Stock Selection Based on Cluster and Outlier Analysis

Steve Craighead
Bruce Klemesrud
Nationwide Financial
One Nationwide Plaza
Columbus, OH 43215
USA

Abstract

In this paper, we study the selection and active trading of stocks by the use of a clustering algorithm and time series outlier analysis.

The Partitioning Among Mediods (PAM) clustering algorithm of Kaufman and Rousseeuw(1990) is used to restrict the initial set of stocks. We find that PAM is effective in its ability to specify nonuniform stock series from the entire universe. We are pleasantly surprised that the algorithm eliminated the bankrupt Enron and Federal Mogul stock series, without our intervention.

We use outlier analysis to define two separate active trading strategies. The outliers within a time series are determined by the use of a Kalman Filter/Smoother model developed by de Jong and Penzer(1998).

Weekly trading in stocks with an initial $30,000 with a closed stock portfolio from 1993 to 2001, we obtained a 17.8% annual return on a cash surrogate passive strategy, 18.1% on a passive strategy using all the stocks in our restricted asset universe, 20.2% on a combined cash protected and outlier active strategy, and 23.3% using the outlier active strategy only.

Comparing these results to the passive strategy being entirely invested in the S&P 500 Large Cap index with at 9.9% return, we find that under this stock portfolio any of our strategies are superior to that of a purely passive index strategy.

1 Introduction

The process of actively managing a stock portfolio is more an art than a science. The industry irritation is that elementary school children tend to pick stocks with better performance than those of the professional. Also, to add insult to injury, it is reputed that stock portfolios chosen randomly from Rolodexes by monkeys perform better than the students. Even though we might be competing with our youth and various other simians, we believe that our experience and two newer statistical tools may still allow us to make some well reasoned decisions in active stock management.

There are at least three difficulties in active trading. The first is the selection process. Here one must decide which stocks to add to the portfolio and which to remove from the

portfolio. Secondly, the size of the trade must be considered. Third, the issue that many consider the most difficult, is when to move from one position to another.

In classic portfolio theory the initial choice of assets is based on a risk/return tradeoff using quadratic programming (or from a CAPM approach comparing various β values). We, however, are interested in the stock price series and we realize that the change in the level of the stock price is masked if we only use the stock return series. This leads us to use the Partitioning Among Mediods (PAM) algorithm. This algorithm is introduced by Kaufman and Rousseeuw(1990) in [2]. PAM is designed to take a collection of vectors and obtain the best representatives for a specific number of clusters. We use the algorithm only to reduce the initial asset universe.

Classic portfolio theory is a short period decision process and though it can be used to determine which assets best optimize the current portfolio, one must deal with issues of portfolio drift and rebalancing. However, we want a strategy that is able to monitor the market and make specific movement recommendations on the specific assets. This leads us to use a time series outlier algorithm developed by de Jong and Penzer(1998) in [1]. Their work is based on using a single pass of a Kalman Filter/Smoother to produce an outlier statistic they call τ^2. We use the change of this statistic to determine when to actively move in and out of various stocks.

In the next section, we will discuss the data collection and selection process.

In Section 3, we will discuss the use of τ^2 to indicate the change of a market paradigm.

In Section 4, we describe the strategies that we use to make our investment decisions.

In Section 5, we examine the results of our strategies.

In Section 6, we discuss our conclusions, model limitations, and possible future research.

In Appendix A, we give an outline of the PAM cluster algorithm.

In Appendix B, we briefly outline the formulation of the τ^2.

2 Data

We start with an initial universe of 138 stocks from many separate sectors and indices. For each of the 138 stocks, we use a stock price history of 54 different times from February 1998 to December 2001. We obtain the average and the standard deviation of the prices for each of the series. We detrend each price series by subtracting the mean and dividing by the standard deviation. This results in 138 vectors of length 54. We use the PAM algorithm to find five representative clusters. We examine each cluster to determine if there is only one asset in that cluster, assuming that those assets are aberrations. This eliminates Enron. Reprocessing the remaining 137 stocks in the same way, we eliminate Federal Mogul. Once eliminating Federal Mogul, the PAM algorithm returns five clusters with several assets in each cluster. Note: We used the L_1 Norm (or the Manhattan distance) to define the distances in the algorithm to reduce the influence of the outliers upon the selection process.

We then removed stocks that didn't have a history longer than nine years. (The choice

of nine years will be discussed below). Finally, we relied upon our investment experience to reduce to the final asset universe displayed in Table 1. We will give a more extensive summary of the reasons backing these choices in Section 5.

We use two stocks (specifically JNJ and XOM) as cash surrogates. We define a cash surrogate stock to be a stock that will replace the use of a highly secure asset such as a Treasury Bill in portfolio selection and analysis. A cash surrogate stock is usually a Blue Chip which is large, well diversified, highly liquid and has minimal price volatility when compared to the overall market.

We use the prior twenty years (if available) of weekly data (from January 1, 1982 to December 31, 2001) prices from Yahoo! Finance (chart.yahoo.com). These prices are adjusted for stock splits and dividends. The outlier statistics are then determined upon these prices. Note: These prices are not detrended as above in the use of PAM.

We then use a nine year data period to set up the historical trading strategies. We did this for two reasons. The first is that we wanted to develop the trading strategies on the middle third of the data and use the other thirds to back and forward validate the strategy. The second reason reflects our view of constantly changing paradigms; in fact companies in existence for longer periods are not the same. We believe that data becomes stale after a given period and that there are not many companies under a new market paradigm in existence for a long period of time. However, we decided that nine years is a good compromise between the historical statistics and the current market paradigm.

3 Implementation

Using the Kalman Filter/Smoother method briefly described in Appendix B, we obtain the outlier τ^2 statistic for each time for each series. Examples of τ^2 are plotted in Figure 1. In Table 1, statistics of τ^2 for each stock are listed. The τ^2 statistics are approximately chi-square, and can provide a means to judge the significance of the values.

We believe that each stock price series contains specific information that is both market and company specific. We assume that the market is fairly efficient and that the price of a stock changes to reflect new information. However, we also believe that there are also complex interchanges between the market and a stock's value, not the least that of market psychology. This leads us to contemplate that there is the possibility that there is additional information contained within the series that has not yet been reflected by the market. Since high values of the τ^2 imply that the stock price has moved away from status quo and has become an outlier, we believe that the statistic can be a good indicator of any and all new information. We may not know the specific reason of the paradigm change, however, we assume that the outlier statistic reveals that a change is occurring. In the next section, we assume that new information is strengthening while the statistic is increasing. However, we assume that as the statistic falls that the majority of new information has already entered, and the price series begins to revert to a status quo. In the next section, we construct two

Symbol	Name	Size ($ Bil)	Index	Sector	Industry
JNJ	Johnson & Johnson	188.4	S&P500/Dow Ind	Healthcare	Major Drugs
XOM	Exxon Mobil Corp	271.8	S&P500/Dow Ind	Energy	Oil & Gas - Integrated
AEP	American Electric Power	14.6	S&P500/Dow Util	Utilities	Electric Utilities
AIG	American International Group	178.7	S&P500	Financial	Insurance (P&C)
AMAT	Applied Materials	41.9	S&P500/Nasdaq 100	Technology	Semiconductors
BAC	Bank of America	115.6	S&P500	Financial	Money Center Banks
CAH	Cardinal Health	30.4	S&P500	Healthcare	Biotechnology & Drugs
D	Dominion Resources	18.8	S&P500/Dow Util	Utilities	Electric Utilities
EK	Eastman Kodak	9.8	S&P500/Dow Ind	Consumer Cyclical	Photography
HWP	Hewlett Packard	60.5	S&P500/Dow Ind	Technology	Computer Hardware
ITW	Illinois Tool Works	21.6	S&P500	Capital Goods	Misc. Capital Goods
IVC	Invacare Corporation	1.2	S&P 600 (SmallCap)	Healthcare	Medical Eqpt & Supplies
LANC	Lancaster Colony	1.5	S&P 400 (MidCap)	Consumer Non-Cyclical	Food Processing
MCD	McDonald's Corporation	38.7	S&P500/Dow Ind	Services	Restaurants
MDT	Medtronic, Inc	52.9	S&P500	Healthcare	Medical Eqpt & Supplies
MO	Philip Morris	119.7	S&P500/Dow Ind	Consumer Non-Cyclical	Tobacco
MRK	Merck & Co	128.3	S&P500/Dow Ind	Healthcare	Major Drugs
MSFT	Microsoft Corporation	285.3	S&P500/Nasdaq 100	Technology	Software & Programming
RPM	RPM Inc	1.94	S&P 400 (MidCap)	Basic Materials	Chemical Manufacturing
SBC	SBC Communications	107.1	S&P500/Dow Ind	Services	Communications Services
USAUX	USAA Aggressive Growth Fund	NA	Mutual Fund		Aggressive Growth
WOR	Worthington Industries	1.3	S&P500	Basic Materials	Iron & Steel

Table 1: Selected Stock Series (Source: YAHOO! Finance)

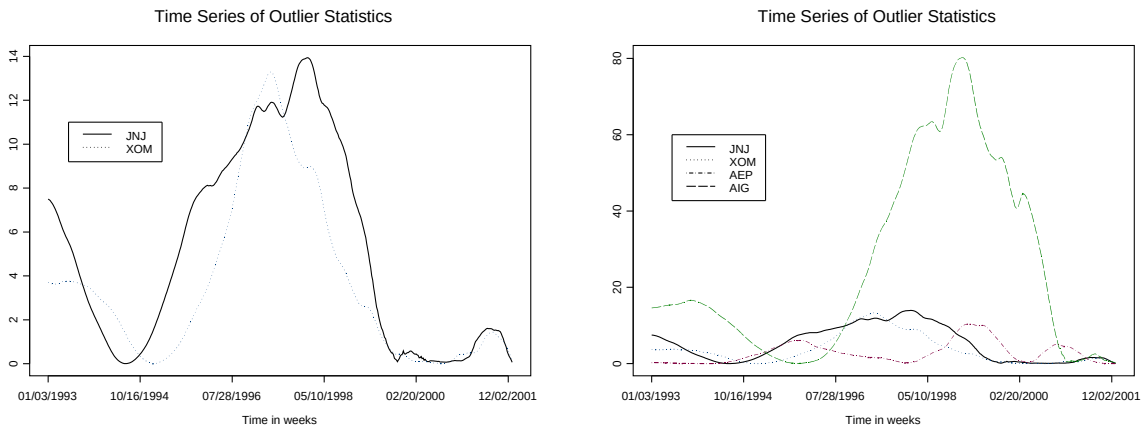

Figure 1: Time Series of Outlier Statistics

active strategies that sell when τ^2 is falling.

4 Model Description

We examine five separate historical trading strategies.

The first we call the "S&P 500" strategy, which is a passive index strategy where we invest the initial amount into the S&P 500 index and make no changes in the investment for the entire investment period.

The second we call the "Cash Surrogate" strategy. This is where we place the initial amount equally split between our cash surrogates, and we do not make any other changes in the investment over the investment period.

The third we call the "Passive" strategy. This strategy we place two thirds of the initial amount evenly in the cash surrogates and the remaining third equally distributed in the other twenty stocks. No other changes are made in the investment over the investment period.

Before introducing the fourth and fifth strategies, we want to examine Figure 2. Here we have two time series. The lower series is a hypothetical price series and the upper series is the corresponding τ^2 series. The two vertical bands in the figure are regions where both series are decreasing. In active trading, we would like to enter a stock position when price is low and exit when the price is high before it turns around. However, we might give up the desire to enter low if we can preserve the value of the portfolio in the event of a downturn.

We use the τ^2 statistic to indicate the strength of information entering the series. We make the assumption that when the price series falls and the τ^2 series is falling that the stock has entered a downturn and will begin to seek status quo. Using this we now develop our two active strategies.

The fourth strategy we call 'Active" and we distribute the initial investment to all 22 stocks as in the "Passive" strategy, but we use the above sell strategy to move between the various

Stock Symbol	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum
JNJ	0.0015	0.3085	1.1220	3.0788	5.4090	13.9371
XOM	4.90E-6	0.09437	1.1091	2.1219	3.0110	13.295
AEP	0.00059	0.06599	0.3933	1.4001	1.66202	10.3763
AIG	3.06E-6	1.7712	9.6778	16.163	15.8791	80.177
AMAT	0.00102	2.07258	6.36806	20.9459	11.33510	176.3947
BAC	0.00029	1.47846	5.22012	17.3243	22.44295	85.1243
CAH	0.00023	1.82930	5.42177	8.1376	11.31191	46.93333
D	0.00096	0.14277	0.88906	4.37229	5.27433	39.54997
EK	3.07E-6	1.08909	10.41518	29.7587	50.461	122.615
HWP	8.47E-5	0.21473	3.79574	11.33841	16.45391	103.3401
ITW	0.00016	4.37575	9.67202	15.71565	21.47767	64.40019
IVC	2.83E-7	0.29823	1.59093	2.87665	4.65674	10.11641
LANC	7.75E-5	0.47648	1.74143	3.64980	4.79727	14.59669
MCD	0.00031	0.09142	1.23688	3.71250	5.88821	22.90411
MDT	0.00006	0.27945	2.25513	5.10800	6.57462	25.67335
MO	0.00006	0.27604	0.95651	3.61067	5.13000	19.80898
MRK	0.00020	0.71763	2.58939	13.73230	20.42116	77.65404
MSFT	0.00286	7.60743	21.54652	49.58955	55.25148	252.13553
RPM	0.00142	0.13271	0.22657	0.51394	0.36017	5.96583
SBC	0.00030	0.40280	1.40575	4.64279	4.89330	32.38622
USAUX	0.00093	0.45461	3.05818	5.94627	5.79714	46.52636
WOR	0.00089	0.39012	1.64675	2.04949	3.71074	5.76535

Table 2: Basic Statistics on Stock Outlier Statistic

stocks. Specifically, if $\tau_t^2 - \tau_{t-1}^2$ for a stock S is negative and the price of S at time t minus the its price at time $t-1$ is negative execute the order to sell one half of the stock position in S into cash surrogates and make the cash available for other investments. Otherwise, if cash is available, execute a buy order of stock S by the bitesize. We define bitesize as the acceptable trading size that we wish to enter at an initial commitment and we consistently use \$500.

The fifth strategy we call the "Restricted" strategy. In this strategy we distribute the initial investment amount to all 22 stocks as in the "Passive" strategy, and we use the "Active" movement strategy with a cash surrogate restriction. The restriction is if the total value of the cash surrogates of the portfolio at a specific time is less than 35% of the total portfolio value, no money will be moved into the other stocks.

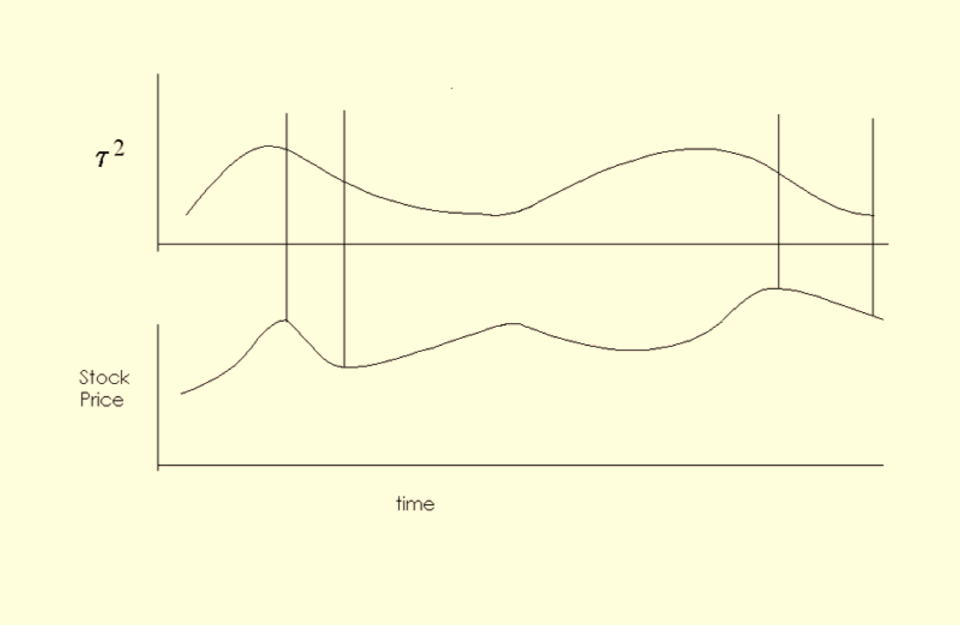

Figure 2: Time Series of the Portfolios

5 Results

We will start with an initial investment of $30,000. We will not add any additional moneys to the portfolios.

In the final three strategies, $15,000 is initially invested in each cash surrogate and $500 in the other stocks.

We initially conducted the study from 1996 to 1998, and found that our active strategies were sound.

Over the nine year period we see in Figure 3 the performance of each of the strategies. The basic statistics on these portfolios are in Table 3. In Figure 3a, we see that all of the strategies exceed that of the "S&P 500" passive strategy, and that in the long term the buy and hold positions of the "Cash Surrogate" and "Passive" approach each other even though most of the time the "Passive" strategy exceeds that of the "Cash Surrogate". Note how the "Active" and "Restricted" exceed that of the "Cash Surrogate" in Figure 3b. Except for some high volatility in the second and third week of March 2001, the "Active" and "Restricted" portfolios seem to have reasonable volatility and performance.

In Figure 4a one can observe the cash surrogate portion of the "Active"strategy with that of the overall portfolio performance. Notice how quickly the strategy moves out of the cash surrogates in the early years and dramatically moves into the safer cash surrogates after the 2000 market downturn in Tech stocks. From 1998 to 2000, the cash surrogates are somewhat volatile, because of the market shifts due in part to difficulties in Russia and the Pacific Rim. Note that the same results are in the "Restricted" strategy except the post 2000 shift into the cash surrogates are not as steep, and the cash surrogate volatility is not as extreme from

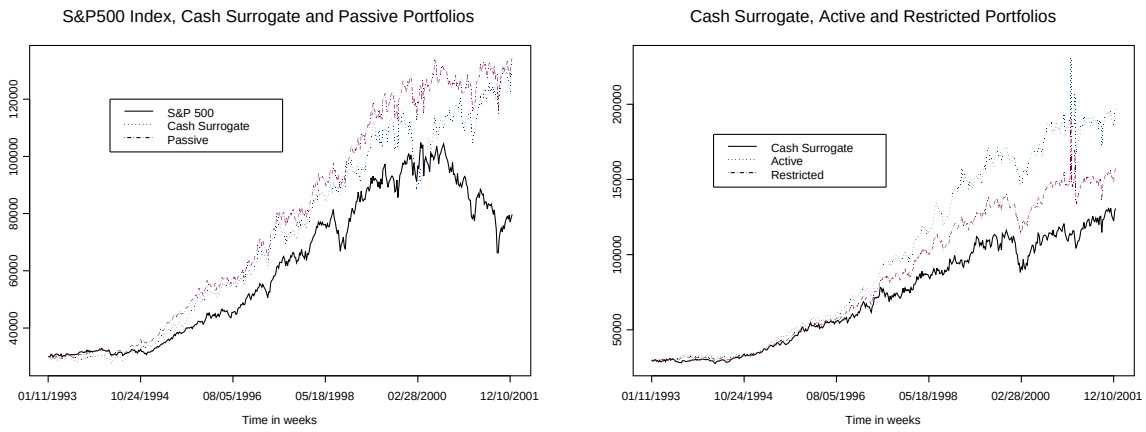

Figure 3: Time Series of the Portfolios

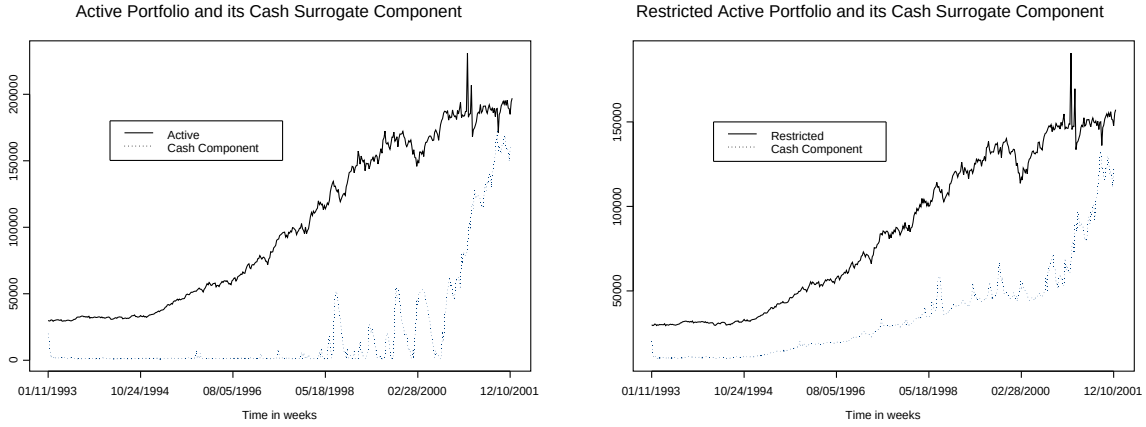

Figure 4: Time Series of Active Portfolios with their Cash Surrogates

1998 into 2000.

The only difficulty is the March 2001 volatility in both of these strategies. In March 2001, we saw the beginning of an increase in unemployment, and a severe drop in the DJIA and the NASDAQ. The perfect active strategy would have moved more into cash surrogates when the market value moved down so drastically. Possibly if we used these strategies on a daily basis, we might have eliminated the volatility, since we only move 50% of a stock's position in one week, but daily trading would have increased the overall volatility of the strategies.

Security diversification is extremely important for defensive purposes. We purposefully selected stocks in Section 2 from various sectors/industries and different indices. Each sector/industry and index behaves differently. This defensive behavior is a benefit, derived from the use of the PAM algorithm, which protects the portfolio from downside risk and also produces extra upside potential for the portfolio.

Notice in Table 4 how the stocks chosen in Section 2 are diversified by sector and in Table 5

Portfolio	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum
S&P 500	$29,799	$34,860	$61,773	$ 61,311	$ 84,997	$104,824
Cash Surrogate	27,546	37,514	72,098	72,098	105,460	130,984
Passive	29,199	39,822	77,263	77,978	118,468	134,501
Active	29,381	38,715	92,503	99,832	160,380	230,961
Restricted	29,502	37,572	83,202	84,728	128,558	190,692
Passive Cash	18,364	25,009	48,065	48,066	70,307	87,322
Active Cash	1,004	1,259	1,443	21,879	17,888	171,706
Restricted Cash	9,768	13,074	28,929	35,547	46,937	132,560

Table 3: Basic Statistics on Portfolios

by size.

Sector	Count
Healthcare	5
Energy	1
Utilities	2
Financial	2
Technology	3
Consumer Cyclical	1
Capital Goods	1
Consumer Non Cyclical	2
Services	2
Basic Materials	2

Table 4: Stock Diversification by Sector

Stock selection is the nebulous "value" that the experienced stock investor provides to the mix. Beyond the data is a "feeling" that a specific portfolio will do well in any environment. Investors view the market as a undulating surface that is reacting to economic and psychological stimuli, including paradigm shifts. Stocks perform singly and in concert based on that undulation. Investors select stocks based on available understanding of the current surface and various paths on that surface. Notice how our active strategies execute optimal buying and selling orders at points of inflection upon that surface.

Size	Count
Dow Industrials	8
Dow Utilities	2
S&P 500	6
NASDAQ 100	2
S&P 400 (MidCap)	2
S&P 600 (SmallCap)	1
Aggregate Growth	1

Table 5: Stock Diversification by Size

6 Conclusions and Further Research

We found that the PAM algorithm was extremely valuable when determining which stocks should be included in the portfolio from the original 138 stocks. It was fascinating that the clustering algorithm specified that Enron and Federal Mogul were unique.

With the initial $30,000, we obtained a 17.8% annual return on the cash surrogate passive strategy, 18.1% on the passive strategy, 20.2% on the combined cash protected and active strategy, and 23.3% using the active strategy only.

Comparing these results to the passive strategy being entirely invested in the S&P 500 Large Cap index with a 9.9% return, we find that all of our strategies are superior to that of a purely passive index strategy.

Our successful use of PAM in restricting the asset universe, leads us to believe that this data discovery tool (or her large dataset cousin CLARA also described in [1]) may be of use in asset management.

Our analysis assumed that there are no transaction costs. Realizing that this does not truly reflect the real world and we need to examine their effects. This may demonstrate that a portion of our 5% pickup may very well disappear.

The model that we have set up assumes that there are no taxes on capital gains. This is valid for endowments and IRA accounts, but the impact of taxes needs to be considered for more general active portfolio management.

The model also assumes that no new money is available for investment. We need to add a new money strategy in our models.

τ^2 is determined over the entire history of the price series. It is possible that decisions made at time t are influenced by the stock prices used in the τ^2 calculation at times greater than t. The asset selection process is also problematic in that when we use the PAM algorithm, we use only the most current history and restrict our asset universe to only those who have stock histories for nine years. The entire study needs to be conducted assuming that PAM is used and τ^2 is found only up to time t.

De Jong and Penzer also discuss methods to determine the type of outliers within the time

series. We use the τ^2 results only to specify a paradigm shift, no matter the type of outlier. Using their other techniques, we could add additional strategies based on the outlier types as well.

The active strategy executes a sell if the change in τ^2 and price are both negative, without regard to the magnitude of the change. Another strategy might be to consider movement only if the price change is above a certain threshold.

Other areas of future research would be to examine the sensitivity of the strategies to varying trading frequencies, bitesizes, trading limits, and cash protection limits.

Acknowledgement:

Steve Craighead: I would like to thank my wife Patricia and my children Sam, Michelle, Bradley, Evan and Carl for their patience. I would also like to thank Stephen Sedlak, Vice President of Corporate Actuarial at Nationwide for his ongoing support and encouragement in our various research endeavors.

Bruce Klemesrud: I would like to thank my wife and children for bearing with my investing errors.

A Partitioning Around Mediods(PAM)

The purpose for the partitioning of a data set of objects into k separate clusters is to find clusters whose members show a high degree of similarity among themselves but dissimilarity with the members of other clusters. The PAM algorithm searches for k representative objects among the data set. These k objects represent the varying aspect within the structure of the data. These representatives are called the mediods. The k clusters are then constructed by assigning each member of the data set to one of the mediods.

Using the notation of Kaufman and Rousseeuw we denote the distance between the objects i and j as $d(i, j)$. This can be any acceptable metric, such as Euclidean or the Manhattan distance.

Denote a dissimilarity as a nonnegative number $D(i, j)$ which is near zero when objects i and j are "near" and is large when i and j are "different." Usually $D(i, j)$ meets all of the metric requirements except the triangular inequality. Various candidates are discussed in [2].

The PAM algorithm consists of two parts. The first build phase follows the following algorithm:

1. Consider an object i as a candidate.

2. Consider another object j that has not been selected as a prior candidate. Obtain its dissimilarity D_j with the most similar previously selected candidates. Obtain its dissimilarity with the new candidate i. Call this $D(j, i)$. Take the difference of these two dissimilarities.

3. If the difference is positive, then object j contributes to the possible selection of i. Calculate $C_{ji} = \max(D_j - D(j,i), 0)$.

4. Sum C_{ji} over all possible j, $\sum_j C_{ji}$. This gives the total gain obtained by selecting i.

5. Choose the object i that maximizes the sum of C_{ji} over all possible j.

Repeat the process until k objects have been found.

The second step attempts to improve the set of representative objects. This does so by considering all pairs of objects (i, h) in which i has been chosen but h has not been chosen as a representative. Next it is determined if the clustering results improve if object i and h are exchanged. To determine the effect of a possible swap between i and h we use the following algorithm:

1. Consider an object j that has not been previously selected. We calculate its swap contribution C_{jih}:

   (a) If j is further from i and h than from one of the other representatives, set C_{jih} to zero.

   (b) If j is not further from i than any other representatives ($d(j,i) = D_j$), consider one of two situations:

      i. j is closer to h than the second closest representative and $d(j,h) < E_j$ where E_j is the dissimilarity of j and the second most similarly representative. Then $C_{jih} = d(j,h) - d(j,i)$. Note: C_{jih} can be either negative or positive depending on the positions of j, i and h. Here only if j is closer to i than to h is there a positive influence that implies that a swap between object i and h are a disadvantage in regards to j.

      ii. j is at least as distant from h than the second closest representative, or $d(j,h) \geq E_j$. Let $C_{jih} = E_j - D_j$. The measure is always positive, because it not wise to swap i with a h further away from j than with the second closest representative.

   (c) If j is further away from i than from at least one of the other representatives, but closer to h than to any other representative, $C_{jih} = d(i,h) - D_j$ will be the contribution of j to the swap.

2. Sum the contributions over all j. $T_{ih} = \Sigma_j C_{jih}$. This indicates the total result of the swap.

3. Select the ordered pair (i, h) which minimizes T_{ih}.

4. If the minimum T_{ih} is negative, the swap is carried out and the algorithm returns to the first step in the swap algorithm. If the minimum is positive or 0, the objective value cannot be reduced by swapping and the algorithm ends.

B Filter/Smoother Model

De Jong and Penzer in [1] obtain the outlier statistic τ^2 by creating two hypothetical models of the data. The τ^2 is a measurement at a specific time of how the data does not match the null hypothesis model. We will use their notation and set up the necessary notation to obtain their τ^2 statistic.

Using their notation, let the data be represented by $y = (y_1^{'}, y_2^{'}, \ldots, y_n^{'})'$ for time $t = 1, \ldots, n$. Assume that the null model of y has mean 0 and has a covariance matrix $\sigma^2\Sigma$. The Σ gives the serial correlation of the data series y. We will represent the null model as $y \sim (0, \sigma^2\Sigma)$. We want to determine if there are departures from the null model and this is modelled by the addition of an intervention variable $D = (D_1^{'}, D_2^{'}, \ldots, D_n^{'})'$, and the alternative hypothesis model will be denoted as $y \sim (D\delta, \sigma^2\Sigma)$, which reduces to the null model if $\delta = 0$.

If D and Σ are known, the intervention parameter δ can be estimated using generalized least squares and is $\hat{\delta} = S^{-1}s$ with $cov(\hat{\delta}) = \sigma^2 S^{-1}$, where $s = D'\Sigma^{-1}y$, and $S = D'\Sigma^{-1}D$. The test of the hypothesis of no shock, $\delta = 0$ is based on $\hat{\delta}'\{cov(\hat{\delta})\}^{-1}\hat{\delta} = \sigma^2 s' S^{-1} s$.

Frequently σ^2 is replaced by the maximum likelihood estimate, $\hat{\sigma}^2 = (y'\Sigma^{-1}y)/n$, which yields the test statistic $\tau^2 = \hat{\sigma}^2 s' S^{-1} s$. τ^2 has an approximate χ_p distribution where p is the rank of the matrix S.

De Jong and Penzer create state-space models for y_t. The null state-state form of y_t is

$$\begin{aligned} y_t &= Z_t\alpha_t + G_t\epsilon_t \qquad \text{and} && \text{(B.1)} \\ \alpha_{t+1} &= T_t\alpha_t + H_t\epsilon_t, \quad t = 1, \ldots, n, && \text{(B.2)} \end{aligned}$$

where $\epsilon_t \sim N(0, \sigma^2 I), \alpha_t \sim (\alpha_1, \sigma^2 P_1)$, and ϵ_t and α_1 are mutually uncorrelated. The matrices Z_t, T_t, G_t, and H_t are deterministic but could vary over time. For $t = 1, 2, \ldots, n$ let

$$\begin{aligned} v_t &= y_t - Z_t a_t && \text{(B.3)} \\ F_t &= Z_t P_t Z_t^{'} + G_t G_t^{'}, && \text{(B.4)} \\ K_t &= (T_t P_t Z_t^{'} + H_t G_t^{'}) F_t^{-1}, && \text{(B.5)} \\ a_{t+1} &= T_t a_t + K_t v_t, \qquad \text{and} && \text{(B.6)} \\ P_{t+1} &= T_t P_t L_t^{'} + H_t J_t^{'}, && \text{(B.7)} \end{aligned}$$

where $L_t = T_t - K_t Z_t$ and $J_t = H_t - K_t G_t$. Now using the Kalman Smoother, we take the results of the Kalman Filter and initialize the Smoother with $r_n = 0$ and $N_n = 0$. Then for $t = n, \ldots, 1$

$$
\begin{aligned}
u_t &= F_t^{-1} v_t - K_t^{'} r_t, && \text{(B.8)} \\
M_t &= F_t^{-1} + K_t^{'} N_t K_t, && \text{(B.9)} \\
r_{t-1} &= Z_t^{'} u_t + T_t^{'} r_t, \qquad \text{and} && \text{(B.10)} \\
N_{t-1} &= Z_t^{'} F_t^{-1} Z_t + L_t^{'} N_t L_t. && \text{(B.11)}
\end{aligned}
$$

De Jong and Penzer set up the alternative model as

$$
\begin{aligned}
y_t &= X_t \delta + Z_t \alpha_t + G_t \epsilon_t \qquad \text{and} && \text{(B.12)} \\
\alpha_{t+1} &= W_t \delta + T_t \alpha_t + H_t \epsilon_t, && \text{(B.13)}
\end{aligned}
$$

where X_t and W_t are called the shock design matrices and δ is the shock magnitude.

They go on to state that for a given time t and null state-space model the maximum of $\rho_t^2 = s_t' S_t^{-1} s_t$, with respect to the X_t and W_t is

$$
\rho_t^{*2} = v_t^{'} F_t^{-1} v_t + r_t^{'} N_t^{-1} r_t, \qquad \text{(B.14)}
$$

where v_t, F_t, r_t, and N_t are computed with the Kalman Filter Smoother applied to the null model. The maximum is attained when $X_t = v_t$ and $W_t = K_t v_t + N_t^{-1} r_t$.

Finally de Jong and Penzer show that τ^2 has a maximum value at $\tau_t^{*2} = \sigma^{-2} \rho_t^{*2}$ and that a plot of τ_t^{*2} against t reveals when the shock design is significant at time t.

We use this τ^{*2} as our outlier statistic in our active strategies.

References

[1] P. de Jong and J. Penzer (1998), "Diagnosing Shocks in Time Series", *JASA* 93, no. 442, 796-806.

[2] Kaufman, L. and Rousseeuw, P.J. (1990), "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley, New York.