Mean-Variance Portfolio Rebalancing with Transaction Costs*

Philip H. Dybvig
Washington University in Saint Louis

First draft: Aug 17, 2003
This draft: January 12, 2005

Summary

Transaction costs can make it unprofitable to rebalance all the way to the ideal portfolio. A single-period analysis using mean-variance theory provides many interesting insights. With fixed or variable costs, there is a non-trading region within which trading does not pay. With only variable costs, any trading is to the boundary of the non-trading region, while fixed costs induce trading to the interior. With costly trading in futures and underlying, it might be optimal to use a synthetic equity strategy or an asymmetric futures overlay strategy that takes better advantage of extra expected return than traditional futures overlays.

*I am grateful for support from NISA Investment Managers LLC and for helpful comments from Dave Brown, Hong Liu, and Bill Marshall. Any errors or opinions are my own.

Optimal portfolio rebalancing given transaction costs is a complex problem. Even with only two assets, solving for the optimal strategy in a continuous-time model involves a free boundary problem (see, for example, Davis and Norman [1990], Dumas and Luciano [1991], Liu and Loewenstein [2002], and Taksar, Klass, and Assaf [1988]). When there are more securities, the multi-asset continuous time model has been solved only in the extreme case of uncorrelated returns and constant absolute risk aversion (Liu [2004]) or with numerical or heuristic approximations (Leland [2000] or Donohue and Yip [2003]). In this article, we study the single-period rebalancing problem in a mean-variance framework that permits exact solutions.

Mean-variance analysis was originated by Markowitz [1952, 1959], who described the basic formulations and the quadratic programming tools used to solve them. The theory was further described by Tobin [1958], who focused on macroeconomic implications of the theory. Early discussions of transaction costs often focused on the intuition that small investors who face high costs will choose a smaller and less diversified portfolio than will a large investor with smaller costs. This intuition has been formalized by a constraint on the number of securities in the portfolio (Jacob [1974]), a fixed cost for each security included in the portfolio (Brennan [1975], Goldstein [1979], and Mayshar [1979, 1981]), or a study of benefits of adding securities without modeling the costs (Mao [1970, 1971]). Unfortunately, this type of assumption tends to produce a somewhat messy combinatoric problem looking at all possible subsets to include, and their static perspective does not seem suited to questions about rebalancing.

The current analysis differs in two important ways from the traditional mean-variance literature on transaction costs. First, the traditional literature considered the purchase of a portfolio from scratch, while the current analysis considers rebalancing from any starting portfolio. Second, we focus primarily on variable costs rather than fixed costs. (Variable costs and other institutional features were included in choice problems of Pogue [1970], but without any analysis of the solution.) In both respects, the current analysis is closer to the continuous-time models we should ideally be using. Therefore, we can have some confidence in the economic understanding we obtain from this model, keeping in mind that no single-period analysis (including the one in this paper or the heuristic analysis in Grinold and Kahn [1995]) can be expected to approach the level of gains that could be obtained from a

continuous-time model.

While most of the current analysis concerns variable costs, there are also models with fixed costs, both security-specific and overall. All the solutions feature the concept of a non-trading region, a set of portfolio positions from which there is no trade that would justify the cost of trading. In the variable cost models, it is optimal to trade only to the boundary of the non-trading region, since trading further would incur additional costs that are not justified.[1] With proportional costs, the cost of trading is additive (if all trades are in the same direction) or less than additive (if the second trade reverses the first trade in some securities). If a candidate trade does not take us to the non-trading region, we could add on the additional trade we would make from that point and be better off. Or, if a candidate trade takes us beyond the boundary of the trading region, is better to trade along the line to the boundary because the part of the trade beyond the boundary is not justified. These arguments do not work for fixed costs, because they rely implicitly on costs being additive for sequential trades along a line, and on costs being no more than additive for sequential trades that are not along a line.

In the models with fixed costs, any trade moves to inside the non-trading region if it is optimal to trade at all. With only an overall fixed cost, any nonzero trade moves to an ideal portfolio that would be held absent costs. This ideal portfolio is in the interior of the non-trading region because the value of trading from nearby is too small to cover the fixed cost. With security-specific fixed costs, any trade will take us to the interior of the non-trading region. However, different starting portfolios will cause us to trade to different target portfolios. Given the traded subset, we can think of this an an optimization, including hedging demand, given the positions in the other securities.[2]

[1] Masters [2003] contains a mean-variance-style analysis with a single risky asset and variable trading costs in which it is claimed that it is not optimal to trade to the boundary of the non-trading region. However, this is because paper computed the non-trading region incorrectly as the set of portfolios from which it would be worth trading to the ideal point that would be chosen absent costs. The error is that there are portfolios from which it pays to trade partway to the ideal portfolio but not all the way.

[2] Although this case is not included in this paper, a model with both fixed costs and variable costs would also involve trade to different points interior to the non-trading region depending on the starting point. Also, it is possible to construct models with fixed costs with trade to the boundary of the non-trading region at least some of the time, if there

If all trade is in individual stocks and takes place through cash, which is held as a residual asset, the optimal non-trading region is a parallelogram or its higher dimensional equivalent. If, in addition, asset returns are uncorrelated, the parallelogram becomes a rectangle. More generally, if the residual portfolio is a fixed-income investment with nonzero correlation with various assets, the sides of the non-trading region will still be linear but may no longer be parallel.

The analysis in this article can accommodate multiple risky assets, trading of individual securities or bundles or pairs, and trading futures or swaps as well as stocks. The Appendix uses matrix algebra to analyze the general model that can accommodate all these features. Section I presents a graphical analysis of the gives a graphical description of the solution for two risky assets, along with intuition and a discussion of the economics. Section II provides analysis of the case of a single risky asset, a simple case that makes it easy to understand our assumptions and analysis without any matrix algebra.

I Numerical Results

Our analysis is based on mean-variance analysis[3] with an additional term (could be zero) penalizing tracking error relative an index. The Appendix contains the statement and solution of a formal model using matrix algebra that can accommodate multiple securities, as well as contracts such as futures and swaps that require no initial investment and are not described by returns (since the denominator of the return calculation is zero). Transaction costs can vary by security, and indeed trades may include bundles trading or swapping one security for another. A simplified analysis for a single security requiring no matrix analysis is given in Section II as an aid to understanding the analysis and intuition. This section uses graphs to illustrate the solution

are constraints such as limits on individual security or industry holdings. In this case, the best choice to trade to within the non-trading region might be a boundary point at which the limit on individual security holdings is binding.

[3]Of course, the means and variances used in the analysis are the conditional mean and variance given our information, and as has been well known at least since the work of Bawa, Brown, and Klein [1979]

of the model and discuss the intuition.

Variable Costs

When there are multiple risky assets, each of which can be bought or sold for cash, each risky asset has a non-trading region. Typically, the non-trading region for one risky security depends on the positions in the other securities due to correlations among the asset returns. For example, Figure 1 illustrates the solution for security-specific variable costs of trading two correlated risky securities. If the initial allocation θ^0 is in the non-trading region, shaded in yellow, then there is no trade whose benefit covers the cost and it is best not to trade. The right boundary of the non-trading region is part of the line along which we are just indifferent about selling Security 1, and it is optimal to sell Security 1 if we start to the right of this boundary. The left boundary of the non-trading region is part of the line along which we are just indifferent about purchasing Security 1, and it is optimal to purchase Security 1 if we start from left of this boundary. The boundaries for purchasing and selling are different because the costs put a wedge between the marginal valuations at market prices and valuation the prices net of costs.

Symmetric to Security 1, we sell Security 2 if we start above the top boundary and we sell Security 2 if we start below the bottom boundary. If we start in the corners (not directly to the right, left, top, or bottom of any of the sides of the non-trading region), then we trade in both securities.

It may not be obvious why the trades are as shown in Figure 1 and do not go to other points on the boundary of the non-trading region. For example, could there be some points in the region above the top boundary of the non-trading region from which we trade to the upper right corner of the region? The answer is no, because if we buy Security 1 at all, we must end up on the corresponding boundary. In this case, any net purchase of Security 1 must be to the left boundary (not the right boundary). Note that while we always traded to the nearest point in the non-trading region in the case of a single risky asset studied in the previous section, that is not a general result when there is more than one risky asset.

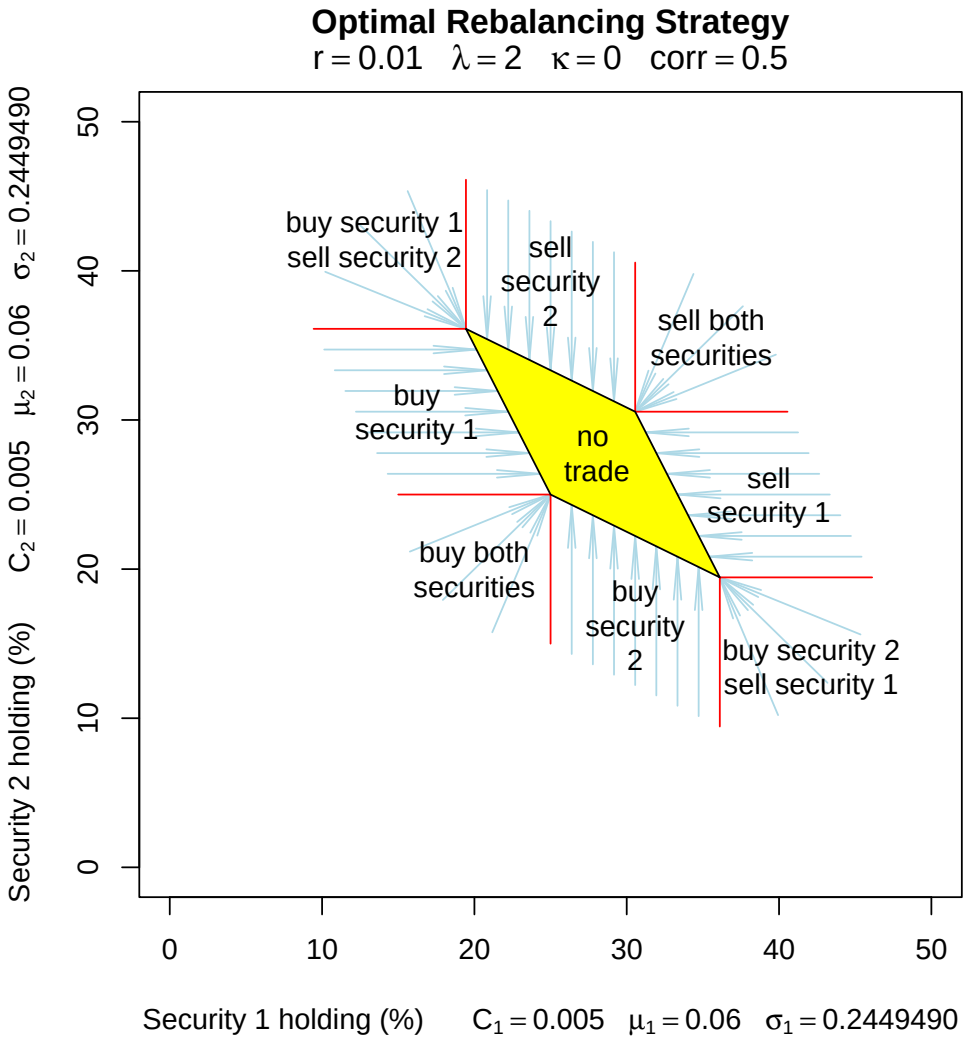

Figure 1: With proportional costs, the non-trading region (shaded yellow) is a parallelogram. Outside the non-trading region, it is optimal to trade (along the blue arrows) to the boundary of the non-trading region. If returns were uncorrelated, then the non-trading region would be a square with sides parallel to the axes. In this example, returns are correlated and the two securities are substitutes and over-weighting in one security is less likely to result in a trade if we are under-weighted in the other security.

Absent the positive correlation between the returns on the two securities, the non-trading region would have been a square (or a rectangle given diverse cost per unit variance across securities) with sides parallel to the axes. With positive correlation (or a positive weight on benchmark deviations and positive correlations with the benchmark), the two securities are substitutes, and over- or under-weighting in one security is more serious if we have the same over- or under-weighting in the other security. This is why the non-trading region is larger along the $-45°$ direction in which the over- and under-weightings cancel than along the $45°$ direction in which the over- and under-weightings are reinforced.

In the first example, there was a penalty ($\lambda = 2$) for taking risk but no penalty ($\kappa = 0$) for tracking error. Adding a penalty for tracking error ($\kappa > 0$) does not change the mathematical form of the problem and is actually equivalent to changing means and covariances. Figure 2 is similar to Figure 1 but with a penalty ($\kappa = 1$) for tracking error. This additional penalty makes the non-trading region smaller and shrinks it towards the benchmark portfolio, which has 40% weight in the first risky asset, 20% weight in the second, and the remaining 40% weight in cash.

The larger the trading costs, the larger the benefits have to be for trade to be justified, and the larger the non-trading region. This is true for either or both securities. In Figure 1, the two securities are symmetric, but in Figure 3 we have increased the cost of trading Security 2. This increases the vertical size of the non-trading region. Indeed, in this case the optimal trade of an agent starting with all cash (indicated by "×") purchases only Security 1 (moving along the green arrow from "×"). A similar result could obtain if Security 2 had more idiosyncratic risk than Security 1.

Futures Overlay

It is increasingly common for plan sponsors to use futures as well as (or instead of) equities for managing exposure to market risk. One popular example of a transaction-cost-aware strategy is to use futures as an inexpensive way of keeping effective asset allocation in line with a benchmark or ideal allocation. For example, if we think the ideal weighting in equities is 60%,

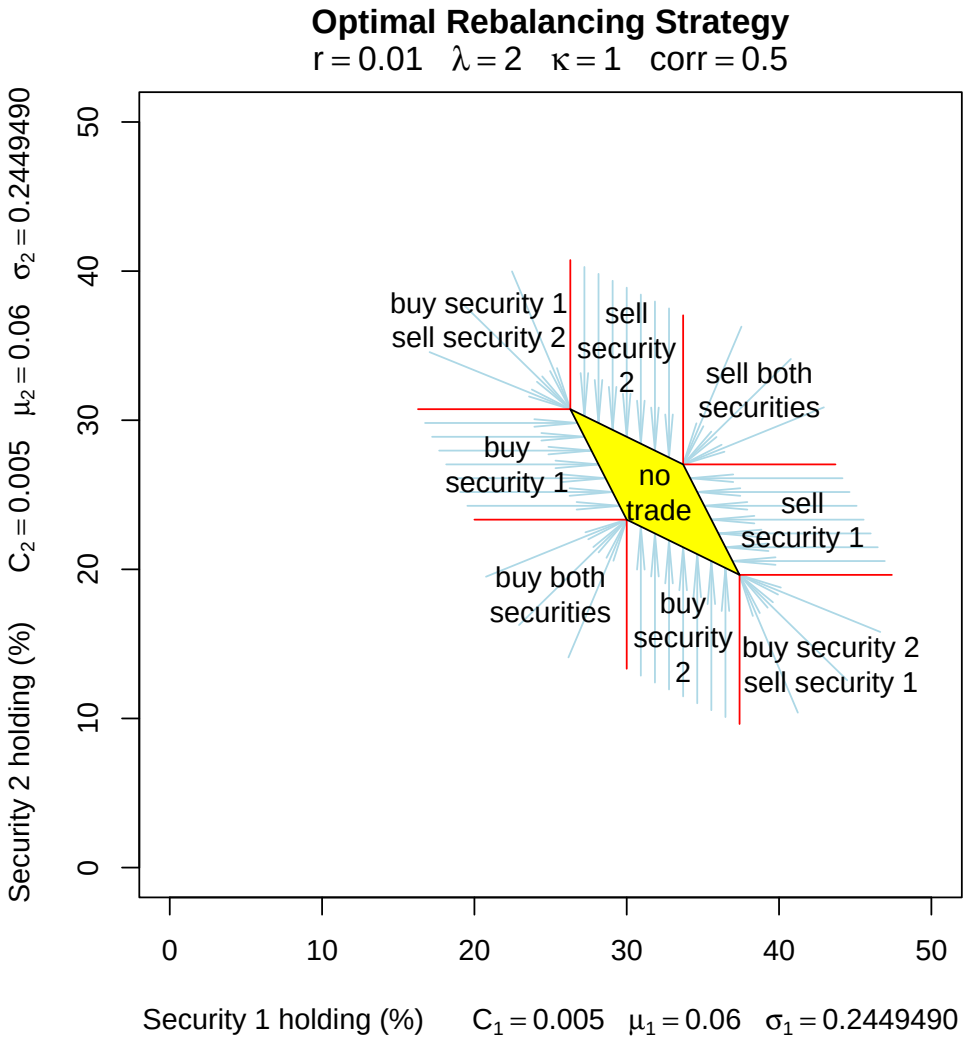

Figure 2: If there is a penalty in the objective function for deviations from a benchmark, the non-trading region shrinks and moves towards the benchmark (here having 40% weight in the first risky asset and 20% weight in the second). Putting a penalty on deviations from a benchmark would not happen in an ideal world, but may improve incentives or coordination for managers and are common in practice.

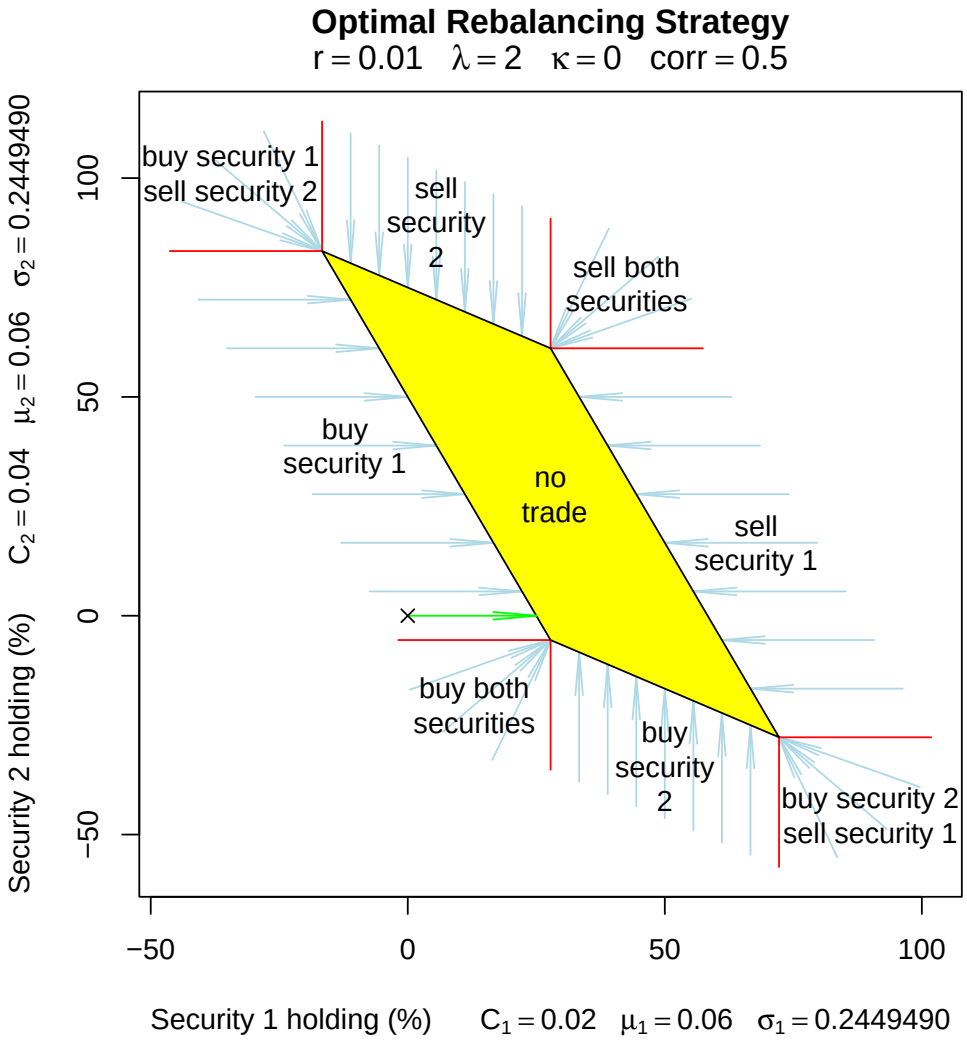

Figure 3: The larger the cost of trading, the larger the non-trading region. In this case, the cost of trading Security 2 is high and the vertical size of the non-trading region is large. Starting with all cash (at the "×"), the optimal trade forgoes the expensive Security 2 and purchases only Security 1 (along the green arrow).

then as the market rises we become overweighted and as the market falls we become under-weighted (since the fixed-income part of the portfolio moves less than proportionately with moves in the equity market). Maintaining a weighting near the ideal weighting by trading equities is very expensive. A futures overlay might correct for minor deviations from the ideal weighting by trading in futures, which are highly correlated with equities but much cheaper to trade. Perhaps futures are used to keep the exposure to equities to within 3% of the ideal allocation with trading in actual shares of stock only when the allocation gets more than 10% out of line. The analysis here confirms the value of using a futures overlay but suggests an improvement on the traditional form of the strategy.

For a futures strategy, we normally think that the correlation between the equity position and futures is close to one, so that holding futures and bonds is a close substitute for trading the underlying equities. We also usually believe that futures are much less expensive to trade the underlying, which is why it is appealing to consider substituting futures trades for trades in equities. The expected returns ("alphas") are not usually discussed much, but they turn out to be very important.

Figure 4 illustrates a case in which the investor gets nearly the same expected return on the underlying and on "synthetic equity" composed of futures and investing in the risky asset. In this case, the additional expected return from holding equities is too small to justify the additional transaction costs. Consequently, if we are over- or under-weighted in equities initially (with no futures), we correct our position by selling or buying futures. Thus, in the absence of some additional return to holding the underlying equities, a synthetic equity strategy would be optimal, and there would seem to be little reason to hold the underlying equities in the first place, or to trade them once we own them.

Generally, we might expect the return on the underlying equities to be higher than the return on synthetic equity due to the benefits of active management. Or, moving somewhat outside the model, the extra return on the underlying may be due to the cost of rolling the futures or the tax-timing advantages to equity. Figure 5 illustrates an example in which equities have a significantly higher expected return than the synthetic equity strategy using futures. In this case, there is a trade-off between transaction costs and expected return

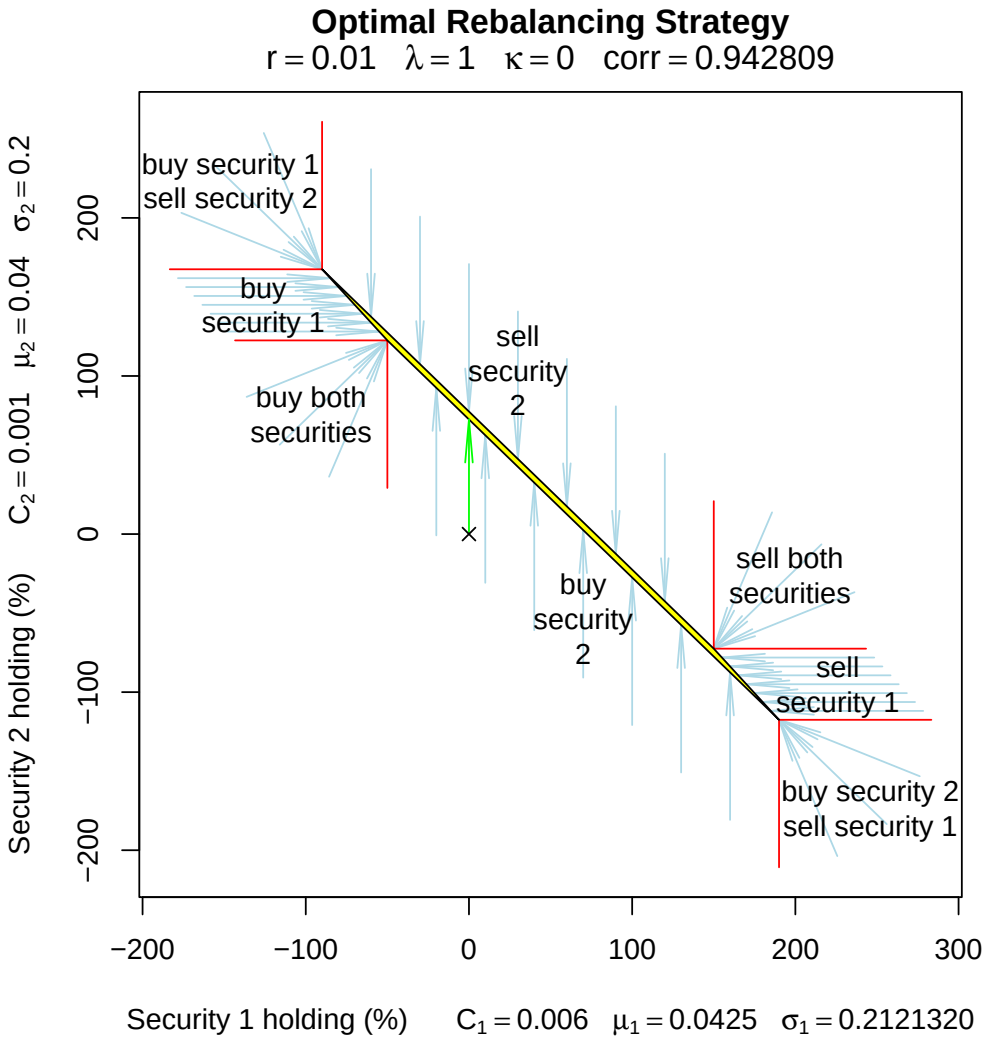

Figure 4: This example considers holdings of equities (Security 1) and futures (Security 2) that are highly correlated in a case in which the equity-equivalent mean returns μ_i are nearly the same. In this case, the cheaper trading costs of futures are decisive. Consequently, it is optimal to pursue a "synthetic equity" strategy and (at least in the normal range of starting points) do all trading in futures. In particular, starting from all cash (the "+"), the optimal trade is the buy futures only.

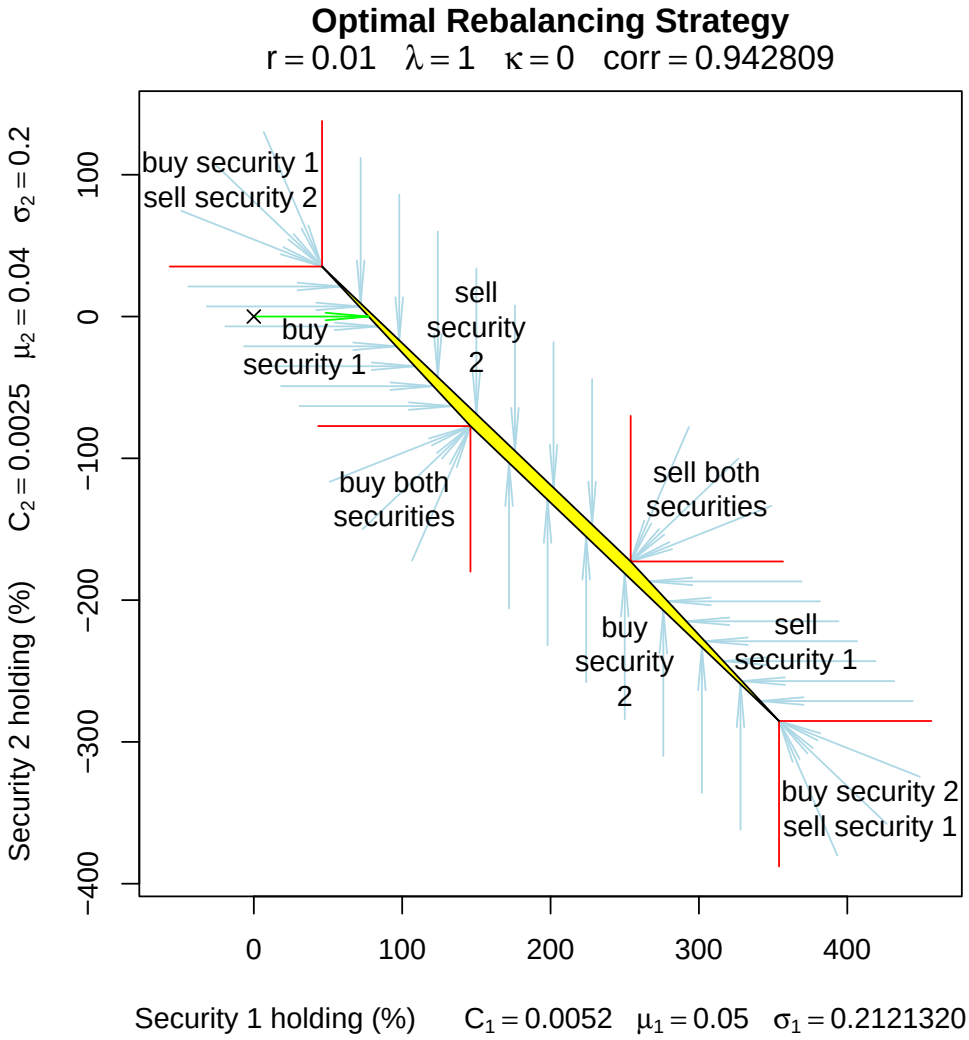

Figure 5: This second example with futures has a significantly higher expected return ("alpha") on actual equities (Security 1) than on synthetic equities composed of futures (Security 2) and cash. The optimal strategy is an asymmetric "futures overlay" strategy typically selling futures to correct for overexposure to market risk but buying underlying equities to correct for underexposure to market risk. This asymmetry is due to the fact that selling futures allows us to keep the alpha on the exposure we are eliminating, while buying equities allows us to gain alpha on the exposure we are taking on.

and it is optimal to use a futures overlay strategy of using futures to substitute for some trading in the underlying. In practice, most plan sponsors use a "symmetric" futures overlay that uses futures to the same extent for correcting over- and under-exposure to the market. However, the analysis here prescribes an asymmetric strategy that makes good economic sense. If the market exposure must be reduced, we sell futures,[4] which allows us to keep the extra return on the underlying equities. On the other hand, if the market exposure must be increased, we buy equities, which have the extra return, rather than futures, which don't.

Bundles Trading

If it is possible to purchase a bundle of securities more cheaply than its constituent securities, then there could be a new non-trading boundary in the problem. Figure 6 illustrates an example that is the same as in Figure 1 except with the additional opportunity of buying or a 50-50 mix of the two securities with a transaction cost of .0035. This leads to new sides of the non-trading boundary. Adjacent the region in which we sell the bundle, there are regions where we sell the bundle and one of the securities. Similarly, adjacent the region in which we buy the bundle, there are regions where we buy the bundle and one of the securities. To keep the graph simple, we have considered a case with only two underlying securities, but bundles trading would obviously be more useful and more interesting with many securities. For example, trading a bundle might be a cost-effective way of aligning market or sector exposures with a target. With many securities in the bundle, the trading pattern could be more complex, for example, with simultaneous buys and sells of different individual securities to compensate for imbalances caused by trading the bundle.

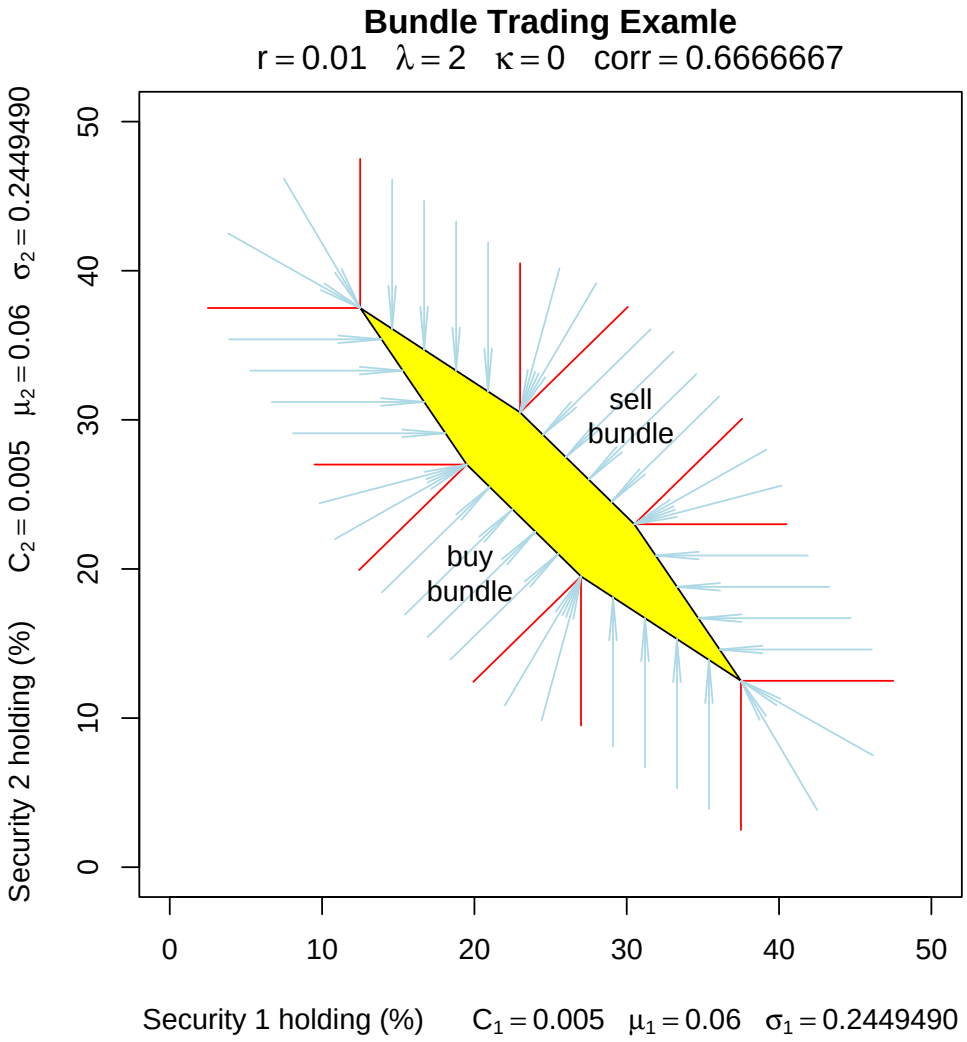

Figure 6: If it is possible to trade a bundle of securities at a favorable cost, there are additional sides to the non-trading region. This is the same case as in Figure 1 but with an additional opportunity to trade a portfolio of the two assets at a cost of .0035.

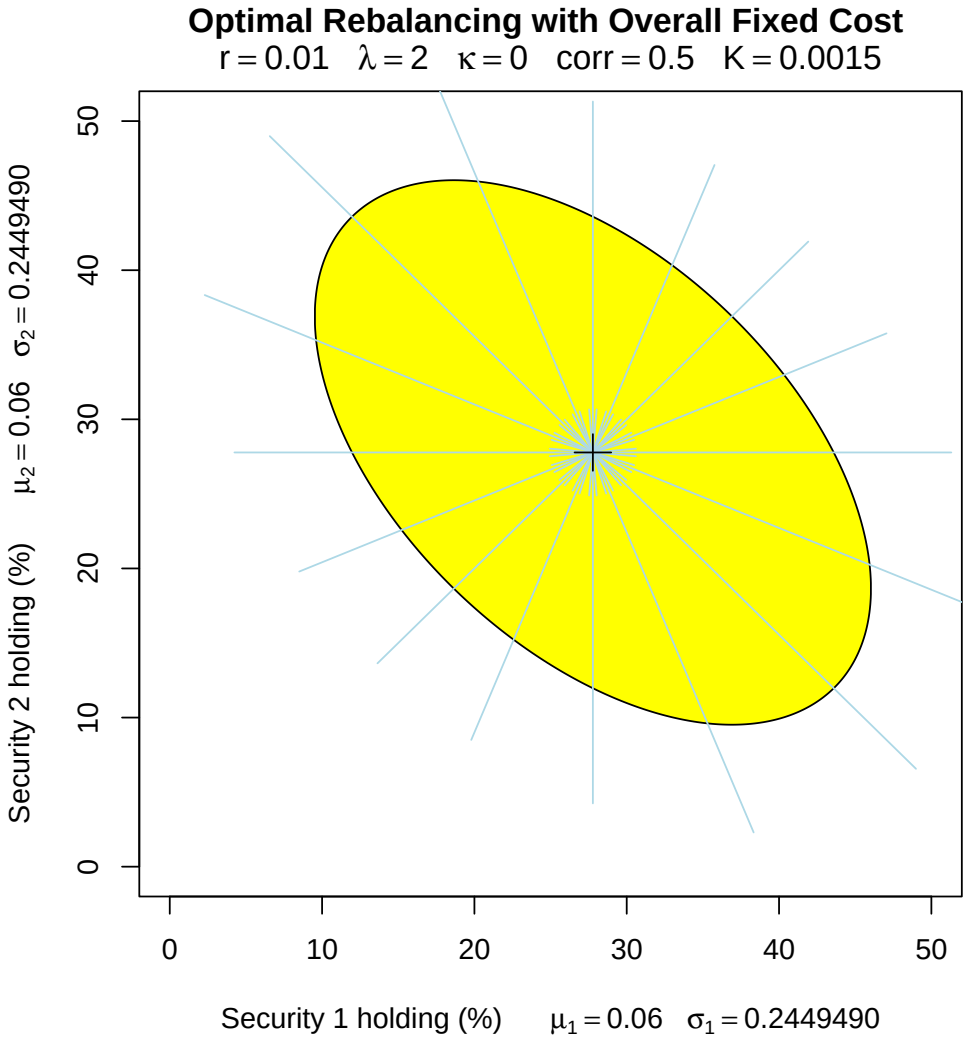

Figure 7: With an overall fixed cost, either there is trade immediately to the ideal point (indicated by "+") or it is not worth trading at all. The non-trading region (shaded yellow) is an ellipse. As with proportional costs, correlation between the assets implies that it is more damaging (and more likely to do trade) when both asset positions are out of line in the same direction.

Overall Fixed Costs

If there is an overall fixed cost, then if we trade the cost is the same whatever trade we make. Therefore, if we are going to trade, we always trade to the same ideal portfolio. Consequently, the choice problem reduces to one of comparing the utility of not trading with the utility of trading to the ideal point but incurring the fixed cost. It is optimal to trade when outside the indifference curve (an ellipse given our mean-variance assumptions) and not trading inside. The size of the trading region is proportional to the square root of the trading cost, since the utility loss in any direction is proportional to the squared distance from the optimum.

The overall fixed cost is illustrated in Figure 7. The non-trading boundary is bounded by an ellipse. From outside this region, it is optimal to trade to the ideal point (indicated by "+"), since it is no more costly to trade to the ideal portfolio than to trade to a less-preferred portfolio. If the asset returns were uncorrelated with symmetric covariances, the non-trading region would be a circle. In this example, everything is symmetric but there is correlation. The correlation means that the two assets are substitutes and it is not so bad if we have too little of one asset if we have too much of the other. As in the case of proportional costs, this is why we are quicker to trade if we are over-weighted in both assets than if we are over-weighted in one and under-weighted in the other.

Security-Specific Fixed Costs

What may be more plausible than an overall fixed cost is a fixed cost for each security. Arguably, a security-specific fixed cost comes from a due diligence requirement to monitor or document any security in the portfolio, although a serious consideration of this motivation probably leads us to informational or strategic considerations outside the current framework.[5] In general, security-

[4]This is the normal situation but extreme cases may be different. For example, in Figure 5 we would sell equities rather than futures if we found ourselves 325% long equities and 225% short futures.

[5]For example, why would we have to monitor a position unless new information arrival is possible and subsequent trade is possible? Perhaps a regulator requires documentation

specific fixed costs face us with a complex combinatorial problem, since each possible set of included portfolios gives a different piece of the overall non-concave objective function. However, in a particular small example it is not so hard to solve, since the set of boundary points where two subsets are equally preferred is a conic section.

Figures 8, 9, and 10 illustrate the various trading regions. Near the ideal point ("+") in the middle, shaded yellow, is the non-trading region. The upper left and lower right boundaries are on the same ellipse, and the other boundaries are linear (see the Appendix for more on this). From the regions in the corner, we trade both securities to the ideal point. From the regions on the right and left, we trade Security 1 but not Security 2 to a line going through the ideal point. This would be a vertical line if we had no correlation, but has negative slope in our case. Similarly, from the regions above and below, we trade only Security 2 to a different line running through the ideal point. Figure 8 illustrates the trades in both securities, Figure 9 illustrates the trades in Security 1 alone, and Figure 10 illustrates the trades in Security 2 alone. This example is consistent with Brennan [1975] or Goldstein [1979], proportional to the number of securities traded, or implicitly Jacob [1974], where we are given exogenously a small number of securities that can be bought. This analysis is more general mostly in that it does not assume the starting position in in cash alone.

II Theory: Single Risky Asset Example

Here is the choice problem for our case with a single risky asset:

Problem 1 *(proportional costs, single risky asset) Choose purchases* $P \geq 0$

of the trade and a due diligence study of the firm issuing each share of stock we hold, even though we know we are not going to learn anything from the exercise. Another question is why we don't have to do monitoring or due diligence on a stock we already hold and choose not to sell. Or, it may be that our broker offers to make any trade in a single maturity, whatever the size, for the same fixed price. It seems much easier to make an argument for why there are variable costs, or perhaps for variable costs plus fixed costs.

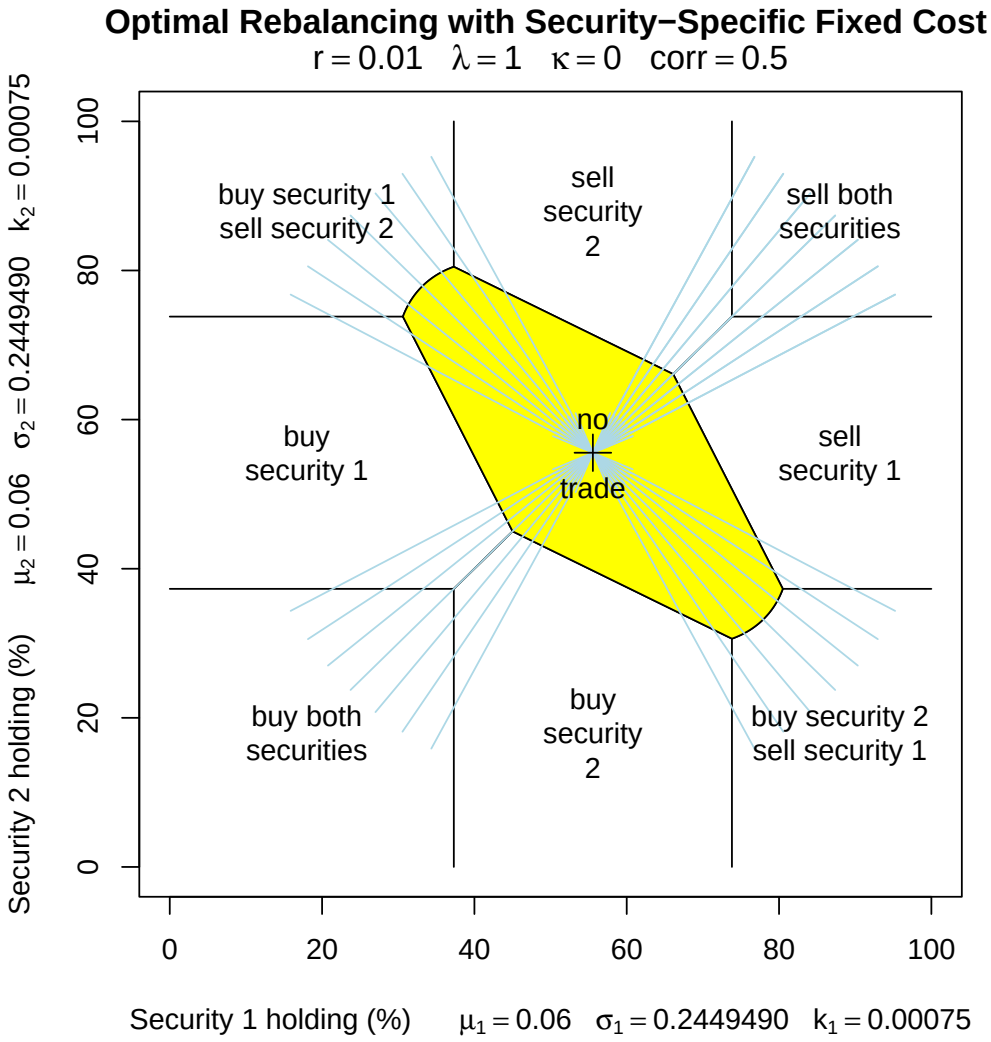

Figure 8: This figure shows the various trading regions for an example with security-specific fixed costs. With security-specific costs, it only pays to trade whatever security or securities is significantly out of line with the ideal allocation ("+"). The blue arrows indicate optimal trades to the ideal point when both securities are traded.

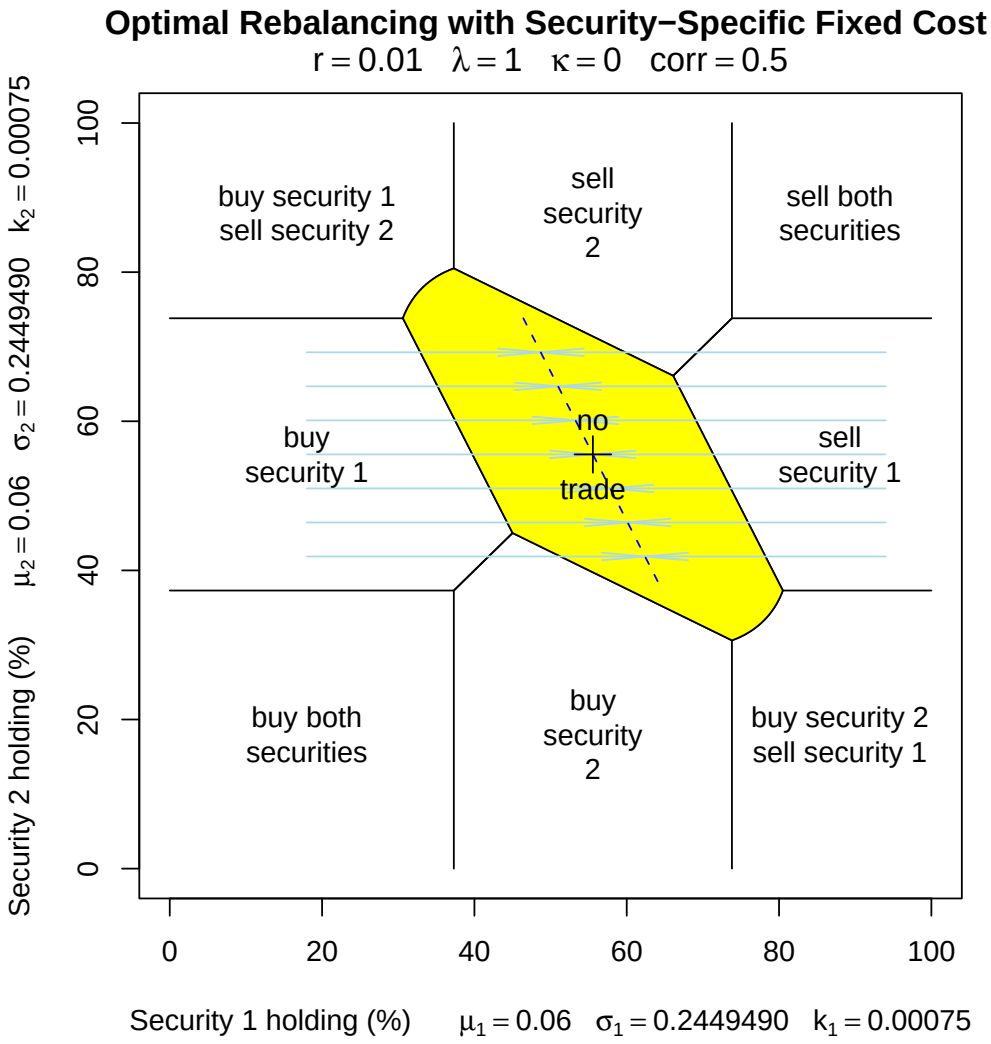

Figure 9: Here, the blue arrows indicate optimal trades to the ideal point when only Security 1 is traded.

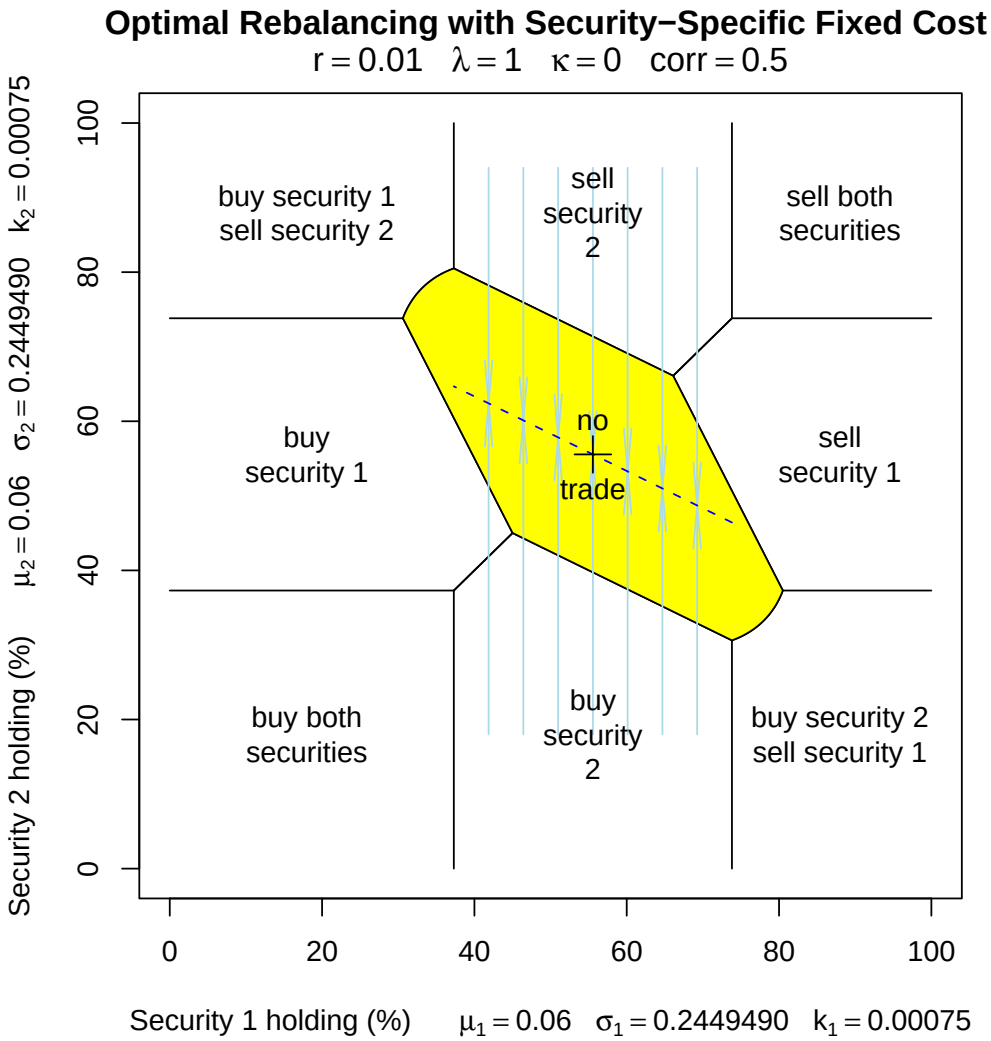

Figure 10: Now, the blue arrows indicate optimal trades to the ideal point when only Security 2 is traded.

and sales $S \geq 0$ of the risky asset to maximize the utility function:

$$(1) \quad r + \theta(\mu - r) - \frac{\lambda}{2}(\theta\sigma)^2 - \frac{\kappa}{2}((\theta - \theta^B)\sigma)^2 - C^P P - C^S S.$$

where $\theta \equiv \theta^0 + P - S$ is the risky portfolio position after trade and where we use the following notation:

$P \geq 0$*: purchase (as a fraction of initial wealth) of the risky asset*
$S \geq 0$*: sale (as a fraction of initial wealth) of the risky asset*
r*: risk-free rate of interest*
μ*: risky asset expected return (net of any liquidation costs)*
$\theta(\mu - r)$*: mean excess return on the risky portfolio*
$\lambda > 0$*: absolute risk aversion parameter*
$\kappa > 0$*: tracking error aversion parameter*
$\sigma > 0$*: standard deviation of the risky asset return*
$(\theta\sigma)^2 > 0$*: variance of the portfolio return*
$((\theta - \theta^B)\sigma)^2 > 0$*: variance of deviations from the benchmark*
θ^0*: initial risky asset holding*
θ^B*: benchmark portfolio's risky asset holding*
$C^P > 0$*: proportional transaction cost for purchases, end-of-period units*
$C^S > 0$*: proportional transaction cost for sales, end-of-period units.*

The number θ gives the portfolio proportion to be chosen implicitly through sales S and purchases P that are adjustments to the initial holding θ^0. In practice, we could add nonnegativity constraints for portfolio positions, a no-borrowing constraint or constraints on proportions in the stock, or risk limits.

In the utility function, the first three terms are standard for mean-variance optimization. The first two terms r and $\theta(\mu - r)$ give the expected return; r is the risk-free rate of return we would obtain from just holding the risk-free asset, and $\theta(\mu - r)$ is the net change in expected return from holding a nontrivial risky portfolio. The third term $-(\lambda/2)(\theta\sigma)^2$ is the disutility of taking on variance. The constant $\lambda > 0$ is the coefficient of risk aversion; the larger the value of λ, the more reluctant the investor is to take on risk

in exchange for return, and $(\theta\sigma)^2$ is the portfolio's variance. Including 2 in the denominator makes the units the same as absolute risk aversion in a multivariate normal model with exponential utility over returns, and also cancels when we look at the first-order conditions.

The fourth term $-(\kappa/2)((\theta-\theta^B)\sigma)^2$ is a penalty for tracking error. This term is probably controversial for academics because it depends on the benchmark θ^B and not just on the distribution of returns. The dependence on the benchmark would be unnecessary and probably damaging in an ideal world, but does arise in practice and should be familiar to practitioners. One theoretical argument for using deviations from the benchmark, but without the mean term, is that we can write mean-variance preferences (excluding costs) as a constant (the utility at the no-cost optimum) plus λ times the variance of the deviation from the no-cost optimum. If the benchmark is the no-cost optimum, this is just a penalty for deviations from the benchmark. If the benchmark is uninformed (as is typical) but the manager has (or believes he has) information about expected returns, there would be the benchmark term plus an "alpha" term giving returns in excess of the uninformed returns. Another argument for penalizing deviations from the benchmark is that this portfolio might be just a piece (say large-cap equities) of a larger portfolio, and the overall portfolio manager does not want this sub-portfolio to deviate too far from its intended role in the whole portfolio. If the benchmark penalty is not desired, the analysis can be conducted with $\kappa = 0$ and $\lambda > 0$.

The final two terms represent the transaction costs. There can be different costs for purchasing and selling.[6] Costs are paid at end-of-period (or equivalently are converted to future values). Since utility is also measured in end-of-period return units, marginal utilities and costs are in identical units. Problem 1 assumes variable costs; we also consider fixed costs.

As in traditional mean-variance analysis, we can map out the frontier either by varying the parameters in Problem 1 or by fixing mean or variance and solving for the other: the first-order conditions are the same. Maximizing expected return subject to variance or tracking error may be especially at-

[6]Actually, having different costs for purchasing and selling is an unnecessary luxury, since we can always make the two equal by redefining the purchase price to be the midpoint between the all-in cost of purchasing and the net proceeds from selling.

tractive to practitioners who have trouble quantifying their preferences but may have an idea how much volatility or tracking error is acceptable.

Solution of Problem 1

The solution below to Problem 1 is characterized by a non-trading region $[\underline{\theta}, \overline{\theta}]$. If our allocation θ^0 is within the non-trading region (i.e. if $\underline{\theta} \leq \theta^0 \leq \overline{\theta}$), then the cost of trading is not justified, and consequently $\theta = \theta^0$ and $P = S = 0$. Outside the non-trading region it is optimal to trade to the boundary of the region. For $\theta^0 < \underline{\theta}$, it is optimal to trade to $\underline{\theta}$ and we have $P = \underline{\theta} - \theta^0$ and $S = 0$. Similarly, for $\theta^0 > \overline{\theta}$, it is optimal to trade to $\overline{\theta}$ and we have $S = \theta^0 - \overline{\theta}$ and $P = 0$. It is never optimal to have $P > 0$ and $S > 0$ at the same time, since that means incurring both C^P and C^S on the round-trip. Instead, it is always better to execute the net transaction.

Having a non-trading region (instead of a single ideal point) follows because locally the cost of trading (which is linear or first-order) is larger than the cost of having the wrong portfolio (which is quadratic or second-order). It is conceivable that there could be second-order costs of trading (for example from price pressure not modeled in this article), but there are probably also proportional costs (e.g. from a bid-ask spread or fixed fee) and perhaps fixed costs, both of which would be more important than the benefits of rebalancing nearby the ideal point.

It is not hard to compute the endpoints $\underline{\theta}$ and $\overline{\theta}$ that characterize the non-trading region. Leaving the calculations to the Appendix, here are the formulas. The ideal portfolio θ^* that would be chosen in the absence of costs is

$$(2) \quad \theta^* = \frac{\lambda}{\lambda + \kappa} \frac{\mu - r}{\lambda \sigma^2} + \frac{\kappa}{\lambda + \kappa} \theta^B,$$

which is a weighted average of the traditional mean-variance portfolio $(\mu - r)/(\lambda\sigma^2)$ that would be chosen for $\kappa = 0$ and the benchmark portfolio θ^B. The distance from this ideal portfolio θ^* to the non-trading boundary is

proportional to the cost of trading:

$$(3) \quad \underline{\theta} = \theta^* - \frac{C^P}{(\lambda+\kappa)\sigma^2}$$

$$(4) \quad \overline{\theta} = \theta^* + \frac{C^S}{(\lambda+\kappa)\sigma^2}.$$

This is different from typical continuous-time models in which the size of the region is more than proportional to costs when costs are small (so that even a small cost induces a significant non-trading region), as noted by Constantinides [1986]. In continuous-time models, it is not so urgent to trade immediately because it is possible to wait and perhaps a market move will do the trade for free. If not, then the trade will still be available soon (in the continuous-time model but not in a single-period model). At the optimum, the impact of transaction costs on optimal utility is of first-order in a single-period model, which is different from traditional continuous-time models but similar to models with jumps (or perhaps more-to-the-point first-order changes in the ideal portfolio weights) as in Jang, Koo, Liu, and Loewenstein [2004].

It is interesting to consider transaction costs that are fixed rather than proportional to the size of the trade. The following choice problem is an example of a problem with fixed costs.

Problem 2 *(fixed cost, single risky asset) Choose a quantity θ of the risky asset to maximize the utility function:*

$$(5) \quad r + \theta(\mu - r) - \frac{\lambda}{2}(\theta\sigma)^2 - \frac{\kappa}{2}((\theta - \theta^B)\sigma)^2 - \iota(\theta \neq \theta^0)C.$$

with the same notation as in Problem 1 except for

$C > 0$*: fixed cost incurred if trading*
$\iota(\theta \neq \theta^0)$*: indicator which is* 1 *if trade occurs and* 0 *if not.*

As in the variable-cost problem, the fixed cost problem also has a non-trading region $[\underline{\theta}^f, \overline{\theta}^f]$ within which it is optimal not to trade. However, trading from outside the region is not to the boundary: given that we are incurring the fixed cost, we may as well trade all the way to the ideal point θ^*. If $\theta < \underline{\theta}^f$, trading to θ^* corresponds to $P = \theta^* - \theta^0$ and $S = 0$, while if $\theta > \overline{\theta}^f$, trading to θ^* corresponds to $S = \theta^0 - \theta^*$ and $P = 0$. The ideal portfolio is still given by (2), and the non-trading boundaries are given by

$$\underline{\theta}^f = \theta^* - \frac{1}{\sigma}\sqrt{\frac{2C}{\lambda+\kappa}} \quad (6)$$

$$\overline{\theta}^f = \theta^* + \frac{1}{\sigma}\sqrt{\frac{2C}{\lambda+\kappa}} \quad (7)$$

In this case, the size of the non-trading region is more than proportional to costs when costs are small. This is because the cost of being away from the ideal portfolio is second-order (quadratic), while the benefit is zero-order (constant).

If there are both fixed and proportional costs of trading, then any trading will typically be to the interior of the non-trading region (due to the fixed cost), although to different positions depending on the starting position (due to the variable part of the cost). It is also possible to use the same sort of analysis as in this paper to consider models in which there is a per-security cost of holding assets (e.g. the cost of due diligence to follow news in the company).

III Conclusion

We have used a mean-variance analysis of portfolio rebalancing given transaction costs to illustrate a number of important economic features in a context that is simple to understand and solve completely. The single-period case is suggestive of good strategies in more realistic cases, and is a useful benchmark for comparisons.

Appendix

Derivation of Single-Risky-Asset Results

The solution of the single-period results in Section II can be derived using the Kuhn-Tucker conditions (the first-order conditions in the presence of inequality constraints), but given the quadratic objective function it is possible (and probably more instructive) to analyze the solution using algebra by completing the square in the objective function.

The ideal portfolio maximizes the objective (1) absent costs, which is

$$(8) \quad r + \theta(\mu - r) - \frac{\lambda}{2}\theta^2\sigma^2 - \frac{\kappa}{2}(\theta - \theta^B)^2\sigma^2.$$

Simple algebra shows that this is the same as

$$(9) \quad r + \frac{\lambda + \kappa}{2}\theta^{*2}\sigma^2 - \frac{\kappa}{2}\theta^{B2}\sigma^2 - \frac{\lambda + \kappa}{2}(\theta - \theta^*)^2\sigma^2,$$

where θ^* is given by (2). In this expression, everything is constant except the last term, which is 0 if $\theta = \theta^*$ and negative otherwise. Therefore, the objective absent costs is maximized if $\theta = \theta^*$.

Simultaneously buying and selling (choosing both $S > 0$ and $P > 0$) can never be optimal because making the net trade moves to the same portfolio but has lower cost. Specifically, switching to $P_{new} = P - \min P, S$ and $S_{new} = S - \min P, S$ gives the same portfolio $\theta_{new} = \theta^0 + P_{new} - S_{new} = \theta^0 + (P - \min(P, S)) - (S - \min(P, S)) = \theta^0 + P - S = \theta$. And, it reduces the cost (and increases the utility (1)) by $C^P P + C^S S - (C^P(P - \min(P, S)) + C^S(S - \min(P, S))) = (C^P + C^S)\min(P, S)$. This shows it is never optimal to buy and sell at the same time.

If purchasing but not selling, then $S = 0$ and since $\theta = \theta^0 + P - S$, we can

infer $P = \theta - \theta^0$, which allows us to rewrite the utility (1) as

$$(10)\quad r + \theta(\mu - r) - \frac{\lambda}{2}\theta^2\sigma^2 - \frac{\kappa}{2}(\theta - \theta^B)^2\sigma^2 - C^P(\theta - \theta^0).$$

Simple algebra shows this is the same as

$$(11)\quad r + C^P\theta^0 + \frac{\lambda + \kappa}{2}\underline{\theta}^2\sigma^2 - \frac{\kappa}{2}\theta^{B2}\sigma^2 - \frac{\lambda + \kappa}{2}(\theta - \underline{\theta})^2\sigma^2,$$

where $\underline{\theta}$ is given by (3).

General Model

Here we consider a general analysis that allows multiple assets, futures (which are purchased at zero price so payoff per dollar invested is not defined), trade in bundles, trade through cash when cash is not a reasonable investment (e.g. when it is dominated by another fixed-income investment), and security swaps. The approach is to think about every potential net trade as an activity to be undertaken at any nonnegative intensity.

Problem 3 *(proportional costs, general trading activities) Choose a vector $q \geq 0$ of intensities of net trade activities to maximize the utility function:*

$$\theta'\mu - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta^B)'\mathbf{V}(\theta - \theta^B)$$

where $\theta = \theta^0 + \mathbf{T}'q$ is the vector of asset holdings after trade, and where we use the following notation:

q: vector of trading activity quantities
$\mathbf{T}$: matrix transforming trading quantities into asset position changes
μ: vector of expected security payoffs
λ: absolute risk aversion parameter

κ: *tracking error parameter*
$\mathbf{V}$: *covariance matrix of returns*
$\theta'\mathbf{V}\theta$: *variance of the portfolio return*
$(\theta - \theta^B)'\mathbf{V}(\theta - \theta^B)$: *variance of deviations from the benchmark*
θ^0: *initial asset holdings*
θ^B: *benchmark portfolio.*

This notation works in payoffs rather than returns, and the riskless asset, if any, is treated as just another security. Net trades are represented as rows of the trading opportunities matrix $\mathbf{T}$. For example, if we have variable costs with trading through cash, we can let the first element of θ be the cash position. There would be a separate row of $\mathbf{T}$ for purchase and sale of each other asset. The row of $\mathbf{T}$ corresponding to the activity for buying security i would have a positive entry for the number of shares (or other units) of the security and a negative entry in the first column for the cash paid. The ratio of the cash paid to the number of shares would be the all-in price including the quoted price plus any transaction costs. A separate row of $\mathbf{T}$ corresponding to the activity of selling security i would have entries of the opposite sign and a different trade-off.

For example, we might let

$$\text{(12)} \quad \mathbf{T} = \begin{bmatrix} -51 & 1 & 0 \\ 49 & -1 & 0 \\ -26 & 0 & 1 \\ 24 & 0 & -1 \end{bmatrix}$$

where the first security is cash, indicating the second security has price 50 with a transaction cost of 1 for buying or selling and the third security has price 25 and also a transaction cost of 1 for buying or selling. Then we might set

$$\text{(13)} \quad \mu = \begin{bmatrix} 1.01 \\ 55 \\ 28 \end{bmatrix}$$

giving a risk-free rate of 1%, a mean return of $10\% = (55 - 50)/50$ on the second security, and a mean return of $12\% = (28 - 25)/25$ on the third asset. (Here mean returns are computed using the arguably arbitrary spread midpoint prices.) The covariance matrix might look like

$$(14) \quad \mathbf{V} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 400 & 100 \\ 0 & 100 & 100 \end{bmatrix}$$

which says the first asset is riskless, the other assets both have standard deviations of $40\% = \sqrt{400}/50 = \sqrt{100}/25$, and the two risky assets have correlation $50\% = 100/\sqrt{400 * 100}$.

First-order conditions The Kuhn-Tucker first-order conditions[7] are, for all activities i, that the marginal benefit M_i of each activity,

$$(15) \quad M_i \equiv \mathbf{T}_i(\mu - \lambda \mathbf{V}\theta - \kappa \mathbf{V}(\theta - \theta^B)),$$

(where $\mathbf{T}_i$ is the ith row of $\mathbf{T}$) is not positive,

$$(16) \quad M_i \leq 0,$$

and is subject to the complementarity slackness condition,

$$(17) \quad q_i M_i = 0.$$

Existence of a solution depends on no-arbitrage conditions that say there is no riskless way of increasing final payoff. In this model, an arbitrage opportunity is a trade $q \geq 0$ with an increase in mean $q'\mathbf{T}\mu > 0$ and zero variance $q'\mathbf{T}\mathbf{V}\mathbf{T}'q = 0$.

[7]To derive the first-order conditions, substitute the definition $\theta \equiv \theta^0 + \mathbf{T}'q$ into the objective function of Problem 3 and apply the condition for "The Case of No Inequality Constraints" from Section 4.2 of Intriligator [1971].

Example: futures strategies To study the futures strategies, we assume the first security is a risk-free asset paying 1 for sure, the second security is investment in a portfolio of shares, and the third security is the futures contract. We let

$$\mu = \begin{bmatrix} 1 \\ 1+\mu^E \\ \mu^F - r \end{bmatrix}, \tag{18}$$

where μ^E is the mean return on equities and μ^F is the equity-equivalent mean return on futures (so $\mu - F - r$ is the mean futures return, which is the mean percentage change in futures price). The equity-equivalent mean futures return μ^F is the mean return on a synthetic equity portfolio investing \$1 in the risk-free asset and buying \$1 nominal exposure in futures.

The covariance matrix is assumed to be

$$\mathbf{V} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & {\sigma^E}^2 & \rho\sigma^E\sigma^F \\ 0 & \rho\sigma^E\sigma^F & {\sigma^F}^2 \end{bmatrix}, \tag{19}$$

where σ^E is the standard deviation of equity, σ^F is the standard deviation of futures, and ρ is their correlation.

The feasible trades are

$$\mathbf{T} = \begin{bmatrix} -1-r-c^E & 1 & 0 \\ 1+r-c^E & -1 & 0 \\ -c^F & 0 & 1 \\ -c^F & 0 & -1 \end{bmatrix}, \tag{20}$$

namely, purchase and sale of equities and purchase and sale of futures, respectively, where c^E is the proportional (end-of-period) cost of trading equities and c^F is the corresponding cost of trading futures.

For a futures strategy, we normally think that the correlation is high (ρ is close to one), so that futures trading is a close substitute for trading the underlying equities. We also usually believe that futures are much less expensive to trade the underlying, $c^F << c^E$, which is why it is appealing to consider substituting futures trades for trades in equities. The expected returns ("alphas") are not usually discussed much, but they turn out to be very important.

Bundles Trading and Trading Through Cash that is Not an Investment

The general activity model can be used to model trading in bundles. By trading a bundle, we mean that it is possible to make a trade in a portfolio of securities as if it is a single stock. Trading the bundle is a useful addition only if trading the bundle is less expensive than trading the individual shares, for example if we can buy one share of each of ten securities for less than ten times the cost of buying one share of each separately.

Consider the two-risky-security tableau of trades in two equities in (12). If we can trade in a portfolio of one share of the first risky asset and two shares of the second for a cost of \$2 per share, the tableau from (12) becomes

$$\text{(21)} \quad \mathbf{T} = \begin{bmatrix} -51 & 1 & 0 \\ 49 & -1 & 0 \\ -26 & 0 & 1 \\ 24 & 0 & -1 \\ -102 & 1 & 2 \\ 98 & -1 & -2 \end{bmatrix},$$

where the last two rows represent the bundle trades. In this case, the no-trade region would have two additional sides corresponding to the bundle trades.

The model of this section is also useful if we cannot hold cash in the portfolio and instead there is a risky fixed-income security. This might be because the risky fixed-income return has very little risk and superior return, or because the investment policy specifies a strict cash discipline. To see how this works,

consider the covariance matrix

$$(22)\quad \mathbf{V} = \begin{bmatrix} .005 & .25 & .125 \\ .25 & 400 & 100 \\ .125 & 100 & 100 \end{bmatrix},$$

which is similar to (14) except that the first security is now risky. The mean return might be

$$(23)\quad \mu = \begin{bmatrix} 1.02 \\ 55 \\ 28 \end{bmatrix}.$$

If purchase or sale of either risky security costs a dollar and purchase or sale of the risk-free asset cost half a percent, the activity matrix is

$$(24)\quad \mathbf{T} = \begin{bmatrix} -51 & 1 & 0 \\ 49 & -1 & 0 \\ -26 & 0 & 1 \\ 24 & 0 & -1 \\ -102 & 1 & 2 \\ 98 & -1 & -2 \end{bmatrix},$$

References

Bawa, Vijay S., Stephen J. Brown, and Roger W. Klein, 1979, Estimation Risk and Optimal Portfolio Choice, North-Holland: New York.

Brennan, Michael, 1975, The Optimal Number of Securities in a Risky Asset Portfolio When There are Fixed Costs of Transacting: Theory and Some Empirical Results, *Journal of Financial and Quantitative Analysis* **10**, 483–496.

Constantinides, George M., 1986, Capital Market Equilibrium with Transaction Costs, *Journal of Political Economy* **94**, 842–862.

Davis, M. H. A., and Norman, A. R., 1990, Portfolio Selection with Transaction Costs, *Mathematics of Operations Research* **15**, 676–713.

Donohue, Christopher, and Kenneth Yip, 2003, Optimal Portfolio Rebalancing with Transaction Costs: Improving on Calendar- or Volatility-Based Strategies, *Journal of Portfolio Management* 29, 49–63.

Dumas, Bernard, and Elisa Luciano, 1991, An Exact Solution to a Dynamic Portfolio Choice Problem under Transaction Costs, *Journal of Finance* **46**, 577–595.

Goldstein, Abraham, 1979, *A Model of Capital Asset Pricing with Transaction Costs*, Ph.D. dissertation, Yale University.

Grinold, Richard C. and Ronald N. Kahn, 1995, *Active Portfolio Management*, Probus: Chicago.

Intriligator, Michael D., 1971, *Mathematical Optimization and Economic Theory*, Prentice-Hall: Englewood Cliffs, NJ.

Jacob, Nancy L., 1974, *A Limited-Diversification Portfolio Selection model for the Small Investor*, *Journal of Finance* **29**, 847–856.

Jang, B.G., H.K. Koo, H. Liu, and M. Loewenstein, 2004, "Transaction Cost Can Have A First-Order Effect on Liquidity Premium", working paper, Washington University.

Leland, Hayne, 2000, Optimal portfolio implementation with transactions costs and capital gains taxes, working paper, University of California at Berkeley.

Liu, Hong, 2004, Optimal Consumption and Investment with Transaction Costs and Multiple Risky Assets, *Journal of Finance* **59**, 289–338.

Liu, Hong, and Mark Loewenstein, 2002, Optimal Portfolio Selection with

Transaction Costs and Finite Horizons, *Review of Financial Studies* **15**, 805–835.

Mao, James C. T., 1970, Essentials of Portfolio Diversification Strategy, *Journal of Finance* **25**, 1109–1121.

Mao, James C. T., 1971, Security Pricing in an Imperfect Capital Market, *Journal of Financial and Quantitative Analysis* **6**, 1105–1116.

Markowitz, Harry, 1952, Portfolio Selection, *Journal of Finance* **7**, 77–91.

Markowitz, Harry, 1959, *Portfolio Selection*, New York: John Wiley.

Masters, Seth J., 2003, Rebalancing: Establishing a Consistent Framework, *Journal of Portfolio Management* **29**, 52–57.

Mayshar, Joram, 1979, Transaction Costs in a Model of Capital Market Equilibrium, *Journal of Political Economy* **87**, 673–700.

Mayshar, Joram, 1981, Transaction Costs and the Pricing of Assets, *Journal of Finance* **36**, 583–597.

Pogue, G. A., 1970, An Extension of the Markowitz Portfolio Selection Problem to Include Variable Transactions' Costs, Short Sales, Leverage Policies and Taxes, *Journal of Finance* **25**, 1005–1027.

Taksar, Michael, Michael J. Klass, and David Assaf, 1988, A Diffusion Model for Optimal Portfolio Selection in the Presence of Brokerage Fees, *Mathematics of Operations Research* **13**, 277–294.

Tobin, James, 1958, Liquidity Preference as Behavior Towards Risk, *Review of Economic Studies* **25**, 65–86.