2005-04-26

Portfolio Construction using Clustering Methods

Zhiwei Ren
Worcester Polytechnic Institute

Portfolio Construction Using Clustering Methods

by

Zhiwei Ren

A Thesis

Submitted to the Faculty of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Professional Masters Degree in Financial Mathematics

May 2005

APPROVED:

Professor Arthur C. Heinricher, Thesis Advisor

Professor Bogdan Vernescu, Head, Mathematical Sciences Department

Abstract

One major criticism about the traditional mean-variance portfolio optimization is that it tends to magnify the estimation error. A little estimation error can cause the distortion of the whole portfolio.

Two popular ways to solve this problem are to use a resampling method[1] or the Black-Litterman method (Bayesian method).

The clustering method is a newer way to solve the problem. Clustering means we group the highly correlated stocks first and treat the group as a single stock. After we group the stocks, we will have some clusters of stocks, then we run the traditional mean-variance portfolio optimization for these clusters. The clustering method can improve the stability of the portfolio and reduce the impact of estimation error. In this project, we will explain why it works and we will perform tests to determine if clustering methods do improve the stabilities and performance of the portfolio.

[1] Introduced by Jorion (1992), Financial Analyst Journal.

Contents

1. Introduction

1.1. A short introduction of Portfolio Optimization

There are many books and papers dedicated to the study of mean-variance portfolio optimization method. All are based on the fundamental paper [3,4] by Harry Markowitz. We will just give a short description here.

Here are the notations we use:

— μ the vector of expected return of stocks, $\mu_i = E(R_i)$, where R_i is the return for stock i

— C: Covariance matrix of the stocks. $C_{i,j} = E[(R_i - \mu_i)(R_j - \mu_j)]$

— h: the vector of weights assigned to each stock, h_i is the fraction of total wealth invested in stock i.

— λ: risk aversion coefficient. [1]

Markowitz's key contribution to portfolio analysis was the definition of risk as standard deviation of return. With this definition, he went on to define as *efficient* any portfolio that maximized return for a given level of risk or minimized risk for a given level of return.

There are several equivalent mathematical formulations of this problem.

[1] Risk aversion coefficient is a parameter used to measure how much risk the person wants to take for his investment.

We will focus on the following formulation.

Problem A: Maximize the utility

$$U(h) = \mu^T h - \frac{h^T C h}{2\lambda} \tag{1.1}$$

subject to some set of constraints. One standard constraint is that the portfolio is fully invested:

$$\sum_{i=1}^{N} h_i = 1$$

and another would be that short-selling is forbidden:

$$h_i \geq 0 \quad \text{for all} \;\; i.$$

We can choose different λ for people with different levels of risk aversion. Then maximizing the function $U(h)$ gives a vector h recording the optimal weight for each stock.

Two other formulations do not use the risk aversion coefficient λ.

Problem B: Maximize the portfolio return

$$U(h) = \mu^T h \tag{1.2}$$

subject to a constraint on the level of risk

$$\begin{aligned} h^T C h &= \sigma_0^2 \\ h^T 1 &= 1 \end{aligned}$$

Problem C: Minimize the portfolio risk

$$U(h) = h^T C h \tag{1.3}$$

subject to a constraint on the level of return

$$\mu^T h = \mu_0$$
$$h^T 1 = 1$$

Each of these three problems determines efficient portfolios. The entire efficient frontier is obtained by solving the problem for different values of λ in the first case, different levels of risk in the second case, and different levels of return in the third case.

2. Major Criticism of Mean-Variance Portfolio Optimization

One of the main criticisms of mean-variance portfolio optimization is that it tends to magnify the effect of estimation error. The optimization process is extremely sensitive to the input expected return vector and covariance matrix. This often leads to poorly diversified portfolios. This is perhaps the main reason why the Markowitz theory is still widely ignored in portfolio management.

Two relatively popular solutions to mean-variance sensitivity are resampling methodologies and Bayesian asset allocation.

Resampling methods generate simulated price data based on the stocks' price history, then build the portfolios based on these simulated prices. Portfolios built using resampling tend to be more diversified and more stable[1]. However, it also has it drawbacks: The computational load of resampling is heavy if we have a large universe of stocks. What's more:

— Suboptimal estimation error in the original inputs is reflected in the final asset allocation;

— It isn't based on economic theory;

— It implicitly assumes that the investor has abandoned the maximum expected

[1] A stable portfolio means: a little change in the expected return or covariance matrix will not change the composition of the portfolio dramatically.

utility framework; and, if you have a benchmark, the procedure leads to unwanted active risk.

Bayesian asset allocation is also called the Black-Litterman method. The main motivation behind it is to somehow combine fundamental analysis with quantitative analysis. It introduced a new parameter for the stock: The fund manager's view. This view is based on the fund manager's experience and his fundamental analysis.

In practice, Black-Litterman method has yielded better performed, more stable portfolio. [2]

Compared to these two methods, the clustering method is a more sophisticated method. The main motivation for the clustering method comes from the following:

1. Portfolio optimization over a universe of uncorrelated, negatively correlated or very low correlated stocks usually gives a better diversified, more stable, less risky portfolio when compared to portfolio optimization over a group of highly correlated stocks. We assume that short selling is not allowed here, which is true for most mutual funds and pension funds. [3]
2. Some stocks tend to have strong correlation with each other. For example, UPS and Fedex, US airway and Delta, Ford and GM. This fact enables us to group these highly correlated stocks into one group.

[2] Based on presentation of Dr.Charvey from Duke University: Consistent Return Estimates in the Asset Allocation Process

[3] Just a digression: Hedge fund is allowed to short sell. Most mutual fund is not allowed to do so because people think it is too risky to short sell. But actually, a good short-long investment is a great way to reduce risk and increase performance.

3. After we group highly correlated stocks, the correlation between these groups will be small. If we think of these groups as a new universe of stocks, all these stocks will have low correlation.
4. It is always harder to estimate expected return than correlation [4].

To summarize, the clustering method takes the following approach:

1. Group highly correlated stocks.
2. Run the mean-variance optimization upon these groups to get the portfolio weights.

In this project, we will focus on the clustering method.

[4] Martin, Linda J. (1985), "Uncertain? How do you Spell Relief?", Journal of Portfolio Management.

3. Clustering Method

In portfolio management, there are two main statistics used: the expected return μ and the covariance matrix C. We estimate them from the price history of stocks. Since these parameters are estimated, it's inevitable that there is estimation error.

A lot of past work and research[1] has shown that: **It is much easier to estimate correlations between stocks than the estimate expected return**s. The values of the covariance matrix are much more accurate and reliable than the values of expected returns.[2]

When we build our portfolio, we want to make more use of information in the covariance matrix (good, more reliable information) and less use of the information in the expected returns (bad, less reliable information). We want good, more accurate information to contribute more to the way we build the portfolio.

What does it mean that good information will contribute more to the way we build the portfolio? It means that when there is some change in the good information, the composition of the portfolio will change more dramatically than when there is some change in the bad information. Thus, the composition of the portfolio is less sensitive to the change of bad information and more sensitive to the change of good information. Since the estimated expected return is

[1] Bernd Scherer (2004), Portfolio Construction and Risk Budgeting, Risk Books.
[2] Bernd Scherer (2004), Portfolio Construction and Risk Budgeting, Risk Books.

the information of less quality, we want to make our portfolio composition less sensitive to it. The way to achieve it is to group our stocks into clusters, so that these clusters have *high intergroup* correlation and *low intragroup* correlation. By grouping stocks in this way, we can dramatically reduce the impact of expected return upon the composition of the portfolio. The reason behind it can be explained in a simplified two stock example:

If we have two stocks, expected return are respectively equal to μ_1 and μ_2. The covariance matrix is: $C = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$ and our risk aversion ratio is λ. Then the optimal weight will be:

$$h_1 = \lambda \frac{\mu_1 \sigma_{22} - \mu_2 \sigma_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}}$$

$$h_2 = \lambda \frac{\mu_1 \sigma_{21} - \mu_2 \sigma_{11}}{\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}}$$

and we know $\sigma_{12} = \sigma_{21}$, $\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}\sigma_{22}}$, and all $\sigma_{ii} \geq 0$.

To measure how sensitive the weight is to the estimated expected return, we take the partial derivative of weights with respect to expected return:

$$\frac{\partial h_1}{\partial \mu_1} = \lambda \frac{\sigma_{22}}{\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}} = \lambda \frac{1}{\sigma_{11} - \rho_{12}^2 \sigma_{11}} = \lambda \frac{1}{\sigma_{11}(1 - \rho_{12}^2)}$$

where h_1 is the weight for stock 1 , μ_1 is the expected return for stock 1, ρ_{12} is the correlation between stock 1 and stock 2, and λ is the risk tolerance.

To minimize the sensitivity of portfolio weight to expected return, we can see that we would want the right side as small as possible, which means we want $\sigma_{11}(1 - \rho_{12}^2)$ as large as possible. When ρ_{12}, the correlation between two stocks is equal to 0, we will have a portfolio what its composition is the least sensitive to the change of expected return. In reality, the smaller ρ is, the better. This is the reason why we will group stocks so that these groups have the high intergroup correlation and low intragroup correlation correlation.

4. Implementing the Clustering Method

The clustering method can be implemented in the following three steps:

1. Take a history price data for 5000 stocks. We pick n stocks and run the statistical analysis. For those having a strong correlation with each other (high correlation number ρ_{ij}), we group them. In the test, We use $\rho = 0.2$ as the conditions to determine high correlation. This means if $\rho_{ij} \geq 0.2$, we will group them in one cluster, if the correlation between them is smaller than 0.2, we will not group them in one cluster. (Later, We will use ρ=0.15 to repeat the test).[1]
2. After we determined the clusters, we evenly weighted them and treat the cluster as a new stock. We evenly weight because it is easy to implement in this way and it gives good diversification in many cases. However, this is not necessarily the best way, some future study is needed here. Also, there exist stocks that are not highly correlated to any other stock, in this case, we just treat the stock as a cluster of one stock.
3. Use the portfolio optimization for this new universe of "stocks". (Note some "stock" here is actually a cluster of evenly weighted stocks. We use the average of the history price of member stock in each cluster as the historical price for the cluster, in this way, these clusters are not different from stocks.)

[1] There is still research to do about how to pick the best ρ to determine highly correlated or not. We use $\rho = 0.2$ just for our tests. It is not necessarily the best choice for ρ.

One other benefit of the clustering methodis that we dramatically reduce the size of covariance matrix. For example, in test 1, we reduced covariance matrix size from 150 by 150 to 69 by 69. In test 2, we reduced covariance matrix size from 150 by 150 to 56 by 56. This will dramatically reduce the computational load for the quadratic programming algorithm when we deal with a portfolio composed of 5000 or more stocks.

5. Estimation Error Sensitivity Test

We have the daily history price for 5000 stocks available. We randomly choose 150 from them and build a portfolio. (The reason we didn't use all 5000 stocks is mainly because of the computation load is too heavy). We did two tests.

Before we proceed to analyze our test results, we want to consider a question here: How do we determine which stocks are highly correlated to each other and which stocks are not?

If we use fundamental analysis, we can group stock in the same industry group as a cluster. Usually stocks in the same industry move together.

The second way is pure quantitative method: get the correlation between each two stocks and set a value ρ^*. Two stocks with correlation greater than ρ^* are considered highly correlated and those with correlation less than it are considered not highly correlated. One more question for this quantitative method is what value is appropriate for ρ^*. We used the quantitative method in this project. And we picked $\rho^* = 0.2$, 0.15 for two tests.

Figure 5.1 and 5.2 give the histogram for the correlation between each two stocks for the two tests we did.

Figure 5.1. Histogram for correlation distribution in Test 1

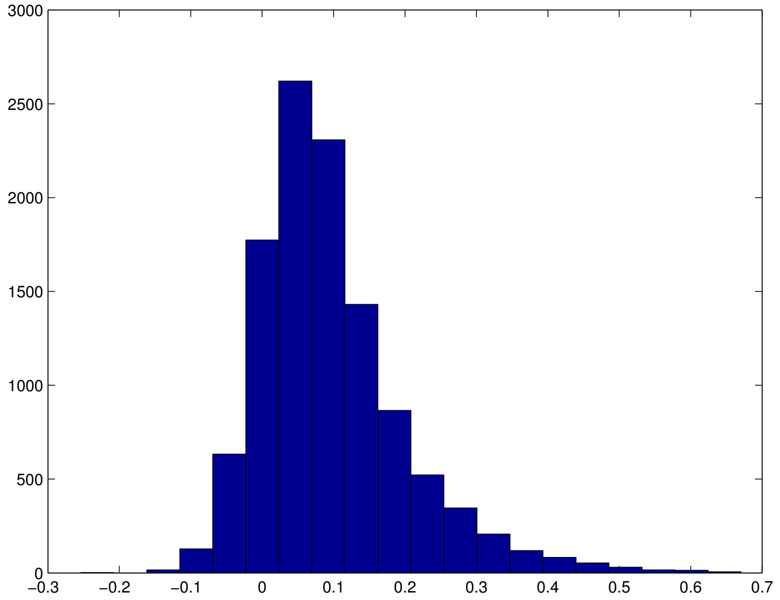


Figure 5.2. Histogram for correlation distribution in Test 2

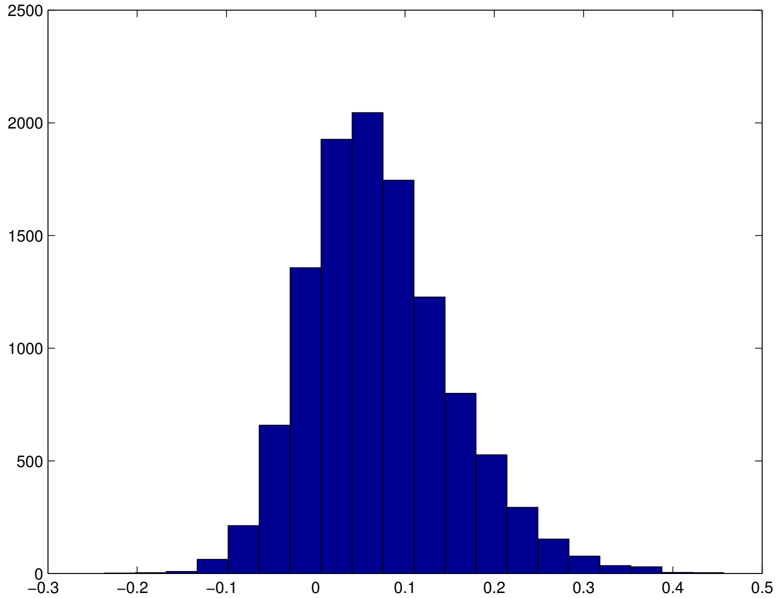
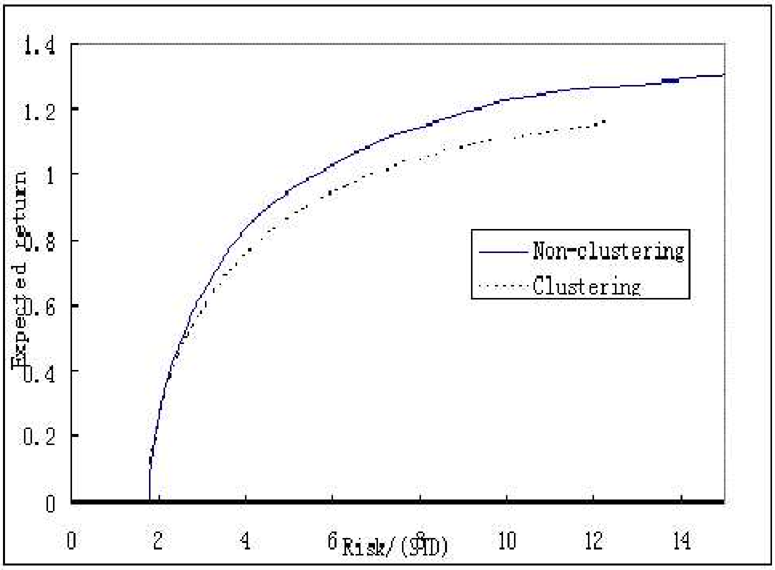


Figure 5.3. Comparison of efficient frontier with/without clustering in Test 1



5.1. First Estimation Error Sensitivity Test

Since there is no clear definition of "high correlation", in this test, we use $\rho = \mathbf{0.2}$ as the standard to test our results.

When $\rho = 0.2$, the efficient frontiers with and without clustering are shown in Figure 5.3. (Note: expected return and risk are both based on the annual data.)

Take a quick look at these two efficient frontiers, we are not impressed: it seems we can achieve higher return with the same risk before we use clustering method. However, a lot of times, we cannot trust this efficient frontier too much. Why?

We have to remember the main criticism against portfolio optimization: It tends to magnify estimation error. Imagine that we mistakenly overestimate/underestimate the expected return for a certain stock, if by chance this stock has low volatility, it will be certain that the optimization will put a lot of weight (long/short position) on this stock, "high return, low risk, who doesn't want it?" This is dangerous because we are betting on poor estimation. If the reality goes against this estimate, our portfolio will perform very poorly.

Clustering method can relieve this problem. Recall the purpose of clustering method is to reduce the impact of expected return estimation and use more information from the covariance matrix. We know there is far more estimation error in expected return than in the covariance matrix, so the clustering method will make our portfolio more resistant to estimation error.

By now, we should realize that a higher efficient frontier doesn't mean you can really get higher return in real life because that the efficient frontier may be based on some inaccurate estimation. And we also should realize that the real advantage of the clustering method is not about the efficient frontier, but its resistance to the estimation error.

To test how sensitive the portfolio is to the estimation error, we did two tests for $\rho = 0.2$ ($\rho > 0.2$ is treated as highly correlated and will be grouped together, $\rho < 0.2$ is treated as low correlation):

— Sensitivity Test 1.1: We randomly pick 15% of the stocks and change their expected return by $\pm 10\%$, then we rebuild the portfolio to see how much it is different from the original portfolio before this change.

— Sensitivity Test 1.2: We randomly pick 15% of the stocks and change their expected return by $\pm 10\%$, In addition, we change the expected return for the leading stock in the portfolio by $\pm 10\%$, and then we rebuild the portfolio to see how much it differs from the portfolio before this change. The reason we did this test is to see what will happen if we have estimation error for the stock that has the largest weight in our portfolio.

What we really did in these two tests is to introduce some artificial estimation

Table 5.1. Test 1.1, Comparing the weight of new portfolio with original portfolio

Points On EF	Change of weight (no clustering)	Change of weight (clustering)
1	3%	0
2	4%	0
3	5%	1%
4	1%	1%
5	1%	1%
6	1%	0
7	1%	0
8	2%	0
9	2%	0
10	2%	0

error into our problem and we want to test how sensitive the portfolio we built is to these estimation errors. We check the change of portfolio composition before and after we introduced the estimation error to measure how sensitive they are to the estimation error. The way we check the change of portfolio composition is as follows:

Let $h_{old,i}$ be the weight for stock i in the original portfolio before we introduced estimation error, and let $h_{new,i}$ be the weight for stock i in the portfolio after we introduce the estimation error. Then we compute $\sum_{i=1}^{N}(|h_{new,i} - h_{old,i}|)$, where N is the number of stocks in the portfolio. We use this sum to quantify the change of portfolio.

5.1.1. Results for Sensitivity Test 1.1.

Table 5.1 is the results for sensitivity test 1.1. We picked 10 points on the efficient frontier and track the change of their composition:

5.1.2. Results For Sensitivity Test 1.2

Table 5.2 is the results for sensitivity test 1.2. We can see that when we use clustering method, the portfolio will rebalance less frequently. This means

Table 5.2. Test 1.2, Comparing the weight of new portfolio with original portfolio

Points on EF	Change of weight (no clustering)	Change of weight (clustering)
1	7%	8%
2	10%	12%
3	13%	14%
4	12%	17%
5	15%	19%
6	16%	22%
7	53%	32%
8	73%	30%
9	114%	15%
10	200%	40%

that when we have some estimation error in expected return, which is very likely, the portfolio built using clustering method will perform much better than those portfolios built without using clustering.

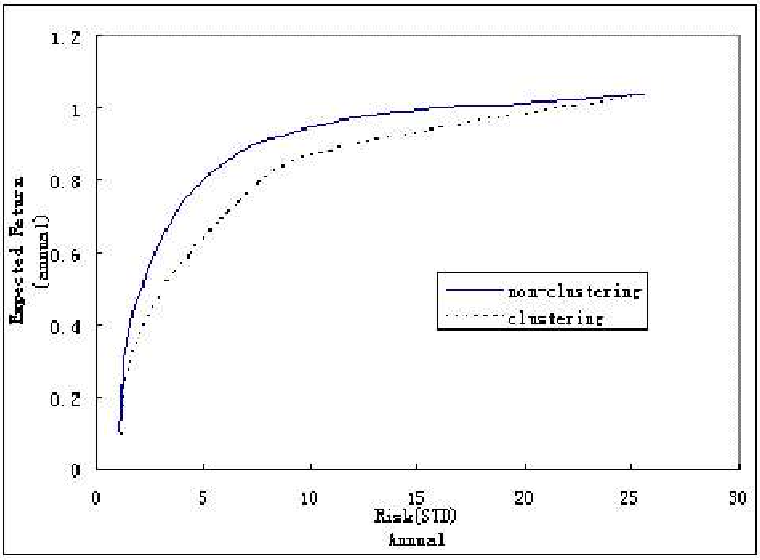
5.2. Second Estimation Error Sensitivity Test

In order to verify that clustering can reduce the impact of estimation error in expected return. We picked another 150 stocks from our stock universe and ran the sensitivity test as following:

— Sensitivity Test 2.1: We randomly pick 15% of the stocks and change their expected return by 10% (up or down), then we rebuild the portfolio to see how much it is different from the portfolio before this change. This is the same as the test 1 in section 5.1.

— Sensitivity Test 2.2: We randomly pick 15% of the stocks and change their expected return by 10% (up or down), And, we change the expected return for the top 2% stock in the portfolio by 10% (in the test 2 of section 5.1, we only changed the weight of one stock of the most weight), then we

Figure 5.4. Comparison of Efficient Frontiers with/without clustering in Test 2



rebuild the portfolio to see how much it is different from the portfolio before this change. The reason we did this test is to see what will happen if we have estimation error for the stock that has the most weight in our portfolio.

The other change is we pick ρ =0.15 as the standard to determine highly correlated stock. The efficient frontier without clustering and the efficient frontier after we used clustering are shown in figure 5.4:

Also, we want to see what will happen to the portfolio composition after we introduce estimation error. We compared the weights of portfolio before and after we introduced estimation error, get the sum of absolute value of difference of portfolio weights.

5.2.1. Results for Sensitivity Test 2.1

Table 5.3 is the results of test 2.1. From Table 5.3, We can not tell much because the change of weight for both portfolios are very small.

Table 5.3. Test 2.1, Comparison of portfolio weights

EF	Change of weight (no clustering)	Change of weight (clustering)
1	0.567%	0.27%
2	0.824%	0.38%
3	0.417%	0.474%
4	0.323%	0.078%
5	0.213%	0.089%
6	0.201%	0.093%
7	0.198%	0.111%
8	0.21%	0.112%
9	0.155%	0.138%
10	0.11%	0.215%

Table 5.4. Test 2.2

EF	Change Of Weight(nonclustering)	Change of weight(clustering)
1	2.3%	2.9%
2	3.5%	3.5%
3	5.8%	5.3%
4	7.4%	6.2%
5	10.1%	8.7%
6	13.6%	10.8%
7	16.4%	8.4%
8	22.3%	11.4%
9	11.9%	2.5%
10	1.4%	1.1%

5.2.2. Results for Sensitivity Test 2.2

In the second test, we changed the expected return for the top 2% of holdings of each portfolio by 10%(up or down). Then we compare the change of weights for each portfolio. The results are in Table 5.4.

This test shows the advantage of clustering method: If we had some estimation error for expected return (especially on the stocks that seems to have excellent return and low risk), the portfolio we build using clustering method will be much closer to the real optimal portfolio than the portfolio without using clustering method. We can see that after we introduce the artificial estimation error, we don't have to rebalance as often to make the portfolio using clustering method reach the new optimal. But for the portfolio that is built without using the clustering method, we have to do a lot more rebalancing to reach the new optimal level.

5.3. Conclusions from the Above Two Tests

From the above two tests, we can see the advantage of clustering method: It can dramatically reduce the impact of estimation error. The data in three tables can show it clearly: With clustering methods, even if we have estimation error, the portfolio we built is still much closer to the optimal portfolio.

After we used clustering method, the portfolio we build will be much more robust, much more immune to the estimation error in expected returns. Again, we want to emphasize the reason why we focus on reducing the impact of estimation error from expected returns: It is because it is usually much easier to estimate covariance matrix than expected return! So there is much more estimation error in the expected return than in covariance matrix.

Robustness is a nice thing to have. But what we care more is how it will affect the performance of the portfolio. Next, we want to test how those portfolios built using clustering methods perform compared with those portfolios built using normal portfolio optimizations.

6. Performance Tests

To test the performance of the portfolio we build using clustering method, we compare the cluster portfolio with the portfolio built without using clustering methods.

Since there are so many portfolios available on the efficient frontier, to compare them, we make sure the portfolios we used to compare have the same risk. We test the performance in the following way:

First, construct the portfolio using two methods, one method is the normal optimized portfolio without clustering, the other method method groups the stocks first then runs portfolio optimization, ie. we build the portfolio using the clustering method. Then, we compare the first month and the second month return for these two portfolios.

6.1. First Performance Test

Table 6.1 and 6.2 contain the results for the performance test. We use the 150 stocks from test 1 in section 5.1.

Look at these two table, the first thing we notice is portfolio built using clustering performs much better than the portfolio built without clustering method. For example, for nonclustering, portfolio 3 and 4 has -54% and -32.4% return in the second month, but the portfolio built using clustering, none of the monthly

Table 6.1. Portfolio performance (non clustering)

Portfolio	Risk	Return in 1st month	Return between 1st and 2nd month
1	2.03	-0.1%	-6.5%
2	2.54	0.6%	-17.1%
3	4.33	3.7%	-54.0%
4	8.72	12.5%	-32.4%

Table 6.2. Portfolio performance (clustering)

Portfolio	Risk	Return in 1st month	Return between 1st and 2nd month
1	2.04	-12.0%	-9.1%
2	2.56	-0.3%	-20.3%
3	4.33	-23.4%	-8.1%
4	8.75	19.2%	-8.8%

return fluctuate more than 25%. This is a very good property for the portfolio because extreme volatility is not welcome.

Looking in more detail, we see that neither portfolio performs well. To get some better idea about how all the 150 stocks perform during the period, we built the evenly weighted portfolio and computed the expected return for the first and second month. The result is in Table 6.3. We also computed the market return during the same period, the result is in table 6.4.

Comparing our portfolio with the evenly weighted portfolio and the market. We see that the evenly weighted portfolio outperforms both portfolios we built.

Considering the DOW and NASDAQ charts during this period first (the historical data we used includes 01/02/2001 to 08/08/2002). They are in figure 6.1 and 6.2. From this two charts, the first thing we notice is that the market has

Table 6.3. Evenly weighted portfolio performance

Portfolio	Return in 1st month	Return between 1st-2nd month
Evenly weight nonclustering	-2.7%	-6.2%
Evenly weight clustering	0.07%	-7.8%

Table 6.4. Market return

Portfolio	Return in 1st month	Return between 1st-2nd month
Dow	-14.1%	9.6%
NASDAQ	-12.1%	-3.8%

Figure 6.1. Dow Chart between 2001 and 2005

been going down during 2001 and 2002 periods. Actually, between 01/02/2001 and 08/08/2002, Dow went down by 30.96% and NASDAQ lost 110.6%.

In June 2002, there is a sharp downside trend, then right after it, there is a sharp uptrend, after the uptrend, there is another sharp downtrend.

These three sharp trend changes explained why our "optimized" portfolio performed so poorly. When we build the portfolio, we used past return data to forecast future return. There is obvious danger here: past good performer might not perform well in the future. If there is a big change in the direction of the market, it is certain that this kind of portfolio will perform very badly.

Actually this test is especially bad compared our test 2 (next section), the main reason is this 150 stocks are mainly NASDAQ stocks such as: MSFT, DELL, EBAY, IBM, etc. And there stocks are the biggest loser in 2002, but in

Figure 6.2. NASDAQ Chart between 2001 and 2005

Table 6.5. Performance Test For nonclustering method

Port	Risk	Return in 1st month	Return between 1st-2nd month
1	1.12	-0.61%	-4.35%
2	1.18	-0.29%	-4.53%
3	2.70	5.3%	-9.34%
4	7.05	11.36%	-2.15%

test 2, the 150 stocks we used are mainly from Dow and few of them are high-tech stocks, so the performance of test 2 is much better than this one.

6.2. Second Performance Test

For the second test, the results are showin in Table 6.5 and 6.6. We use the 150 stocks from test 2 in section 5.2. Also, we want to get some better idea how the market performs during this period, so we constructed the evenly weighted portfolio. The results are in table 6.7.

In this test, we can see that the portfolio built using the clustering method outperforms the portfolio without using clustering method. These two tests

Table 6.6. Performance Test for Clustering method

Port	Risk	Return in 1st month	Return between 1st-2nd month
1	1.12	-0.9%	-1.83%
2	1.18	-0.88%	-1.19%
3	2.70	3.41%	-1.09%
4	7.05	11.72%	-1.98%

Table 6.7. Evenly weighted portfolio performance

Portfolio	Return in 1st month	Return between 1st and 2nd month
clustering	1.49%	-6.05%
Non clustering	2.57%	-10.6%

showed that when compared with traditional portfolio optimization, the clustering method can improve the performance of the portfolio.

7. Summary and Some Future Work

These tests indicate that the clustering method is a good way to improve the robustness and performance of the portfolio. When compared to the resampling method, it is less computational intensive. Because the clustering method is not as widely used as the resampling method, there is not much work available to review. To make a compare, we need to do more tests to measure the effectiveness of clustering method.

There are still a few questions to be answered:

— How show we group stocks? In the project, we used pure quantitative approach. However, fundamental analysis method(group in industry sectors) is another choice also. Which one is better?

— What ρ should we choose? We used 0.2 and 0.15 in this project, but what value is the optimal? This is still question remaining to be answered.

— When will the quantitative method yield good result? In Section 6.1, we saw that when the trend of the market suddenly changes, the quantitative method will perform very poorly because it basically assumes that past winner will be the future winner. When the trend changes, those stocks with historical high return might yield negative return and those that don't perform well probably will begin to perform better. Quantitative method will not be able to adjust to this. So we should do more research to understand this better.

One fix is combine quantitative method with fundamental analysis. In a lot of situations, the macro economic data can give us some signals about the change of the market trend. We, as quantitative analysts, have to try to understand both quantitative analysis and the economic indicator to improve the performance of portfolios.

— In this project, in each cluster, we used evenly weighted portfolios for each cluster. For example, if there are 5 stocks in one cluster, we use 20% as the weight for each stock. We did this to simplify the computation. Are there any other ways to build the portfolio for each cluster?

In summary, the clustering method can improve the stability and performance of the portfolio based on the tests we did. However, before we apply it to real work, more work should be done to verify its advantages. This project is just a start for future work.

References:

[1] Richard Grinold and Ronald Kahn (1999), *Active Portfolio Management*,McGraw-Hill.

[2] Bernd Scherer (2004), Portfolio Construction and Risk Budgeting, Risk Books.

[3] Markowitz,H.M, (1971) Portfolio Selection: *Efficient Diverfication of Investments.* Yale University Press.

[4] Markowitz,H.M, (2000) *Mean-Variance Analysis and Portfolio Choice in Capital*, Frank J. Baboozi Associates.

[5] Yahoo Finance `http://finance.yahoo.com`.

[6] Philippe Jorion (1992), Portfolio Optimization in Practice, Financial Analyst Journal.

[7] Martin, Linda J (1985) Uncertain? How do you Spell Relief, Journal of Portfolio Management.

[8] Compell Harvey's website `http://www.duke.edu/~charvey/`