investor who can sell securities short as well as buy them long, provided that certain conditions are satisfied. One might think that ever-faster computers obviate the need for such fast algorithms. However, analyses with large numbers of securities, users waiting for answers in real time, Monte Carlo simulation runs that require many reoptimizations, and simulation experiments requiring many simulation runs, make speedy computation of efficient frontiers still prized. (Parkinson's Law continues to outpace Moore's law.)

A *feasible* portfolio is one that meets specified constraints. A mean-variance *efficient* portfolio is one that provides minimum variance among feasible portfolios with a given (or greater) expected return, and maximum expected return for given (or less) variance. The expected return and variance provided by an efficient portfolio is called an efficient mean-variance (EV) combination. The set of all efficient EV combinations is called the *efficient frontier*.

The critical line algorithm (CLA) traces out a piecewise linear set of efficient portfolios that provide the efficient frontier, subject to any system of linear equality or weak inequality constraints. In general, the inputs to the CLA are constraint parameters, the means and variances of securities, and the covariances between pairs of securities.

The CLA is especially fast if the covariances between securities are described by a "factor model." A factor model assumes that the return on a security depends linearly on the movement of one or more factors common to many securities (e.g., a general market factor, industry factors, a flight-to-quality factor) plus the security's independent "idiosyncratic" term. The use of a factor model not only accelerates computation, it also reduces input requirements. Furthermore, factor model inputs (including regression coefficients of security returns against factors, and variances of underlying factors) are more easily understood, and more easily adjusted to reflect changing conditions, than are the coefficients of a full covariance matrix.

In fast efficient-set algorithms using factor models, "fictitious securities" are introduced into the model, one for each common factor (see Sharpe 1963, Cohen and Pogue 1967). The "amount invested" in each fictitious security is constrained to be a linear combination of the investments in the real securities. With the model thus augmented, the covariance matrix becomes diagonal, or nearly so, and the equations for the pieces of the efficient set become much easier to solve.

Scenario models provide an alternative to factor models for describing the relationships among security returns. A scenario model enumerates different scenarios that can occur in the future and estimates the mean and variance of each security's return under each scenario. Fast efficient-set algorithms using scenario models are similar to those using factor models. Fast algorithms (albeit not quite as fast) also exist that combine factor and scenario models of covariance.

Fast computational methods are also available for covariances computed from historical returns with many more securities than observations. Applicable cases encountered in practice include ones with thousands of securities, but only dozens of months or hundreds of days worth of observations.

This paper presents fast algorithms for tracing out efficient sets when factor, scenario, or certain historical models are assumed, and when the investor is allowed to short securities.

Some capital asset pricing models (CAPMs) assume, in effect, that one can sell a security short without limit and use the proceeds to buy securities long. This is a mathematically convenient assumption for hypothetical models of the economy, but it is unrealistic. Actual constraints on long-short portfolios change over time and, at a given instant, vary from broker to broker and from client to client. Thus, the portfolio analyst charged with generating an efficient frontier for a particular investor must model the specific constraints to which that investor's choice is subject, including constraints the investor itself imposes as a matter of policy. To our knowledge all such constraints—whether imposed by regulators, brokers, or self-imposed—are expressible as linear equalities or weak inequalities and therefore can be incorporated into the general portfolio selection model. Later, we will give examples of current real-world constraints, but our results are not restricted to some particular constraint set.

A portfolio optimization with n securities, which can be bought long or sold short, may be set up as a model with n variables representing long positions and another n variables representing short positions. The types of constraints noted in the preceding paragraph are easily expressed in terms of the $2n$ variables. However, even if a factor or scenario model holds for the n securities held long, it does not hold for the $2n$-variable model representing short and long positions. Specifically, the $2n$-variable long-short model violates the assumption that the idiosyncratic terms are uncorrelated. Nevertheless, under certain assumptions, if the requisite information (e.g., regression coefficients and idiosyncratic variances for the factor model for the $2n$ variables) is fed into the appropriate factor or scenario program, a correct efficient frontier results.

The principal result of this paper is a sufficient condition that assures that an existing (originally long-only) factor or scenario code will compute the correct answer to the long-short problem. We refer to this condition as "Property P." Property P does not hold in general for an arbitrary long-short portfolio selection model, but it appears to be widely satisfied in practice. When a factor or scenario model of covariance is assumed and Property P is satisfied, a fast algorithm for the long-short model is readily at hand. No new programming is needed. The long-only program produces the correct answer to the $2n$-variable long-short problem, despite the "error" in assumption. Also, the fast algorithm for historical covariance matrices (when the number of securities greatly exceeds the number of observations) produces correct answers to the $2n$-variable long-short problem, whether or not Property P holds.

The results reported in this paper generalize a result due to Alexander (1993) and Kwan (1995). Their results apply to the Elton et al. (1976) algorithm. The Elton et al. algorithm assumes only one constraint equation—namely, a budget constraint—and makes special assumptions about the factor structure of a factor model.

Section 2 defines the "general" mean-variance problem. Section 3 summarizes its solution by CLA. Section 4 describes how the covariance matrix can be (almost) diagonalized if a factor, scenario, or historical model of covariance is used. Section 5 outlines short sales in the real world. Section 6 presents notation for portfolio optimization with short sales and a diagonizable model of covariance. Section 7 derives fast methods for solving the latter problem. Section 8 illustrates the results. Section 9 summarizes.

2. The General Mean-Variance Problem

Suppose that the return R_P on the portfolio over some forthcoming period is a weighted sum of the n security returns $R=[r_1, r_2, \ldots, r_n]'$,

$$R_P = R'X, \tag{1}$$

where the weights $X=[X_1, \ldots, X_n]'$ are chosen by the investor. Assuming that the r_i are random variables with finite means and variances,

$$E_P = \sum_{i=1}^{n} \mu_i X_i = \mu' X, \tag{2}$$

$$V_P = \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_{ij} X_i X_j = X'CX, \tag{3}$$

where E_P and V_P are the expected return and variance of the portfolio, $\mu=[\mu_1, \ldots, \mu_n]'$ are the expected returns on the n securities, σ_{ij} is the covariance between r_i and r_j, and C is the covariance matrix (σ_{ij}). Markowitz (1959) assumes that X is chosen subject to the following constraints:

$$AX = b, \tag{4}$$

$$X \geqslant 0, \tag{5}$$

where A is an $m \times n$ constraint matrix and b an m component "right-hand side" vector.

As in linear programming, constraints (4) and (5) can represent weak linear inequalities ($\geqslant$ or $\leqslant$) by use of slack variables. For example,

$$\sum_j a_{ij} X_j \leqslant b \tag{6}$$

is written as

$$\sum_j a_{ij} X_j + X_s = b, \quad X_s \geqslant 0. \tag{7}$$

Also, a variable X_i not required to be nonnegative is handled in (4) and (5) by substituting for it

$$X_i = X_{ip} - X_{in}, \quad X_{ip} \geqslant 0,\ X_{in} \geqslant 0, \tag{8}$$

where X_{ip} and X_{in} are the positive and negative parts of X_i.

It is not required that the covariance matrix C in (3) be nonsingular. This is essential, because X may include risk-free securities, slack variables, and pairs of securities representing short and long positions. Also, sometimes C is estimated from historical returns with less periods than there are securities. Any of these circumstances will result in $\det(C) = 0$. In addition, it is desirable for the computational procedure not to fail if A in (4) is not of full rank.

A portfolio X is said to be *feasible* if it satisfies constraints (4) and (5). A pair of real numbers (E_P, V_P) is said to be a feasible EV combination if E_P and V_P satisfy (2) and (3) for some feasible portfolio X. A feasible (E_P, V_P) pair is *inefficient* if some other feasible pair (E_P^*, V_P^*) dominates it; that is, has higher expected return, $E_P^* > E_P$, but no higher variance, $V_P^* \leqslant V_P$; or, has lower variance, $V_P^* < V_P$, but no lower expected return, $E_P^* \geqslant E_P$. If (E_P, V_P) is not thus dominated, it is called an *efficient* EV combination. A feasible portfolio X is efficient or inefficient according to whether its (E_P, V_P) is efficient or inefficient.

The general (single-period) mean-variance portfolio selection problem is to find all efficient EV combinations, and feasible portfolios that yield these, for all possible A, b, μ, and C in (2), (3), (4), and (5). Problems with weak linear inequalities and variables not required to be nonnegative can be converted into this form.

3. Solution to the General Problem

It is possible that, for a given A and b, the model is infeasible, that is, no portfolio X satisfies (4) and (5). It is also possible for a model to be feasible and yet have no mean-variance efficient portfolios. In this case, if $\widehat{X}$ is feasible with minimum V_P and with expected return E, there is another feasible portfolio X^* with the same V and with $E^* > E$. This can occur if C is singular and the constraint set unbounded. Below, we assume that the model is feasible and has efficient portfolios.

Next, we summarize (without proof) certain properties and formulas of efficient sets.[2] The set of efficient EV combinations is piecewise parabolic. In general, there may be more than one efficient portfolio X for a given efficient EV combination. When the set of efficient portfolios is unique—with only one feasible portfolio X for any given efficient EV combination—the set of efficient portfolios is piecewise linear. The formula for an efficient segment (of the piecewise linear efficient set) is given below. When the set of efficient portfolios is not unique there is nevertheless a "complete, nonredundant" set of efficient portfolios that satisfy the equations below. By "complete, nonredundant" we mean a set of efficient portfolios with one and only one X for each efficient EV combination. The CLA provides such a complete, nonredundant set of efficient portfolios whether or not the set of efficient portfolios is unique.

The Lagrangian expression for the general model is

$$L = V/2 + \sum_{k=1}^{m} \lambda_k \left(\sum_{j=1}^{n} a_{kj} X_j \right) - \lambda_E \sum_{i=1}^{n} \mu_i X_i. \tag{9}$$

Let

$$\eta = \frac{\partial L}{\partial X} = [C \quad A' \quad \mu] \begin{bmatrix} X \\ \lambda \\ -\lambda_E \end{bmatrix}, \tag{10}$$

where $\lambda = [\lambda_1, \ldots, \lambda_m]'$. For the moment, to develop a definition, arbitrarily select a nonempty subset of $\{1, 2, \ldots, n\}$ and designate this subset as the IN variables, and its complement as the OUT variables. Let

$$M = \begin{bmatrix} C & A' \\ A & O \end{bmatrix} \tag{11}$$

and let M_{IN} be the M matrix with the rows and columns that correspond to OUT variables deleted. Similarly, let μ_{IN} and X_{IN} be the μ and X vectors with OUT components deleted, and 0_{IN} be a zero vector of the same size as μ_{IN}. If M_{IN} is nonsingular, we say that the arbitrarily chosen IN set has an associated *critical line* satisfying

$$X_i = 0 \quad \text{for } i \in \mathrm{OUT} \tag{12}$$

and

$$M_{\mathrm{IN}} \begin{bmatrix} X_{\mathrm{IN}} \\ \lambda \end{bmatrix} = \begin{bmatrix} 0_{\mathrm{IN}} \\ b \end{bmatrix} + \begin{bmatrix} \mu_{\mathrm{IN}} \\ 0 \end{bmatrix} \lambda_E. \tag{13}$$

Multiplying through by M_{IN}^{-1} solves (13) for X_{IN} and λ as linear functions of λ_E:

$$\begin{bmatrix} X_{\mathrm{IN}} \\ \lambda \end{bmatrix} = \alpha_{\mathrm{IN}} + \beta_{\mathrm{IN}} \lambda_E. \tag{14}$$

If we substitute (14) into (10), we find that the η vector is also a linear function of λ_E:

$$\eta = \gamma_{\mathrm{IN}} + \delta_{\mathrm{IN}} \lambda_E. \tag{15}$$

Conditions (13) imply

$$\eta_i = 0 \quad \text{for } i \in \mathrm{IN}. \tag{16}$$

In light of (12) and (16), if a point on the critical line also satisfies

$$X_i \geqslant 0 \quad \text{for } i \in \mathrm{IN}, \tag{17}$$

$$\eta_i \geqslant 0 \quad \text{for } i \in \mathrm{OUT}, \tag{18}$$

$$\lambda_E > 0, \tag{19}$$

then the point is efficient, by the Kuhn-Tucker theorem. If any point on the critical line is efficient, then there will be an interval of that line (possibly open-ended) all of whose points are efficient. We refer to such an interval as an *efficient segment*.

Because there are $2^n - 1$ nonnull subsets of $\{1, \ldots, n\}$, it is impractical to enumerate them all, determine which have nonsingular M_{IN}, among these determine which contain efficient segments, then piece these together to form a complete, nonredundant set of efficient portfolios. The CLA produces such a set without searching among irrelevant IN sets.

The CLA proceeds as follows.[3] It traces out the efficient set from high to low λ_E. At a typical step, we have in hand a critical line with an efficient segment, and with IN-set IN_t; we also have in hand the corresponding M_{IN} and M_{IN}^{-1}. We can then solve for α_{IN}, β_{IN}, γ_{IN}, and δ_{IN}, from which it is easy to determine which occurs first as λ_E is reduced:

$$\text{an } X_i \downarrow 0 \quad \text{for } i \text{ IN}, \tag{20}$$

$$\text{an } \eta_i \downarrow 0 \quad \text{for } i \text{ OUT}, \tag{21}$$

$$\text{or } \lambda_E \downarrow 0. \tag{22}$$

In case $\lambda_E \downarrow 0$ first, we have reached the efficient portfolio with minimum feasible V, and the algorithm stops.[4]

If $X_i \downarrow 0$ first, then i moves from IN to OUT on the next ("adjacent") efficient segment. It is shown that η_i will increase on this next segment. On the other hand, if $\eta_i \downarrow 0$ first, then i moves from OUT to IN in the new IN set, IN_{t+1}, and X_i will increase on the new segment.[5] If the algorithm has not stopped, because $\lambda_E \downarrow 0$ has not been reached, the new M matrix, $M_{\text{IN}(t+1)}$, is nonsingular. It is obtained from the old by adding or deleting one column and the corresponding row. This allows us to update M^{-1} relatively inexpensively, and use it to solve for α, β, γ, δ, etc., as before. The algorithm ends, with $\lambda_E \downarrow 0$, in a finite number of iterations.[6]

4. Diagonizable Models of Covariance

Factor Models. In the introduction, we referred to "fast algorithms" based on certain models of covariance. In this section, we summarize such algorithms for the factor and scenario models and for models with historical covariance matrices when there are more securities than observations. In problems with a large number of securities, computation time may differ by orders of magnitude between using a dense covariance matrix and using the diagonal or nearly diagonal covariance matrices permitted by the aforementioned models of covariance.

For the present, we are concerned with portfolios of long positions, which we denote as

$$X^{\mathbb{L}} = \left[X_1, \ldots, X_\nu, \ldots, X_n\right]'.$$

We write its constraints as

$$A^{\mathbb{L}} X^{\mathbb{L}} = b^{\mathbb{L}}, \tag{23}$$

$$X^{\mathbb{L}} \geqslant 0. \tag{24}$$

The portfolio may include zero-variance "securities" such as cash or dummy variables. We assume that

$$\begin{aligned} V_i > 0 \quad & \text{for } i \in [1, \nu], \\ V_i = 0 \quad & \text{for } i \in [\nu+1, n]. \end{aligned} \tag{25}$$

If $\nu = n$, then $[\nu+1, n]$ is empty.

A factor model of covariance assumes that security returns are related to each other because they are related to common underlying factors. Specifically, it assumes that

$$r_i = \alpha_i + \sum_{k=1}^{K} \beta_{ik} f_k + u_i, \quad i = 1, \ldots, n, \tag{26}$$

where K is the number of common factors, f_k is the kth common factor, and u_i is an idiosyncratic term assumed uncorrelated with f_k, $k = 1, \ldots, K$, and all u_j for $i \neq j$. In matrix notation,

$$R = \alpha + BF + U, \tag{27}$$

where $\alpha = [\alpha_1, \ldots, \alpha_n]'$, $B = [\beta_{ik}]$ is $n \times K$, $F = [f_i, \ldots, f_k]'$, and $U = [u_1, \ldots, u_n]'$. From (27) and (1), we see that

$$R_P = \alpha' X^{\mathbb{L}} + F' B' X^{\mathbb{L}} + U' X^{\mathbb{L}}. \tag{28}$$

Because F and U are uncorrelated, the above implies

$$V_P = (X^{\mathbb{L}})' B Q_f B' X^{\mathbb{L}} + (X^{\mathbb{L}})' Q_u X^{\mathbb{L}}, \tag{29}$$

where Q_f and Q_u are the covariance matrices of F and U, respectively. By assumption, Q_u is diagonal with ith diagonal term $V(u_i)$. Q_f is not necessarily diagonal.

Define K "fictitious" investments in terms of "real" investments,

$$B' X^{\mathbb{L}} - \begin{bmatrix} X_{n+1} \\ \vdots \\ X_{n+K} \end{bmatrix} = 0. \tag{30}$$

We let $X^{\mathbb{L}\mathbb{A}} = [X_1, \ldots, X_{n+K}]'$ and

$$A^{\mathbb{L}\mathbb{A}} X^{\mathbb{L}\mathbb{A}} = b^{\mathbb{L}\mathbb{A}} \tag{31}$$

be constraints (23) with (30) appended. We may write (29) as

$$V_P = (X^{\mathbb{L}\mathbb{A}})' C^{\mathbb{L}\mathbb{A}} X^{\mathbb{L}\mathbb{A}}, \tag{32}$$

where

$$C^{\mathbb{LA}} = \begin{bmatrix} Q_u & 0 \\ 0 & Q_f \end{bmatrix}.$$

The original problem may be restated as finding EV-efficient $X^{\mathbb{LA}}$ subject to (31) and (24) with portfolio variance defined as in (32). The M-matrix now is

$$M^{\mathbb{LA}} = \begin{bmatrix} C^{\mathbb{LA}} & (A^{\mathbb{LA}})' \\ A^{\mathbb{LA}} & 0 \end{bmatrix}. \tag{33}$$

Because $X_i \geqslant 0$ is not required for the fictitious securities, $i \in [n+1, n+K]$, it is convenient to permit $X_i < 0$ for these variables (rather than separate them into positive and negative parts as in (8)). Then, for $i > n$, X_i is IN on all critical lines. We refer to the portfolio selection model with constraints (31) and covariance matrix (32) as the "diagonalized version" of the factor model. (Strictly speaking, we mean "almost diagonalized" because Q_f is not necessarily diagonal.) It is assumed that all risky securities, $i \in [1, \nu]$, have positive idiosyncratic risk:

$$V(u_i) > 0, \quad i \in [1, \nu]. \tag{34}$$

Typically, $\nu \gg n + K - \nu$; therefore $M^{\mathbb{LA}}$ is quite sparse and well structured. This is the basis for fast CLAs for factor models.[7] If Q_f is known to be diagonal, the algorithm can be further streamlined.

Scenario Models. A scenario model analyzed by Markowitz and Perold (1981a, b) assumes that one of S mutually exclusive scenarios will occur with probability P_s, $s = 1, \ldots, S$. If scenario s occurs, then the return on the ith security is

$$r_i = \mu_{is} + u_{is}, \tag{35}$$

where $\mathrm{E}(u_{is}) = \mathrm{cov}(u_{is}, u_{js}) = 0$ for $i \neq j$. Let $V_{is} = \mathrm{E}(u_{is}^2 \mid s)$. The expected return E of the portfolio is still given by (2), provided that the μ in (2) are computed as follows:

$$\mu_i = \sum_{s=1}^{S} P_s \mu_{is}. \tag{36}$$

These μ_i can be computed in advance of the optimization calculation. Let

$$X_{n+s} = \sum_{i=1}^{n} X_i(\mu_{is} - \mu_i) \quad \forall s \in [1, S]. \tag{37}$$

This equals the expected value E_s of the portfolio, given that scenario s occurs, less portfolio grand mean E. The variance of the portfolio is

$$V_P = \mathrm{E}(R_P - E_P)^2 = \sum_{s=1}^{S} P_s \mathrm{E}(R_P - E_s + E_s - E)^2$$
$$= \sum_{i=1}^{n+S} X_i^2 \widetilde{V}_i, \tag{38}$$

where

$$\widetilde{V}_i = \sum_{s=1}^{S} P_s V_{is}, \quad i = 1, \ldots, n,$$

$$\widetilde{V}_{n+s} = P_s, \quad s = 1, \ldots, S.$$

Thus, V_P can be expressed as a positively weighted sum of squares in the n original variables and S new, fictitious variables that are linearly related to the original variables by (37).

Apart from notation (e.g., using S for K and (37) for (30)), the scenario model is formally the same as the factor model with Q_f diagonal. That is, the meanings of the coefficients are different but, with change of notation, the portfolio selection problem with a scenario model of covariance has an $M^{\mathbb{LA}}$ matrix as in (33), with diagonal Q_f. We refer to the portfolio selection problem with constraints (37) appended to the given constraints, and variance expressed as in (38), as the diagonalized version of the scenario model (35).[8]

Historical Covariance Matrices. Consider the case in which T historical periods (e.g., months or days) are used to estimate covariances among n securities. Let

$$X_{n+t} = \sum_{i=1}^{n} X_i(r_{it} - m_i), \quad t = 1, \ldots, T, \tag{39}$$

where r_{it} is the return on the ith security during period t, and m_i is the ith security's historical average return:

$$m_i = \frac{1}{T} \sum_{t=1}^{T} r_{it}.$$

The m_i do not necessarily equal the estimated expected return μ_1 in (2). Then, X_{n+t} is the difference between portfolio return in the tth period and the portfolio's average return. Therefore, the historical variance of a portfolio is a constant times

$$V_P = \sum_{t=1}^{T} X_{n+t}^2. \tag{40}$$

This is a sum of squares in new, fictitious securities that are linearly related to the old. Once again, the problem can be expressed as a portfolio selection problem with $M^{\mathbb{LA}}$ matrix as in (33). In the present case, we have

$$Q_u = 0. \tag{41}$$

$M^{\mathbb{LA}}$ is again sparse and well structured but, because of (41), requires different handling than in the fast algorithms for the factor and scenario models.[9]

We refer to the portfolio selection model with constraints (39) appended and variance expressed as in (40) as the diagonalized version of the historical covariance model. We refer to the three models described in this section as "diagonalizable" models. For large problems, the above models afford a reduction in computation requirements roughly proportional to the reduction in the number of nonzero entries between $M^{\mathbb{L}}$ and $M^{\mathbb{LA}}$.

5. Short Sales in Practice

Capital asset pricing models (CAPMs) frequently assume that the investor chooses a portfolio subject only to the constraint

$$\sum_{i=1}^{n} X_i = 1, \tag{42}$$

without constraint on the sign of X_i. Negative X_i are interpreted as short positions. In particular, (42) permits $(x, 1-x, 0, \ldots, 0)$ as feasible for all real x. For example, (42) would permit an investor to deposit $1,000 with her broker, short $1,000,000 of Stock A, and use the proceeds plus the original deposit to purchase $1,001,000 of Stock B. This is not how short positions work in fact.

No single constraint set applies to all long-short investors. The portfolio analyst must model the specific constraint set for the particular client. To illustrate what this may involve, we outline a few real-world short-sale constraints (see also Jacobs and Levy 2000).

To sell short for any customer, a broker must borrow the stock to be sold, and actually sell it. The brokerage firm may borrow the stock from itself, typically from customer stock held "in street name" in margin accounts. Alternatively, the broker may borrow the stock from another investor, typically a large institutional investor. Some intermediary may facilitate the process of bringing together demand and supply of stock-to-lend. Sometimes a lender cannot be found for the desired stock. In this case, the stock cannot be sold short. Furthermore, the lender retains the right to call back the stock; if he does, and another lender is not promptly available, the investor must cover the short position (i.e., buy back the stock) and deliver it to the lender.

The proceeds of a short sale are used as collateral for the lender of the stock. In fact, if the stock is borrowed from another investor, the broker must put up more than 100% of the proceeds of the sale as collateral, usually about 105%. (Note that this is required of the broker to protect the stock lender, as opposed to the requirement on the short seller discussed in the next paragraph.) The proceeds of the stock sale are invested in "cash instruments" such as short-term Treasury bills. The broker and the stock lender retain a portion of the interest earned on the proceeds. A large institutional investor that shorts stock typically receives a portion of the interest (referred to as a "short rebate"). A small retail customer who sells short typically receives no part of the interest.

The short seller is subject to Regulation (Reg) T. Reg T covers common stock, convertible bonds, and equity mutual funds; securities such as U.S. Treasury bonds or bond funds and municipal bonds or bond funds are exempt from Reg T. Reg T requires that the sum of the long positions plus the sum of the (absolute value of) short positions must not exceed twice the equity in the account. If we normalize so that "1" represents the equity in the account, then Reg T requires

$$\sum_{i=1}^{2n} X_i \leqslant H, \tag{43}$$

where X_i represents a long position for $i \in [1, n]$, a short position for $i \in [n+1, 2n]$, and currently Reg T specifies $H = 2$. This inequality, of course, can be converted to an equality by introduction of a slack variable.

As a matter of policy, the broker or investor may set H at a lower level. There may be additional constraints on the choice of

$$X^{\mathbb{LS}} = [X_1, \ldots, X_{2n}]'. \tag{44}$$

For example, some securities are hard to borrow. The broker may therefore limit the amount of the short position or not permit short positions in the particular security.

Constraint (43) with $H = 2$ is referred to as a "50% margin requirement" on both short and long positions. In practice, the nature of this margin requirement is different for short and long positions. In the case of long positions, the customer may borrow as much as 50% of the value of the position from the broker. In the case of a short position, the customer does not borrow money from the broker; the margin requirement is a collateral requirement. Furthermore, the Reg T requirements are for "initial margin"—the equity required in the account to establish initial positions. It does not constrain the value of the positions maintained after they are established. However, there are "maintenance margin" requirements imposed by securities exchanges and by brokers. Consequently, one motive of the investor in setting her or his own H in (43) is to reduce the probability of needing additional cash for maintenance margin.[10]

Reg T can be circumvented in several ways. For example, hedge funds often set up offshore accounts, which are not subject to Reg T. Alternatively, a large hedge fund can set up as a broker-dealer, with a "real" broker-dealer acting as the "back office." In this case, the hedge fund, as broker-dealer, is subject to broker-dealer capital requirements rather than Reg T requirements. This permits much more leverage than Reg T. In the extreme, the only constraint is what the broker imposes on the hedge fund's portfolio to assure that, in the case of unfavorable market movements, the broker is secure. A hedge fund could also circumvent Reg T by having a broker set up a proprietary trading account of its own that is managed by the fund. Gains and losses in the proprietary trading account are transferred to the hedge fund via prearranged swap contracts. The only constraint imposed by this arrangement is the broker's own capital requirements, plus whatever constraints the broker imposes.[11]

Also lying outside Reg T are certain arrangements that allow the investor to use noncash collateral, including existing long positions, to collateralize the shares borrowed to sell short, freeing up the proceeds from short sales to be used for further purchases and short sales. In all these cases, the broker-dealer imposes its own requirements for its own security. The portfolio analyst must model the situation as she or he finds it.[12]

6. Modeling Short Sales

We assume that the choice of $X^{\mathbb{LS}}$ is subject to some system of linear constraints in nonnegative variables

$$A^{\mathbb{LS}}X^{\mathbb{LS}} = b^{\mathbb{LS}}, \tag{45}$$

$$X^{\mathbb{LS}} \geqslant 0. \tag{46}$$

Portfolio return is

$$R_P = \sum_{i=1}^{n} r_i X_i + \sum_{i=n+1}^{2n} (-r_{i-n})X_i + r_c \sum_{i=n+1}^{2n} h_{i-n}X_i. \tag{47}$$

The first term on the right of Equation (47) represents the return contribution of the securities held long. The second term represents the contribution of the securities sold short. The third term represents the short rebate, where

$$h_i \leqslant 1, \quad i = 1, \ldots, n. \tag{48}$$

Usually, $h_i \geqslant 0$, but this condition is sometimes violated for hard to borrow stocks, and is not required for our results.[13] r_c is the return on "cash" or "collateral." Cash is also a risk-free security that can be held long, i.e., $c \in [\nu+1, n]$. In particular, we assume that $\nu < n$.[14] Let

$$R^{\mathbb{LS}} = \begin{bmatrix} \tilde{r}_1 \\ \vdots \\ \tilde{r}_{2n} \end{bmatrix} = \begin{bmatrix} R^{\mathbb{L}} \\ -R^{\mathbb{L}} + hr_c \end{bmatrix}, \tag{49}$$

$$\mu^{\mathbb{LS}} = \mathrm{E}(R^{\mathbb{LS}}) = \begin{bmatrix} \mu^{\mathbb{L}} \\ -\mu^{\mathbb{L}} + hr_c \end{bmatrix}. \tag{50}$$

The expected return and variance of the long-short portfolio are

$$E = (\mu^{\mathbb{LS}})' X^{\mathbb{LS}}, \tag{51}$$

$$V_P = (X^{\mathbb{LS}})' C^{\mathbb{LS}} X^{\mathbb{LS}}, \tag{52}$$

where

$$C^{\mathbb{LS}} = \begin{bmatrix} C^{\mathbb{L}} & -C^{\mathbb{L}} \\ -C^{\mathbb{L}} & C^{\mathbb{L}} \end{bmatrix}, \tag{53}$$

and where $C^{\mathbb{L}}$ is the long-only covariance matrix.

If we assume a multifactor model with returns r_i given by (26) and (27), then (49) implies

$$R^{\mathbb{LS}} = \begin{bmatrix} \alpha \\ -\alpha \end{bmatrix} + \begin{bmatrix} B \\ -B \end{bmatrix} F + \begin{bmatrix} U \\ -U \end{bmatrix} + \begin{bmatrix} 0_n \\ h \end{bmatrix} r_c, \tag{54}$$

where 0_n is an n-vector of zeros. Hence, the covariance matrix of $R^{\mathbb{LS}}$ is

$$C^{\mathbb{LSA}} = \begin{bmatrix} B \\ -B \end{bmatrix} Q_f \begin{bmatrix} B \\ -B \end{bmatrix}' + \begin{bmatrix} Q_u & -Q_u \\ -Q_u & Q_u \end{bmatrix}. \tag{55}$$

Thus, if we define

$$(X^{\mathbb{LSA}})' = [(X^{\mathbb{L}})', (X^{\mathbb{S}})', (X^{\mathbb{A}})'], \tag{56}$$

where

$$X^{\mathbb{A}} = [X_{2n+1}, \ldots, X_{2n+K}]'$$

are portfolio betas, then $X^{\mathbb{LSA}}$ is chosen subject to constraints (46) and

$$A^{\mathbb{LSA}} X^{\mathbb{LSA}} = b^{\mathbb{LSA}}. \tag{57}$$

The latter are constraints (45) with

$$X^{\mathbb{A}} = [B, -B] X^{\mathbb{LS}}$$

appended.

Define the vector

$$\delta = \begin{bmatrix} B \\ -B \end{bmatrix}' \begin{bmatrix} X^{\mathbb{L}} \\ X^{\mathbb{S}} \end{bmatrix} = B'X^{\mathbb{L}} - B'X^{\mathbb{S}}. \tag{58}$$

The kth entries of the vectors $B'X^{\mathbb{L}}$ and $B'X^{\mathbb{S}}$ are the contributions of the long and short portions of the portfolio, respectively, to the kth "fictitious security." Thus, δ is a vector of the differences between the contributions of the long and short portions of the portfolio. Using Equations (55), (58), and definitions, we obtain portfolio variance as

$$V_P = \delta' Q_f \delta + \begin{bmatrix} X^{\mathbb{L}} \\ X^{\mathbb{S}} \end{bmatrix}' \begin{bmatrix} Q_u & -Q_u \\ -Q_u & Q_u \end{bmatrix} \begin{bmatrix} X^{\mathbb{L}} \\ X^{\mathbb{S}} \end{bmatrix}. \tag{59}$$

Recalling that Q_u is diagonal, this can be written as

$$V_P = \delta' Q_f \delta + \sum_{i=1}^{2n} X_i^2 V_i - 2\sum_{i=1}^{n} X_i X_{n+i} V_i. \tag{60}$$

7. Solution to Long-Short Model

Equation (60) is the same as the diagonalized form (32) of the diagonalizable models of §4 *except* for the inclusion of the last sum of cross-product terms. Fortunately, for certain models the sum of cross products can be ignored. For these models, a portfolio optimizer that assumes that variance is

$$V_P'(X) = \delta^T Q_f \delta + \sum_{i=1}^{2n} X_i^2 V_i \tag{61}$$

instead of that given in Equation (60), will still produce a correct mean-variance efficient frontier.

Note that if

$$X_i X_{n+i} = 0, \quad i = 1, \ldots, n \tag{62}$$

holds (i.e., the investor is not long and short the same security), then V_P in (60) equals V_P' in (61). We shall refer to a portfolio that satisfies (62) as *trim*; otherwise it is *untrim*. We will refer to the portfolio selection model with E, V_P, and constraints given by (51), (60), (57), and (46) as the *original* model; and that with (60) replaced by (61) as the *modified* model. In this section, we consider conditions under which the efficient set for the modified model is an efficient set for the original model.

Clearly, $V_P'(X) \equiv V_P(X)$ if $V_i = 0$ for all $i \in [1, 2n]$, as is the case for the diagonalized historical model. We will denote this simple but useful result as a theorem.

Theorem 1. *An efficient set for the modified historical model provides an efficient set for the original historical model.*

Proof. See the preceding paragraph. □

Theorem 1 makes no assumption concerning the constraint set or expected returns other than the background assumptions that the model is feasible and has efficient portfolios. The other diagonalizable models of §4 require a further assumption to reach a similar conclusion. The following assumption is sufficient.

Property P. If in the original model X is a feasible portfolio with $X_i X_{n+i} > 0$ for some specific i, then there is a feasible portfolio Y with

$$\begin{aligned} &Y_i = X_i - \theta_i,\\ &Y_{n+i} = X_{n+i} - \theta_i,\\ &Y_j = X_j, \quad j \neq i, n+i,\ j \in [1, \nu] \cup [n+1, n+\nu] \end{aligned} \tag{63}$$

for $\theta_i = \min\{X_i, X_{n+i}\}$. Also, Y has the same or greater mean as X.

In other words, if X has a positive long and a positive short position in the same security, it is feasible to subtract the above θ_i from both positions, keeping all other risky securities unchanged, adjusting only zero-variance X_i, without reducing portfolio expected return. Note that $Y_i Y_{n+i} = 0$.

While Property P is not necessarily true, it does hold for a wide variety of constraint sets met in practice. Suppose, for example, that choice of a long-short portfolio is subject to any or all of the following constraints: (A) a Reg T type of constraint as in (43), perhaps with $H > 2$ for an investor not subject to Reg T; (B) upper bounds on individual long or short positions; (C) the requirement that the value of long positions be close to the value of short positions—specifically,

$$\tau \geqslant \sum_{i=1}^{\nu} X_i - \sum_{i=1}^{\nu} X_{n+i} \geqslant -\tau \tag{64}$$

for some given tolerance level τ;[15] as well as the nonnegativity requirement (46), and a budget constraint

$$\sum_{i=1}^{\nu} X_i + X_c - X_b \leqslant 1, \tag{65}$$

where X_c is a cash balance and X_b is an amount borrowed. (Note that the sum in (65) is through ν, i.e., it includes risky *long* positions only. Recall that, unlike investors in CAPMs with (42) as their only constraint, Reg T-constrained investors do not get to spend the proceeds from selling short, although they may share the interest collected on these proceeds.) If X is any feasible portfolio (i.e., meets each of the above constraints) with $X_i X_{n+i} > 0$, then Y with

$$\begin{aligned} &Y_i = X_i - \theta_i,\\ &Y_{n+i} = X_i - \theta_i,\\ &Y_c = X_c + \theta_i,\\ &Y_j = X_j \quad \text{for } j \in \big\{[1, \nu] \cup [n+1, n+\nu]\big\} \backslash \{i, n+i\},\\ &\theta_i = \min\{X_i, X_{n+i}\} \end{aligned} \tag{66}$$

meets the constraints. When the constraints are written as equalities, as in (7), then zero-variance slack variables are adjusted to maintain the equalities. Also, from (47) and (48),

$$E_Y = E_X + \theta_i (1 - h_i) r_c \geqslant E_X. \tag{67}$$

Thus, a constraint set consisting of (46), (65), and some or all of (A), (B), and (C) does satisfy Property P. Note that Property P only requires Y to be feasible, not necessarily efficient; thus we need not be concerned, in checking Property P that, say, Y might be improved by reducing X_b rather than increasing X_c in case $X_b > 0$.

On the other hand, if there is an upper bound on the holding of cash,

$$X_c \leqslant u_c, \tag{68}$$

then Property P may not be satisfied. If, for example, there are no upper bounds on the other X_i, then

$$\begin{aligned} &X_1 = 1 - u_c,\\ &X_{n+1} = 1 - u_c,\\ &X_c = u_c,\\ &X_i = 0 \quad \text{otherwise} \end{aligned} \tag{69}$$

is feasible, but X_1 and X_{n+1} cannot be reduced by adjusting zero-variance variables, including X_c in (65), in the manner required by Property P without violating (68).

Theorem 2. *If Property* P *holds in the original model, then for each efficient* (E, V_P) *combination there is one and only one trim portfolio* Y *with the same* (E, V_P).

(There may also be untrim efficient portfolios with this (E, V_P).)

Proof. Because (E, V_P) is feasible, there is a portfolio X that provides it. If X is untrim, successive transformations (63) for each i in turn with $X_i X_{n+i} > 0$ yields a trim, feasible Y with the same or greater E than X and, from Equations (58) and (59), the same V_P. If Y has greater E, then X could not be efficient, whereas if Y has the same E as X, then Y too is efficient. Thus, for any efficient (E, V_P) combination, there exists a trim feasible Y that provides it.

To show that Y is unique, let us suppose that another trim feasible (therefore efficient) portfolio Z supplies (E, V_P). Let

$$W = \xi Y + (1-\xi)Z. \tag{70}$$

If we show that V_P as a function of ξ is strictly convex, then $(1/2)Y + (1/2)Z$ is feasible (because the constraint set is convex), has the same E (because E is linear), and less V_P than Y or Z, contradicting the assumption that Y and Z are efficient. To see that V_P is a strictly convex function of ξ, first confirm that the first term on the right-hand side of (60) is constant or a convex function of ξ (because Q_f is positive semidefinite and δ is linear in ξ). Next, note that the last two terms of Equation (60) may be obtained by substituting $\widetilde{X}_i = X_i - X_{n+i}$ into $\sum_{i=1}^{n} V_i \widetilde{X}_i^2$. As a function of ξ, this is a sum of terms that are either constant (i.e., for those i with $\widetilde{Y}_i = \widetilde{Z}_i$, which implies $Y_i = Z_i$ because Y and Z are trim) or strictly convex. It follows that V_P is a strictly convex function of ξ provided $Y \neq Z$. □

From (60) and (61), we see that

$$V_P' - V_P = 2\sum_{i=1}^{n} X_i X_{n+i} V_i. \tag{71}$$

Thus, $V_P' = V_P$ for trim portfolios, and $V_P' > V_P$ for untrim ones.

Theorem 3. *If Property* P *holds in the original model, then the modified model has the same set of efficient* (E, V_P) *combinations as does the original model. Also, it has a unique set of efficient portfolios* (*one for each efficient* (E, V_P) *combination*) *that is the same as the unique set of trim efficient portfolios in the original model.*

Proof. First, we show that all efficient (E, V_P) combinations in the original model, and all trim efficient portfolios in the original model, are efficient (E, V_P') combinations and portfolios for the modified model. Then, we show that no additional portfolios or (E, V_P') combinations are efficient for the latter model. Because $V_P = V_P'$ for trim portfolios, and each efficient (E, V_P) combination in the original model can be supplied by a trim portfolio X, each efficient (E, V_P) combination of the original model is feasible in the modified model. It will also be efficient in the modified model unless some other feasible portfolio Y dominates it in that model (i.e., has greater E for the same or less V_P', or less V_P' for the same or greater E). Because the models have the same feasible sets and expected returns, and because $V_P(Y) \leqslant V_P'(Y)$, $V_P'(X) = V_P(X)$, if Y dominated X in the modified model (e.g., with $\mathrm{E}(Y) \geqslant \mathrm{E}(X)$ and $V_P'(Y) < V_P'(X)$), then it would also dominate it in the original model, contradicting the hypothesis that X is efficient in the original model. Thus, all trim portfolios that are efficient in the original model are efficient in the modified model.

We now show that no other portfolios are efficient in the modified model. If the constraint set is bounded, therefore compact as well as closed, then the efficient (E, V_P) combinations of the original model span a closed interval $[\underline{E}, \bar{E}]$ of expected returns, where $\bar{E}$ is the maximum feasible expected return and $\underline{E}$ is the expected return of the efficient portfolio with minimum V_P. According to Theorem 2, if X is a trim, efficient portfolio and Y is another efficient portfolio with the same (E, V_P), then Y is untrim. Therefore, (71) implies $V_P'(Y) > V_P(Y) = V_P'(X)$. Thus, Y is not efficient in the modified model. This, plus the fact that $V_P' = V_P$ for trim portfolios, and the uniqueness statement in Theorem 2, implies that the efficient set for the modified model is unique for $E \in [\underline{E}, \bar{E}]$. Nor can the modified model have an efficient portfolio with E outside $[\underline{E}, \bar{E}]$, for then the modified model will either have a feasible portfolio with greater E than $\bar{E}$, which is impossible because the two models have the same feasible sets and expected returns, or have smaller V than the minimum feasible V in the original model, which is impossible because $V_P' \geqslant V_P$.

If E is not bounded above, then the preceding argument applies except that it is unnecessary to check for an efficient portfolio in the modified model with $E > \bar{E}$. □

Theorem 3 assures us that we can naively use a factor or scenario portfolio optimizer, ignoring the negative correlation between u_i and u_{n+i}, and get a correct answer to the long-short portfolio selection problem when Property P holds. This is not necessarily the case if Property P does not hold. For example, consider any diagonalized model with a Reg T constraint (with $H = 2$), a budget constraint (65), and an upper bound ($u_c < 1.0$) on cash. Assume that $V_i > 0$ for all $i \in [1, \nu]$. In the original model, consider the portfolio with

$$X_1 = X_{n+1} = 1, \qquad X_i = 0 \quad \text{otherwise}.$$

This portfolio is feasible and has zero variance. Thus, zero variance is feasible; therefore, some portfolio (not necessarily the above portfolio) has zero variance and is efficient. However, the modified version of this model has no feasible zero-variance portfolios: The upper bound on cash implies that $X_i > 0$ for some $i \in [1, \nu]$, which implies $V_P' > 0$, because $\mathrm{cov}(r_i, r_j) = 0$ for risky securities in the modified model. Thus, absent some assumption such as Property P, it is possible that an efficient set for the modified model may not be an efficient set for the original model.

8. Example

Tables 1 through 8 illustrate the content and purpose of the theorems of the preceding section. We consider a three-security, one-factor model subject only to Reg T, the budget constraint, and nonnegativity constraints. In this case, (26) may be written as

$$r_i = \alpha_i + \beta_i f + u_i, \quad i = 1, 2, 3. \tag{72}$$

Table 1 presents inputs to such a model for three hypothetical securities. In all the tables, long positions in the three securities are labeled 1L, 2L, and 3L. Table 1 shows for each of these long positions, the expected return μ_i, beta β_i, idiosyncratic variance $V_i = \text{var}(u_i)$, and rebate fraction h_i. The latter is needed to compute the expected return of the corresponding short position. Table 1 also shows the lending rate, the borrowing rate, and the variance of the underlying factor f.

The betas of the securities, their idiosyncratic variances, and the variance of the underlying factor could be used to compute the covariances among the long positions according to the formulas

$$\text{cov}(r_i, r_j) = \beta_i \beta_j V(f), \quad i \neq j, \tag{73a}$$

$$V(r_i) = \beta_i^2 V(f) + V(u_i), \quad i = 1, 2, 3. \tag{73b}$$

The result of this calculation for the present example is shown in Table 2.

As Sharpe (1963) explains for a long-only portfolio analysis, the covariance matrix for a one-factor model can be transformed into a sum of squares by introducing a new variable constrained to be the portfolio beta, as in (30). Table 3 contains the covariance matrix for this four-security version of the three-security single-factor model. The algorithm presented in Sharpe (1963) takes advantage of the fact that the covariance matrix is diagonal, with nonzero entries on the diagonal, rather than a dense arbitrary covariance matrix (i.e., an arbitrary positive, semidefinite matrix) as the general critical line algorithm permits. Sharpe's diagonalized version of the n-security one-factor model is frequently referred to as *the* diagonal model.

The advantage of thus diagonalizing the covariance matrix increases with the number of securities in the portfolio analysis. Column 2 of Table 4 presents the number of input coefficients required by the diagonal model of covariance: namely n betas, n idiosyncratic variances, and one factor f variance. The third column of Table 4 presents the number of unique covariances needed by a computation expecting an arbitrary covariance matrix, namely $n(n+1)/2$. Specifically, with three securities there are actually more coefficients in the diagonal model than in the nondiagonalized version. With 5,000 securities, the diagonal model works with about 10,000 coefficients, whereas the 5,000-by-5,000 covariance matrix of the general model has over 12 million unique covariances (counting $\sigma_{ij} = \sigma_{ji}$ as *one* covariance). Both versions of the model also need n expected returns.

Both versions of the model will go through the same number of iterations and come out with the same efficient frontier. The work per iteration depends on how many securities are IN as well as the total number of securities. For moderate to large-size analyses, much less work is required by the diagonal model per iteration.[16]

Table 5 presents the expected returns, betas, and idiosyncratic variances for both the long and short securities corresponding to the long securities in Table 1. Short positions are labeled 1S, 2S, 3S. The expected returns for the short positions are computed according to Equation (50). The betas of the short position are the negative of those for the long position, whereas the idiosyncratic variances are the same for the short position as for the corresponding long position.

Table 1. Illustrative three-security one-factor model.

Security i	Expected return $\mu(i)$	Beta $\beta(i)$	Idiosyncratic variance $V(i)$	Rebate fraction $h(i)$
1L	0.10	0.80	0.0768	0.5
2L	0.12	1.00	0.1200	0.5
3L	0.16	1.25	0.1875	0.5
Lend	0.03	0.00	0.0000	NA
Borrow	0.05	0.00	0.0000	NA
Variance of factor		0.0400		

Notes. This table shows inputs to a three-security, one-factor long-short model. These consist of the expected return, beta against the factor, and idiosyncratic variance of each long position. Also needed are the rebate fraction of each security (for computing expected returns of short positions), the rates at which the investor can borrow and lend, and the variance of the underlying factor.

Table 2. Covariances among long positions.

Security	1L	2L	3L
1L	0.1024	0.0320	0.0400
2L	0.0320	0.1600	0.0500
3L	0.0400	0.0500	0.2500

Notes. This table shows covariances among long positions, computed from their betas, idiosyncratic variances, and the variance of the underlying factor.

Table 3. Covariances when dummy security is included.

Security	1L	2L	3L	PB
1L	0.0768	0	0	0
2L	0	0.1200	0	0
3L	0	0	0.1875	0
PB	0	0	0	0.0400

Notes. In a model with long positions only, the introduction of "portfolio beta" as a fourth (dummy) "security" diagonalizes the covariance matrix. An added equation is needed to constrain PB to equal portfolio beta.

Table 4. Number of unique coefficients required by model of covariance.

Number of securities	With dummy security	Without dummy security
3	7	6
20	41	210
100	201	5,050
500	1,001	125,250
1,000	2,001	500,500
3,000	6,001	4,501,500
5,000	10,001	12,502,500

Notes. This table shows the number of coefficients needed to characterize the covariance structure when the dummy variable of Table 3 is or is not added to the model. Since $\text{cov}(i, j) = \text{cov}(j, i)$ these are counted only once.

The covariances among long and short positions, presented in Table 6, are derived from Table 2 using Equation (53). We could compute an efficient frontier for the short-long model using the expected returns in Table 5, and the covariance matrix in Table 6, using a general portfolio analysis program that permits an arbitrary covariance matrix. If we perform the Sharpe (1963) trick of expressing return as a linear function of amount invested in the factor, plus amounts invested in the idiosyncratic terms as in our Equations (54) and (57), then the covariance matrix for the long-short model is as presented in Table 7. Note that the covariance matrix is no longer diagonalized because, for example, the 1L idiosyncratic term has a −1.0 correlation with 1S.

If we present the data in Table 5 to the Sharpe (1963) algorithm, it will assume that the covariance matrix is in fact diagonal, such as that in Table 8. Theorem 3 assures us that the efficient frontier computed assuming the diagonal covariance matrix in Table 8 is the same as the efficient frontier computed using the correct covariance matrix in Table 7. It also assures us that, for any number of securities, we get the correct result if we ignore the correlations among the idiosyncratic terms for a many-factor model, scenario model, or a mixed factor and scenario model. It further assures us that the efficient frontier is correctly computed if additional constraints are imposed on the choice of portfolio, provided that the constraint set satisfies Property P. In particular, we may present the requisite parameters to the Markowitz-Perold (1981a, b) algorithm for the scenario model or mixed-scenario models, ignoring the correlation between the short and long idiosyncratic terms, for any system of constraints that satisfies Property P.

In the case of the n-security one-factor model, the advantage of using the diagonal model (as permitted by Theorem 3) rather than a general model is again given by Table 4 and Endnote 13, except that now an n-security long-short model has $2n$ "securities." For example, if there are 500 securities in the universe, then the diagonal model will be told that there are 1,001 securities whose covariance structure is described by 2,001 coefficients, whereas the general model will require 500,500 unique (arbitrary, as far as it knows) covariances.

Table 5. Illustrative three-security one-factor model with long (L) and short (S) positions.

Security i	Expected return $\mu(i)$	Beta $\beta(i)$	Idiosyncratic variance $V(i)$
1L	0.100	0.80	0.0768
2L	0.120	1.00	0.1200
3L	0.160	1.25	0.1875
1S	−0.085	−0.80	0.0768
2S	−0.105	−1.00	0.1200
3S	−0.145	−1.25	0.1875
Lend	0.030	0.00	0.0000
Borrow	0.050	0.00	0.0000
Variance of factor		0.0400	

Notes. The table shows properties of a three-security one-factor long-short model, derived from Table 1. Here the expected returns of the short positions are the negative of those of the corresponding long positions, plus short rebate interest on the proceeds. The betas of the short positions are the negative of the long positions; the idiosyncratic variances are the same as those of the long positions. Not noted in the table is the fact that the covariances between the idiosyncratic terms of nS and nL are not zero.

Table 6. Covariances among long and short positions.

Security	1L	2L	3L	1S	2S	3S
1L	0.1024	0.0320	0.0400	−0.1024	−0.0320	−0.0400
2L	0.0320	0.1600	0.0500	−0.0320	−0.1600	−0.0500
3L	0.0400	0.0500	0.2500	−0.0400	−0.0500	−0.2500
1S	−0.1024	−0.0320	−0.0400	0.1024	0.0320	0.0400
2S	−0.0320	−0.1600	−0.0500	0.0320	0.1600	0.0500
3S	−0.0400	−0.0500	−0.2500	0.0400	0.0500	0.2500

Notes. The table shows covariances among short and long positions. These entries are of the same magnitude as the long-only covariances in Table 2, with the same sign in case of long-long or short-short covariances, and opposite sign in case of long-short or short-long covariances.

9. Summary

CAPMs frequently assume, in effect, that an investor can sell a security short without limit and invest the proceeds of the short sale in some other stock. In fact, this is not the case. This paper describes some actual short-sale arrangements. However, short-sale requirements vary from time to time, broker to broker, and investor to investor. Thus, the portfolio analyst must model the sale requirements of the specific client as she or he finds them.

The CLA traces out a piecewise linear set of efficient portfolios subject to any finite system of linear equality or inequality constraints, for any covariance matrix and expected return vector. Because the covariance matrix is arbitrary, the CLA can trace out efficient sets for long-short

Table 7. Covariances when dummy security is included.

Security	1L	2L	3L	1S	2S	3S	PB
1L	0.0768	0	0	−0.0768	0	0	0
2L	0	0.1200	0	0	−0.1200	0	0
3L	0	0	0.1875	0	0	−0.1875	0
1S	−0.0768	0	0	0.0768	0	0	0
2S	0	−0.1200	0	0	0.1200	0	0
3S	0	0	−0.1875	0	0	0.1875	0
PB	0	0	0	0	0	0	0.0400

Notes. Above are the covariances among long, short, and the dummy security, PB, when portfolio beta is introduced as a seventh dummy security. An equation is added to constrain PB to be portfolio beta. Unlike the long-only case in Table 3, the covariance matrix is no longer diagonal.

portfolio selection problems provided that the constraints on choice of portfolio are linear equalities or weak inequalities. Examples of such constraints include a budget constraint, the Reg T "margin requirement constraint," upper bounds on long or short positions in individual or groups of assets, or the requirement that the sum (or a weighted sum) of long positions not differ "too much" from the sum (or the weighted sum) of short positions.

While the CLA may be applied to an arbitrary covariance matrix, it is especially fast for models in which covariances are implied by a factor or scenario model. In this case, an equivalent model can be written, including new "fictitious" securities whose magnitudes are linearly related to the magnitudes of the "real" securities, so that the covariance matrix becomes diagonal or almost so. Special programs exist to exploit the resultant sparse, well-structured efficient-set equations.

A portfolio selection problem in which securities can be held short or long can be modeled as a $2n$-security problem, in which a first n represents long positions, and another n short positions, and all $2n$ are required to have nonnegative values. Even if long positions in n securities satisfy the assumptions of the factor or scenario model, the $2n$-variable long-short model does not satisfy these same assumptions, because idiosyncratic terms are not uncorrelated. Nevertheless, if the information for the $2n$ variables is fed into a factor or scenario program, a correct answer is computed—provided that a certain condition ("Property P") holds.

Property P essentially requires that if a portfolio with short and long positions in the same stock is feasible, then it is also feasible to reduce both positions, keeping the holdings of all other risky stocks the same; and this reduction in both the short and long positions in the same stock does not decrease the expected return of the portfolio. When this condition is met, then the $2n$-variable version of the long-short problem can be run on the appropriate factor or scenario model program. The correct answer is produced despite the violation of the assumption that the idiosyncratic terms are uncorrelated.

A fast CLA also exists for the situation in which historical covariances are used, but there are many more securities than time periods. This algorithm produces the correct answer when applied to the $2n$-variable version of the long-short problem, whether or not Property P holds.

The speed-up in computation that results from the use of "diagonalized" versions of factor, scenario, or historical models is approximately equal to the ratio of nonzero coefficients in the equations of the two models. For large problems, this timesaving can be considerable.

Table 8. Covariances based on Theorem 3.

Security	1L	2L	3L	1S	2S	3S	PB
1L	0.0768	0	0	0	0	0	0
2L	0	0.1200	0	0	0	0	0
3L	0	0	0.1875	0	0	0	0
1S	0	0	0	0.0768	0	0	0
2S	0	0	0	0	0.1200	0	0
3S	0	0	0	0	0	0.1875	0
PB	0	0	0	0	0	0	0.0400

Notes. If the data in Table 5 are presented to a standard factor model portfolio optimizer, the program will assume that the model has the covariance structure in this table, with a diagonal covariance matrix, rather than the correct one, that in Table 7. Theorem 3 assures us that the optimizer will nevertheless compute the efficient frontier correctly. Theorem 3 further assures us that this is so for a many-factor model, a scenario model, or a mixed factor-scenario model of covariance; and remains true for any system of linear equality or (weak) inequality constraints that satisfy Property P.

Endnotes

1. The results reported in this paper were first circulated in a Jacobs Levy Equity Management working paper (Jacobs et al. 2001).

2. For proofs and further details, see Markowitz (1959), Appendix A, Perold (1984), Markowitz (1987), or Markowitz and Todd (2000).

3. See Markowitz and Todd (2000), Chapter 8, for how to get a first critical line.

4. If C is singular, there may be more than one portfolio with minimum feasible V. Because the V-minimizing portfolios may have different Es, they may not all be efficient, but it is shown that the portfolio reached by the CLA when $\lambda_E \downarrow 0$ is efficient as well as V-minimizing.

5. See Markowitz and Todd (2000), Chapter 9, for what to do in case of ties.

6. Note that the CLA as presented in Markowitz (1956) is an example of a linear complementary algorithm as defined in Wolfe (1959).

7. See Sharpe (1963) in particular, and Markowitz and Perold (1981b) in general, for details.

8. For models that combine both scenarios and factors, see Markowitz and Perold (1981a, b).

9. For details, see Markowitz et al. (1992).

10. See Fortune (2000) for details on initial and maintenance margin requirements for long and short positions on exempt and nonexempt securities. Also see www.federalreserve.gov/regulations, 12 CFR 220, Credit by Brokers and Dealers (Regulation T). See Jacobs and Levy (1993) on margin requirements and cash needed for liquidity.

Equation (43) can also be written as

$$\sum_{i=1}^{2n} 0.5X_i \leqslant 1, \tag{10.1}$$

reflecting a 50% margin on short and long positions. Actually, the Reg T initial short margin requirement is stated as 150%—of which 100% out of the 150% is supplied by the proceeds of the sale of the borrowed stock. Constraint (10.1) is a special case of

$$\sum_{i=1}^{2n} m_i X_i \leqslant 1, \tag{10.2}$$

where m_i here represents the net (after proceeds, where applicable) margin requirement of the ith position. Constraint (10.2) is more general than (43) or (10.1) in that it permits, in particular, (a) a net short margin requirement that differs from the long margin requirement, and (b) securities that are exempt from Reg T requirements. We use (43) in examples, but Theorems 1, 2, and 3 apply to *any* system of constraints (4) and (5) with properties specified in theorems.

11. Rule 15c3-1 of the Securities Exchange Act of 1934 governs capital requirements for broker-dealers, including the provision that indebtedness cannot exceed 1,500% of net capital (800% for 12 months after commencing business as a broker or dealer).

12. Noncash collateral typically consists of letters of credit or securities. It is usually 100% to 105% of the amount borrowed. The gains and losses on the collateral belong to the borrower, and the lender is generally paid a fee. The collateral is marked to market and augmented by the borrower if necessary.

13. Usually $h_i < 1$. However, the case of $h_i = 1$ is conceivable, and is covered by our theorems. Large institutional investors often perform mean-variance analysis at an asset class level and then implement the asset class allocations using either index funds or using internal or external fund managers. If, say, an internal market neutral fund borrows shares from, say, an internal large-cap or small-cap fund, the allocation of interest on the proceeds between borrowing fund and lending fund is arbitrary. The institution's policy might allocate all the interest to the borrowing fund, because the institution's policy might prohibit external stock lending, so that the particular interest income would not exist except for the internal market neutral fund's activities.

If no zero-variance variable is ever held short, we may write (47) as

$$R_P = \sum_{i=1}^{n} r_i X_i + \sum_{i=n+1}^{n+\nu} (-r_{i-n}) X_i + r_c \sum_{i=n+1}^{n+\nu} h_{i-n} X_i. \tag{13.1}$$

Alternatively, we can leave it *as is* in (47) and assume that (45) contains equations of the form

$$X_{n+i} = 0 \tag{13.2}$$

for $i \in [\nu + 1, n]$. Generally, if a security cannot be sold short (e.g., because it cannot be borrowed), then this can be represented either by including a constraint of the form (13.2) or by omitting $n + i$ from the analysis. The latter approach is advisable in practice; the former is notationally convenient here.

14. Equation (47) does not include tax considerations, and therefore would be applicable to tax-exempt organizations such as university endowments and corporate pension plans.

15. Jacobs et al. (1998, 1999) address the conditions under which optimal portfolios that are constrained to hold roughly equal amounts in long and short positions are equivalent to optimal portfolios without this constraint. In practice, long-short portfolios are often managed in this "market-neutral" fashion.

16. If n_I securities are IN, then the Sharpe (1963) algorithm requires a few more than $3n + 7n_I$ multiplications and divisions plus $3n + 5n_I$ additions, whereas the general algorithm requires $2n_I n + 5n + 2n_I^2 - n_I$ multiplications and divisions, and $2n_I n + 3n + 2n_I^2 - 2n_I$ additions. Thus, if $n = 1{,}000$ and $n_I = 10$, as at the high end of the frontier,

or $n_I = 100$ as might occur at the low end of the frontier, then the diagonal model requires 3,070 or 3,700 multiplications and divisions for the iteration, whereas the general algorithm requires 25,190 or 269,900.

References

Alexander, Gordon J. 1993. Short selling and efficient sets. *J. Finance* **48** 1497–1506.

Cohen, K. J., J. A. Pogue. 1967. An empirical evaluation of alternative portfolio selection models. *J. Bus.* **40** 166–193.

Elton, Edwin J., Martin J. Gruber, Manfred W. Padberg. 1976. Simple criteria for optimal portfolio selection. *J. Finance* **31** 1341–1357.

Fortune, Peter. 2000. Margin requirements, margin loans, and margin rates: Practice and principles. *New England Econom. Rev.* **2000** (September/October) 19–44.

Jacobs, Bruce I., Kenneth N. Levy. 1993. The generality of long-short equitized strategies: A correction. *Financial Analysts J.* **49**(March/April) 22.

Jacobs, Bruce I., Kenneth N. Levy. 2000. *Equity Management: Quantitative Analysis for Stock Selection*. McGraw-Hill, New York.

Jacobs, Bruce I., Kenneth N. Levy, David Starer. 1998. On the optimality of long-short strategies. *Financial Analysts J.* **54**(March/April) 40–51.

Jacobs, Bruce I., Kenneth N. Levy, David Starer. 1999. Long-short portfolio management: An integrated approach. *J. Portfolio Management* **25**(Winter) 23–32.

Jacobs, Bruce I., Kenneth N. Levy, Harry M. Markowitz, David Starer. 2001. Optimization and neutrality of long-short portfolios. Jacobs Levy Equity Management, Florham Park, NJ.

Kwan, Clarence C. Y. 1995. Optimal portfolio selection under institutional procedures for short selling. *J. Banking Finance* **19** 871–889.

Markowitz, Harry M. 1956. The optimization of a quadratic function subject to linear constraints. *Naval Res. Logistics Quart.* **3** 111–133.

Markowitz, Harry M. 1959. *Portfolio Selection: Efficient Diversification of Investments*. John Wiley and Sons, New York, and 1991 2nd ed., Basil Blackwell, Cambridge, MA.

Markowitz, Harry M. 1987. *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. Basil Blackwell, Cambridge, MA.

Markowitz, Harry M., André F. Perold. 1981a. Portfolio analysis with factors and scenarios. *J. Finance* **36** 871–877.

Markowitz, Harry M., André F. Perold. 1981b. Sparsity and piecewise linearity in large portfolio optimization problems. I. S. Duff, ed. *Sparse Matrices and Their Uses*. Academic Press, London, UK, 89–108.

Markowitz, Harry M., Peter Todd. 2000. *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. Revised reissue of Markowitz (1987) with chapter by Peter Todd, Frank J. Fabozzi Associates, New Hope, PA.

Markowitz, Harry M., Peter Todd, Gan Lin Xu, Yuji Yamane. 1992. Fast computation of mean-variance efficient sets using historical covariances. *J. Financial Engrg.* **2** 117–132.

Perold, André F. 1984. Large-scale portfolio optimization. *Management Sci.* **10** 1143–1160.

Sharpe, William F. 1963. A simplified model for portfolio analysis. *Management Sci.* **9** 277–293.

Wolfe, Philip. 1959. The simplex method for quadratic programming. *Econometrica* **3** 382–398.