Chapter 5

Efficient Portfolio Selection

Benyounes Amjoun[2], Dan Calistrate[1], Myriam Caprioglio[3], Brenda Hawkins[4], Mounia Kjiri[3], Tamara Koziak[4], Vincent Lemaire[3], Jason McVean[5], Miro Powojowski[1], Bill Reed[6], Satoshi Tomoda[1], Julie Zhou[6].

Report prepared by Min Tsao[6],

with assistance from Rita Aggarwala[1], Hassan Aurag[3], and Marc Paulhus[1].

5.1 Introduction

The problem described in this report deals with finding an efficient or optimal portfolio from many possible portfolios each of which consists of a subset of a large set of projects. The problem was brought to the PIMS workshop by Merak Project Ltd. of Calgary, a developer of oil and gas software for the petroleum industry. If portfolios made up of a selection of petroleum projects are plotted on a graph of expected value versus risk, there is an upper boundary above which no portfolios are found. The portfolios on this boundary are said to be efficient. The collection of these efficient portfolios is known as the efficient frontier.

For a portfolio to be efficient, it must be the case that no other portfolio has more value while having the same or less risk, and there is no other portfolio that has less risk while having the same or more value. Harry Markowitz revolutionized the field of portfolio theory with his pioneering work in the 1950s (Markowitz, 1952, 1959). His approach to efficient frontier analysis uses matrix algebra to determine an analytical expression for the curve representing the efficient frontier.

Merak develops economic software for the petroleum industry. In applying efficient frontier theory to the realm of the petroleum industry, Merak has taken a different approach from the traditional Markowitz technique. Using previously generated Monte Carlo results randomly selected portfolios are generated and plotted on an efficient frontier graph. As more portfolios are plotted, it quickly becomes apparent that there is an upper boundary. The efficient frontier is, therefore, implied by the upper boundary, but a curve is not explicitly drawn.

Both of these approaches have strengths and weaknesses. For instance, Merak's approach lacks the analytical certainty regarding the efficiency of promising portfolios that the Markowitz approach has. Even though there may seem to be no portfolios above a particular portfolio on the graph, it is always possible that the next randomly generated portfolio will be better.

On the other hand, the Markowitz approach has some severe limitations when applied to the petroleum industry:

[1]University of Calgary
[2]Ocean and Coastal Environmental Sensing Inc.
[3]University of Montreal
[4]University of Alberta
[5]Merak Project Ltd.
[6]Univerity of Victoria

- The simplification of describing a risk profile with only a mean and variance will lead to inaccuracies in the efficient frontier.
- It may not be possible to participate in a project at an arbitrarily fine level of granularity. Some projects may be such that they require 100% investment or they cannot be done at all.
- The constraints that determine which portfolios are valid can be complicated in the petroleum industry. Constraints like "If A then also B, C, and D" or "If E then not F or " or "At least 2 of H, I, J, and K" cannot be easily expressed as a linear equation, which is required for the Markowitz approach.

Merak brought this problem to the workshop in the hope that we could find a way to address the weaknesses of the two approaches, potentially by combining or partially combining them. An efficient frontier analysis method that combines the robustness of the Monte Carlo approach with the confidence of the Markowitz approach would indeed be a powerful tool for any industry. However, it soon became clear to us that there are other ways to address the problem which do not require a Monte Carlo component. Members of our group formed three subgroups and each subgroup developed a different approach for solving the problem. The first is the Portfolio Selection Algorithm Approach where we try to develop a practical searching algorithm which will lead us to the efficient portfolio without having to examine each and every possible portfolio. The second approach is the Statistical Inference Approach where we discuss statistical estimation and inference of the efficient portfolio. This approach provides a solution to the weakness of the Monte Carlo method of Merak. The third approach is the Integer Programming Approach where we try to find the exact efficient portfolio by setting up the problem as an integer programming problem and then solving it.

The rest of this report is organized as follows: Sections 5.2, 5.3 and 5.4 cover the three approaches, respectively. Section 5.5 contains a short summary.

5.2 The Portfolio Selection Algorithm Approach

Members of the subgroup which developed this approach are Hassan Aurag, Myriam Caprioglio, Mounia Kjiri and Vincent Lemaire, all of the University of Montreal. This approach is motivated by the fact that exhaustive computation of portfolios cannot be undertaken when the number of projects exceeds 30. Thus we have to propose a selective method. In Merak's Monte Carlo method, portfolios are generated using a Monte Carlo technique. Since the portfolios are selected randomly, the best ones may not appear in the graph. Can we find an algorithm that would eventually select a representative set of portfolios such that none of the best portfolios are missed? In the following, we describe a portfolio selection algorithm for this purpose. We will explain the algorithm using an example data set.

5.2.1 The Data

The data consists of a set of 10 projects. For each project, we had two sets of numbers (of length 200 each) corresponding to Net Present Value (NPV) and Capital Investment (CI). The expected value and risk of a project are respectively the mean and standard deviation of the NPV's. The expected value and risk of a portfolio are respectively the mean and standard deviation of the sum of its component projects.

5.2.2 The Constraints

The first constraint is on total capital investment. Then one has to deal with some petroleum specific constraints, e.g., constraints 2-4 below. In all our tests, we used the following set of rules:

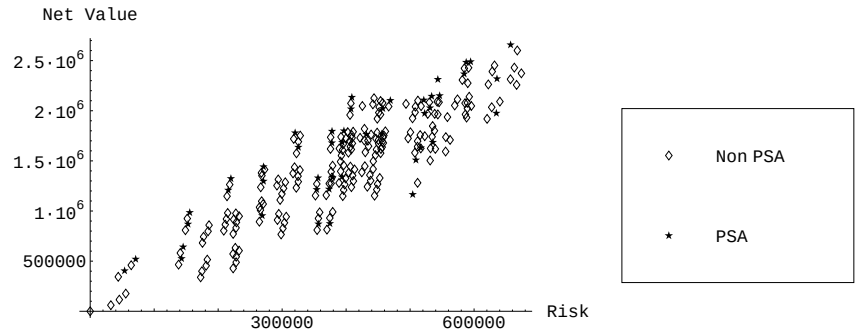
1. The maximum investment is $2,000,000.
2. Projects 1 and 4 may not belong to the same portfolio.
3. If a portfolio contains project 9, it must also contain one and only one of projects 7 and 10.
4. If a portfolio contains project 4 then it must contain project 8.

5.2.3 Portfolio Selection Algorithm (PSA)

The idea is to classify the set of all portfolios by the number of projects they contain. One then starts by picking a random project among the subset of best portfolios. At the next stage, you consider a sampling of portfolios containing two projects with the constraint that one of them must be the one we started with. In all subsequent stages, we keep adding a sampling of projects. Moreover, at each stage, we will require our portfolios to be valid and we only keep those considered best.

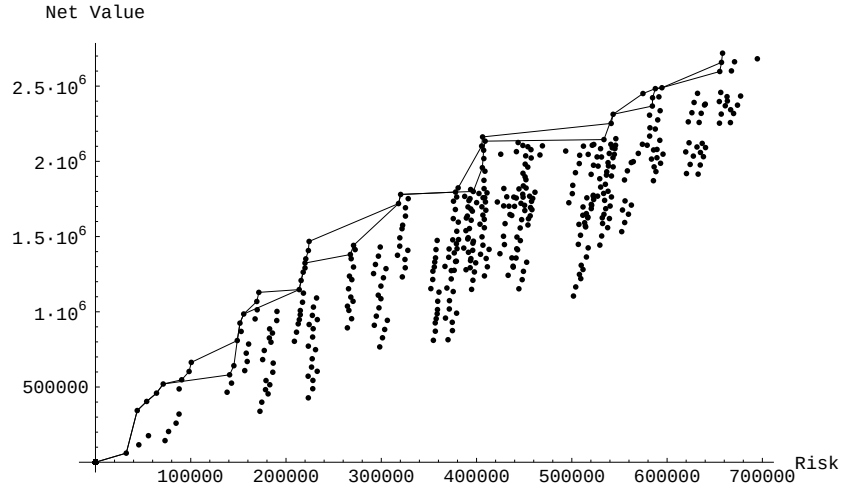
Notice that we cannot always compare two portfolios. Hence, in selecting the best portfolios we will only keep those portfolios that satisfy the following: Its NPV is greater than the NPV of all portfolios that have less or equal risk.

5.2.4 PSA Results



5.2.5 Results Analysis

Initial results show that adding boolean type constraints does not affect the efficient frontier very much. Using our approach, we have noticed that we didn't miss the best portfolios. The plot below further illustrates this point.



In the plot the dots represent all portfolios not all of which satisfy the boolean type constraints. One of the two lines represents the efficient frontier for all portfolios and the other represents that for those that satisfy the boolean type constraints. The two lines are very close to each other.

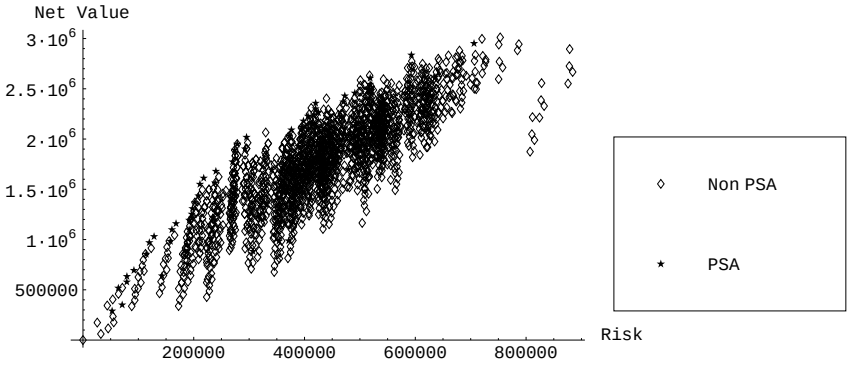
It should be noted that the data given to us by Merak was not strongly correlated.

5.2.6 Second Case: a 15 projects database

Here we used the original 10 projects and made up another 5 of the original ones using linear combinations, in order to obtain higher correlation between some of the projects. The correlation matrix of the set of 15 projects is a 15 by 15 matrix, too large to be shown in full here. The first 10 columns are:

$$\begin{pmatrix}
1.0000 & -0.0210 & 0.0290 & -0.0360 & 0.0610 & -0.1100 & -0.0067 & 0.0051 & 0.1200 & 0.0370 & \dots \\
-0.0210 & 1.0000 & -0.1200 & -0.0620 & 0.0022 & -0.0320 & 0.0190 & -0.0570 & 0.0640 & 0.0770 & \dots \\
0.0290 & -0.1200 & 1.0000 & 0.0720 & -0.0620 & 0.0940 & -0.0270 & -0.0980 & 0.0730 & -0.0860 & \dots \\
-0.0360 & -0.0620 & 0.0720 & 1.0000 & -0.1300 & 0.0066 & -0.1000 & -0.1700 & 0.0130 & -0.1300 & \dots \\
0.0610 & 0.0022 & -0.0620 & -0.1300 & 1.0000 & -0.1100 & -0.0340 & 0.1200 & -0.0071 & 0.2400 & \dots \\
-0.1100 & -0.0320 & 0.0940 & 0.0066 & -0.1100 & 1.0000 & 0.0290 & 0.0150 & -0.0460 & -0.0460 & \dots \\
-0.0067 & 0.0190 & -0.0270 & -0.1000 & -0.0340 & 0.0290 & 1.0000 & 0.0730 & 0.0420 & 0.1000 & \dots \\
0.0051 & -0.0570 & -0.0980 & -0.1700 & 0.1200 & 0.0150 & 0.0730 & 1.0000 & 0.0150 & 0.0047 & \dots \\
0.1200 & 0.0640 & 0.0730 & 0.0130 & -0.0071 & -0.0460 & 0.0420 & 0.0150 & 1.0000 & -0.0100 & \dots \\
0.0370 & 0.0770 & -0.0860 & -0.1300 & 0.2400 & -0.0460 & 0.1000 & 0.0047 & -0.0100 & 1.0000 & \dots \\
0.1400 & -0.0590 & -0.0930 & -0.1700 & 0.1300 & 0.0004 & 0.0720 & 0.9900 & 0.0320 & 0.0096 & \dots \\
-0.0220 & 0.9900 & -0.1200 & -0.0760 & -0.0024 & -0.0270 & 0.1500 & -0.0460 & 0.0680 & 0.0900 & \dots \\
-0.0960 & -0.0550 & 0.2900 & 0.0210 & -0.1200 & 0.9800 & 0.0230 & -0.0056 & -0.0300 & -0.0620 & \dots \\
0.0500 & -0.0150 & -0.0410 & 0.1500 & 0.9600 & -0.1100 & -0.0630 & 0.0740 & -0.0033 & 0.2000 & \dots \\
0.0340 & -0.0340 & -0.0140 & 0.4500 & 0.8200 & -0.0960 & -0.0890 & 0.0140 & 0.0013 & 0.1400 & \dots
\end{pmatrix}$$

Again, one obtains similar results as in the case of the 10 original projects. The graph below represents the results of our algorithm.



5.2.7 Conclusions

We would like, as a conclusion, to raise some questions and give our views regarding their answers.

1. When we have a large number of projects, we need to start choosing a random number of portfolios at each step. Do we still get the accuracy obtained in our case?

2. It is not necessary to start with 1-project portfolios. Can we for example start with r-projects portfolios where $r < N$, and N is the number of projects in the company's database?

3. Is it possible to merge this approach with Merak's current Monte-Carlo approach?

We believe the answers to these questions might be of great help in achieving better accuracy for the portfolio analysis software. In the short time we had for this study, we could not make all the necessary tests. Thus, we must say that the answer to the first question is unknown to us.

Depending on our goal, the answer to question 2 can be yes, if an approach such as ours is used to refine Merak's current model. We can easily imagine a situation where a *manager* would be able to select certain portfolios and ask for a refinement of the result using our PSA. In this case, we would just start from the given r.

Concerning our last point, if the answer to the first is negative, we believe we could still use our approach as an optional path for the end-user (e.g., to refine results), as proposed above. The user would supply the projects and we would build the portfolios using our approach. In this case, the user should be aware that no more than a certain number of projects can be selected; this number depending on the computational cost of our approach.

Finally, we would like to mention that all our simulations were done using Mathematica software. Even though Mathematica is a good software for simulation, it is certainly true that much faster code can be produced using C for instance. Mathematica was able to handle, exhaustively, up to 15 projects in 30 minutes without any optimization. In general, this means that one can go up to 20 projects, exhaustively, in less than 2 minutes, without counting expert programmer's optimizations.

5.3 The Statistical Inference Approach

Members of the subgroup which developed this approach are Rita Aggarwala of the University of Calgary, Brenda Hawkins and Tamara Koziak of the University of Alberta, and Bill Reed, Min Tsao and Julie Zhou of the University of Victoria. This approach explores the question of what constitutes *adequate sampling* of portfolio values in the sense that the client's *desired confidence* in obtaining a solution *close to* the true optimal portfolio value can be specified and satisfied if adequate sampling is carried out. The ideas of confidence and close to the true optimal will also be discussed more thoroughly.

Every portfolio is composed of a number of projects, each of which may be associated with various distributional and constraint assumptions. For example, a particular project may involve drilling an oil well. The results (which an attempt is usually made to quantify, for example, the net return) of such a drill will not be known prior to the drilling, however an idea of the various possibilities or result outcomes, may be adequately expressed by applying a distributional assumption to these possibilities. In addition, legal issues may dictate constraints on the drilling of the well. The total portfolio itself may also be subject to assumptions and constraints, such as budgetary and resource considerations, and/or restrictive relationships between possible projects.

Since each portfolio is usually composed of projects with uncertain returns, the return on any particular portfolio itself is also uncertain and therefore associated with some probability distribution. Present practice assigns each portfolio a *value*, based on the *expected* return of the portfolio, and a *risk* in choosing the portfolio, based on the differences between the *possible* returns for the portfolio and the expected return. These two quantities are calculated using the distributional assumptions and constraints discussed above, which are supplied by a client wishing to decide upon a suitable portfolio. As a general rule, portfolios with higher values tend to have higher corresponding risk.

As can be gathered from this preliminary discussion, there are quite a few assumptions and approximations involved in computing any single portfolio value and risk. Therefore, although it is very easy to become absorbed in the search for the theoretically absolute optimal portfolio, it should be kept in mind that even if the theoretical optimal were to be defined as in the following discussions, and found (which in most cases is either not possible, not feasible or both at present), it would be based on some approximated, albeit experience-based inputs given by the client. The conclusion that this portfolio is, in reality, *the* optimal one is almost certainly incorrect. However, given the information available is the best that we have, the ability to find the theoretically optimal portfolio, and perhaps others in that region, would be useful in making an intelligent decision. Based on this reasoning, being in the *region* of the theoretical optimum with high confidence is a very reasonable target for a client.

Let us suppose that there are M possible portfolios meeting the required constraints, and therefore M associated probability distributions, giving M value and risk pairs. Quite often, M is unknown, however we are usually able to put at least some rough constraints on M. In the simple case where there are k possible projects, and each project is either opted in to or opted out of, we can say that $M \leq 2^k$. In practice, some projects may also be opted in to at less than full commitment, for example some projects may allow commitment levels of 0%, 50%, or 100%. Constraints added to the portfolio will determine how many portfolios are eliminated from the collection represented by the upper bound on M. In situations where the bound on M is not too large for exhaustive enumeration of portfolio values and risks (and simulated probability densities for returns), exhaustive enumeration is the method taken, and statistical sampling and inference techniques need not be applied.

Notice that the collection of values (expected returns) itself can be seen as M measurements on the population of all M possible portfolios. Since each portfolio value is the expected value of some distribution which is arrived at through different (possibly continuous) distributional assumptions, constraints, etc, it is highly unlikely that any two of the M possible values will be the same. The standard procedure, in the case where the number of possible portfolios is too large to enumerate exhaustively and too complicated to attack analytically is to

randomly sample a pre-determined number m of unique portfolios from the population of possible portfolios and obtain their m corresponding values, that is, to obtain a random sample (without replacement) of size m from the probability mass function for portfolio values. The portfolio decided upon could then be the largest of the m sampled. In addition to the cardinality M of the probability mass function being unknown, the support of this probability mass function (the range of allowable portfolio values) is also unknown. Similar reasoning can be applied towards the collection of portfolio risks.

Our approach was to define an optimal portfolio as that which gave the maximum value or expected return for a specified *risk bin*. A risk bin is taken to be a range of quantified risk which the client is comfortable with and indifferent to. We will assume that N is the number of allowable portfolios in the specified risk bin, and that a *random* sample of n values can be attained from this risk bin. We will also assume that this random sampling approach will be applied in cases where N is prohibitively large to enumerate all portfolio values and risks in the specified risk bin exhaustively, and an analytic solution for the maximum value is not attainable. Thus, a statistical approach to deciding upon a suitable value of n given a desired confidence and range of risk is required. It should be noted that unless we are able to (internally) fit a suitable probability distribution to individual clients' samples of n portfolio values and use parametric inference from that point on, it is important to develop robust procedures.

The subgroup of investigators explored the following specific questions from a number of angles:

- What is the confidence associated with being in the top $100\alpha\%$ of possible portfolios values, if n portfolios are randomly sampled?
- Can a confidence interval be specified for the estimation of the theoretically optimal (highest possible) portfolio value?

5.3.1 Percentile Estimation

Robust techniques for percentile estimation have been explored in many books on order statistics, see for examples, David (1981) and Arnold, Balakrishnan and Nagaraja (1992). The application of these techniques will apply to this problem, as well, provided we assume that the random sample of size n of portfolio values from a risk bin is much smaller than the risk bin population N. (This is in order that we may justify the assumption of approximately sampling from an infinite population). This assumption is actually a conservative one, in that if it does not hold, the results of this subsection are generally stronger than stated.

A simple and robust probabilistic argument may be used to determine the probability β that the r^{th} largest value in the sample will be in the top $100\alpha\%$ of possible portfolio values based on a sample of size n, where α is generally taken to be a small proportion. Once the sample has been taken, the term *probability* must be replaced by *confidence level*. Conversely, for fixed r, α, and desired confidence level β, the required sample size n can be determined. The expressions given hold for continuous distributions of portfolio values. Since we are actually dealing with a discrete distribution of portfolio values, the probabilities β obtained here will be approximations. They will be very close approximations if it may be assumed that (at least the top $100\alpha\%$ of) the distribution of portfolio values can be approximated by a continuous distribution. Notice that this implies that the theoretical max does not "stand alone." This will be important in our discussion of approximating the actual maximum possible portfolio value in the next section. From examination of typical plots and discussions with the industrial mentor, these are not unreasonable assumptions. The probability β is given as follows. We will denote the i^{th} order statistic (ordered value) in the random sample of size n by $Y_{i:n}$.

$$\begin{aligned}
\beta &= P\left(\begin{array}{c}\text{at least } r \text{ of the } n \text{ observed values are in}\\ \text{the top } 100\alpha\% \text{ of possible portfolio values}\end{array}\right)\\
&= \sum_{i=r}^{n}\binom{n}{i}\alpha^i(1-\alpha)^{n-i}\\
&= 1-\sum_{i=0}^{r-1}\binom{n}{i}\alpha^i(1-\alpha)^{n-i}.
\end{aligned}$$

It is easily seen that for $r = 1$, where we are only interested in the probability of the largest observed value being in the top $100\alpha\%$ of possible values, the above expression for β reduces to $1 - (1 - \alpha)^n$. By specifying β and α it is also easy to determine the required sample size n in the case $r = 1$. For larger values of r, solving for n will involve polynomial root finding, which is handled with ease using any simple mathematical software tool. Recall that, in view that all calculated portfolio values are based on approximations themselves, it may be of greater practical interest to consider a few feasible portfolios in the top percentiles of possible values rather than a single one, and perhaps base the final selection of portfolios on other considerations such as convenience.

A table displaying selected values of n, r, α and β follows.

n	100	100	100	200	200	200	500	500	500
r	1	1	3	1	1	5	1	3	10
α	.01	.05	.05	.01	.05	.05	.01	.01	.05
β	.634	.994	.882	.866	1.00	.974	.993	.877	1.00

Thus, for example, if 200 points are sampled from the desired risk bin, we will have approximately 97% confidence that the top 5 observed portfolio values are within the top 5% of all possible portfolio values, and almost 100% confident that the highest observed portfolio value is within the top 5% of all possible portfolio values.

5.3.2 Efficient Frontier Estimation

The previous section on percentile estimation discussed robust methods for determining the probability of observing one or more portfolio values in the top $100\alpha\%$ of all possible values of portfolios, for a specified risk bin. However, the true maximum possible portfolio value was never assumed nor estimated. In fact, to use the techniques of the previous section, no true maximum needs to exist. In the present context, there is a theoretical true maximum portfolio value θ which the client desires to be at or close to. In this section we will explore point and interval estimation of the value θ based on a random sample of size n of the N possible portfolio values in a risk bin.

We will again assume that n is much smaller than N, for if this were not the case, all possible portfolios would be enumerated. We will also assume that the distribution of portfolio values can be approximated by some (perhaps piecewise disjoint) continuous distribution. Recall that this implies that the theoretical maximum does not "stand alone." Again, this seems to be a reasonable assumption, however, for this section, the concept of not standing alone is clarified and quantified.

Specified Intervals

In this approach, the client would want to know the probability that their maximum observed portfolio value $Y_{n:n}$ is within δ of the true maximum possible portfolio value, θ, where δ is a number (perhaps a percentage of the largest observed value) specified by the client. It should be noted that an appropriate sample size may be chosen by methods in the section on percentile estimation. The probability which we will estimate here will be easily computed after the sample has been observed, as a post-hoc analysis, and therefore will be viewed as a confidence level.

We would like to estimate $P(Y_{n:n} \in [\theta - \delta, \theta])$. Since we have a random sample of observations, the probability that any observed value lies in the interval $[\theta - \delta, \theta]$ is the same, say α. Therefore,

$$P(\text{at least one observation is in } [\theta - \delta, \theta]) = 1 - (1 - \alpha)^n .$$

Notice that this expression looks very similar to the binomial sum expression for $r = 1$ in the previous section on percentile estimation. However, in this case, α is not specified by the client, rather it is something which must be estimated from the data in order to arrive at an approximate probability. Assuming the distribution of values behaves similarly in $[Y_{n:n} - \delta, Y_{n:n}]$ and $[\theta - \delta, \theta]$, we may estimate α by

$$\frac{\text{number of observations in } [Y_{n:n} - \delta, Y_{n:n}]}{n}.$$

This approximation would be reasonable if the distribution of portfolio values were assumed to behave uniformly in $[Y_{n:n} - \delta, \theta]$. Choosing a sample size large enough that there is a reasonable probability of sampling a few values in the higher percentiles by using the techniques discussed in the previous section should elicit a good idea of the shape of the distribution, even in the tails, since the sampling is random.

Likelihood Intervals

The idea of likelihood intervals is explored, for example, in Kalbfleisch (Volume 2, 1985) and Royall (1997). In the present situation, likelihood point estimation is very intuitive, and therefore likelihood intervals are a natural way to estimate θ. The interpretation of likelihood intervals can be compared to that of traditional confidence intervals for large sample sizes, for example, the authors mentioned above discuss the near equivalence of a 14.7% likelihood interval with a 95% confidence interval, and a 3.6% likelihood interval with a 99% confidence interval for regular distributions from which large samples have been drawn. As a general rule, points inside a 10% likelihood interval are labeled as "plausible values" for the parameter θ, and points outside a 1% likelihood interval as "very implausible values." It is important to realize here that lower percentage likelihood intervals are more desirable, whereas with traditional confidence intervals, higher percentage confidence intervals are desired. This is simply due to the construction of the intervals.

In general, the Maximum Likelihood Estimate of a parameter is the value of the parameter which maximizes the likelihood function of the observed data. The likelihood function is simply the product of individual probability functions for the observed data when the data are a random sample. We will seek robust estimates and intervals, with minimal assumptions made on probability functions. Specifically, we will assume that the distribution of portfolio values can be approximated by a truncated (possibly interval piecewise) continuous distribution. Following discussions with the industrial mentor and examination of some sample data, it seems that it is quite common for the density of portfolio values to appear to be truncated at θ. This may be due to one or more of the constraints associated with a problem.

We assume the truncated distribution takes the following form:

$$\begin{aligned} f(x) &= \frac{g(x)}{G(\theta)},\ 0 \le x \le \theta \\ &= 0,\ \text{otherwise,} \end{aligned}$$

where $g(x) = \frac{d}{dx}G(x)$, and $G(x)$ is a valid, differentiable cumulative distribution function. The lower bound on the support need not be 0.

The likelihood function for a random sample of n values from this distribution is then

$$\begin{aligned} L(\theta) &= \frac{\prod_{i=1}^{n} g(y_i)}{[G(\theta)]^n},\ \theta \ge y_{n:n} \\ &= 0,\ \theta < y_{n:n}. \end{aligned}$$

Since $G(\theta)$ must be an increasing function, this likelihood will take its maximum value at $\theta = y_{n:n}$, and thus $y_{n:n}$ is the maximum likelihood estimate of θ here.

In general a $100\gamma\%$ likelihood interval for a parameter θ is given by

$$\{\theta : L(\theta) > \gamma L(\theta^*)\}$$

where θ^* is the maximum likelihood estimate of θ. A $100\gamma\%$ likelihood interval for θ in this problem is therefore

$$\left\{\theta : G(y_{n:n}) \le G(\theta) \le \gamma^{-1/n} G(y_{n:n})\right\}.$$

If one can assume a parametric form for $G(\theta)$ (for example, by internally fitting a truncated distribution to the observed random sample), one can solve this inequality. For example if the distribution of portfolio values can be assumed to be uniformly distributed on $(0, \theta)$, then the likelihood interval for θ is

$$\left\{\theta : y_{n:n} \le \theta \le \gamma^{-1/n} y_{n:n}\right\},$$

whereas if the distribution of portfolio values can be assumed to come from a truncated exponential distribution with mean $1/\lambda$, the likelihood interval is

$$\left\{\theta : y_{n:n} \leq \theta \leq -\ln\left[1-\gamma^{-1/n}\left(1-\exp\left(-\lambda y_{n:n}\right)\right)\right]/\lambda\right\}.$$

Notice that each of these examples assumes a tidy parametric form for the density $f(x) = g(x)/G(\theta)$ over the entire range of feasible portfolio values, $0 \leq x \leq \theta$. It is possible that the data would allow such a density to be fit to portfolio values, however, we will give a few ideas in the likely event that this is not the case.

Firstly, a Taylor's series expansion of $\ln G(\theta)$ about $\theta = y_{n:n}$ will give us the interval

$$\left\{\theta : \frac{g(y_{n:n})}{G(y_{n:n})}(\theta - y_{n:n}) + \cdots \leq -\frac{1}{n}\ln\gamma\right\}.$$

One may proceed to obtain the first order likelihood interval for θ

$$\left\{\theta : y_{n:n} \leq \theta \leq y_{n:n} - \frac{\ln\gamma}{nf(y_{n:n})}\right\}$$

where a non-parametric estimate of the density $f(\cdot)$ at $y_{n:n}$ may be substituted for $f(y_{n:n})$ provided n is large enough to obtain a good estimate of this value. Similarly, a second order likelihood interval may be obtained by solving the quadratic equation arising from the Taylor's series expansion above. This will involve non-parametric estimates of $f(\cdot)$ and $f'(\cdot)$ at $y_{n:n}$.

Another possible approach is to argue that since the likelihood function is non-zero only for $\theta \geq y_{n:n}$, this is the region in which a parametric form for $G(\cdot)$ (and as a result, $f(\cdot)$) will be needed in obtaining the desired interval. Unfortunately, we have no data in this interval! If we consider using a few, say r of the upper observed ordered values in order to estimate the shape of $G(\theta)$ in that region, it will be sufficient to assume that the *upper tail* of the distribution, from $Y_{n-r:n}$ to θ can be approximated by a continuous distribution on one interval segment, and the sample size of observed values should be chosen large enough that at least r values of the n sampled will be in this upper tail with high probability (the more robust techniques of the previous section may be used to determine a large enough sample size for this latter condition; If, for example, it is felt that it is safe to assume the top $100\alpha\%$ percent of values can be approximated by a continuous distribution, a sample size may be chosen so that the corresponding probability β discussed in the previous section on percentile estimation is high for some reasonable value of r). Then, the assumed parametric form for $G(\cdot)$ in this tail region, which includes $\theta \geq y_{n:n}$ may be used to solve the likelihood interval

$$\left\{\theta : G(y_{n:n}) \leq G(\theta) \leq \gamma^{-1/n}G(y_{n:n})\right\}.$$

This approach may be most reasonable if, for example, the sample of portfolio values does not seem to be continuous on one interval, but rather over interval segments. The industrial mentor did indicate that quite often, data is observed in "clumps".

Finally, it should be noted that if parametric forms are assumed for the probability density of portfolio values, classical confidence intervals may also be explored for θ. Care should be taken that these confidence intervals make sense, in that they do not include values of $\theta < y_{n:n}$. As one can see from the likelihood interval above, this is not a concern for the likelihood intervals discussed here, since $G(\cdot)$ is an increasing function.

We close with the following example: suppose it is reasonable to assume that $f(\cdot)$ can be approximated by a uniform distribution on $[Y_{n-r:n}, \theta]$ with high probability. Then $G(\cdot)$ will be linear in this interval. This could be quite a reasonable assumption for large enough n. Thus, $G(x) = ax + b$ and the likelihood interval is given by

$$\left\{\theta : \theta \leq \alpha^{-1/n}y_{n:n} + \frac{b}{a}\left(\alpha^{-1/n} - 1\right)\right\}.$$

The quantity $\frac{b}{a}$ may be approximated by drawing a line of best fit through the largest r values of the empirical cumulative distribution function for $f(\cdot)$. Thus, since $f(x) = g(x)/G(\theta)$, and $F(x) = G(x)/G(\theta)$, the slope

of the line will be an estimate for $a/G(\theta)$ and the intercept will be an estimate for $b/G(\theta)$, and an estimate of b/a can be obtained using the ratio of the fitted intercept to the fitted slope.

An interesting observation arising from this example can be made. If we do assume a uniform (or even some other) distribution of portfolio values in the upper tail of the density, we may consider the following: a useful property of order statistics is that given $Y_{n-r:n} = y_{n-r:n}$, the remaining larger order statistics $Y_{n-r+1:n}, \ldots, Y_{n:n}$ behave as a random sample of size r from the same distribution (uniform or some other) left truncated at $y_{n-r:n}$. A likelihood interval may then be determined directly for θ from first principles, but only using the r values in the tail as discussed earlier. Notice that this approach does not require the entire distribution of portfolio values to be of a truncated form. In the case of the uniform assumption for the upper tail, the likelihood interval for θ becomes

$$\left\{\theta : y_{n:n} \leq \theta \leq \frac{y_{n-m:n} - \alpha^{-1/r} y_{n:n}}{1 - \alpha^{-1/r}}\right\}.$$

For the uniform distribution tail assumption, a traditional confidence interval for θ is also easily obtained by using the fact that order statistics from uniform distributions behave as Beta random variables. A $100\,(1-\alpha)\,\%$ upper confidence interval for θ in this case is given by

$$\left\{\theta : y_{n:n} \leq \theta \leq y_{n-r:n} + \frac{y_{n:n} - y_{n-r:n}}{(1-\alpha)^{1/r}}\right\}.$$

The advantage in both of these cases is that the form of the upper bound on θ is very simple. A disadvantage is that the number of values r is very important in the resulting width of the interval, and a larger sample size n will be necessary to ensure observation of enough of these tail values with high probability.

Deciding upon the best method to employ (or if other methods should be sought) should involve experiments and simulations with typical projects and portfolios.

5.4 The Integer Programming Approach

Members of the subgroup which developed this approach are Benyounes Amjoun of the Ocean and Coastal Environmental Sensing Inc. and Marc Paulhus, Miro Powojowski and Satoshi Tomoda of the University of Calgary. This approach provides a rigorous method for solving a large class of portfolio selection problems. We now give a detailed description of this approach, beginning with some background material.

The problem is to find an efficient portfolio of projects under some constraints. If risk is measured by a single parameter, and the company is assumed to be *rational* (prefer more wealth to less wealth and less risk to more risk) then "efficient" is easy to define. If we plot all available portfolios on a standard expected return versus risk graph, then a portfolio P is *efficient* if there are no portfolios both above and to the left of P. See Figure 5.1.

We will assume that there are n projects which are available to a company, which we will label $0 \ldots n-1$. The company is not free to enter these projects at any level of granularity, indeed, we shall assume that a company is either invested in a project or not invested in a project (in Section 5.4.1 we explain how to relax this slightly). Hence, for project $i \in \{0, \ldots, n-1\}$ we can associate the variable x_i, such that

$$x_i = \begin{cases} 0 & \text{if project } i \text{ is not included} \\ 1 & \text{if project } i \text{ is included.} \end{cases}$$

Thus the vector $X = (x_0, \ldots, x_{n-1})'$ is a vector of zeros and ones which defines a portfolio.

Further associate with project i:

- a dollar cost c_i (let $C = (c_0, \ldots, c_{n-1})'$),
- an expected net return μ_i (let $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_{n-1})'$).

In our approach the project costs are taken to be fixed. The generalization of our approach where costs are stochastic might be found in the literature on *stochastic programming*. The interested reader could start with [4].

Let $\Sigma = (\sigma_{ij})$ be the covariance matrix, that is σ_{ij} is the covariance of projects i and j and σ_{ii} is the variance (not the standard deviation) of the return of project i. We will make use of the fact that Σ is non-negative definite.

Our method requires two assumptions:

1. The measure for risk is the standard deviation,
2. All the constraints are expressible as linear or parabolic equations.

The next section will show how the second assumption might not be too restrictive.

5.4.1 Constraints

In the last section we mentioned that for our approach to work all the constraints would have to be expressible as linear or parabolic equations.

Naturally there will be a budgetary constraint, perhaps the company must spend less than M_h dollars and more than M_l on this portfolio. Hence

$$M_l \leq X'C \leq M_h.$$

There also might be other constraints such as the choice of one project forbids the option to choose another project. Many of these types of constraints can be written linearly. For example:

- "If project a then also projects b, c and d" can be described as

$$3x_a \leq x_b + x_c + x_d.$$

- "If project e then not projects f or g" can be described as

$$x_f + x_g \leq 2 - 2x_e.$$

- "At least two of h, i, j and k" can be described as

$$x_h + x_i + x_j + x_k \geq 2.$$

- "If project l then not both projects m and n" can be described as

$$x_m + x_n \leq 2 - x_l.$$

Of course it is possible to construct constraints which cannot be written as a linear equation, for example

- "Exactly one or exactly three of projects a, b and c."

But it is simple enough to consider the two different feasible regions corresponding to

- "Exactly one of projects a, b and c" which can be described as

$$x_a + x_b + x_c = 1,$$

- "Exactly three of projects a, b and c" which can be described as

$$x_a + x_b + x_c = 3.$$

Then solve the problem over each of these regions and compare the solutions.

In the introduction we stated that the company can either be fully invested in a project or not invested at all. In reality this might not be true. Perhaps, for a particular project A for example, not only can the company be either completely in the project or completely out of the project, but they might be able to invest in $1/2$, $1/3$ or $2/3$ of the project. In this case we define four new projects:

- Project a: invest in $1/2$ of project A.
- Project b: invest in $1/3$ of project A.
- Project c: invest in $1/3$ of project A.
- Project d: invest in $1/3$ of project A.

Note that projects b,c and d are identical. The constraint is

$$x_b + x_c + x_d \leq 3 - 3x_a.$$

Hence, if we choose to invest in project a ($1/2$ investment in A) then we are forbidden to invest in any of projects b, c and d. Otherwise we are free to invest exactly one of projects b, c and d ($1/3$ investment in A), exactly two of projects b, c and d ($2/3$ investment in A) or invest in all three of projects b, c and d (100% investment in A). Perhaps an interesting generalization would be the corresponding mixed-integer problem where some projects can be included at a continuum of levels.

Define S to be the set of all the constraints. In what follows we assume that S consists of only linear and parabolic equations.

A portfolio X will be called *feasible* if it satisfies all of the constraints in S. Plotting all the feasible portfolios on a graph of expected return versus risk we would get a graph similar to Figure 5.1. Note that due to the finite granularity constraint there will only be a finite number of possible portfolios. Hence, unlike the traditional Markowitz portfolio problems [9] where a continuous "efficient frontier" is expected, the efficient and feasible portfolios from this problem will form a set of "efficient fenceposts". In Figure 5.1 the efficient portfolios (fenceposts) are shown as x's and the dominated portfolios are shown as o's. One portfolio, labeled P_0, is the global minimum for risk and is an efficient fencepost.

Under the assumption that the standard deviation is the measure of risk we can locate point P_0 simply by minimizing the objective function $X'\Sigma X$ under the constraint set S. The next section will describe how to do this. Section 5.4.3 will describe how to locate the other fenceposts.

5.4.2 The Integer Programming Problem

The problem is to find the X which minimizes $X'\Sigma X$ subject to a set S of linear and parabolic equations. Following the outline of a method described in [5] we will describe how to transform this quadratic problem into an equivalent problem which can be solved.

Definition 5.4.1 *A parabolic constraint of rank k is one which can be put into the form*

$$a_{00} - L_0(X) - b_1(L_1(X))^2 - \cdots - b_k(L_k(X))^2 \geq 0,$$

where

$$L_s(X) = a_{s1}x_1 + \cdots + a_{sn}x_n, \qquad s = 0, 1, \cdots, k$$

are a set of $k+1$ linearly independent homogeneous linear forms of n variables and

$$b_i \geq 0, \qquad i = 1, \cdots, k.$$

What we do is transform the objective function $X'\Sigma X$ into a new objective function z and add a parabolic constraint

$$z - X'\Sigma X \geq 0$$

to S. The variable z is called a slack variable. This means that we have to express $z - X'\Sigma X \geq 0$ in the form

$$a_{00} - L_0(X) - b_1(L_1(X))^2 - \cdots - b_k(L_k(X))^2 \geq 0$$

as in Definition 5.4.1. To this end, consider

$$A = \sigma_{00}x_0^2 + 2\sigma_{01}x_0x_1 + \cdots + 2\sigma_{0n-1}x_0x_{n-1}$$

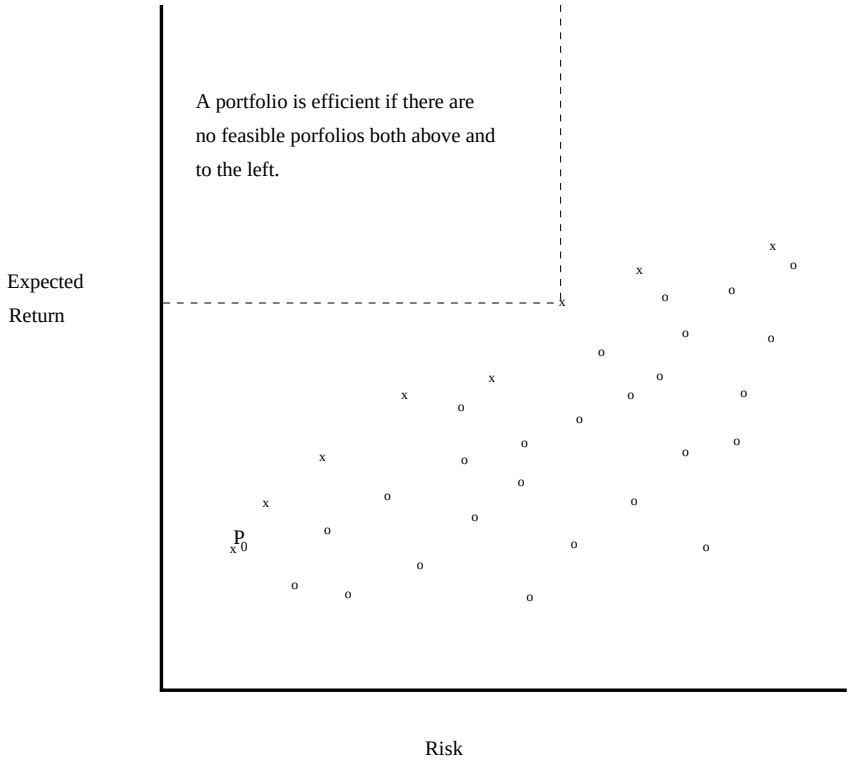


Figure 5.1: A plot of all feasible solutions. Efficient fenceposts are shown as x's. Dominated portfolios are shown as o's. The fencepost corresponding to the globally minimal risk is labeled P_0.

which is all the terms with the variable x_0 in the expansion of $X'\Sigma X$ (Note that the covariance matrix Σ is symmetric). We can rewrite A as

$$A = \frac{1}{\sigma_{00}}(\alpha - (\beta + \gamma))$$

by completing the square, where

$$\begin{aligned} \alpha &= (\sum_{j=0}^{n-1} \sigma_{0j} x_j)^2, \\ \beta &= \sum_{j=1}^{n-1} (\sigma_{0j} x_j)^2, \text{ and} \\ \gamma &= \sum_{1 \le i < j}^{n-1} 2\sigma_{0i}\sigma_{0j} x_i x_j. \end{aligned}$$

Note that $X'\Sigma X - A$, β and γ have no x_0 terms. Thus, since Σ is non-negative definite, $X'\Sigma X$ is a non-negative definite quadratic form (let k be its rank) and therefore the resulting expression $X'\Sigma X - A - \frac{1}{\sigma_{00}}(\beta + \gamma)$ is a non-negative definite quadratic form (with rank $k-1$). We can repeat this process of completing the squares for each variable. Clearly, the resulting expression $z - (\sum_{0 \le i < j}^{n-1} a_{ij} x_j)^2 \ge 0$ satisfies the conditions of the parabolic constraint defined in Definition 5.4.1. After this transformation, our problem can be stated as follows:

$$\min z$$

subject to the original constraint set, S, described in Section 5.4.1 together with the parabolic constraint

$$z - (\sum_{0 \le i < j}^{n-1} a_{ij} x_j)^2 \ge 0$$

where a_{ij} is an appropriate coefficient derived from completing the square for the variable x_j.

The algorithm to solve this transformed problem is rather lengthy and shall be omitted. The interested reader is invited to explore [5], page 277. A number of commercial software packages exist which should solve the transformed problem. An example might be the "Professional Linear Programming System" available from Sunset Software Technology[7] [11].

5.4.3 Finding the Efficient Fenceposts

In the last section we described how to locate the efficient fencepost P_0. We are left with the task of locating the other fenceposts. Associated with each fencepost $P_0 \dots P_m$ there exists an associated expected return $r_0 \dots r_m$ as shown in Figure 5.2. Once we have located P_0 we know the value of r_0. Thus, if we can minimize the objective function $X'\Sigma X$ under the constraint set S plus a further constraint $X'\boldsymbol{\mu} \ge r_0 + \epsilon$ (ϵ is less than a penny and is included to insure our feasible space is closed), then the solution will be P_1. Since the new constraint is linear, the process described in Section 5.4.2 can be used. Iterating this process (until our algorithm returns a "no solution" result) will locate all the efficient fenceposts and hence solve the problem given.

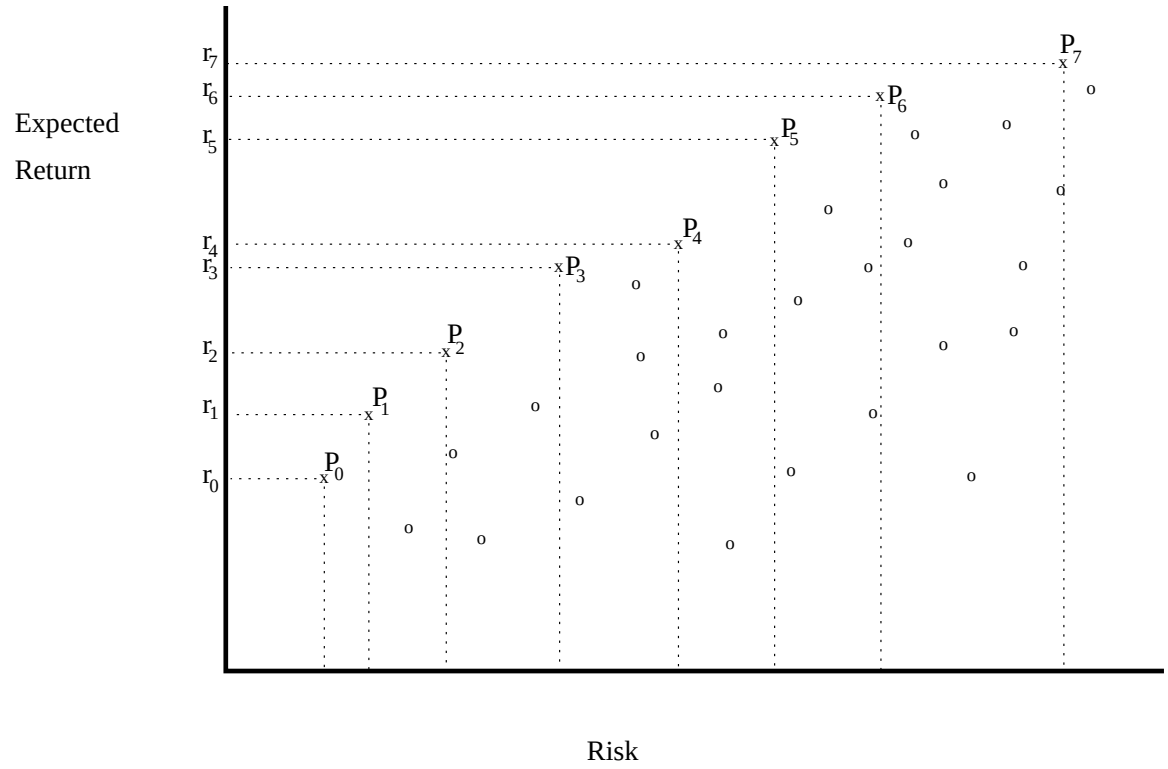


Figure 5.2: Associated with each efficient fencepost $P_0 \dots P_7$ is an expected return $r_0 \dots r_7$.

5.4.4 Conclusion

We have presented a rigorous method for solving the portfolio problem under some assumptions.

[7] www.sunsetsoft.com

The assumption that risk is measured by standard deviation can not be easily relaxed. Although the standard deviation is a common measure for risk, for some reasonable utility functions and non-normal payoff distributions the standard deviation might not be a good measure for comparing portfolios. There are many advantages to using a single parameter risk measure, but, in the case where standard deviation is thought to be inadequate perhaps the concept of *stochastic dominance* should be used. Stochastic dominance is a partial-order relation used to compare payoff distributions. Second-degree stochastic dominance is exactly the concept that an investor is rational. We refer the reader to any graduate text in finance, for example [6], or to [2] for some results which simplify the computations necessary to compare distributions.

We argue in Section 5.4.1 that the assumption that all constraints must be expressible as linear or parabolic equations might not be too restrictive.

The final point we have not yet addressed is the question of computational efficiency. The method described requires multiple solutions to potentially large integer programming problems. These problems can be solved very efficiently but it is possible, especially when there are many projects and many constraints, that the method might be too computationally demanding for commercial applications.

Note that the problem has some nice properties: the variables are binary and the covariance matrix is symmetric. It might be possible to exploit these properties to make the method even more efficient.

In any case, this method can be used to solve a large class of interesting and non-trivial portfolio problems. If the general case does not satisfy the assumptions required by this method or is too computationally demanding for commercial applications, then perhaps this method can be used to test and benchmark more heuristic approaches.

5.5 Summary

Each of the three approaches has its advantages and disadvantages. The Portfolio Selection Algorithm approach is based on the appealing idea of trying to find the efficient portfolio with the minimum amount of effort. It is easy to implement and it works well on examples we have considered. If problems raised in Section 5.2.7 can be successfully resolved, this approach will provide a valuable solution to Merak's problem.

The Statistical Inference Approach is also easy to use. Its main disadvantage is that it does not give the (exact) efficient portfolio. On the other hand, it has the advantage that it can be used in any situation, regardless of the number of projects and the nature of the constraints. In the absence of a universally applicable method which will always find the efficient portfolio, this approach provides a practical solution to Merak's problem.

The practicality of the Integer Programming Approach depends on the number of projects and the nature of the constraints. Nevertheless, it has the advantage that it gives the efficient portfolio for problems where it is applicable. Since most clients of Merak will likely want to know the efficient portfolio if it can be found, for problems where this method is applicable, it is the most preferred approach.

Bibliography

[1] Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992). *A First Course in Order Statistics*, John Wiley & Sons, New York.

[2] Calistrate, D., Paulhus, M. and Sick, G. (1998). "Using Real Options to Manage Risk", *Proceedings of The 2nd Real Options Conference, Ernst & Young*, Northwestern University.

[3] David, H. A. (1981). *Order Statistics*, Second Edition, John Wiley & Sons, New York.

[4] Dempster, M. A. H. (1980). *Stochastic Programming*, Academic Press.

[5] Hu, T. C. (1969). *Integer Programming and Network Flows*, Addison-Wesley Publishing Company, Inc.

[6] Ingersoll, J. E. (1987). *Theory of Financial Decision Making*, Rowman & Littlefield.

[7] Kalbfleisch, J. G. (1985). *Probability and Statistical Inference Volume 2: Statistical Inference*, Springer-Verlag, New York.

[8] Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, Vol. VII, No. 1, pp. 77-91.

[9] Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investments*, Blackwell Publishers, Cambridge MA and Oxford UK, second ed., 1997.

[10] Royall, R. (1997). *Statistical Evidence*, Chapman and Hall, New York.

[11] Sunset Software Technology (1987). "Professional Linear Programming System", 1613 Chelsea Road, Suite 153. San Marino, CA 91108, USA.